ОГЛАВЛЕНИЕ

Часть 1. Тематический выпуск «Современные методы создания компьютерных ресурсов для автоматической обработки языка»

А. И. Кацнельсон, О. Н. Ляшевская
АВТОМАТИЧЕСКОЕ РАСПОЗНАВАНИЕ РЕЧИ: ФАКТОРЫ СНИЖЕНИЯ ТОЧНОСТИ ПРИ РАБОТЕ С РЕГИОНАЛЬНЫМ ВАРИАНТОМ РУССКОГО ЯЗЫКА (НА МАТЕРИАЛЕ ДАГЕСТАНА) **2058–2079**

К. С. Клокова, М. А. Кронгауз, В. А. Шульгинов, Т. А. Юдина
ПРАГМАТИЧЕСКИЕ ГРАНИЦЫ ПРИВЕТСТВИЙ И ПРОЩАНИЙ: ОЦЕНКА LLM НА МАТЕРИАЛЕ МУЛЬТИМЕДИЙНОГО КОРПУСА РЕЧЕВОГО ЭТИКЕТА **2080–2104**

Е. А. Козлова, А. М. Баюк, М. А. Петрова
О ПОВЫШЕНИИ КАЧЕСТВА RAG ПРИ ВЫЯВЛЕНИИ ПРОТИВОРЕЧИЙ В СУДЕБНЫХ ПОСТАНОВЛЕНИЯХ **2105–2133**

А. В. Колмогорова, Е. В. Явшиц, М. С. Майорова, С. В. Дмитриева
ПРОЕКТИРОВАНИЕ БЕЗНАСИЛЬСТВЕННОЙ БОЕВОЙ СИСТЕМЫ ДЛЯ ДЕТСКОЙ RPG-ИГРЫ **2134–2157**

А. А. Некесова, Е. А. Сидорова, Ж. Б. Ламашева, М. А. Самбетбаева, А. С. Еримбетова, М. Ж. Калдарова, А. А. Назымхан
ОТ РАЗМЕЧЕННОГО КОРПУСА К ОНТОЛОГИИ: ФОРМАЛЬНАЯ МОДЕЛЬ МНОГОУРОВНЕВОЙ СЕМАНТИЧЕСКОЙ РАЗМЕТКИ ФЕЙКОВЫХ НОВОСТЕЙ В КАЗАХСКО-РУССКОМ МЕДИАПРОСТРАНСТВЕ (KazFakeCorpus/KazFakeOnto) **2158–2187**

Ф. А. Садковский, Р. Р. Насырова, А. А. Сорокин
RusHallu-RAG: КОМПЛЕКСНЫЙ ПОДХОД К ОБНАРУЖЕНИЮ ГАЛЛЮЦИНАЦИЙ В РУССКОЯЗЫЧНЫХ RAG-СИСТЕМАХ **2188–2214**

А. С. Шатских, А. А. Сорокин
АВТОМАТИЧЕСКИЙ МОРФОЛОГИЧЕСКИЙ АНАЛИЗ ДЛЯ ОСЕТИНСКОГО ЯЗЫКА НА МАТЕРИАЛЕ УСТНОГО КОРПУСА 2215–2239

П. В. Ярошенко, А. А. Петрова, Н. В. Лукашевич
МОРФОЛОГИЧЕСКИЕ ФОРМЫ ГЛАГОЛА В ЗАДАЧЕ РАСПОЗНАВАНИЯ ЭМОЦИЙ: ХАРАКТЕРИСТИКИ ДАННЫХ И РЕЗУЛЬТАТЫ КЛАССИФИКАЦИИ 2240–2264

Часть 2. Оригинальные статьи

А. А. Агафонов, А. Р. Хусаинов, Н. А. Прокопьев
PROMPTALIGN: АВТОМАТИЧЕСКОЕ ИТЕРАТИВНОЕ ВЫРАВНИВАНИЕ БОЛЬШОЙ ЯЗЫКОВОЙ МОДЕЛИ ПОД ЭКСПЕРТНУЮ РАЗМЕТКУ БЕЗ ДООБУЧЕНИЯ 2265–2292

Е. И. Александров, И. Н. Александров, Д. В. Беляков, Н. А. Давыдова, П. В. Зрелов, Л. А. Калмыкова, М. А. Любимова, Т. Ф. Сапожникова, Т. С. Сыресина, А. В. Яковлев
ВОЗМОЖНОСТИ РАЗРАБОТАННОЙ ИНФОРМАЦИОННОЙ СИСТЕМЫ СОПРОВОЖДЕНИЯ ЛИЦЕНЗИЙ 2293–2306

К. В. Гарев
ТРЕБОВАНИЯ К АРХИТЕКТУРЕ СИСТЕМЫ УПРАВЛЕНИЯ НАУЧНЫМИ ДАННЫМИ И ВЫБОР РЕШЕНИЙ ДЛЯ РЕАЛИЗАЦИИ ЕЕ КОМПОНЕНТОВ 2307–2329

А. Р. Гатиатуллин
АРХИТЕКТУРА ЛИНГВИСТИЧЕСКОЙ ПЛАТФОРМЫ TurkLang 2330–2356

Ф. М. Гафаров, П. Н. Устин, К. С. Николаев
ПРОГНОЗИРОВАНИЕ ИЗМЕНЕНИЙ В ПОВЕДЕНИИ ЛИЧНОСТИ ПРИ ВОЗДЕЙСТВИИ СОБЕСЕДНИКА 2357–2370

П. О. Гафурова
УПРАВЛЕНИЕ И ОРГАНИЗАЦИЯ МЕТАДАННЫХ НАУЧНЫХ СТАТЕЙ В ЦИФРОВЫХ МАТЕМАТИЧЕСКИХ БИБЛИОТЕКАХ **2371–2389**

Р. А. Гильмуллин
БАЗА СЕМРОНОВ ТАТАРСКОГО ЯЗЫКА КАК ЦИФРОВОЙ ЛИНГВИСТИЧЕСКИЙ РЕСУРС ДЛЯ ОБЪЯСНИТЕЛЬНОГО ИСКУССТВЕННОГО ИНТЕЛЛЕКТА **2390–2409**

М. С. Дьяченко, А. Г. Леонов
АВТОМАТИЧЕСКАЯ КОРРЕКЦИЯ ИНДИВИДУАЛЬНОЙ ОБРАЗОВАТЕЛЬНОЙ ТРАЕКТОРИИ В АДАПТИВНЫХ СИСТЕМАХ ОБУЧЕНИЯ НА ОСНОВЕ КОМПОЗИТНОГО ПОДХОДА **2410–2432**

В. С. Еременко, В. В. Наумова
МУЛЬТИАГЕНТНАЯ СИСТЕМА ДЛЯ ЦИФРОВОЙ ИНТЕЛЛЕКТУАЛЬНОЙ СФЕРЫ ГЕОЛОГИЧЕСКИХ ЗНАНИЙ **2433–2445**

Ю. А. Загорулько, И. Р. Ахмадеева, А. С. Серый, Е. А. Сидорова
ПОДХОД К ОЦЕНКЕ ДОСТОВЕРНОСТИ ИНФОРМАЦИИ В НАУЧНО-ПОПУЛЯРНОМ ДИСКУРСЕ НА ОСНОВЕ АНАЛИЗА АРГУМЕНТАЦИИ И НАУЧНОГО ФАКТ-ЧЕКИНГА **2446–2481**

Д. С. Коптев, Ю. А. Халин, А. С. Сизов
ИНФОРМАЦИОННО-АНАЛИТИЧЕСКАЯ СИСТЕМА ОЦЕНКИ КОГНИТИВНЫХ ФУНКЦИЙ АВИАЦИОННОГО ПЕРСОНАЛА НА ОСНОВЕ НЕЧЕТКОЙ ЛОГИКИ **2482–2507**

А. Г. Леонов, К. А. Мащенко
ЦИФРОВАЯ ОБРАЗОВАТЕЛЬНАЯ ПЛАТФОРМА МИРЕРА **2508–2539**

С. В. Масюченко, А. О. Щирый
МЕТОДИКА, РЕЗУЛЬТАТЫ И СРАВНИТЕЛЬНЫЙ АНАЛИЗ ПРИМЕНЕНИЯ НЕЙРОСЕТЕВЫХ МОДЕЛЕЙ MASK-RCNN И YOLO ДЛЯ АВТОМАТИЧЕСКОГО ВЫДЕЛЕНИЯ ТРЕКОВ МОД НА ИОНОГРАММАХ НАКЛОННОГО РАДИОЗОНДИРОВАНИЯ ИОНОСФЕРЫ **2540–2554**

К. С. Николаев
ОБРАБОТКА СКАНИРОВАННЫХ МАТЕМАТИЧЕСКИХ PDF-ДОКУМЕНТОВ В СЕРВИСЕ СЕМАНТИЧЕСКОГО ПОИСКА ФОРМУЛ **2555–2572**

А. С. Савинчева, З. А. Еникеева, К. Ш. Еникеев
МЕТОД СЕМАНТИЧЕСКОГО АНАЛИЗА ИЗМЕНЕНИЙ ТРЕБОВАНИЙ К ПРОГРАММНОМУ ОБЕСПЕЧЕНИЮ В ВЕРСИЯХ ТЕХНИЧЕСКОЙ ДОКУМЕНТАЦИИ **2573–2587**

А. М. Симаков, А. О. Щирый
МЕТОДИКА И ПРЕДВАРИТЕЛЬНЫЕ РЕЗУЛЬТАТЫ СИНТЕЗА ИОНОГРАММ ВЕРТИКАЛЬНОГО РАДИОЗОНДИРОВАНИЯ ИОНОСФЕРЫ ПОСРЕДСТВОМ ФИЗИЧЕСКИ ИНФОРМИРОВАННЫХ НЕЙРОСЕТЕЙ **2588–2600**

М. В. Шматко, Е. В. Евдущенко
КОГНИТИВНОЕ МОДЕЛИРОВАНИЕ ПОЛЬЗОВАТЕЛЬСКОГО ВЗАИМОДЕЙСТВИЯ В МУЛЬТИРОЛЕВЫХ ВЕБ-ПРИЛОЖЕНИЯХ С 3D-ФУНКЦИОНАЛОМ **2601–2630**

УДК 004.934

АВТОМАТИЧЕСКОЕ РАСПОЗНАВАНИЕ РЕЧИ: ФАКТОРЫ СНИЖЕНИЯ ТОЧНОСТИ ПРИ РАБОТЕ С РЕГИОНАЛЬНЫМ ВАРИАНТОМ РУССКОГО ЯЗЫКА (НА МАТЕРИАЛЕ ДАГЕСТАНА)

А. И. Кацнельсон[1] [0009-0009-2808-3243], О. Н. Ляшевская[2] [0000-0001-8374-423X]

[1, 2]*Национальный исследовательский университет «Высшая школа экономики», г. Москва, Россия*

[1]akatsnelson@hse.ru, [2]olesar@yandex.ru

Аннотация

Работа посвящена сравнительному анализу качества автоматического распознавания стандартного русского языка и его дагестанского регионекта (дагестанского русского) – контактного идиома, формирующегося в условиях многоязычия Дагестана. Десять конфигураций ASR-систем (без дообучения), представляющих три архитектурных подхода (Conformer+RNNT, Conformer+CTC, Encoder-Decoder Transformer), протестированы на корпусах стандартного русского (GOLOS, RuDevices, примерно 20 и 90 тыс. высказываний) и дагестанского русского (DaGRuS, примерно 40 тыс. высказываний). Все модели демонстрируют падение метрики WER на 33–50% при переходе к дагестанскому русскому, что указывает на проблему сильного доменного сдвига. Анализ структуры ошибок выявил три основных механизма, вызывающих снижение WER: это русификация – систематическая замена дагестанских форм фонетически близкими стандартными русскими словами; удаление дискурсивных маркеров и OOV-лексики (этнонимов, топонимов); галлюцинации генеративных моделей на непонятных фрагментах. Установлено, что сквозные архитектуры-архитектуры деградируют слабее каскадных (33–35% против 40–42%), однако ни одна из моделей не достигает приемлемого качества (минимум WER около 43%). Сформулированы рекомендации для проектирования ASR-систем, устойчивых к региональной вариативности русского языка.

Ключевые слова: *автоматическое распознавание речи, ASR,* дагестанский регионект русского языка*, регионект, региональные варианты речи, WER,*

© А. И. Кацнельсон, О. Н. Ляшевская, 2026.
Данная статья распространяется на условиях международной лицензии Creative Commons License Attribution 4.0 International (CC BY 4.0).

языковая вариативность, доменный сдвиг, Conformer, Whisper, GigaAM, галлюцинации.

ВВЕДЕНИЕ

Современные системы автоматического распознавания речи (Automatic Speech Recognition, ASR) достигают значения WER (Word Error Rate, коэффициент ошибок на уровне слов) ниже 5% на эталонных корпусах стандартного русского языка. Русский язык существует не как единый стандарт, а как спектр региональных вариантов, формирующихся под влиянием контактных языков. По данным переписи 2020-2021 годов, более 20 млн жителей России говорят на языках коренных народов наряду с русским, что структурно влияет на их русскую речь.

Дагестан представляет собой крайний случай языкового разнообразия: в республике функционирует более 30 автохтонных языков. Русский выступает языком межнационального общения и подвержен влиянию на всех уровнях [1]. Дагестанский регионолект русского языка (далее тире дагестанский, или региональный, русский) характеризуется специфическими консонантными фонетическими чертами, оригинальной лексикой (топонимы, этнонимы) и особыми дискурсивными структурами. В результате голосовые помощники и системы транскрипции работают для таких говорящих существенно хуже.

Целью настоящей работы было количественно оценить и лингвистически охарактеризовать разрыв в качестве ASR между стандартным и региональным русскими. Десять конфигураций ASR протестированы на корпусе DaGRuS; в качестве контрольных выбраны GOLOS и RuDevices. В качестве задач для поставленной цели рассмотрены следующие: выявление конкретных лингвистических контекстов деградации, анализ типов ошибок и соотнесение их с архитектурными свойствами моделей.

СМЕЖНЫЕ РАБОТЫ

Устойчивость к региональной вариативности привлекает все большее внимание в литературе по ASR. В работе [2] продемонстрированы значительные различия WER для шотландского, индийского и нигерийского английских в коммерческих системах. В [3] показано, что дообучение на региональных данных снижает WER вплоть до 35% относительно исходного уровня. Для языков, отличных

от английского, многоязычные self-supervised-модели, протестированные на типологически разнообразных малоресурсных языках, сохраняют устойчивый разрыв в точности [4].

Для русского языка региональная вариативность в ASR остается практически неисследованной. В существующих бенчмарках, таких как GOLOS [5], преобладает стандартная речь. Корпус DaGRuS [6] представляет собой уникальную тестовую площадку для оценки устойчивости ASR к местным вариантам. В обзоре [7] отмечено, что архитектура декодера существенно определяет типы ошибок при сдвиге распределения, что согласуется с гипотезами, проверяемыми в настоящей работе.

Для смягчения таких доменных сдвигов в промышленных приложениях все чаще применяют федеративное обучение [8]. Однако анализ ошибок на диалектах показывает, что даже при использовании продвинутых методов по-прежнему трудно разработать единые модули нормализации, стабильно работающие на всех региональных вариантах.

ДАННЫЕ И МЕТОДИКИ

Корпуса и лингвистический профиль

Корпус DaGRuS [6] – это собрание спонтанных бытовых нарративов носителей дагестанских языков на русском, записанных в полевых условиях (39851 высказывание, 243 тыс. токенов[1]). Информанты – носители аварского, даргинского, кумыкского, лакского, лезгинского и других языков. Высказывания DaGRuS снабжены богатой ручной разметкой, включая технические события, неразборчивые фрагменты и нестандартные произносительные варианты с указанием канонических эквивалентов.

В качестве контрольных корпусов стандартного русского языка выступили GOLOS (части crowd и farfield) [5] и RuDevices (93241 высказывание). Для надежной оценки перед расчетом метрик применялись корпусно-специфичные конвейеры нормализации (разрешение разметки, стандартизация числительных и т. д.). Подробные описания нормализаторов представлены в Приложении А.

[1] Корпус DaGRuS доступен для исследовательских целей по запросу к авторам. Снимок корпуса, скрипты нормализации текста и конфигурации гиперпараметров инференса, использованные в настоящем исследовании, будут опубликованы в открытом репозитории.

В табл. 1 приведены основные лингвистические характеристики корпусов. DaGRuS демонстрирует «тяжелый хвост редкой лексики» (наклон Ципфа −1.175), представленный главным образом местными этнонимами и топонимами, отсутствующими в обучающих выборках ASR. Кроме того, в нем существенно выше доля частиц (13.0%) и междометий, что отражает спонтанный разговорный характер корпуса.

Табл. 1. Лингвистические характеристики корпусов.

Характеристика	GOLOS-crowd	GOLOS-far	RuDevices	DaGRuS
Высказывания	9.896	1.915	93.241	39.851
Токены	50.394	8.382	732.589	243.840
Средняя длина (слов)	5.09	4.38	7.86	6.12
MSTTR	0.876	0.689	0.813	0.659
Наклон Ципфа	−0.857	−0.747	−1.047	−1.175
Частицы (%)	1.2	1.3	10.8	13.0
Междометия (%)	0.2	0.5	2.0	2.1
Местоимения (%)	3.7	2.3	10.4	8.5
Гапаксы/словарь (%)	66.8	68.4	55.8	51.2
80% покрытие (леммы)	1.189	525	970	356

Предобработка и нормализация текста

Надежный расчет WER требует, чтобы эталонный текст и гипотеза модели были представлены в сопоставимом формате. Поскольку корпуса различаются конвенциями разметки, а модели ASR – форматом вывода, перед расчетом метрик применяют корпусно-специфичные конвейеры нормализации. Все нормализаторы имеют общую базу: приведение к нижнему регистру, унификация «ё → е» (большинство моделей ASR не различает эти символы), удаление пунктуации и схлопывание пробелов. Подробные описания корпусно-специфичных нормализаторов даны в Приложении А.

Метрики оценки

Основная метрика – Word Error Rate (WER), определяемая как отношение суммарного числа операций редактирования на уровне слов (замены *S*, удаления *D*, вставки *I*) к числу слов в эталоне: WER = (*S* + *D* + *I*)/*N*. Дополнительно используется Character Error Rate (CER), вычисляемая аналогично на уровне символов и менее чувствительная к выбору токенизации. WER разложена на три компоненты: долю удалений, долю вставок и долю замен – для характеристики качественной структуры ошибок.

Статистическая значимость различия WER между корпусами оценивается с помощью U-критерия Манна – Уитни [9]. Доверительные интервалы для среднего WER получены бутстрэп-ресамплингом [10] (1000 итераций, 95%-ные процентильные интервалы).

Модели ASR и гипотезы

Протестированы десять конфигураций на основе семи моделей (табл. 2). Выбор покрывает архитектурное разнообразие: Conformer+RNN-Transducer (авторегрессионный), Conformer+CTC (неавторегрессионный) и Encoder-Decoder Transformer (генеративный seq2seq). Представлены разные масштабы (от ~100 млн до 809 млн параметров) и режимы обучения (каскадные и сквозные (end-to-end)).

Табл. 2. Характеристики оцениваемых моделей ASR.

Модель	Разработчик	Архитектура	Параметры, млн
GigaAM-v2/v3 RNNT/CTC	Sber	Conformer + RNNT/CTC	~600
GigaAM-v3-e2e RNNT/CTC	Sber	E2E Conformer, joint CTC/Att	~600
Whisper-large-v3	OpenAI	Enc-Dec Transformer	~809
T-One	T-Bank	Conformer + Transducer	~100

Все модели оценивались «из коробки» в режиме zero-shot, без какого-либо целевого дообучения. Для Whisper-large-v3 (openai/whisper-large-v3) принудительно задавался языковой токен русского языка и использовалось «жадное декодирование» по умолчанию (beam size 1, temperature 0.0) для обеспечения детерминированности вывода.

Проверялись три гипотезы.

H1 (влияние декодера): RNN-T склонен к удалениям на входах с низкой уверенностью, CTC – к заменам, а seq2seq «галлюцинирует».

H2 (совместный E2E и каскадные AM+LM): каскадными мы называем модели с раздельным декодированием акустической и языковой моделей (например, стандартный GigaAM RNN-T), в отличие от чисто совместных сквозных архитектур. Мы предполагаем, что совместные E2E-модели формируют более гибкие представления, слабее привязанные к стандартным фонемам, и тем самым смягчают деградацию.

H3 (доменная генерализация): узкодоменные модели (например, T-One) деградируют сильнее, чем модели широкого домена.

РЕЗУЛЬТАТЫ: СТАНДАРТНЫЙ И ДАГЕСТАНСКИЙ РУССКИЙ

Общая картина: градиент деградации

На рис. 1 представлены результаты по четырем корпусам. Наборы данных можно рассматривать как последовательность: стандартная речь (GOLOS-crowd), стандартная речь в шумных условиях (GOLOS-farfield), спонтанная стандартная речь (RuDevices) и спонтанная речь местных вариантов (DaGRuS). Хотя WER растет на каждом шаге, переход к дагестанскому русскому дает резкий скачок, непропорциональный предшествующим приростам.

Все восемь моделей, протестированных и на GOLOS-crowd, и на DaGRuS, демонстрируют статистически значимую деградацию (критерий Манна – Уитни, $p < 0.001$ для всех пар; размер эффекта $r = 0.52 – 0.76$). Доверительные интервалы (95%-ный бутстрэп) не пересекаются ни для одной пары.

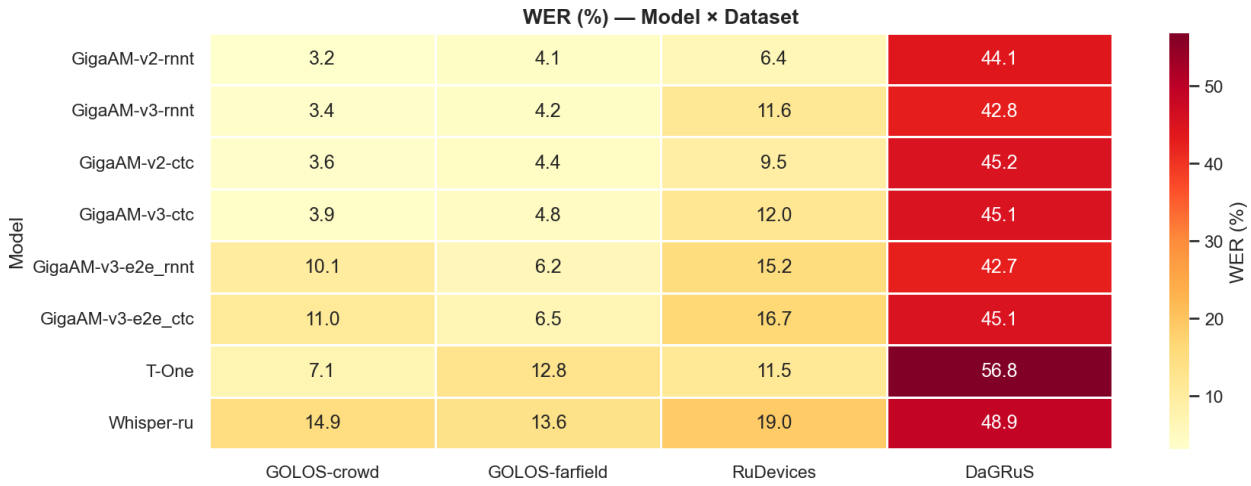


Рис. 1. Матрица WER (%) по корпусам и моделям. Прочерки обозначают отсутствие данных.

Переход от GOLOS-crowd к RuDevices (оба корпуса представляют стандартный русский, но последний – спонтанная речь, записанная на различных устройствах) дает умеренный рост WER: +6–7% для каскадных моделей GigaAM, +4–6% для e2e-моделей и +4% для Whisper. Однако переход от RuDevices к DaGRuS вызывает дополнительный скачок еще на 28–36%. Таким образом, даже с учетом эффектов спонтанности и шума основным фактором падения качества – является региональный фонетический компонент дагестанского русского.

Принципиально важно, что ранжирование моделей заметно изменяется в зависимости от набора данных. На GOLOS-crowd лучшие результаты показали каскадные GigaAM v2/v3 (WER 3–4%), на RuDevices лидирует GigaAM-v2-ctc (9.5%), а на DaGRuS – GigaAM-v3-e2e RNN-T (42.7%). Коэффициент ранговой корреляции Спирмена между WER на GOLOS и DaGRuS для восьми моделей равен 0.12 ($p > 0.05$): качество на стандартном русском не является предиктором качества на дагестанском русском.

Более высокий WER моделей GigaAM-v3 E2E на GOLOS-crowd (10–11% против 3–4% у каскадных версий) объясняется встроенной текстовой нормализацией и пунктуацией: эти модели выводят числительные словами и вставляют знаки препинания (запятые и точки), что создает рассогласование с форматом эталона. Однако на DaGRuS, где речь спонтанна, а числительные естественно произносятся словами, этот эффект компенсируется, и E2E-модели демонстрируют наименьший абсолютный WER (42.7–45.1%).

Эволюция структуры ошибок по корпусам

Разложение ошибок на удаления (Del), вставки (Ins) и замены (Sub) выявляет качественный сдвиг в характере ошибок при переходе к дагестанскому русскому (табл. 3).

Табл. 3. Профиль ошибок (Del/Ins/Sub) в процентах от числа слов эталона по корпусам.

Модель	GOLOS-crowd			RuDevices			DaGRuS		
	Del	Ins	Sub	Del	Ins	Sub	Del	Ins	Sub
GigaAM-v3-e2e RNN-T	1.49	0.69	5.42	1.98	2.36	6.60	8.11	3.60	10.17
GigaAM-v3 RNN-T	0.28	0.19	1.79	2.23	1.36	5.37	8.58	3.24	9.87
GigaAM-v3 CTC	0.27	0.23	2.10	2.00	1.49	5.78	7.70	3.48	12.81
GigaAM-v2 CTC	0.25	0.21	1.88	1.54	1.04	4.92	7.90	2.51	11.59
T-One	0.45	0.63	3.51	–	–	–	11.96	2.78	15.29
Whisper-ru	1.22	1.85	7.22	2.86	2.21	8.23	6.95	4.18	14.70

На GOLOS-crowd удаления минимальны (0.2–1.5%), доминируют замены. На RuDevices удаления возрастают до 1.5–2.9%: модели начинают пропускать слова в спонтанной речи. На DaGRuS удаления скачкообразно растут до 7–12%: модели систематически «молчат» на незнакомых фрагментах. Замены также увеличиваются в два-три раза (с 2–7% до 10–15%). Это означает, что на дагестанском русском модели не просто ошибаются чаще, они ошибаются качественно иначе.

Профиль ошибок на RuDevices сохраняет пропорции, наблюдаемые на GOLOS (доминирование замен, умеренные удаления). На DaGRuS эти пропорции нарушаются: доля удалений в общем числе ошибок возрастает с 10–15% до 30–40% для большинства моделей. Это говорит о том, что региональная фонетика создает принципиально иной тип трудности по сравнению с обычной спонтанной речью.

АНАЛИЗ ОШИБОК

Части речи

Анализ ошибок по частям речи (ЧР), размеченным с помощью pymorphy2 [11], позволяет локализовать конкретные зоны падения качества. В табл. 4 приведены доли неверно распознанных слов для двух моделей: GigaAM-v3-e2e RNN-T и Whisper-ru – на корпусах GOLOS-crowd и DaGRuS.

Табл. 4. Доля ошибок (%) по частям речи: GOLOS-crowd → DaGRuS.

ЧР	Описание	GigaAM GOL	GigaAM DaG	Whis. GOL	Whis. DaG
VERB	Глагол	4.2	54.3	11.6	59.4
PREP	Предлог	6.5	47.3	12.1	48.6
INFN	Инфинитив	3.3	49.3	13.3	55.6
NOUN	Существительное	30.3	63.0	31.5	64.7
NUMR	Числительное	37.4	56.6	24.6	81.7
INTJ	Междометие	33.3	62.3	24.1	65.4
NPRO	Местоимение	2.2	38.4	10.2	43.0
UNKN	Неизв.	84.9	97.6	81.2	95.4

Наиболее выраженная деградация наблюдается у глаголов (+50% у GigaAM, +48% у Whisper), инфинитивов и предлогов. Глаголы в дагестанском русском часто демонстрируют нестандартное управление и морфологию из-за интерференции родного языка (L1) говорящих [12]. Существительные, уже имеющие высокую базовую долю ошибок на GOLOS (~30%), достигают 63–65% на DaGRuS; эту категорию преимущественно наполняют внесловарные (Out-Of-Vocabulary, OOV) этнонимы и топонимы.

Особенно показателен случай числительных: у GigaAM-e2e ошибки растут лишь на +19%, тогда как у Whisper – на +57%. Причина состоит в архитектурной особенности: при декодировании Whisper предпочитает цифровую запись (например, «1961» вместо «тысяча девятьсот шестьдесят первого»). Напротив,

модели GigaAM E2E благодаря встроенной нормализации выводят числительные словами, что соответствует формату эталона DaGRuS.

Категория UNKN – слова, не распознанные морфологическим анализатором, – дает 95–98% ошибок на DaGRuS у всех моделей. Это прежде всего дагестанизмы, не имеющие русских эквивалентов: топонимы, этнонимы и бытовая лексика кавказского происхождения.

«Русификация» как основной паттерн замен

Анализ конкретных замен выявляет систематический паттерн, который мы называем «русификацией»: модели заменяют дагестанские формы фонетически ближайшими стандартными русскими словами. Это предсказуемая стратегия декодера, обученного на стандартном русском: столкнувшись с незнакомой фонетической последовательностью, он ищет ближайшее соответствие в пространстве известной лексики. Иллюстративные примеры приведены в табл. 5.

Табл. 5. Примеры «русификации»: эталон и гипотезы двух моделей. Символ ✓ обозначает верную транскрипцию.

Эталон	GigaAM-v3-e2e RNN-T	Whisper-ru
…к грузинам	к грузинам ✓	к кукурузинам
…относились к кумуху	к муху	к кумуху ✓
наше село	наше село ✓	наши силу
…через село и вниз	через село и вниз ✓	через силу и вниз
…наших овец ведь мало	наших овец ведь мало ✓	наших везде мало

Русификация проявляется по-разному в зависимости от архитектуры. Как генеративная модель, Whisper активно «достраивает» последовательность до правдоподобного текста, подбирая фонетически близкие русские слова (*село* → *силу*, *овец* → *везде*, *грузинам* → *кукурузинам*). GigaAM-e2e, использующий символьный/BPE-словарь, чаще выдает приблизительные незнакомые формы (например, *кумуху* → *муху*), что снижает семантическое искажение, но также

прибегает к «русификации» (*коршинцы → корчинцы*). Обе стратегии ведут к потере смысла, однако генеративный подход порождает особенно опасные «правдоподобные» ошибки (галлюцинации).

Дискурсивные маркеры как источник ошибок

Маркеры обратной связи (*угу, ага, мм, ээ*) составляют около 15% токенов DaGRuS, но практически отсутствуют в обучающих данных стандартных корпусов. В результате модели справляются с ними плохо: например, слово *угу* (2668 вхождений) имеет долю ошибок свыше 50% у Whisper. Ключевой проблемой является сочетание этих маркеров с региональной фонетикой: стандартное русское *угу* артикулируется иначе, чем его дагестанский аналог. Кроме того, оценка этих ошибок осложняется неоднозначностью акустической реализации и непоследовательностью корпусной разметки. Подробный разбор наиболее частых удалений маркеров и проблем разметки приведен в Приложении Б.

Сегменты с нулевой точностью: где не справляется ни одна модель

«Слепые зоны» – это высказывания, на которых все десять моделей дают WER выше 50%, – составили 23105 примеров (58% DaGRuS). Характерная лексика этих сегментов включает такие единицы, как рикванинский, лезгины, тюркский, ногайцы, Янгикент, табасаранский, полулак, Джамал, Тажутдин. Это дагестанские этнонимы, топонимы и антропонимы, для которых в обучающих данных моделей нет ни акустических, ни языковых соответствий.

На DaGRuS наблюдается отрицательная корреляция между длиной высказывания и WER (ρ Спирмена ≈ −0.25, $p < 0.001$ для всех моделей), тогда как на GOLOS-crowd корреляция близка к нулю, а на RuDevices – слабо положительная (см. рис. 2). Этот «парадокс длины» объясняется природой коротких реплик в DaGRuS: высказывания из 1–3 слов часто представляют собой изолированные дагестанизмы, топонимы или маркеры обратной связи, где одна ошибка дает WER 100%. В более длинных фразах выше доля стандартной русской лексики, которая «разбавляет» общую долю ошибок.

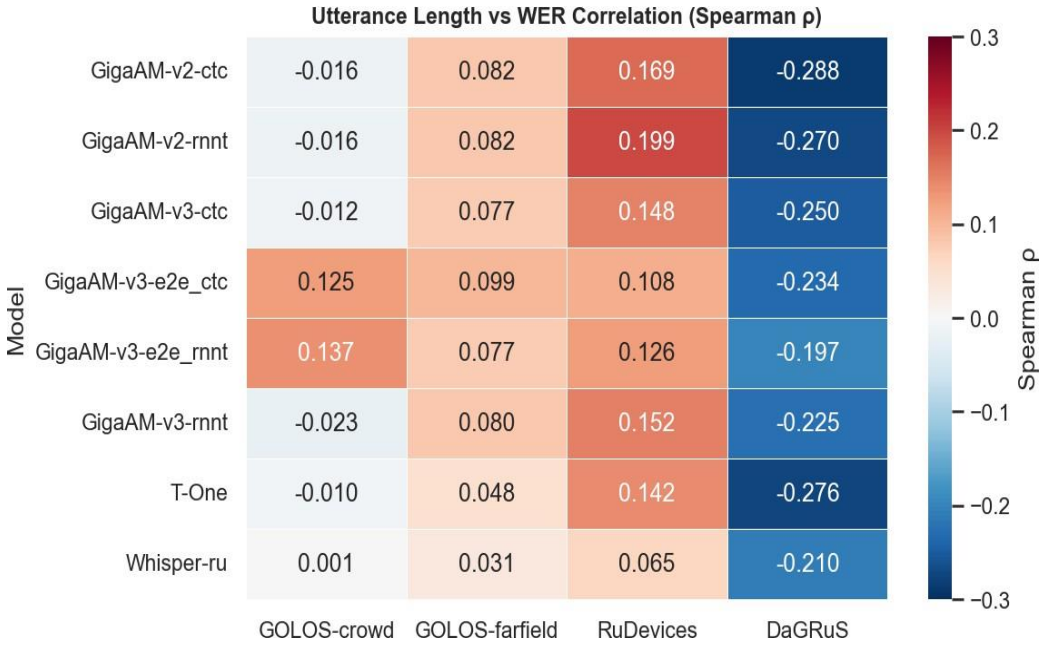


Рис. 2. Корреляция Спирмена между длиной высказывания и WER по моделям и корпусам.

Стратегические контрасты: расхождение архитектур

Сравнение гипотез, порожденных разными моделями на одних и тех же примерах, выявляет контрастные стратегии обработки неразборчивого акустического входа. Расхождения наиболее выражены между семейством Conformer (GigaAM) и генеративным Whisper. GigaAM точнее работает с числительными (записывая их словами), топонимами и дискурсивными маркерами. Whisper, напротив, лучше справляется с контекстной реконструкцией в длинных высказываниях, состоящих из стандартной лексики.

Для систематического анализа этих ошибок необходимо различать фонетические замены и истинные галлюцинации. Следуя концептуальной рамке, предложенной для генеративных моделей [13], заменой мы считаем ошибочный вывод, фонетически мотивированный акустическим входом (например, замена *грузинам* на *кукурузинам*). Галлюцинацией ASR мы считаем порождение правдоподобно звучащего текста или повторяющихся циклов, которым во входном сигнале вообще не соответствуют акустические события.

Примечательной особенностью Whisper является окно обработки по умолчанию в 30 с, обеспечивающее широкий контекст. Это способствует контекстной реконструкции в длинных высказываниях со стандартной лексикой, однако на

неразборчивых региональных фрагментах генеративный декодер склонен к галлюцинациям. Особенно показательны следующие расхождения:

- *наши спускались в алазанскую долину к грузинам*» – GigaAM: верно; Whisper: *к кукурузинам*;
- *вскоре после прихода Советской власти мы относились к кумуху*» – Whisper: верно; GigaAM: *к муху*;
- *около трехсот пятисот овец* – GigaAM: словесная запись верна; Whisper: 300500;
- *Отгоняли* – GigaAM: верно; Whisper: «*отгорели*» – фонетическая, но семантически отклоняющаяся замена.

Паттерны вставок также существенно различаются. На DaGRuS модель GigaAM-e2e часто вставляет дискурсивные биграммы вроде *да да* (483 вхождения) или *э э* (393 вхождения), по сути удваивая уже имеющиеся маркеры. Whisper, напротив, порождает повторяющиеся числовые последовательности (например, 1988 1988 встретилось почти 17000 раз) и шаблонные фразы вроде *продолжение следует*, характерные для зацикливания seq2seq-декодера. Отсутствие этих паттернов в GOLOS и RuDevices подтверждает, что они запускаются именно неразборчивым акустическим входом.

Эти паттерны вставок не случайны: они, по-видимому, восходят к артефактам обучающих данных моделей. В обучающей выборке GigaAM, вероятно, много разговорных данных с *да да*, тогда как Whisper мог обучаться на транскриптах с повторяющимися датами или техническими заглушками.

ОБСУЖДЕНИЕ

Разрыв в качестве между стандартным и дагестанским русскими (33–50%) существенно превышает разброс между контрастными архитектурами, оцененными на одном и том же корпусе. Даже наиболее устойчивая конфигурация (GigaAM v3-e2e) останавливается на WER 42.7% на дагестанском русском: одних лишь архитектурных оптимизаций недостаточно, чтобы решить задачу региональной языковой вариативности.

Полученные результаты указывают на ряд необходимых шагов по повышению устойчивости ASR. Во-первых, обучающие выборки необходимо расши-

рить тремя категориями, сейчас отсутствующими в основных корпусах: региональной фонетикой (акцентное произношение стандартных русских слов); региональной лексикой (этнонимы, топонимы и бытовые заимствования из кавказских и тюркских языков); дискурсивными маркерами спонтанной речи (сигналы обратной связи, хезитации и заполненные паузы). Во-вторых, предпочтительны сквозные (E2E) архитектуры: они дают снижение деградации на 7–8% по сравнению с каскадными системами на том же энкодере. Символьные или BPE-словари превосходят словарные (на уровне слов), поскольку позволяют модели частично транскрибировать незнакомые формы, а не удалять их целиком. В-третьих, генеративные (seq2seq) модели при обработке нестандартной речи требуют специальных мер против галлюцинаций – контроля температуры или штрафов за повторы. В-четвертых, узкодоменные модели (телефония, голосовые команды) критически уязвимы к речи местных вариантов; перед развертыванием в региональных контекстах, например в телефонной аналитике на Северном Кавказе, им необходима устойчивая доменная адаптация.

Отметим, что «приемлемое качество» зависит от приложения. Однако для большинства практических задач – голосовых помощников или телефонной аналитики – WER выше 15–20% делает систему практически непригодной [2]. Наблюдаемый здесь минимум WER в 43% явно выходит за границы практической полезности.

У работы есть ряд ограничений. Во-первых, DaGRuS представляет лишь один региональный вариант русского; полученные выводы не обязательно переносятся на другие контактно обусловленные разновидности (например, среднеазиатский, недагестанский кавказский или сибирский русский). Кроме того, GOLOS состоит преимущественно из прочитанной речи, тогда как DaGRuS содержит спонтанные диалоги. Часть наблюдаемой деградации может быть связана со стилистическими различиями, а не только с акцентом. RuDevices (стандартная спонтанная речь) дает промежуточные значения WER (9–19%), что помогает оценить стилистический вклад, однако строгое разделение этих факторов требует дополнительного контролируемого эксперимента. Во-вторых, наш конвейер нормализации намеренно удаляет нерусские сегменты (например, аварские или даргинские слова). Анализ того, как системы ASR опознают переключение кодов

или галлюцинируют на таких фрагментах, является важной задачей, однако мы сосредоточились на изолировании фонетического сдвига от стандартного к региональному русскому. В-третьих, проверка точного эффекта декодера путем сравнения моделей с одинаковыми энкодерами, но с разными декодерами (например, Whisper v3 и Whisper Turbo) выходила за рамки настоящей работы и представляет перспективное направление дальнейших исследований. Наконец, в отношении дискурсивных маркеров (Приложение Б) предварительный анализ показал, что флаг `remove_fillers` снижает базовый WER на всех корпусах, однако пропорциональный разрыв между стандартным и региональным русским остается тем же.

ЗАКЛЮЧЕНИЕ

Все десять протестированных конфигураций ASR демонстрировали значительное падение качества на дагестанском русском: WER возрастает на 33–50% по сравнению со стандартным русским. Ни одна модель не достигла значения WER ниже 43%. Различия статистически значимы при большом размере эффекта ($r = 0.52 – 0.76$).

Ошибки моделей являлется неслучайными; выявлены три систематических механизма сбоев: «русификация» (замена дагестанских форм фонетически близкими стандартными русскими словами), удаление дискурсивных маркеров и внесловарной лексики, а также генеративные галлюцинации на неразборчивых фрагментах. Примечательно, что 58% высказываний DaGRuS попадают в «слепую зону», где все модели дают WER выше 50%.

Архитектура декодера определяет тип ошибки (RNN-T склонен к удалениям, CTC – к заменам, seq2seq – к галлюцинациям), но не снижает общий масштаб проблемы. Сквозное обучение уменьшает деградацию на 7–8% по сравнению с каскадным подходом, а узкодоменные модели оказываются наиболее уязвимыми (например, +50% у T-One). Однако главным фактором выступает отсутствие региональных данных в обучающих выборках: разрыв между корпусами в несколько раз превышает разрыв между архитектурами.

Чтобы обеспечить инклюзивность цифровых сервисов, обучающие данные ASR необходимо целенаправленно расширять региональными вариантами рус-

ского языка, маркерами спонтанной речи и методами доменной адаптации к региональной фонетике. Корпус DaGRuS может служить бенчмарком для оценки устойчивости ASR-систем к региональной языковой вариативности.

СПИСОК ЛИТЕРАТУРЫ

1. *Добрушина Н.Р.* Многоязычие в Дагестане, или зачем человеку три языка // Социологический журнал. 2007. № 1. С. 103–127.

2. *Markl N.* Language variation and algorithmic bias: Understanding algorithmic biases in British English automatic speech recognition // Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. 2022. P. 521–534.

3. *Feng S., Kudina O., Halpern B.M., Scharenborg O.* Quantifying bias in automatic speech recognition // arXiv:2103.15122. https://doi.org/10.48550/arXiv.2103.15122

4. *Conneau A., Baevski A., Collobert R., Mohamed A., Auli M.* Unsupervised cross-lingual representation learning for speech recognition // Proceedings of Interspeech 2021. 2021. P. 2426–2430.

5. *Karpov N., Denisenko A., Minkin F.* Golos: Russian dataset for speech research // Proceedings of Interspeech 2021. 2021. P. 1419–1423.

6. *Dobrushina N., Daniel M., von Waldenfels R., Maisak T., Panova A.* Corpus of Russian spoken in Daghestan. 2018. URL: http://www.parasolcorpus.org/dagrus/ (дата обращения: 24.02.2026).

7. *Hinsvark A., Delworth N., Del Rio M. et al.* Accented speech recognition: A survey. 2021 // arXiv:2104.10747. 2021. https://doi.org/10.48550/arXiv.2104.10747

8. *Ramesh G.V. et al.* Federated representation learning for automatic speech recognition // Proc. 3rd Symposium on Security and Privacy in Speech Communication. 2023. P. 29–33. https://doi.org/10.21437/SPSC.2023-5

9. *Mann H.B., Whitney D.R.* On a test of whether one of two random variables is stochastically larger than the other // The Annals of Mathematical Statistics. 1947. Vol. 18, No. 1. P. 50–60.

10. *Efron B.* Bootstrap methods: Another look at the jackknife // The Annals of Statistics. 1979. Vol. 7, No. 1. P. 1–26.

11. *Korobov M.* Morphological analyzer and generator for Russian and Ukrainian languages // Analysis of Images, Social Networks and Texts. Communications in Computer and Information Science. Vol. 542. Cham: Springer, 2015. P. 320–332.

12. *Naccarato C., Panova A., Stoynova N.* Word-order variation in a contact setting: A corpus-based investigation of Russian spoken in Daghestan // Language Variation and Change. 2021. Vol. 33, No. 3. P. 387–411.

13. *Guerreiro N.M., Voita E., Martins A.F.T.* Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation // Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. 2023.

ПРИЛОЖЕНИЕ А. КОНВЕЙЕРЫ НОРМАЛИЗАЦИИ ТЕКСТА

DagrusNormalizer. Эталонный текст DaGRuS содержит полевую разметку, которую необходимо разрешить перед оценкой. Нормализатор работает следующим образом. (1) Маркеры технических событий, заключенные в знаки решетки (#laughs#, #speaks another language#), удаляются. (2) Аннотированные пары «слово{каноническая*}» разрешаются к каноническому варианту (например, «щас{сейчас*} → сейчас»). Аннотации с маркером «%» (нестандартные формы без строгого канонического эквивалента) снимаются, сохраняется только поверхностная форма. (3) Отдельные маркеры неразборчивости «нрзб» удаляются. (4) Непомеченные технические фразы, случайно оставленные аннотаторами без знаков решетки (например, говорит на другом языке), удаляются точным сопоставлением со закрытым списком.

Для гипотез ASR, не содержащих разметки, нормализатор применяет отображение разговорных форм на канонические: словарь из 30 однозначных пар («щас → сейчас», «че → что», «тыща → тысяча») обеспечивает, что и произнесенная, и каноническая формы считаются эквивалентными при выравнивании. Например, если в эталоне стоит сейчас (разрешенное из разметки «щас{сейчас*}»), а модель выдает «щас», отображение преобразует гипотезу в сейчас, и сопоставление считается верным.

GolosNormalizer. В эталонных текстах GOLOS все числительные записаны словами (например, двадцать пять), тогда как некоторые модели ASR, в частности Whisper, выводят арабские цифры (25). Без нормализации каждое такое числительное засчитывается как ошибка. GolosNormalizer переводит цифры в гипотезе в русские слова с помощью библиотеки num2words, после чего применяется стандартная очистка.

SimpleNormalizer. Для RuDevices и других корпусов без особых конвенций разметки базовый нормализатор применяет только общий конвейер: нижний регистр, ё → е, удаление пунктуации и схлопывание пробелов.

Важное проектное решение касается дискурсивных заполнителей: «угу», «ага», «мм», «э». По умолчанию мы их не удаляем: если модель удаляет заполнитель, присутствующий в эталоне, это регистрируется как ошибка удаления, что лингвистически информативно. Для абляционных экспериментов предусмотрен необязательный флаг `remove_fillers`.

ПРИЛОЖЕНИЕ Б. ПРОФИЛЬ ОШИБОК НА ДИСКУРСИВНЫХ МАРКЕРАХ И ПРОБЛЕМЫ РАЗМЕТКИ

Показатели WER для дискурсивных маркеров в DaGRuS несут неустранимую измерительную неопределенность. Разметка сигналов обратной связи в корпусе непоследовательна: одни аннотаторы систематически транскрибировали бэк-канальные сигналы (*угу*, *мм*), другие опускали их или включали в соседние реплики. Если маркер отсутствует в эталоне, но есть в аудио и порожден моделью, метрика фиксирует ложную вставку; напротив, если маркер есть в эталоне, но едва ли был отдельным речевым актом, его удаление завышает счетчик удалений. Граница между заполненной паузой и сигналом обратной связи часто неоднозначна даже для слушателя-человека.

Несмотря на эти проблемы разметки, для ряда биграмм наблюдается явная тенденция к систематическим удалениям. Наиболее часто удаляемые биграммы показывают склонность моделей отбрасывать повторенные дискурсивные заполнители:

- *да да* – 366 удалений у Whisper, 297 у GigaAM;
- *м м* – 266/234;
- *угу угу* – 141/95.

Таким образом, приводимые доли ошибок на дискурсивных маркерах представляют верхнюю оценку сбоев моделей и нижнюю оценку истинной трудности явления.

AUTOMATIC SPEECH RECOGNITION: FACTORS REDUCING ACCURACY WHEN WORKING WITH A REGIONAL VARIANT OF THE RUSSIAN LANGUAGE (USING MATERIAL FROM DAGESTAN)

A. I. Katsnelson[1] [0009-0009-2808-3243], O. N. Lyashevskaya[2] [0000-0001-8374-423X]

[1, 2]*HSE University, Moscow, Russia*

[1]akatsnelson@hse.ru, [2]olesar@yandex.ru

Abstract

We evaluate the quality of automatic speech recognition (ASR) for Standard Russian and its Dagestani regional variety – a contact idiom formed in the multilingual environment of Dagestan. We tested ten off-the-shelf ASR configurations in a zero-shot setting (without fine-tuning) representing three architectural approaches (Conformer+RNN-T, Conformer+CTC, and Encoder-Decoder Transformer) on datasets of Standard Russian (GOLOS, RuDevices; approx. 20K and 90K utterances respectively) and Dagestani Russian (DaGRuS; approx. 40K utterances). All models demonstrate a word error rate (WER) degradation of 33%–50% when evaluated on Dagestani Russian, highlighting a severe domain shift problem. Error structure analysis reveals three primary degradation mechanisms: “Russification” (systematic replacement of Dagestani forms with phonetically similar Standard Russian words); deletion of discourse markers and out-of-vocabulary (OOV) terms such as ethnonyms and toponyms; and generative model hallucinations on incomprehensible audio segments. We found that end-to-end architectures suffer less degradation than cascaded ones (33%–35% vs. 40%–42%), yet no model achieves acceptable performance (minimum WER ≈ 43%). Finally, we formulate recommendations for designing ASR systems robust to the regional variability of the Russian language.

Keywords: *automatic speech recognition, ASR, Dagestani Russian, regiolect, speech of local varieties, WER, language variation, domain shift, Conformer, Whisper, GigaAM, hallucinations.*

REFERENCES

1. *Dobrushina N.R.* Mnogoyazychie v Dagestane, ili zachem cheloveku tri yazyka // Sotsiologicheskii zhurnal. 2007. No. 1. P. 103–127.

2. *Markl N.* Language variation and algorithmic bias: Understanding algorithmic biases in British English automatic speech recognition // Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. 2022. P. 521–534.

3. *Feng S., Kudina O., Halpern B.M., Scharenborg O.* Quantifying bias in automatic speech recognition // arXiv:2103.15122. https://doi.org/10.48550/arXiv.2103

4. *Conneau A., Baevski A., Collobert R., Mohamed A., Auli M.* Unsupervised cross-lingual representation learning for speech recognition // Proceedings of Interspeech 2021. 2021. P. 2426–2430.

5. *Karpov N., Denisenko A., Minkin F.* Golos: Russian dataset for speech research // Proceedings of Interspeech 2021. 2021. P. 1419–1423.

6. *Dobrushina N., Daniel M., von Waldenfels R., Maisak T., Panova A.* Corpus of Russian spoken in Daghestan. 2018. URL: http://www.parasolcorpus.org/dagrus/ (accessed: 24.02.2026).

7. *Hinsvark A. et al.* Accented speech recognition: A survey // arXiv:2104.10747. 2021. https://doi.org/10.48550/arXiv.2104.10747

8. *Ramesh G.V. et al.* Federated representation learning for automatic speech recognition // Proc. 3rd Symposium on Security and Privacy in Speech Communication. 2023. P. 29–33. https://doi.org/10.21437/SPSC.2023-5

9. *Mann H.B., Whitney D.R.* On a test of whether one of two random variables is stochastically larger than the other // The Annals of Mathematical Statistics. 1947. Vol. 18, No. 1. P. 50–60.

10. *Efron B.* Bootstrap methods: Another look at the jackknife // The Annals of Statistics. 1979. Vol. 7, No. 1. P. 1–26.

11. *Korobov M.* Morphological analyzer and generator for Russian and Ukrainian languages // Analysis of Images, Social Networks and Texts. Communications in Computer and Information Science. Vol. 542. Cham: Springer, 2015. P. 320–332.

12. *Naccarato C., Panova A., Stoynova N.* Word-order variation in a contact setting: A corpus-based investigation of Russian spoken in Daghestan // Language Variation and Change. 2021. Vol. 33, No. 3. P. 387–411.

13. *Guerreiro N.M., Voita E., Martins A.F.T.* Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation // Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. 2023.

СВЕДЕНИЯ ОБ АВТОРАХ



КАЦНЕЛЬСОН Артем Игоревич – ML-разработчик, преподаватель Департамента компьютерной инженерии Национального исследовательского университета «Высшая школа экономики» (МИЭМ им. А. Н. Тихонова). Область научных интересов: прикладные системы искусственного интеллекта, мультиагентные системы, RAG-архитектуры, речевая аналитика, машинное зрение, прогнозирование временных рядов.

Artem Igorevich KATSNELSON is a ML developer, lecturer at the Department of Computer Engineering at the National Research University Higher School of Economics (Tikhonov MIEM). He has more than 5 years of experience in the development and implementation of commercial machine learning and AI systems. Research interests: applied artificial intelligence systems, multi-agent systems, RAG architectures, speech analytics, machine vision, time series forecasting. Author of more than 10 scientific publications. Two-time winner of the HSE Electronics Olympiad.

email: akatsnelson@hse.ru
ORCID: https://orcid.org/0009-0009-2808-3243



ЛЯШЕВСКАЯ Ольга Николаевна – профессор Школы лингвистики факультета гуманитарных наук Национального исследовательского университета «Высшая школа экономики», ведущий научный сотрудник Института русского языка им. В. В. Виноградова РАН, кандидат филологических наук. Область научных интересов: компьютерная и корпусная лингвистика, лексическая и грамматическая семантика, грамматика конструкций, компьютерная лексикография. Активно участвует в развитии Национального корпуса русского языка.

Olga Nikolayevna LYASHEVSKAYA is Professor at the School of Linguistics, Faculty of Humanities, National Research University Higher School of Economics, leading research fellow at Vinogradov Russian Language Institute of the Russian Academy of Sciences, Candidate of Philological Sciences.

Her research interests include computational and corpus linguistics, lexical and grammatical semantics, construction grammar, and computational lexicography. She actively participates in the development of the National Corpus of the Russian Language.

email: olesar@yandex.ru

ORCID: https://orcid.org/0000-0001-8374-423X

Материал поступил в редакцию 19 августа 2026 года

УДК 004.81

ПРАГМАТИЧЕСКИЕ ГРАНИЦЫ ПРИВЕТСТВИЙ И ПРОЩАНИЙ: ОЦЕНКА LLM НА МАТЕРИАЛЕ МУЛЬТИМЕДИЙНОГО КОРПУСА РЕЧЕВОГО ЭТИКЕТА

К. С. Клокова[1] [0000-0002-6650-1946], М. А. Кронгауз[2] [0000-0003-2956-0565], В. А. Шульгинов[3] [0000-0003-2217-3677], Т. А. Юдина[4] [0000-0001-6631-2128]

[1, 3, 4]*Московский физико-технический институт, г. Долгопрудный, Россия*

[2]*Российский государственный гуманитарный университет, г. Москва, Россия*

[1]klokova.ks@mipt.ru, [2]mkronhaus@yandex.ru, [3]shulginov.va@mipt.ru, [4]yudina.tatiana.a@mipt.ru

Аннотация

Исследована способность больших языковых моделей выявлять в русскоязычных диалогах последовательности приветствий и прощаний и определять их точные прагматические границы. Рассмотрены YandexGPT 5.1 Pro, GigaChat 2 Max и GPT-5 с целью: классификации диалогов по наличию указанных последовательностей и извлечении отдельных высказываний с учетом их интенциональных границ. На материале 216-ти диалогов из «Мультимедийного корпуса речевого этикета русского языка» проведен эксперимент, позволивший учесть широкий коммуникативный и социальный контекст взаимодействия. GPT-5 продемонстрировал наиболее высокое качество классификации на уровне диалога, тогда как YandexGPT и GigaChat достигли почти идеальной точности за счет консервативных стратегий. При извлечении точных границ качество всех моделей заметно снижается. Анализ ошибок показал, что модели систематически включают лишние контекстуальные невербальные действия и пропускают нестандартные маркеры: жесты, формы обращения, дискурсивные маркеры и импликатуры, указывающие на будущий контакт. Представлены актуализированное описание корпуса, его веб-интерфейс, а также возможности формирования диагностических подкорпусов и бенчмарков. Полученные результаты подчеркивают необходимость более точного учета структуры дискурса и контекстуальной вариативности повседневного общения.

© К. С. Клокова, М. А. Кронгауз, В. А. Шульгинов, Т. А. Юдина, 2026.
Данная статья распространяется на условиях международной лицензии Creative Commons License Attribution 4.0 International (CC BY 4.0).

***Ключевые слова**: большие языковые модели, мультимедийный корпус речевого этикета, приветствие, прощание, речевой этикет, прагматические границы, формулы вежливости, анализ диалога.*

ВВЕДЕНИЕ

Автоматическое распознавание структуры диалога, прежде всего последовательностей, устанавливающих и завершающих коммуникативный контакт, важно как для прагматической теории, так и для создания диалоговых агентов, способных органично включаться в общение. Кроме того, автоматизация анализа таких коммуникативных структур открывает новые возможности для текстовой аналитики и предобработки данных, повышая точность разметки в прикладных задачах. С лингвистической точки зрения это позволяет проводить систематические корпусные исследования того, как базовые паттерны коммуникативного взаимодействия варьируются в зависимости от социального контекста, отношений между участниками и ситуативных параметров – центральных факторов в исследованиях вежливости [1, 2] и конверсационном анализе [3].

Цель настоящего исследования – оценить способность современных больших языковых моделей (англ. Large Language Models; далее – LLM или модели) идентифицировать и интерпретировать на материале русского языка последовательности реплик, инициирующих и завершающих контакт. Мы выходим за рамки структурного анализа, предполагая, что интенциональные границы необязательно совпадают с формальным завершением высказывания.

Поставленная задача относится к области оценки прагматических способностей LLM. При прагматической оценке модели прежде всего тестируются на способность понимать косвенные речевые акты [4, 5], распознавать разговорные импликатуры [6–8] и интерпретировать образный язык [9, 10]. Тесно связанной областью является понимание диалога, в рамках которого LLM применяются для таких задач, как классификация диалоговых актов [11], в том числе в мультимодальных и многосторонних диалогах [12].

Для достижения цели сформулирована двухкомпонентная задача: (1) идентификация и извлечение последовательностей приветствий и прощаний из транскриптов диалогов и (2) классификация диалогов по наличию этих последовательностей. Опираясь на определения открывающих и закрывающих последовательностей, принятые в конверсационном анализе [13], мы рассматриваем

приветствия и прощания как многоходовые, многоактные последовательности, которые могут включать обращения, дискурсивные маркеры, фатические вопросы и невербальные действия, при условии их совпадения с вербальными элементами. Набор данных составили 216 диалогов из «Мультимедийного корпуса речевого этикета русского языка» (далее – корпус) [14], размеченных с указанием границ последовательностей и меток на уровне диалога. Проанализированы три модели – YandexGPT 5.1 Pro, GigaChat 2 Max и GPT-5 – с использованием конвейера на основе промптинга с мажоритарным голосованием и многоуровневой системой метрик.

Результаты выявили существенный разрыв между классификацией на уровне диалога и извлечением последовательностей. Анализ ошибок показал, что модели испытывают трудности с определением границ, а не с распознаванием содержания: модели выделяют материал, относящийся к приветствиям и прощаниям, однако нередко включают лишние токены или пропускают пограничные маркеры. Эти результаты подтверждают, что понимание прагматических последовательностей остаётся нетривиальной задачей для современных LLM, даже несмотря на их достаточно высокое качество в задачах с укрупнённой гранулярностью.

Настоящая статья представляет собой существенно расширенную и переработанную версию доклада, опубликованного в трудах конференции «Диалог – 2026» [21]. В этой версии существенно переработан раздел о «Мультимедийном корпусе речевого этикета»: приведены актуализированные сведения о его составе и разметке, описаны веб-интерфейс, параметры поиска и раздел с бенчмарками. Кроме того, добавлено развернутое обсуждение корпуса как инструмента для диагностических экспериментов, а также расширены разделы, посвященные интерпретации полученных результатов.

Статья имеет следующую структуру. В разд. 1 описаны «Мультимедийный корпус речевого этикета русского языка» и недавние обновления его схемы разметки. В разд. 2 определены задачи идентификации последовательностей и классификации диалогов, а также представлен набор данных. Разд. 3 посвящен экспериментальной установке: дизайну промптов, стратегиям голосования и метрикам оценки. В разд. 4 приведены результаты как по классификации, так и по извлечению. Разд. 5 включает обсуждение ограничений и перспектив дальнейшей работы.

1. ОПИСАНИЕ КОРПУСА И ОБНОВЛЕНИЯ

1.1. Обзор корпуса

«Мультимедийный корпус речевого этикета русского языка» [14] включает в себя фрагменты естественных диалогов из российских фильмов и телесериалов. Каждый фрагмент сопровождается видеозаписью, транскриптом и многоуровневой аннотацией. Основной единицей разметки является фрейм [15] – непосредственно наблюдаемая экстралингвистическая информация о коммуникативных ситуациях. В рамках фреймов выделяются стандартные этикетные ситуации, связанные с интенциями участников коммуникации: приветствия, прощания, извинения, выражения благодарности и просьбы.

Разметка включает в себя: (1) пространственно-временные характеристики; (2) типы фреймов; (3) социальные и реляционные характеристики собеседников; (4) мультимодальные маркеры; (5) формы обращения. Один диалог может содержать несколько этикетных ситуаций и, соответственно, получать несколько меток.

2.2. Недавние обновления и дополнения

Первоначально корпус был ориентирован прежде всего на современные коммуникативные практики и включал медиаматериалы, отражающие дискурс нулевых годов XX в. [14]. В дальнейшем его временные рамки были значительно расширены – в настоящее время корпус охватывает период с 1956 по 2025 годы (рис. 1). Объем данных составляет 95 фильмов и телесериалов, 736 диалогов и 1028 размеченных этикетных ситуаций общей продолжительностью 491.6 минуты.

В схему включены ситуации просьбы, введена возможность разметки различных этикетных ситуаций в рамках одного диалога. Характеристика невербального поведения дополнена описанием действий (например, «вошел в комнату»), способных вызывать этикетную реакцию собеседника. Закрытые перечни профессий и отношений заменены открытыми реестрами, которые отражают фактически представленные в материале категории и допускают дальнейшее пополнение. Важными этапами развития ресурса стали также запуск публичного портала и создание раздела для публикации корпусных бенчмарков.

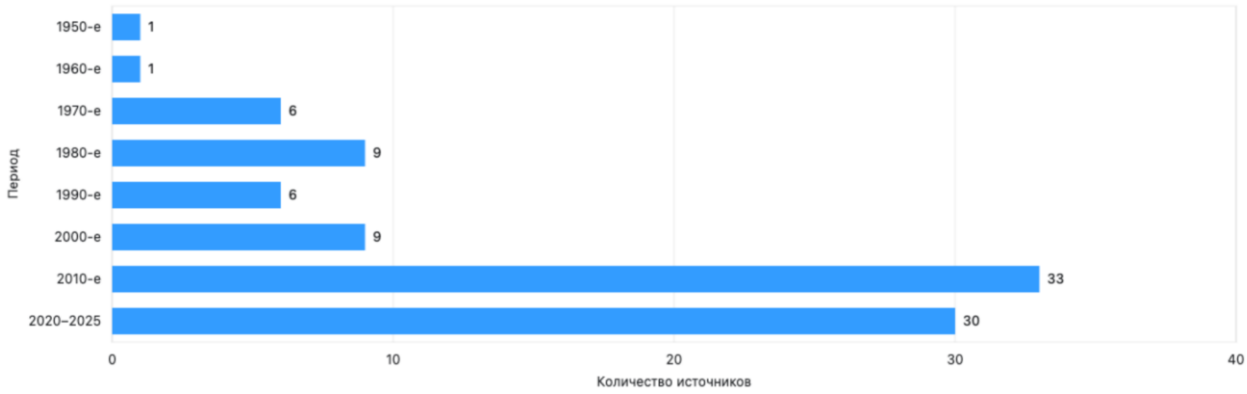


Рис. 1. Распределение количества фильмов/сериалов в корпусе по десятилетиям.

1.3 Веб-интерфейс и возможности поиска

Интерфейс включает в себя главную страницу с описанием проекта, поиск по корпусу, краткую инструкцию, сводную статистику и раздел с исследованием коммуникативных компетенций LLM. Поиск поддерживает два сценария: полнотекстовый запрос с последующим уточнением результатов и отбор только по параметрам, без ввода текста. Внутри одной категории можно выбирать несколько значений (см. рис. 2).

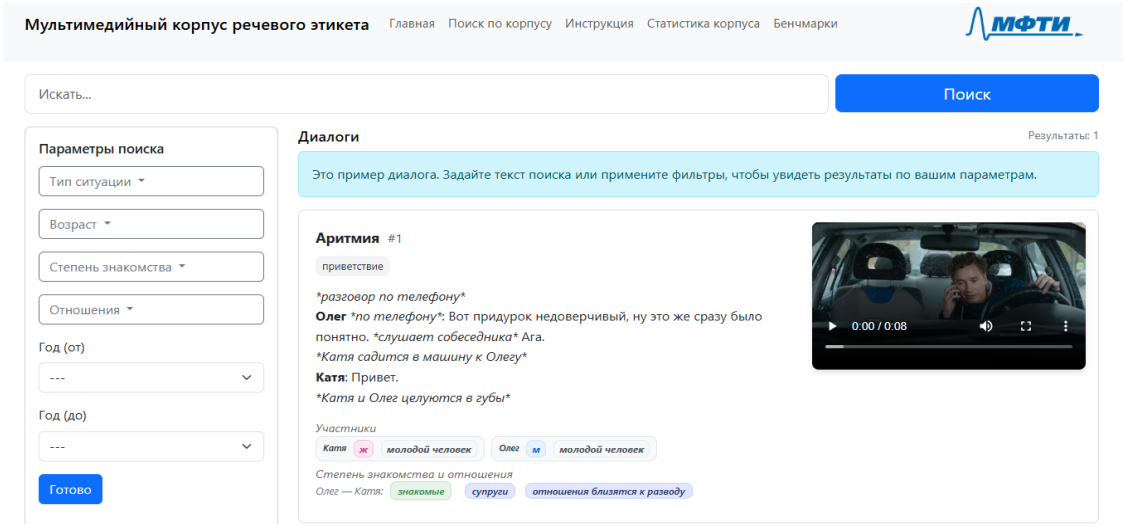


Рис. 2. Страница поиска по корпусу.

Панель поиска включает четыре категориальных фильтра и отдельный выбор временного диапазона. Их состав представлен в табл. 1.

Табл. 1. Параметры поиска.

Параметр	Доступные значения
Тип ситуации	Приветствие, прощание, извинение, благодарность, просьба, знакомство
Возраст	Семь возрастных категорий
Степень знакомства	Незнакомцы, малознакомые, знакомые
Отношения	Открытый детализированный список (например, друзья, супруги, соседи, коллеги, начальник/подчиненный)
Год	Диапазон от 1956 до 2025 г.

Карточка выдачи объединяет полный транскрипт с обозначением паравербальных и невербальных компонентов, соответствующий видеофрагмент, типы этикетных ситуаций и сведения об участниках. Структура корпуса позволяет не просто констатировать ошибки моделей, но и строить эксперименты, учитывающие социальный контекст и характеристики собеседников. Возможность обращаться непосредственно к фрагменту видео также открывает перспективы для более широкого анализа социокультурных и коммуникативных возможностей моделей.

1.4. «Бенчмарки на материале корпуса»

Специальный раздел веб-платформы предназначен для систематизации задач и результатов оценки LLM на материале корпуса. Рассматриваемые в настоящей работе классификация диалогов и извлечение границ последовательностей приветствий и прощаний образуют пилотный сценарий такого тестирования. Инфраструктура нашего проекта может использоваться для построения бенчмарков по другим прагматическим ситуациям, а также для диагностических экспериментов на подкорпусах, сформированных по возрасту, степени знакомства и типу отношений между участниками.

2. ОПРЕДЕЛЕНИЕ ЗАДАЧИ И НАБОР ДАННЫХ

2.1. Постановка задачи

В рамках исследования решались две задачи: (1) идентификация и извлечение последовательностей приветствий и прощаний и (2) классификация диалогов по наличию этих последовательностей. Это позволило провести корпусные исследования паттернов взаимодействия в различных социальных контекстах и способствовало разработке социально ориентированных диалоговых систем.

2.1.1. Идентификация и извлечение последовательностей

Последовательности приветствий и прощаний представляют собой конвенционализированные явления речевого этикета, выполняющие важные фатические функции в структуре разговора. В соответствии с традицией конверсационного анализа приветствия функционируют как открывающие последовательности, устанавливающие коммуникативный контакт в начале взаимодействия, тогда как прощания служат закрывающими последовательностями, отмечающими завершение контакта [13, 16, 17]. Для целей настоящей работы приняты следующие операциональные определения:

- последовательности приветствия – это речевые акты, устанавливающие коммуникативный контакт в начале диалога и включающие конвенциональные формулы (*привет, здравствуйте*), обращения, фатические вопросы и ответные реплики;
- последовательности прощания – это речевые акты, завершающие контакт и включающие формулы прощания (*до свидания, пока*), пожелания (*удачи*), указания на будущий контакт (*созвонимся*) и ответные реплики.

Последовательности могут охватывать несколько речевых актов. Извлечение предполагает выделение полных коммуникативных единиц с сопутствующими обращениями, дискурсивными маркерами и многоактными реализациями. Невербальные действия (рукопожатия, взмахи рукой) включаются при условии их совпадения с вербальными элементами.

2.1.2. Классификация диалогов

В зависимости от наличия или отсутствия выявленных последовательностей каждый диалог относится к одной из четырех взаимоисключающих категорий: только приветствие, только прощание, приветствие + прощание или ни одного.

2.2. Набор данных

Для настоящего исследования из «Мультимедийного корпуса речевого этикета русского языка» было отобрано 216 диалогов с метками стандартных ситуаций (приветствие и/или прощание). Итоговый набор данных включает 54 диалога только с приветствием, 53 только с прощанием, 17 с обоими и 92 без таковых.

В исследовании участвовали три специалиста с опытом разметки прагматических явлений и анализа речевых актов. Два разметчика независимо определяли границы последовательностей в соответствии с критериями из п. 2.1.1. Разногласия разрешались в ходе совместных обсуждений с третьим разметчиком-экспертом, который выносил решения по спорным случаям и устанавливал окончательный эталон разметки. Хотя формальные метрики согласия (каппа Коэна, альфа Криппендорфа) не рассчитывались ввиду предварительного характера исследования, процедура совместного разрешения разногласий обеспечила согласованность итогового набора данных.

Общий объем набора данных составляет 9868 токенов; среднее количество токенов на диалог – 46, максимальное – 425. Среднее количество реплик на диалог – 4, максимальное – 35. Подробная статистика по последовательностям приветствий и прощаний приведена в табл. 2. Относительно небольшая средняя длина диалога отражает структуру корпуса: диалоги фиксируют конкретные стандартные коммуникативные ситуации и интерактивные последовательности, а не развернутые беседы. Примеры размеченных диалогов представлены в Приложении А.

Табл. 2. Распределение количества диалогов, токенов и невербальных маркеров по типам последовательностей.

Тип последовательности	Количество диалогов	Среднее/максимальное число токенов на последовательность	Содержат невербальные маркеры, %
Приветствия	131	3/13	22.14
Прощания	118	4/11	18.64

Подробная инструкция по разметке, включая правила определения границ и обработку сложных случаев, формализована в системном промпте, описанном в п. 3.1.

3. ЭКСПЕРИМЕНТАЛЬНАЯ УСТАНОВКА

Для решения задач идентификации и классификации последовательностей приветствий и прощаний был разработан конвейер последовательной обработки и протестированы три модели: YandexGPT 5.1 Pro, GigaChat 2 Max и GPT-5. Выбор моделей обусловлен тем, что они представляют (1) современные русскоязычные LLM, оптимизированные для понимания русскоязычного дискурса, и (2) одну из ведущих мультиязычных моделей в качестве эталонного разметчика. Каждый диалог обрабатывался каждой моделью трижды. Результаты агрегировались с помощью мажоритарного голосования (п. 3.2), после чего сопоставлялись с эталонной разметкой с использованием метрик, описанных в п. 3.3.

3.1. Дизайн промпта и генерация

Для управления поведением моделей использовался структурированный подход с несколькими примерами (few-shot prompting) [18]. Промпт включал следующие компоненты:

- назначение роли;
- четкие определения последовательностей приветствий и прощаний (п. 3.1.1);
- руководство по определению границ с разбором сложных случаев (невербальные действия, дискурсивные маркеры, многоходовые последовательности);

- алгоритм извлечения последовательностей;
- три контрастных размеченных диалога, иллюстрирующих (1) последовательности приветствия и прощания; (2) отсутствие этикетных актов; (3) только приветствия и невербальные маркеры;
- структурированный формат вывода JSON;
- указания на сложные явления (диминутивы, имплицитные закрытия, телефонные открытия).

Для обеспечения корректного сравнения для всех трех моделей использовалась одинаковая структура промпта. Параметры генерации: температура 0.3; по 3 независимых запуска на диалог; остальные параметры использовались со значениями по умолчанию в соответствии с API провайдера.

3.2. Стратегии голосования

Для учета стохастической природы генерации моделей был реализован механизм мажоритарного голосования на основе точного совпадения, агрегирующий результаты трех независимых запусков для каждого диалога.

Голосование по классификации типа диалога: если одна метка встречалась более чем в одном запуске, она выбиралась в качестве итоговой классификации. В случаях, когда все три запуска давали разные метки, одна из них выбиралась случайным образом.

Голосование по извлечению последовательностей: каждая уникальная извлеченная последовательность (текст диалога + атрибуция говорящего) подсчитывалась по всем трем запускам. Последовательности, встретившиеся в двух и более запусках, включались в итоговый результат с достоверностью, пропорциональной числу голосов.

3.3. Метрики оценки

Для оценки качества классификации использовались стандартные метрики: точность, полнота и F1. Для более полной оценки извлечения была разработана многоуровневая метрика.

Уровень 1: точное совпадение (Exact Matching) основано на сравнении множеств и учитывает только идеальные совпадения. Для истинных последовательностей $T = \{t_1, t_2, \dots, t_n\}$ и предсказанных последовательностей $P =$

$\{p_1, p_2, \ldots, p_n\}$ вычисляем

$$TP = |T \cap P|, \quad FP = |P \setminus T|, \quad FN = |T \setminus P|.$$

Уровень 2: нечеткое совпадение (Fuzzy Matching) основано на сходстве Джаро – Винклера [19, 20] с жадным сопоставлением в режиме «один к одному» и адаптивными порогами (0,88 для строк менее 50 символов; 0,92 для 50–100 символов; 0,95 для более 100 символов) для учета различий в длинах строк. Данный уровень допускает незначительные вариации (опечатки, расхождения в пунктуации, проблемы с нормализацией пробелов).

Уровень 3: уровень токенов (Token–Level) оценивает понимание содержания: совпадение по набору токенов засчитывается как частичный результат даже при наличии ошибок в определении границ высказывания. Метрики рассчитывались на множествах токенов.

4. РЕЗУЛЬТАТЫ

4.1. Классификация типа диалога

В табл. 3 представлены результаты классификации типов диалога для трех моделей (YandexGPT 5.1 Pro, GigaChat 2 Max и GPT-5), сопоставленных по способности распознавать диалоги, содержащие последовательности приветствия, прощания или обеих одновременно. Качество оценивалось с помощью показателей точности (precision), полноты (recall) и F1-меры для каждого типа последовательности и в целом.

Табл. 3. Эффективность классификации типа диалога.

Модель	Тип	Точность	Полнота	F1
YandexGPT 5.1 Pro	Приветствие	1.00	0.62	0.77
	Прощание	1.00	0.57	0.73
	В целом	**1.00**	0.60	0.75
GigaChat 2 Max	Приветствие	1.00	0.63	0.78
	Прощание	0.98	0.60	0.74
	В целом	0.99	0.62	0.76
GPT-5	Приветствие	0.87	0.87	0.87
	Прощание	0.93	0.89	0.91
	В целом	0.90	0.88	0.89

GPT-5 достиг наилучшего общего результата с F1-мерой 0.89, за ним расположились GigaChat 2 Max и YandexGPT 5.1 Pro. У всех моделей точность систематически выше полноты, причем наибольший дисбаланс наблюдается у YandexGPT и GigaChat. Это свидетельствует о том, что эти модели придерживаются консервативных стратегий: когда они идентифицируют последовательности приветствий и прощаний, они почти всегда делают это верно, однако пропускают значительную долю истинных примеров (37–43% ложноотрицательных результатов). Более низкая точность GPT-5 отражает иной компромисс: толерантность к некоторому количеству ложноположительных результатов в обмен на существенно более высокую полноту, что дает лучший общий F1-показатель.

Все модели демонстрируют высокую точность при определении приветствий, что указывает на редкость ложноположительных ошибок. Однако полнота существенно варьируется: 0.62 у YandexGPT, 0.63 у GigaChat и 0.87 у GPT-5. Уровень ложноотрицательных результатов около 38% у русскоязычных моделей свидетельствует об их трудностях с распознаванием нестандартных реализаций приветствий. При обнаружении прощаний наблюдается схожая, но несколько более слабая картина. YandexGPT сохраняет идеальную точность, однако полнота составляет лишь 0.57; GigaChat показывает точность 0.98 при полноте 0.60; GPT-5 вновь демонстрирует более сбалансированные результаты – точность 0.93 и полноту 0.89. Возможные причины низких значений полноты рассматриваются в п. 4.3.

4.2. Извлечение последовательностей приветствий/прощаний

В табл. 4 представлены результаты извлечения последовательностей по трем уровням оценки (п. 3.3.). Качество существенно ниже, чем при классификации (п. 4.1.), что отражает повышенную сложность точного определения границ последовательностей. Примеры извлеченных последовательностей приветствий и прощаний приведены в Приложении А.

Табл. 4. Результаты извлечения последовательностей по типам (T1: точное совпадение, T2: нечеткое совпадение, T3: уровень токенов).

Модель	Тип последовательностити	T1 (P/R/F1)	T2 (P/R/F1)	T3 (P/R/F1)
YandexGPT				

5.1 Pro	Приветствие Прощание	0.43/0.38/0.39 **0.40**/0.36/0.37	0.47/0.41/0.43 0.42/0.38/0.40	0.52/0.58/0.53 0.51/0.52/0.51
GigaChat 2 Max	Приветствие Прощание	0.34/0.34/0.33 0.36/0.35/0.35	0.36/0.35/0.36 0.39/0.38/0.38	0.48/0.59/0.52 0.49/0.55/0.51
GPT-5	Приветствие Прощание	**0.48/0.49/0.47** **0.40/0.40/0.39**	**0.52/0.53/0.52** **0.50/0.49/0.48**	**0.62/0.76/0.66** **0.64/0.78/0.68**

GPT-5 стабильно превосходит другие модели на всех уровнях. Все модели демонстрируют последовательный рост показателей от T1 к T3, при этом F1-мера увеличивается на 10–27%. Это указывает на то, что основная проблема состоит в определении границ, а не в распознавании содержания. Незначительный прирост от T1 к T2 говорит о том, что большинство ошибок представляет собой существенные различия в границах, а не мелкие вариации форматирования. Модели, как правило, извлекают фрагменты общения, релевантные для приветствий и прощаний, но часто включают лишние токены или пропускают пограничные маркеры, что ведет к неточным совпадениям на уровне последовательностей при высоком содержательном пересечении.

На уровне T1 приветствия показывают несколько более высокий F1 у YandexGPT и GPT-5, тогда как у GigaChat разница минимальна. Однако на уровне T3 разрыв сокращается или изменяет направление. Это позволяет предположить, что приветствия имеют более четкие границы, но по сложности содержания сопоставимы с прощаниями.

4.3. Анализ ошибок

Существенный разрыв между качеством классификации на уровне диалога и точностью извлечения последовательностей, отмеченный в пп. 4.1., 4.2, требует более детального рассмотрения паттернов ошибок каждой модели. Анализ выявил систематические трудности в идентификации нестандартных реализаций вежливости, при этом заметно различаются консервативные стратегии YandexGPT и GigaChat и более экспансивный подход GPT-5.

Ошибки в последовательностях приветствия. Одной из характерных ошибок при обнаружении приветствий является избыточное включение элементов, не выполняющих самостоятельной коммуникативной или этикетной функции. Все протестированные модели включали описание ситуативных действий, таких как вход *в квартиру, подход к бару или проход через дверной проем* – в последовательности, обозначающие установку контакта. Модели YandexGPT 5.1 Pro и GigaChat 2 Max ошибочно относили к приветственным жестам такие действия, как *помещение монет в окошко киоска* и *звон бокалов*, тогда как GPT-5 ошибочно интерпретировал *состояние опьянения* или *снятие обуви* как коммуникативные маркеры. Более высокая частота ложноположительных срабатываний у GPT-5 объясняется его общей большей полнотой и числом извлеченных последовательностей, а не принципиально иным профилем ошибок.

Одновременно пропускали жесты, которые в конкретном контексте выполняли функцию приветствия. К таким случаям относятся рукопожатия и объятия за плечи, не распознанные YandexGPT 5.1 Pro и GigaChat 2 Max, а также поцелуи, объятия и похлопывания по спине, пропущенные YandexGPT 5.1 Pro. Таким образом, трудность состоит не в обработке невербальных ремарок как таковых, а в различении ситуативного действия и действия, включенного в этикетную последовательность. Для правильного разрешения необходимо учитывать связь действия и реплики, функцию в установлении контакта, а также реакцию собеседника.

Еще одной проблемой оказалась идентификация форм обращения как маркеров приветствия – особенно диминутивных форм (*Алиночка, доченька*), ласкательных прозвищ (*кисуля*) и нестандартных номинаций (например, *прима-балерина*). Они зачастую пропускались, особенно русскоязычными моделями.

Обнаружение открывающих маркеров в телефонном дискурсе также представляло значительную трудность. Модели часто пропускали нестандартные написания (например, *алё* вместо *алло*), функциональное употребление частицы *да* в начале разговора и вокативные формы (например, *Маш, папань*).

Ошибки в последовательностях прощания. При обнаружении прощаний выявились сходства и различия между моделями: YandexGPT и GigaChat чаще используют консервативную стратегию и не выделяют неканонические сигналы завершения контакта, в то время как GPT-5 включает в границы последовательности дополнительный контекст.

Чаще всего пропускаются дискурсивно-прагматические маркеры в закрывающей позиции: *давай, бывай, всё, спасибо* и их варианты. Эти единицы глубоко конвенционализированы в русской речи как имплицитные сигналы завершения контакта, однако лишены формульной очевидности канонических прощаний *до свидания* или *пока*. Модели также нередко не распознавали ссылки на будущий контакт как закрывающие импликатуры: фразы типа *я буду через полчаса, до понедельника, ждем в четверг* и *увидимся* не интерпретировались как сигналы завершения коммуникации, несмотря на их очевидную функцию.

Экспансивная стратегия GPT-5 проявлялась в добавлении к прощанию лишних невербальных действий и тематически связанных реплик. Среди ложноположительных срабатываний, связанных с действиями, чаще всего встречается включение ремарок об уходе персонажа (например, *уходит, ушла*), которые, будучи прагматически информативными, тем не менее выходят за рамки нашего операционального определения прощания. Такой профиль ошибок позволяет предположить, что модель использует фоновые знания о типичном сценарии завершения встречи, но недостаточно точно воспроизводит операциональные правила аннотации.

В обоих типах последовательностей главным источником ошибок служит обнаружение нестандартных маркеров речевого этикета. У русскоязычных моделей (YandexGPT, GigaChat) эти маркеры нередко остаются функционально невидимыми, что дает высокую точность при низкой полноте. У GPT-5 те же явления провоцируют ложноположительные включения, отражая более низкий порог принятия решений и приводя к большей полноте при более низкой точности. Трудности с распознаванием форм обращения, нестандартных написаний, дискурсивных маркеров (*давай, всё*) и импликатур, отсылающих к будущему, указывают на ограниченность нынешних прагматических компетенций LLM.

Структура корпуса позволяет не просто констатировать ошибки моделей, но и строить эксперименты, учитывающие социальный контекст и личностные характеристики собеседников. Возможность обращаться непосредственно к

фрагментам видео также открывает перспективы для более широкого анализа социокультурных и коммуникативных возможностей моделей.

ЗАКЛЮЧЕНИЕ

В настоящем исследовании были оценены три модели: YandexGPT 5.1 Pro, GigaChat 2 Max и GPT-5 – с точки зрения их способности идентифицировать и извлекать последовательности приветствий и прощаний из русскоязычных диалогов. Двухкомпонентный дизайн исследования выявил существенный разрыв в качестве: GPT-5 достиг высоких результатов в классификации на уровне диалога (F1 = 0.89), однако все модели испытывали трудности с точным определением границ последовательностей. По всей видимости, надежное распознавание интенций при непоследовательной локализации границ характерно не только для речевых паттернов речевого этикета, но и для других задач сегментации дискурса, что указывает на то, что определение границ, а не распознавание содержания, остается узким местом при понимании естественного языка на дискурсивном уровне.

GPT-5 продемонстрировал наиболее сбалансированное соотношение точности и полноты, тогда как YandexGPT и GigaChat придерживались консервативных стратегий с почти идеальной точностью, но с высоким уровнем ложноотрицательных результатов – особенно на нестандартных и имплицитных маркерах. Анализ ошибок выявил систематические трудности трех типов: (1) чрезмерное включение невербальных действий и контекстуальных элементов, выходящих за функциональные границы; (2) пропуск нестандартных маркеров – диминутивов, дискурсивных частиц, импликатур с отсылкой к будущему контакту; (3) непоследовательная обработка телефонных открытий и описаний жестов. Это подчеркивает, что современные LLM хорошо справляются с прототипическими формульными выражениями, однако испытывают трудности с вариативностью живых разговорных открытий и закрытий. Для русскоязычных моделей это существенное ограничение в коммуникативно насыщенных сценариях, так как систематический пропуск имплицитных маркеров способен значительно снизить качество взаимодействия пользователя с диалоговыми системами и агентами на основе LLM.

Высокие показатели при решении задачи классификации демонстрируют перспективность применения обнаружения этикетных фреймов на уровне диалога для фильтрации диалогов и предобработки данных. В то же время более низкие результаты при извлечении границ реплик подтверждают, что полностью автоматическая аннотация последовательностей на данном этапе преждевременна. Значение F1 на уровне токенов (0.65) свидетельствует о том, что полуавтоматические подходы с участием эксперта представляются оптимальным решением при текущем уровне развития моделей.

Обозначим ключевые перспективы дальнейшей работы. В первую очередь необходим диагностический анализ качества моделей на подкорпусах, сформированных с помощью социальных фильтров. Сопоставление диалогов между знакомыми и незнакомыми собеседниками, участниками разных возрастных групп и носителями разных типов отношений позволит установить, в каких социальных контекстах модели чаще пропускают этикетные маркеры или неточно определяют их границы.

Следующим шагом должно стать расширение бенчмарка на другие этикетные ситуации, представленные в корпусе, включая извинение, благодарность и просьбу. Такое расширение позволит проверить, является ли выявленная сложность с определением границ общим свойством обработки этикетных формул или специфична для открывающих и закрывающих последовательностей.

Хотя конкретные показатели будут изменяться с выходом новых версий моделей, обнаруженные типы ошибок указывают на более общие ограничения прагматической компетенции LLM. Регулярный бенчмаркинг на сфокусированных прагматических задачах целесообразен для: (1) ориентирования разработчиков на создание дискурсно-ориентированных архитектур; (2) совершенствования обучающих данных для систем многостороннего диалога; (3) формирования реалистичных ожиданий относительно автоматизации разметки. В числе приоритетов будущих исследований выделим разработку формальных протоколов межаннотаторского согласия, расширение кросс-модельной оценки, а также изучение того, способны ли целевое дообучение или более продвинутые техники промптинга сократить разрыв в качестве между задачами классификации и извлечения.

Благодарности

Авторы выражают благодарность Министерству науки и высшего образования Российской Федерации за поддержку данного исследования в рамках дополнительного соглашения № 075-03-2026-305 (от 16 января 2026 г.), связанного с проектом «Прикладные исследования по внедрению технологий искусственного интеллекта в высшее образование» (шифр научной темы: FSMG-2025-0086).

СПИСОК ЛИТЕРАТУРЫ

1. *Brown P., Levinson S. C.* Politeness: Some Universals in Language. Cambridge University Press, 1987.

2. *Kádár D.Z., Haugh M.* Understanding politeness. Cambridge University Press, 2013.

3. *Schegloff E.A.* Sequence Organization in Interaction: A Primer in Conversation Analysis I. Cambridge University Press, 2007.

4. *Ruis L. et al.* The goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by LLMs // arXiv:2210.14986. 2024. https://doi.org/10.48550/arXiv.2210.14986

5. *Jian M., Narayanaswamy S.* Are LLMs good pragmatic speakers? // arXiv:2411.01562. 2024. https://doi.org/10.48550/arXiv.2411.01562

6. *Hu J., Floyd S., Jouravlev O., Fedorenko E., Gibson E.* A fine-grained comparison of pragmatic language understanding in humans and language models // Proc. of the 61st Annual Meeting of the Association for Computational Linguistics. 2023. Vol. 1. P. 4194–4213.

7. *Lipkin B., Wong L., Grand G., Tenenbaum J.B.* Evaluating statistical language models as pragmatic reasoners // arXiv:2305.01020. 2023. https://doi.org/10.48550/arXiv.2305.01020

8. *Iida A., Okuoka K., Omori T., Nakashima R., Osawa M.* Llm-based evaluation of utterances with implicature understanding: A preliminary study // Proc. of the 13th International Conference on Human-Agent Interaction. 2025. P. 485–487. https://doi.org/10.1145/3765766.3765841

9. *Yerukola A., Vaduguru S., Fried D., Sap M.* Is the pope catholic? Yes, the pope is catholic. Generative evaluation of non-literal intent resolution in LLMs // Proc. of the 62nd Annual Meeting of the Association for Computational Linguistics. 2024. Vol. 2. P. 265–275.

10. *Liu R., Sumers T. R., Dasgupta I., Griffiths T.L.* How do large language models navigate conflicts between honesty and helpfulness? // Proc. of the 41st International Conference on Machine Learning. 2024. P. 31844–31865.

11. *Chen Z., Yang R., Zhao Z., Cai D., He X.* Dialogue act recognition via CRF-attentive structured network. // Proc. of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. 2018. P. 225–234.

12. *Qamar A., Tong J., Huang R.* Do LLMs understand dialogues? A case study on dialogue acts. // Proc. of the 63rd Annual Meeting of the Association for Computational Linguistics. 2025. Vol. 1. P. 26219–26237.

13. *Schegloff E., Sacks H.* Opening up closings // Semiotica. 1973. Vol. 8, No. 4. P. 289–327.

14. *Klokova K., Krongauz M., Shulginov V., Yudina T.* Towards a russian multimedia politeness corpus. // Proc. of the International Conference “Dialogue 2023”. 2023. P. 233–244.

15. *Terkourafi M.* An argument for a frame-based approach to politeness: Evidence from the use of the imperative in Cypriot Greek // Pragmatics and beyond. New series. 2005. Vol. 139. P. 99–116.

16. *Laver J.* Communicative Functions of Phatic Communion // Organization of Behavior in Face-to-face Interaction. 1975. P. 215–238.

17. *Laver J.* Linguistic Routines and Politeness in Greeting and Parting // Conversational routines. 1981. P. 289–305.

18. *Brown T.B. et al.* Language models are fewshot learners // arXiv:2005.14165. 2020. https://doi.org/10.48550/arXiv.2005.14165

19. *Winkler W. E.* String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage // Proc. of the Section on Survey Research. 1990. P. 354–359.

20. *Cahyono S.C.* Comparison of document similarity measurements in scientific writing using jaro-winkler distance method and paragraph vector method // IOP Conference Series: Materials Science and Engineering. 2019. Vol. 662, No. 5. https://doi.org/10.1088/1757-899X/662/5/052016

ПРИЛОЖЕНИЕ А. ПРИМЕРЫ РАЗМЕЧЕННЫХ ДИАЛОГОВ И ПОСЛЕДОВАТЕЛЬНОСТЕЙ, ИДЕНТИФИЦИРОВАННЫХ МОДЕЛЯМИ

(1) Приветствие

Катя входит в комнату и включает свет

Катя: <greeting>*Олег, Олег,*</greeting> *проснись!*

*Олег: *просыпаясь** <greeting>*О, привет!*</greeting> *Ложись!*

- Эталонные последовательности: *Олег, Олег,* | *О, привет!*
- YandexGPT 5.1 Pro: *О, привет!*
- GigaChat 2 Max: *О, привет!*
- GPT-5: **просыпаясь* О, привет!*

(2) Прощание

*Марина: М... А... *на что-то указывает**

Максим Васильев: Тебе идти надо?

Марина: Да, <goodbye>*я пойду.*</goodbye> **встает и начинает собирать вещи**

Марина: Извини... <goodbye>*Пока.*</goodbye>

- Эталонные последовательности: *я пойду.* | *Пока.*
- YandexGPT 5.1 Pro: *Извини... Пока.*
- GigaChat 2 Max: *Да, я пойду. *встает и начинает собирать вещи* Извини... Пока.*
- GPT-5: *Да, я пойду. *встает и начинает собирать вещи* Извини... Пока.*

(3) Приветствие + прощание

У Саяпина звонит телефон

разговор по телефону

*Саяпин: *берет трубку** <greeting>*Да?*</greeting>

Подчиненный Саяпина: Ну что, и куда теперь?

…

Саяпин: Я знаю, что табельные. Спишем. Главное, чтобы тела не нашли, а то запалимся по глупой.

Подчиненный Саяпина: Угу.

*Саяпин: *кивает** <goodbye>*Всё.*</goodbye>

разговор по телефону завершён

- Эталонные последовательности: *Да?* | *Всё.*
- YandexGPT 5.1 Pro: –

- GigaChat 2 Max: –
- GPT-5: **берет трубку* Да?* | **кивает* Всё.*

PRAGMATIC BOUNDARIES OF GREETINGS AND GOODBYES: EVALUATING LLMS ON THE RUSSIAN MULTIMEDIA POLITENESS CORPUS

K. S. Klokova[1] [0000-0002-6650-1946], M. A. Krongauz[2] [0000-0003-2956-0565],

V. A. Shulginov[3] [0000-0003-2217-3677], T. A. Yudina[4] [0000-0001-6631-2128]

[1, 3, 4]*Moscow Institute of Physics and Technology, Dolgoprudny, Russia*

[2]*Russian State University for the Humanities, Moscow, Russia*

[1]klokova.ks@mipt.ru, [2]mkronhaus@yandex.ru, [3]shulginov.va@mipt.ru, [4]yudina.tatiana.a@mipt.ru

Abstract

This study investigates the ability of large language models to detect greeting and goodbye sequences in Russian-language dialogues and determine their precise pragmatic boundaries. YandexGPT 5.1 Pro, GigaChat 2 Max, and GPT-5 are evaluated on two tasks: classifying dialogues based on the presence of these sequences and extracting individual utterances, taking their intentional boundaries into account. The experiment is conducted on 216 dialogues from the Russian Multimedia Politeness Corpus, thereby accounting for the broad communicative and social context of interaction. GPT-5 achieves the strongest dialogue-level classification performance, whereas YandexGPT and GigaChat attain near-perfect precision through conservative prediction strategies. Performance declines markedly for all models when exact boundaries must be extracted. Error analysis shows that the models systematically over-include contextual nonverbal actions and miss non-conventional markers, including gestures, address forms, discourse markers, and implicatures indicating future contact. The article presents an updated description of the corpus and its web interface, as well as facilities for constructing diagnostic subcorpora and benchmarks. The findings underscore the need to model discourse structure and the contextual variation of everyday communication more precisely.

Keywords: *Large Language Models, Russian Multimedia Politeness Corpus,*

greeting, goodbye, speech etiquette, pragmatic boundaries, politeness formulae, dialogue analysis.

REFERENCES

1. *Brown P., Levinson S.C.* Politeness: Some Universals in Language. Cambridge University Press, 1987.

2. *Kádár D.Z., Haugh M.* Understanding politeness. Cambridge University Press, 2013.

3. *Schegloff E.A.* Sequence Organization in Interaction: A Primer in Conversation Analysis I. Cambridge University Press, 2007.

4. *Ruis L. et al.* The goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by LLMs // arXiv:2210.14986. 2024. https://doi.org/10.48550/arXiv.2210.14986

5. *Jian M., Narayanaswamy S.* Are LLMs good pragmatic speakers? // arXiv:2411.01562. 2024. https://doi.org/10.48550/arXiv.2411.01562

6. *Hu J., Floyd S., Jouravlev O., Fedorenko E., Gibson E.* A fine-grained comparison of pragmatic language understanding in humans and language models // Proc. of the 61st Annual Meeting of the Association for Computational Linguistics. 2023. Vol. 1. P. 4194–4213.

7. *Lipkin B., Wong L., Grand G., Tenenbaum J.B.* Evaluating statistical language models as pragmatic reasoners. // arXiv:2305.01020. 2023. https://doi.org/10.48550/arXiv.2305.01020

8. *Iida A., Okuoka K., Omori T., Nakashima R., Osawa M.* LLM-based evaluation of utterances with implicature understanding: A preliminary study // Proc. of the 13th International Conference on Human-Agent Interaction. 2025. P. 485–487. https://doi.org/10.1145/3765766.3765841

9. *Yerukola A., Vaduguru S., Fried D., Sap M.* Is the pope catholic? Yes, the pope is catholic. Generative evaluation of non-literal intent resolution in LLMs // Proc. of the 62nd Annual Meeting of the Association for Computational Linguistics. 2024. Vol. 2. P. 265–275.

10. *Liu R., Sumers T. R., Dasgupta I., Griffiths T.L.* How do large language models navigate conflicts between honesty and helpfulness? // Proc. of the 41st International Conference on Machine Learning. 2024. P. 31844–31865.

11. *Chen Z., Yang R., Zhao Z., Cai D., He X.* Dialogue act recognition via CRF-attentive structured network // Proc. of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. 2018. P. 225–234.

12. *Qamar A., Tong J., Huang R.* Do LLMs understand dialogues? A case study on dialogue acts. // Proc. of the 63rd Annual Meeting of the Association for Computational Linguistics. 2025. Vol. 1. P. 26219–26237.

13. *Schegloff E., Sacks H.* Opening up closings // Semiotica. 1973. Vol. 8, No. 4. P. 289–327.

14. *Klokova K., Krongauz M., Shulginov V., Yudina T.* Towards a russian multimedia politeness corpus. // Proc. of the International Conference “Dialogue 2023”. 2023. P. 233–244.

15. *Terkourafi M.* An argument for a frame-based approach to politeness: Evidence from the use of the imperative in Cypriot Greek // Pragmatics and beyond. New series. 2005. Vol. 139. P. 99–116.

16. *Laver J.* Communicative Functions of Phatic Communion // Organisation of Behaviour in Face-to-face Interaction. 1975. P. 215–238.

17. *Laver J.* Linguistic Routines and Politeness in Greeting and Parting // Conversational routines. 1981. P. 289–305.

18. *Brown T.B. et al.* Language models are few-shot learners // arXiv:2005.14165. 2020. https://doi.org/10.48550/arXiv.2005.14165

19. *Winkler W. E.* String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage // Proc. of the Section on Survey Research. 1990. P. 354–359.

20. *Cahyono S.C.* Comparison of document similarity measurements in scientific writing using jaro-winkler distance method and paragraph vector method // IOP Conference Series: Materials Science and Engineering. 2019. Vol. 662, No. 5. https://doi.org/10.1088/1757-899X/662/5/052016

СВЕДЕНИЯ ОБ АВТОРАХ

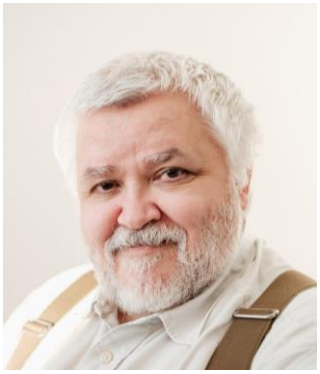


КЛОКОВА Ксения Сергеевна – младший научный сотрудник Центра междисциплинарных исследований МФТИ, специализирующийся на прикладной лингвистике, прагматике и применении методов машинного обучения в обработке естественного языка. Ведет исследования в области оценки дискурсивной компетенции больших языковых моделей и разработки корпусных ресурсов по речевому этикету русского языка. Разрабатывает мультиагентные LLM-системы и RAG-сервисы для задач анализа текста и дискурса.

Ksenia Sergeevna Klokova is a junior researcher at the Center for Interdisciplinary Research of MIPT, specializing in applied linguistics, pragmatics, and machine learning for natural language processing. Conducts research on evaluating the discursive competence of large language models and developing corpus resources on Russian speech etiquette. Develops multi-agent LLM systems and RAG services for text and discourse analysis.

email: klokova.ks@mipt.ru

ORCID: 0000-0002-6650-1946



КРОНГАУЗ Максим Анисимович – доктор филологических наук, профессор, руководитель группы Центра междисциплинарных исследований, главный научный сотрудник Центра прикладных лингвистических исследований и тестирования «ИСТОК» МФТИ, главный научный сотрудник учебно-научной лаборатории социолингвистики РГГУ. Специалист в области структурной и прикладной лингвистики, семиотики, современного русского языка и социолингвистики.

Maxim Anisimovich KRONGAUZ — Doctor of Philology, Professor, head of a research group at the Center for Interdisciplinary Research at MIPT, Chief research fellow at the ISTOK Center for applied linguistic research and testing at MIPT, and chief research fellow at the educational and research laboratory of sociolinguistics at the Russian state university for the humanities (RSUH). His research interests include structural and applied linguistics, semiotics, contemporary Russian, and sociolinguistics.

email: mkronhaus@yandex.ru

ORCID: 0000-0003-2956-0565



ШУЛЬГИНОВ Валерий Александрович – кандидат филологических наук, старший научный сотрудник Центра междисциплинарных исследований, ведущий научный сотрудник Центра прикладных лингвистических исследований и тестирования «ИСТОК» МФТИ, специалист в области прагматики, лингвистической конфликтологии и автоматизации анализа текстовых данных.

Valery Alexandrovich SHULGINOV – Candidate of Sciences in Philology, senior research fellow at the Center for Interdisciplinary Research at MIPT, and leading research fellow at the ISTOK Center for Applied Linguistic Research and Testing. His research interests include pragmatics, linguistic conflict studies, and automated textual data analysis.

email: shulginov.va@mipt.ru

ORCID:0000-0003-2217-3677



ЮДИНА Татьяна Александровна – Младший научный сотрудник Центра прикладных лингвистических исследований и тестирования «ИСТОК» и Центра междисциплинарных исследований МФТИ. Специализируется на прикладной и корпусной лингвистике и лингвистической прагматике, а также на разработке корпусов и баз знаний.

Tatiana Alexandrovna YUDINA is a junior researcher at the Center for Interdisciplinary Research of MIPT and the ISTOK Center for Applied Linguistic Research and Testing. She specializes in applied and corpus linguistics, linguistic pragmatics, and the development of corpora and knowledge bases.

email: yudina.tatiana.a@mipt.ru

ORCID: 0000-0001-6631-2128

Материал поступил в редакцию 20 августа 2026 года

УДК 004.82+004.85+340.115

О ПОВЫШЕНИИ КАЧЕСТВА RAG ПРИ ВЫЯВЛЕНИИ ПРОТИВОРЕЧИЙ В СУДЕБНЫХ ПОСТАНОВЛЕНИЯХ

Е. А. Козлова[1] [0009-0002-5884-126X], **А. М. Баюк**[2] [0009-0002-6721-0005], **М. А. Петрова**[3] [0009-0006-0393-1400]

[1, 3]*DM Solutions, г. Москва, Россия*

[2, 3]*Институт языкознания РАН, г. Москва, Россия*

[1]e.kozlova@dm-solutions.ru, [2]alexandra.m.baiuk@gmail.com,
[3]m.petrova@dm-solutions.ru

Аннотация

Настоящее исследование является продолжением работы над пилотной системой по детекции противоречий в судебных постановлениях, целью которой является выявление противоречий между судебными постановлениями по делам об административных правонарушениях и регулирующими нормами. Представленный ранее пайплайн, включающий правиловую предобработку документов, поиск по репозиторию правовых норм и дообучение LLM с использованием адаптера LoRA, обеспечил приемлемый уровень качества [1], однако имел ряд ограничений, связанных с компонентом RAG. В отличие от LLM, RAG не обучался на юридических текстах, что приводило к потере релевантных для сопоставления фрагментов на этапе поиска сходных текстовых пар, в связи с чем проведен анализ качества RAG и выполнено улучшение RAG-части системы.

Для оценки полноты извлечения данных созданы датасет из 30 документов и автоматические тесты, содержащие все пары релевантных для сопоставления предложений из данного датасета (более ста пар). На этих данных рассчитана полнота для исходной версии модели, использующей эмбеддинги предложений sbert_large_nlu_ru для создания векторной базы и вычисления косинусной близости.

Для повышения полноты извлекаемых данных проведены следующие эксперименты: (1) дообучение исходных эмбеддингов на обучающем датасете; (2) добавление модели-ранжировщика с дообучением и без; (3) комбинирование

© Е. А. Козлова, А. М. Баюк, М. А. Петрова, 2026.
Данная статья распространяется на условиях международной лицензии Creative Commons License Attribution 4.0 International (CC BY 4.0).

вариаций «базовые/дообученные эмбеддинги» и «базовый/дообученный ранжировщик» на нескольких версиях параметра top_k (количество извлекаемых соответствий); (4) замена модели эмбеддингов предложений на другие варианты. В результате существенно повысилось качество RAG – с 72% до 90%, однако увеличение количества извлекаемых фрагментов снизило общую точность системы, что требует ее дальнейшей калибровки. Кроме того, проведены реорганизация базы кодексов, расширение тестовой выборки и улучшение алгоритма фильтрации нерелевантных предложений из постановлений.

Ключевые слова: *выявление противоречий, юридические тексты, обучение LoRA (Low-Rank Adaptation — низкоранговая адаптация), RTE (Recognizing Textual Entailment — распознавание логического следования), RAG (Retrieval-Augmented Generation — генерация с дополнением по результатам поиска).*

ВВЕДЕНИЕ

Настоящая статья является продолжением работы [1], посвященной созданию пилотной модели по выявлению противоречий в судебных постановлениях.

Модель предназначена для выявления противоречий между судебными постановлениями по делам об административных правонарушениях и регулирующими их нормативными положениями. Целью текущей работы является оптимизация предложенного ранее пайплайна, затрагивающая, с одной стороны, оценку масштабируемости созданного прототипа, с другой — решение проблем, выявленных на предыдущем этапе работы, в частности, улучшение качества RAG-компонента системы.

Пилотная модель показала достаточно высокие метрики: precision = 0.96, recall = 0.69, F1-мера = 0.81, однако использованный в предыдущей версии RAG (Retrieval-Augmented Generation*)* не обучался на юридических текстах, что приводило к существенной потере полноты фрагментов, извлекаемых на этом этапе, в связи с чем повышение качества RAG стало центральной задачей дальнейшей доработки.

Выявление противоречий является частью задачи распознавания логического следования в тексте (Recognizing Textual Entailment, RTE) и логического вывода на естественном языке (Natural Language Inference, NLI), представляя собой важный аспект понимания текста.

Число работ, посвященных RTE и, в частности, выявлению противоречий, значительно [2–11], см. также библиографию в них.

Исследования в этой области фокусируются на создании классификаций и типологий противоречий [2, 3, 12], формировании аннотированных текстовых корпусов, содержащих пары противоречащих друг другу предложений [5, 13], и на разработке автоматических систем выявления противоречий. За последние несколько лет для решения этой задачи применялся широкий спектр инструментов – от логики, правил и различных лингвистических средств, таких как WordNet [2] в более ранние годы, до применения моделей на основе трансформеров и больших языковых моделей (Large Language Models, LLM) [7, 11] и нейросимволических подходов [14, 15] в недавних работах.

Выявление противоречий актуально для текстов разной тематики. Особый интерес эта задача представляет для юристов, поскольку обнаружение расхождений между документами и регулирующими их правовыми нормами требует, как правило, существенных временных и материальных затрат. Решение задачи RTE и выявление противоречий в юридических текстах рассматривались в [16–24], материалах конференции COLIEE (Competition on Legal Information Extraction/Entailment) и многих других работах (см. также обзор в [25]).

Генеративные модели вывели задачу выявления противоречий на новый уровень и позволили устанавливать не только противоречия, выражаемые через антонимы или отрицание, но и противоречия, основанные на структурных и лексических контрастах, а также на знаниях о мире (согласно типологии противоречий [3]). LLM успешно находят противоречия между двумя утверждениями или двумя документами и в большинстве случаев предлагают разумное объяснение, почему те или иные утверждения считаются противоречащими друг другу.

Однако, задача сравнения одного документа с крупным репозиторием, включающим законодательные базы, и выявления противоречий между рассматриваемым документом и всей базой данных к настоящему моменту не имеет удовлетворительного решения из-за требуемых ресурсов и сложности

природы противоречий. Лучший результат можно получить, если обучать модель на специальных данных, состоящих из релевантных правовых норм и базы фрагментов с противоречиями, типичными для указанной тематической области [14].

Следуя этой идее, мы разработали пилотную версию системы, предназначенной для выявления противоречий между судебными постановлениями по делам об административных правонарушениях, с одной стороны, и соответствующими правовыми нормами, с другой. Мы также определили набор типов противоречий, наиболее часто встречающихся в подобных юридических документах, и обучили модель их выявлять.

Сценарий работы системы выглядит следующим образом:

- репозиторий документов предварительно заполняется набором необходимых кодексов;
- пользователь загружает один или несколько документов в формате .docx;
- система проверяет, содержат ли загруженные документы утверждения, противоречащие статьям кодексов в репозитории, и, если такие противоречия обнаружены, представляет их пользователю.

Рассмотрим далее описание архитектуры системы, используемые данные, результаты нового этапа доработки системы и дальнейшие планы.

ОБЩАЯ АРХИТЕКТУРА СИСТЕМЫ

Архитектура системы состоит из трех компонентов. Первый компонент включает предварительную обработку документов.

Судебные постановления имеют типовую структуру и включают несколько частей: вводную (содержащую официальные реквизиты и идентификационные данные дела), описательную (содержащую краткое изложение сути заявленных требований), мотивировочную (содержащую логико-правовое обоснование постановления путем установления фактических обстоятельств дела) и резолютивную (итоговую часть судебного акта, содержащую решение суда об удовлетворении или отказе в удовлетворении требований).

Юридически значимая часть каждого судебного решения извлекается с помощью правилового подхода из резолютивной части – эта часть почти всегда

следует за ключевым словом «Постановил». Далее текст разбивается на предложения с использованием Python-пакета razdel[1], а предложения дополнительно фильтруются на предмет удаления заведомо нерелевантных предложений.

В новой реализации улучшен алгоритм такой фильтрации, что позволяет точнее исключать фрагменты, не содержащие потенциальных правовых противоречий, в частности предложения с указанием реквизитов банковских счетов, подписей судей и адресов. Это, с одной стороны, снижает дальнейшую нагрузку на модель, с другой – формально влияет на количественные значения метрик: благодаря фильтрации модель получает на вход меньше простых для анализа предложений, что несколько снижает процент правильных ответов, однако не означает реального ухудшения качества.

С помощью второго компонента мы находим потенциально релевантные фрагменты в репозитории правовых норм для их дальнейшей передачи модели, которая, в свою очередь, будет сравнивать пары вида «гипотеза (фрагмент постановления) – предпосылка (фрагмент кодекса)». В зависимости от эксплицитного указания номера статьи в тексте постановления правовые нормы извлекаются из репозитория либо посредством правил, либо технологией RAG.

Поскольку RAG-обработка занимает больше времени, система сначала пытается извлечь номер релевантных статьи/части/пункта правилами. Правиловое извлечение работает достаточно надежно благодаря четкой структуре постановлений. Если же предложение не содержит номера статьи, система использует семантический поиск. Для этого юридический корпус был преобразован в векторную базу данных с помощью библиотек FAISS[2] и LangChain[3] с использованием модели sentence BERT[4], что позволяет выполнять семантический поиск по косинусному сходству эмбеддингов предложений.

Однако модель эмбеддингов, используемая в предыдущей версии системы, не обучалась специально на юридических текстах, что приводило к суще-

[1] https://pypi.org/project/razdel/
[2] https://github.com/facebookresearch/faiss
[3] https://www.langchain.com/
[4] https://huggingface.co/ai-forever/sbert_large_nlu_ru

ственной потере фрагментов, извлекаемых на этом этапе (до 30%), и делало данный компонент наиболее слабым этапом текстовой обработки, создавая заведомые ограничения для полноты извлекаемых противоречий.

Наконец, извлеченные пары предложений разбиваются на батчи и обрабатываются LLM с адаптером LoRA, специально настроенным для обсуждаемой задачи.

Рассмотрим далее подробно все этапы: подготовку данных, способы повышения качества RAG как наиболее проблемного компонента первоначальной реализации, промпт, а также выбор и обучение моделей.

ПОДГОТОВКА ДАННЫХ

Система включает следующие датасеты: репозиторий релевантных правовых норм, обучающий корпус, валидационную выборку для анализа качества работы моделей по детекции противоречий, тестовый набор документов, представляющих собой полные судебные постановления для оценки качества RAG и всего пайплайна, а также автотесты, сформированные на их основе.

Репозиторий правовых норм

Репозиторий правовых норм, с которыми соотносятся судебные постановления, включает Кодекс Российской Федерации об административных правонарушениях (КоАП РФ), Кодекс административного судопроизводства РФ (КАС РФ), Гражданский кодекс РФ (ГК РФ), Уголовно-процессуальный кодекс РФ (УПК РФ), Гражданский процессуальный кодекс РФ (ГПК РФ), Уголовно-исполнительный кодекс РФ (УИК РФ), Уголовный кодекс РФ (УК РФ), а также инструкции по судебному делопроизводству.

На этапе предобработки из документов удаляются комментарии издателя и технические примечания (такие как «в редакции…»).

В предыдущей версии программы тексты преобразовывались далее в унифицированную структуру вида «Номер статьи. Текст статьи» с разделением статей на части, однако, поскольку постановления часто содержат ссылки не только на статьи и части статей, но и на пункты частей статей (например, «пункт 1 части 2 статьи 23.3»), мы добавили в базу кодексов также разделение по пунктам частей статей, что способствует повышению точности обработки. С этой целью были переработаны предобработанная база кодексов и векторная база.

При наличии в статье пунктов общее содержание излагается в части статьи, являющейся общей смысловой преамбулой, в то время как в пунктах описываются дальнейшие варианты действий. Это означает, что если пункт будет выступать отдельным фрагментом, смысл его будет утерян. В этой связи при каждом пункте хранится также и содержание части, к которой он относится.

Обучающие данные

Поскольку мы рассматриваем решаемую задачу как задачу текстового следования (Textual Entailment), обучающий набор данных был организован как коллекция пар предложений, где первый элемент (предпосылка) – это фрагмент юридической статьи, а второй (гипотеза) – фрагмент судебного постановления. Отношение между предложениями в каждой паре размечено вручную как “entailment” (следствие – одно предложение следует из другого), “contradiction” (противоречие – одно предложение противоречит другому) или “not mentioned” (одно предложение не является ни следствием, ни противоречием другого). Примеры обучающих данных представлены в [1].

Для задачи выявления противоречий в выделении класса entailment нет непосредственной необходимости, достаточно было ограничиться бинарной классификацией вида «противоречие» vs «отсутствие противоречий», где entailment и not mentioned объединены в один класс. Однако мы сохранили данные с тремя классами, чтобы иметь возможность преобразовывать их также в тип «похожие/непохожие» для дальнейшего дообучения модели эмбеддингов предложений.

Процесс создания датасета

В первую очередь были определены наиболее типичные виды противоречий, встречающиеся в рассматриваемых документах. К таковым относятся:

- неверное указание суммы штрафа цифрами и/или прописью;
- отсутствие суммы штрафа в постановлении;
- назначение дополнительного наказания, не предусмотренного регулирующей статьей;
- неверные сроки обжалования постановления или иные процессуальные сроки;

- антонимы («включая случаи» vs «за исключением случаев»);
- неверное назначение места/режима содержания (меры обеспечения исполнения выдворения);
- отрицание («может быть» vs «не может быть»);
- неверные реквизиты закона (номер/дата).

На текущем этапе модель обучена выявлять наиболее характерные противоречия и не нацелена на библиографические/цитатные ошибки, такие как неверный номер статьи/пункта/подпункта или неверная ссылка на правовой кодекс.

На основе этих данных был сгенерирован обучающий датасет, состоящий из пар предложений, где примеры противоречий охватывают вышеупомянутые типы: штрафы, выходящие за допустимый диапазон, неверные сроки, отсутствующие или лишние санкции и т. д. Датасет также включает пары предложений с entailment (где, например, суммы штрафов находятся в пределах законодательно допустимых диапазонов) и not mentioned.

Для защиты персональных данных исходные судебные постановления были анонимизированы путем замены личных идентификаторов на плейсхолдеры, при этом структура юридических документов была сохранена.

Исходный корпус включал около 500 пар, где гипотезам постановлений были вручную сопоставлены релевантные предпосылки кодексов. После этого набор данных был расширен с помощью GPT-5.2: для каждого предложения из постановления мы генерировали пары entailment, contradiction и not mentioned через специфичные для класса промпты и проверяли все сгенерированные примеры. Промпты были разработаны так, чтобы возвращать таблицу, где каждая строка содержала (i) норму кодекса как предпосылку, (ii) фрагмент судебного постановления как гипотезу и (iii) идентификатор цитируемой статьи/части.

Размер датасета и баланс классов

Полученный датасет был разделен на обучающую и валидационную выборки, содержащие описанные выше пары «предпосылка – гипотеза». Следуя предыдущим работам в области доменно-независимого NLI, мы стремились сохранять распределение трех классов сбалансированным, чтобы уменьшить эф-

фект перекоса классов, который, как показано, влияет на обобщающую способность модели (см., в частности, [26]). Эксперименты по оптимизации баланса классов представлены в [1], а итоговые цифры для датасетов с тернарной и бинарной классификациями – в табл. 1 (бинарная классификация включает contradiction vs entailment + not mentioned).

Табл. 1. Итоговый датасет и распределение классовых меток.

Датасет	Всего текстовых фрагментов	Contradiction	Entailment	Not mentioned	Entailment + Not mentioned
Обучающий датасет	2671	971 (36.4%)	832 (31.1%)	868 (32.5%)	1700 (63.65%)
Валидационный датасет	393	134 (34.1%)	129 (32.8%)	129 (32.8%)	258 (65.90%)

Тестовые данные и автотесты

В текущей реализации мы увеличили тестовую выборку документов с 13 до 30 постановлений. Обновленная выборка включает как документы с противоречиями, так и документы без противоречий. В сумме датасет содержит 66 пар с отношением contradiction и 43 пары с отношением entailment.

Новые постановления касаются тех же статей, что и в предыдущей выборке, однако содержат новые формулировки, отсутствующие в обучающем датасете. Анализ данных показал, что в случае близости текста предпосылки к тексту гипотезы система успешно находит и оценивает нужную пару, несмотря на отсутствие формулировки в обучающей выборке. Если же текст гипотезы существенно отличается от предпосылки, например содержит дополнительные фрагменты, то модель чаще классифицирует такие пары, как not mentioned вместо entailment.

На основе выбранных судебных постановлений были сформированы автотесты, представляющие собой excel-таблицы, где для каждой пары вида «гипотеза – предпосылка» указан тег отношения между гипотезой и предпосылкой.

В случаях, когда одному предложению гипотезы соответствуют несколько предложений из предпосылок, предложение входит в датасет более одного раза – по отдельному вхождению для каждой релевантной предпосылки.

Автотесты предназначены для оценки как полноты семантического поиска RAG, так и общего качества работы всего пайплайна. Качество работы непосредственно LLM при этом измеряется на валидационной выборке независимо от работы RAG.

ПОВЫШЕНИЕ КАЧЕСТВА RAG

Работы по оптимизации RAG-компонента включают подготовку датасета для оценки полноты извлеченных предпосылок, создание автоматических тестов для быстрой оценки результатов проводимых экспериментов и проведение экспериментов по повышению качества.

Оценка качества исходного RAG

В качестве тестового корпуса для оценки качества RAG был взят расширенный тестовый корпус для оценки всего пайплайна (см. п. «Тестовые данные и автотесты»), включающий 30 документов, на которые в сумме приходится 109 пар-противоречий и пар-следований. Автотест представляет собой excel-таблицу с данными парами. Для оценки RAG не требуется разделение пар на отношения противоречия и следствия.

Для оценки полноты поиска релевантных предпосылок использовалась метрика micro-averaged Recall@n. Поскольку используемая нами система является гибридной, итоговый Recall@n (*total Recall@n*) характеризует полноту всей системы, тогда как Recall@n для retrieval-компонента (*FAISS Recall@n*) рассчитывается только на подмножестве гипотез, для которых был задействован семантический поиск. При микроусреднении число корректно найденных эталонных ссылок суммируется по всем гипотезам и делится на общее число эталонных ссылок:

$$\text{Recall@n} = \frac{\sum_{i=1}^{N} |G_i \cap R_i^{(n)}|}{\sum_{i=1}^{N} |G_i|},$$

где G_i – множество эталонных релевантных нормативных ссылок для *i*-й гипотезы, а $R_i^{(n)}$ – множество ссылок, найденных системой среди первых *n* результатов.

На начальном этапе оптимизации полнота была расчитана для исходной версии RAG, используемой в [1]. Для создания векторной базы данная версия использует уже упоминавшиеся ранее эмбеддинги предложений sbert_large_nlu_ru. Для каждого предложения-гипотезы система RAG предоставляет некоторый список предложений-предпосылок, упорядоченный по косинусной близости с гипотезой. Первые *k* предпосылок отдаются далее модели для детекции противоречий, этот параметр далее будем называть top_k. В [1] top_k извлекаемых посылок был равен 20.

Метрики качества исходной системы, оцененной на обновленных тестах с учетом улучшенной фильтрации, изменились и для лучшей модели составили precision = 1, recall = 0.59, F1 = 0.74 для класса contradiction.

Наиболее проблемными являются случаи, когда сходные объекты имеют разные наименования в предпосылке и гипотезе. Например:

Гипотеза:

«Исполнение постановления возложить на межрайонный отдел судебных приставов по административному выдворению ***ГУФССП России*** *по г. Москве»*.

Предпосылка:

«Постановление об административном выдворении за пределы Российской Федерации иностранных граждан или лиц без гражданства исполняется:

2) ***федеральным органом исполнительной власти*** *<...>».*

«ГУФССП России по г. Москве» из гипотезы соответствует «федеральному органу исполнительной власти» из предпосылки, однако RAG-компонент подобные пары не находит.

Другой контекст, вызывающий сложности, – это случаи, когда релевантные статьи указываются в мотивационной и отсутствуют в резолютивной части постановления. Например, статьи 3.10 часть 5 и 27.19 КоАП в большинстве случаев указываются только в мотивационной части, поскольку сами по себе не являются статьями обвинения в нарушении, а лишь процессуально обосновывают принудительное содержание во ЦВСИГ.

Наконец, интерес представляют случаи соотношения одной гипотезы с несколькими предпосылками. Практика показывает, что одно предложение постановления может регулироваться несколькими нормами, причем текст может содержать отсылки не на все регулирующие статьи.

В качестве примера приведем следующее постановление:

> ***«До исполнения** административного наказания в виде принудительного выдворения за пределы Российской Федерации, содержать гражданина Рахимова **в Центре временного содержания иностранных граждан и лиц без гражданства** ГУ МВД России по городу Москве на основании **ст. 27.19.1 ч 1** Кодекса РФ об АП **сроком, не превышающем 100 суток** <…>.»*

Верным было бы указать в данном случае три предпосылки: статья 3.10 Часть 5 (право поместить задержанного в ЦВСИГ до выдворения), статья 27.19 Часть 1 (регулирует помещение задержанных в ЦВСИГ до выдворения), статья 27.19.1 КоАП РФ (регулирует срок содержания гражданина в специальном учреждении).

Базовый вариант RAG извлекает в таких случаях не все статьи, а статью без указания номера при указании в тексте других статей не извлекает вообще, так как при текущем описании семантический поиск применяется только при отсутствии эксплицитно указанных в тексте номеров статей. Подобные примеры немногочисленны, однако их наличие свидетельствует в пользу применения семантического поиска не только в случаях отсутствия ссылок на статьи.

Оптимизация исходного RAG

Для повышения качества извлечения семантически релевантных статей были проведены следующие эксперименты:

(1) дообучение эмбеддингов sbert_large_nlu_ru на нашем обучающем датасете;

(2) добавление модели-ранжировщика с дообучением и без;

(3) комбинирование вариаций «базовые/дообученные эмбеддинги sbert» и «базовый/дообученный ранжировщик» с разными параметрами candidate_top_k (количество кандидатов предпосылок, которые предоставляет модель до применения ранжировщика) и final_top_k (количество предпосылок, которые предоставляет ранжировщик);

(4) замена эмбеддингов предложений на эмбеддинги моделей bge-m3[5], multilingual-e5-large-instruct[6] и Qwen3-Embedding-0.6B[7].

Первая серия экспериментов состояла в комбинированном дообучении текущих эмбеддингов предложений (назовем их базовыми) и использовании предобученного либо дообученного на наших данных ранжировщика[8]. В этих вариантах использовались параметры 20:10 и 40:20[9]. Следует отметить, что для правиловой части извлечения предпосылок полнота не изменялась и всегда составляла 0.93. Для обеих версий параметров лучше всех показала себя версия с дообученным ранжировщиком (см. табл. 2).

Табл. 2. Дообучение текущих эмбеддингов предложений и использование ранжировщика.

Эмбеддинги	Ранжировщик	Candidate_top_k	Final_top_k	RAG Recall@n	Total Recall@n
Базовые	отсутствует	20	10	0.55	0.71
		40	20	0.58	0.72
	предобученный	20	10	0.45	0.64
		40	20	0.49	0.67
	дообученный	20	10	0.58	0.72
		40	**20**	**0.63**	**0.75**
Дообученные	отсутствует	20	10	0.63	0.75
		40	20	0.63	0.75
	предобученный	20	10	0.62	0.74
		40	20	0.62	0.74
	дообученный	20	10	0.63	0.75
		40	20	0.63	0.75

Сразу несколько вариантов дали одинаковое качество, поэтому для дальнейших экспериментов был выбран вариант с базовыми эмбеддингами и дообученным ранжировщиком для последующих экспериментов с заменой модели эмбеддингов.

[5] https://huggingface.co/BAAI/bge-m3
[6] https://huggingface.co/intfloat/multilingual-e5-large-instruct
[7] https://huggingface.co/Qwen/Qwen3-Embedding-0.6B
[8] https://huggingface.co/Alibaba-NLP/gte-multilingual-reranker-base
[9] Здесь и далее указываются параметры в виде (candidate_top_k:final_top_k).

Вторая серия экспериментов состояла в исследовании влияния параметров candidate_top_k и final_top_k. Результаты представлены в табл. 3.

Табл. 3. Влияние параметров candidate_top_k и final_top_k.

Вариант	Candidate_top_k	Final_top_k	RAG Recall@n	Total Recall@n
Базовые эмбеддинги + дообученный ранжировщик	20	10	0.58	0.72
	40	20	0.63	0.75
	60	40	0.71	0.8
	80	60	0.75	0.83
	100	10	0.74	0.82
	100	20	0.74	0.82
	100	40	0.74	0.82
	100	**60**	**0.77**	**0.83**
	100	80	**0.77**	**0.83**

Как видно из табл. 3, общая полнота получается наилучшей (и одинаковой) для версий 100:60 и 100:80 (поскольку полнота становится неизменной, очевидно, что увеличивать далее final_top_k нецелесообразно).

В третьей серии экспериментов мы взяли параметры 100:60, учитывая полученные во второй серии результаты, и попробовали заменить модель эмбеддингов предложений. Результаты представлены в табл. 4. При этом модель sbert_large_nlu_ru – классический трансформер, базирующийся на архитектуре BERT, так же, как и bge-m3; e5, представляет собой архитектуру, состоящую только из энкодеров, но принимает на вход также промпт, влияющий на итоговый эмбеддинг предложения, а Qwen3-Embedding – модель, построенная на базе генеративной Qwen. Соответственно, последние две модели при генерации эмбеддингов предложений получают короткий промпт вида «Instruct: Given a Russian legal statement from the operative section of a court decision, retrieve the relevant provision of Russian law. Query:».

Табл. 4. Замена модели эмбеддингов предложений.

Модель	RAG Recall@n	Total Recall@n
sbert_large_nlu_ru	0.77	0.83
bge-m3	0.86	0.89
Qwen3-Embedding-0.6B	**0.88**	**0.9**
multilingual-e5-large-instruct	0.57	0.72

Лучший результат полноты RAG-модуля показала версия с эмбеддингами Qwen3-Embedding-0.6B.

В результате выполненных экспериментов полнота RAG-компонента системы увеличилась на 18%, что достигается в версии RAG-компонента с эмбеддингами QWEN и ранжировщиком 100:60 (табл. 5).

Табл. 5. Итоговые показатели RAG по результатам экспериментов.

Вариант системы	Candidate_top_k	Final_top_k	RAG Recall@n	Правиловый Recall@n	Итоговый Recall@n
Бейзлайн RAG	20	20	0.58	0.93	0.72
RAG-Qwen 100:60	100	60	0.88	0.93	**0.90**
RAG-Qwen 40:20	40	20	0.82	0.93	0.86

РАЗРАБОТКА ПРОМПТОВ

При работе с генеративными моделями важным аспектом является создание промптов. Влияние промптов на модель не менее значимо, чем обучающий набор данных, и даже небольшие изменения в промпте могут оказать существенное влияние на результаты. Промпты, используемые в рассматриваемой системе, а также описания экспериментов по их оптимизации представлены в [1].

ВЫБОР И ДООБУЧЕНИЕ МОДЕЛЕЙ

Вследствие конфиденциальности обрабатываемых документов выбор модели осуществлялся между локальными (on-premise) моделями, соответствующими нашим ресурсным возможностям, вместо облачных версий.

Были выбраны следующие модели с версиями в 8 млрд параметров, потребляющие приблизительно 22 ГБ видеопамяти GPU при дообучении LoRA и

15 ГБ при инференсе: LLAMA 3.1-8B, MINISTRAL-8B-INSTRUCT, QWEN3-8B и T-LITE-IT-2.1. О мотивации выбора можно посмотреть [1, разд. 5.1].

Все модели были протестированы в тернарной и бинарной постановках задачи в режиме zero-shot. Мы рассматриваем Llama 3.1-8B как базовую модель (baseline), поскольку это самая простая модель без явной поддержки русского языка. Затем модели дообучили с помощью адаптера LoRA, используя библиотеку trl, в течение трех эпох с размером батча 2, накоплением градиента 8 и скоростью обучения 2×10^{-4}. Гиперпараметры были намеренно выбраны одинаковыми для всех моделей. Качество моделей оценивалось промежуточно на валидационном наборе, где лучшие метрики оказались у бинарной версии модели T-Lite (F1 = 0.9982), однако метрики на тестовых данных представляют более значительный интерес, поскольку тестовые данные серьезно отличаются от обучающих по распределению классов и более близки к реальным данным.

В табл. 6 приведены метрики на исходном тестовом датасете для моделей с LoRA-адаптерами (это метрики, оценивающие непосредственно качество работы моделей, без учета RAG-компонента).

Табл. 6. Метрики моделей после дообучения LoRA для бинарной и тернарной постановок на тестовом датасете.

	Бинарный вариант						Тернарный вариант					
	contradiction			no			contradiction			not mentioned		
Модель	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Llama 3.1-8B	0.42	0.52	0.46	0.98	0.97	0.97	0.54	0.28	0.37	0.97	0.99	0.98
Ministral-8B-Instruct-2410	0.47	1	0.64	1	0.95	0.98	**0.96**	**1**	**0.98**	**1**	**1**	**1**
Qwen3-8B	0.49	0.96	0.65	1	0.96	0.98	0.52	1	0.68	1	0.96	0.98
T-Lite-it-2.1	0.43	1	0.6	1	0.94	0.97	0.69	1	0.82	1	0.98	0.99

Лучшие результаты довольно неожиданно показала модель Ministral. Вероятно, это связано с тем, что архитектура модели довольно простая, а метод обучения исходной модели максимально близок к методу дообучения адаптера.

ОЦЕНКА И ОГРАНИЧЕНИЯ

Во время инференса результирующая генеративная модель получает промпт, содержащий пару «фрагмент постановления – правовая норма», и выдает метку класса для отношения между гипотезой и предпосылкой. В зависимости от выбора классификации (бинарной или тернарной) модель либо различает contradiction/entailment/not mentioned, либо объединяет entailment и not mentioned в единый класс «no». Далее мы приводим показатели для всех классов тернарной классификации как более наглядной, однако актуальными для текущей задачи являются показатели для contradiction. Вывод результатов пользователю представлен на рис. 1.

	Предложение	Номер статьи	Текст статьи	Результат сравнения
1	Признать Алимова Орхана Илхамовича виновным в совершении административного правонарушения, предусмотренного ч.3 ст.20.20 Кодекса РФ об административных правонарушениях, и назначить ему наказание в виде административного штрафа в размере двух тысяч рублей с административным выдворением за пределы Российской Федерации в форме контролируемого самостоятельного выезда из Российской Федерации.	КоАП РФ: Статья 20.20. п. 3.	КоАП РФ: Статья 20.20. п. 3. Действия, указанные в части 2 настоящей статьи, совершенные иностранным гражданином или лицом без гражданства, - влекут наложение административного штрафа в размере от четырех тысяч до пяти тысяч рублей с административным выдворением за пределы Российской Федерации либо административный арест на срок до пятнадцати суток с административным выдворением за пределы Российской Федерации.	противоречие

Рис. 1. Вывод программы.

Для итогового измерения качества общей системы были взяты три конфигурации:

1) бейзлайн (исходная система с базовым RAG-компонентом, но с обновленными правилами фильтрации);

2) система с RAG-компонентом на основе эмбеддингов Qwen3 с дообученным ранжировщиком, параметры top_k – 100:60;

3) система с RAG-компонентом на основе эмбеддингов Qwen3 с дообученным ранжировщиком, параметры top_k – 40:20 (данная версия была взята для того, чтобы иметь возможность сравнить бейзлайн также с более близкими параметрами top_k).

Результаты представлены в табл. 7. В ней указаны включает оценки качества полного цикла работы системы для следующих параметров: вариант RAG,

тип модели (базовая vs LoRA-адаптер) и название модели. Кроме того, приведены точность (contradiction precision), полнота (contradiction recall) и F-мера (F1) для детекции противоречий, а также отличие данных значений при доработанном RAG от аналогичных значений при базовом RAG: ΔP – дельта по точности, ΔR – дельта по полноте, ΔF1 – дельта по f-мере.

Табл. 7. Оценка качества работы полного цикла системы.

Тип RAG	Тип модели	Имя модели	Contradiction precision	ΔP	Contradiction recall	ΔR	Contradiction F1	ΔF1
Baseline	Базовая LLM	Llama	0.05	0	0.15	0	0.07	0
		Ministral	0	0	0	0	0	0
		Qwen	0.08	0	0.86	0	0.15	0
		T-Lite	0.44	0	0.68	0	0.53	0
	LoRA	Llama	0.82	0	0.14	0	0.23	0
		Ministral	1	0	0.52	0	0.68	0
		Qwen	0.53	0	0.82	0	0.64	0
		T-Lite	0.68	0	0.83	0	0.75	0
QWEN 100:60	Базовая LLM	Llama	0.01	–0.04	0.12	–0.03	0.02	–0.05
		Ministral	0	0	0	0	0	0
		Qwen	0.03	–0.05	0.82	–0.04	0.05	–0.1
		T-Lite	0.16	–0.28	0.8	0.12	0.26	–0.27
	LoRA	Llama	0.57	–0.25	0.12	–0.02	0.2	–0.03
		Ministral	0.97	–0.03	0.56	0.04	0.71	0.03
		Qwen	0.23	–0.3	0.91	0.09	0.37	–0.27
		T-Lite	0.44	–0.24	0.94	0.11	0.6	–0.15
QWEN 40:20	Базовая LLM	Llama	0.02	–0.03	0.08	–0.07	0.03	–0.04
		Ministral	0	0	0	0	0	0

		Qwen	0.07	–0.01	0.79	–0.07	0.13	–0.02
		T-Lite	0.3	–0.14	0.79	0.11	0.43	–0.1
	LoRA	Llama	0.62	–0.2	0.12	–0.02	0.2	–0.03
		Ministral	0.97	–0.03	0.53	0.01	0.69	0.01
		Qwen	0.34	–0.19	0.86	0.04	0.49	–0.15
		T-Lite	0.57	–0.11	0.91	0.08	0.7	–0.05

Как видно из табл. 7, наиболее сильное влияние на итоговую оценку работы системы оказывает тип модели (в данном случае, дообучение LoRA-адаптера). Наилучшее качество для класса contradiction дают версии системы с LoRA-адаптером, при этом повышение параметров top_k улучшает полноту, однако может серьезно ухудшить точность.

Модель, показавшая лучшие результаты в предыдущей работе (Ministral в тернарном варианте), на обновленных тестах в бейзлайне дала P = 1.00, R = 0.52 и F1 = 0.68, проиграв при этом по F1-мере некоторым другим моделям из-за падения полноты. Версии с улучшенным RAG-компонентом несколько подняли ее полноту (до 0.56 в лучшем случае), но ухудшили точность до 0.97.

Лучшие результаты по полноте, полученные с использованием обновленных тестов, представлены в табл. 8, лучшие результаты по F1-мере – в табл. 9.

Табл. 8. Лучшие результаты по полноте полного пайплайна.

Модель	Тип оценки	Precision	Recall	F1
T-Lite	RAG-Qwen: 100:60	0.44	0.94	0.60
	RAG-Qwen: 40:20	0.57	0.91	0.70
Qwen	RAG-Qwen: 100:60	0.23	0.91	0.37

Табл. 9. Лучшие результаты по F1-мере полного пайплайна.

Модель	Тип оценки	Precision	Recall	F1
T-Lite	Бейзлайн RAG	0.68	0.83	0.75
Ministral	RAG-Qwen: 100:60	0.97	0.56	0.71
T-Lite	RAG-Qwen: 40:20	0.57	0.91	0.70

Наилучшую полноту демонстрирует вариант с моделью T-Lite, где полнота выявленных противоречий достигает 94%. Тем не менее в этом случае мы сталкиваемся с проблемой слишком большого количества кандидатов: если модель обрабатывает большее количество пар предпосылка – гипотеза, точность детекции класса contradiction заметно падает. Очевидно, что версия RAG-компонента с эмбеддингами QWEN и ранжировщиком лучше извлекает релевантные предпосылки (полнота собственно RAG-компонента системы прирастает почти на 20%), однако детектор противоречий чувствителен к возникающему из-за большего количества предпосылок шуму. Несмотря на снижение точности, улучшение качества RAG важно для системы, так как с точки зрения сценария работы наиболее критичным является пропуск имеющегося в тексте противоречия, что делает полноту более значимым параметром, чем точность.

ЗАКЛЮЧЕНИЕ

В результате проведенной оптимизации качество RAG существенно возросло, что позволило повысить до 94% полноту извлекаемых системой противоречий, однако увеличение числа рассматриваемых пар снизило точность. Следующей задачей, таким образом, является снижение количества пар-кандидатов при возможном сохранении полноты RAG.

Дальнейшая работа будет развиваться также в следующих направлениях.

Мы планируем провести эксперименты с гиперпараметрами дообучения LoRA, расширить обучающий датасет для увеличения его репрезентативности и улучшения устойчивости к более редким случаям, а также расширить таксономию выявляемых ошибок, а также масштабировать систему на документы других типов.

Благодарности

Исследование поддержано проектом FMNE-2025-0002 «Проблемы прикладной лингвистики и гуманитарного знания в свете современных цифровых технологий», выполняемым в Институте языкознания РАН.

СПИСОК ЛИТЕРАТУРЫ

1. *Kozlova E., Petrova M., Baiuk A.* Contradiction Detection in Administrative Offense Rulings: A Case Study Analysis // Proc. of Conference on Computational Linguistics and Intellectual Technologies Dialog-2026. 2026. P. 294–310.

2. *Harabagiu S., Hickl A., Lacatusu F.* Negation, contrast and contradiction in text processing // Proc. of the 21st National Conference on Artificial Intelligence (AAAI-06). 2006. P. 755–762.

3. *De Marneffe M.-C., Rafferty A.N., Manning C.D.* Finding contradictions in text // Proc. of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-08: HLT). 2008. P. 1039–1047.

4. *Mehdad Y., Negri M., Federico M.* Towards cross-lingual textual entailment // Proc. of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2010). 2010. P. 321–324.

5. *Bowman S., Angeli G., Potts C., Manning C.D*. A large annotated corpus for learning natural language inference // arXiv:1508.05326. 2015. https://doi.org/10.48550/arXiv.1508.05326

6. *Dragos V.* Detection of contradictions by relation matching and uncertainty assessment // Procedia Computer Science. 2017. Vol. 112. P. 71–80.

7. *Pielka M., et al.* A linguistic investigation of machine learning based contradiction detection models: an empirical analysis and future perspectives // arXiv:2210.10434. 2022. https://doi.org/10.48550/arXiv.2210.10434

8. *Putra I M. S., Siahaan D., Saikhu A*. Recognizing textual entailment: a review of resources, approaches, applications, and challenges // ICT Express. 2024. Vol. 10, No. 1. P. 132–155. https://doi.org/10.1016/j.icte.2023.08.012

9. *Shajalal M., et al.* Textual entailment recognition with semantic features from empirical text representation // arXiv:2210.09723v4. 2022. https://doi.org/10.48550/arXiv.2210.09723

10. *Rahimi Z., ShamsFard M.* A knowledge-based approach for recognizing textual entailments with a focus on causality and contradiction // Research Square preprint. 2024. https://doi.org/10.21203/rs.3.rs-3826973/v1

11. *Asif M., Ahmed Khan T., Song W.-C.* Evaluating large language models for optimized intent translation and contradiction detection using KNN in IBN // IEEE Access. 2025. Vol. 13. P. 20316–20327.

12. *Wu X., Niu X., Rahman R.* Topological analysis of contradictions in text // Proc. of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2022). 2022. P. 2478–2483. https://doi.org/10.1145/3477495.3531881

13. *Williams A., Nangia N., Bowman S.* A broad-coverage challenge corpus for sentence understanding through inference // Proc. of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2018). 2018. P. 1112–1122. https://doi.org/10.18653/v1/N18-1101

14. *Camarda A.D., Ianni G.* A study on contradiction detection using a neuro-symbolic approach // Proc. of 40th Italian Conference on Computational Logic (CILC 2025). 2025, paper 08.

15. *Feng X., Hunter A.* Identification of entailment and contradiction relations between natural language sentences: a neurosymbolic approach // arXiv:2405.01259. 2024. https://doi.org/10.48550/arXiv.2405.01259

16. *Walton D., Reed C.* Argumentation schemes and defeasible inferences // Working Notes of the ECAI 2002 Workshop on Computational Models of Natural Argument (CMNA 2002). 2002. P. 45–55.

17. *Ibraim S.S., Mendonça P.C.C.* Book Review: «Argumentation Schemes» // Ensaio Pesquisa em Educação em Ciências. 2013. Vol. 15, No. 3. P. 255–262.

18. *Walton D.* Legal argumentation and evidence // Penn State University Press, 2010. 392 p.

19. *Mantravadi A., et al.* LegalWiz: a multi-agent generation framework for contradiction detection in legal documents // arXiv:2510.03418. 2025. https://doi.org/10.48550/arXiv.2510.03418

20. *Koreeda Y., Manning C.D.* ContractNLI: a dataset for document-level natural language inference for contracts // Findings of the Association for Computational Linguistics: EMNLP 2021. 2021, P. 1907–1919. https://doi.org/10.18653/v1/2021.findings-emnlp.164

21. *Nguyen H.-T., et al.* Transformer-based approaches for legal text processing: JNLP team – COLIEE 2021 // The Review of Socionetwork Strategies. 2022. Vol. 16, No. 1. P. 135–155. https://doi.org/10.1007/s12626-022-00102-2

22. *Nguyen C., et al.* CAPTAIN at COLIEE 2023: efficient methods for legal information retrieval and entailment tasks // arXiv:2401.03551. 2024. https://doi.org/10.48550/arXiv.2401.03551

23. *Aires J.P., Strube de Lima V.L., Meneguzzi F.* Identifying potential conflicts between norms in contracts // 18th International Workshop on Coordination, Organizations, Institutions, and Norms (COIN 2015). 2016. P. 179–188.

24. *Surana S., Dembla S., Bihani P.* Identifying contradictions in the legal proceedings using natural language models // SN Computer Science. 2022. Vol. 3, No. 3. Art. no. 187. https://doi.org/10.1007/s42979-022-01075-3

25. *Ariai F., Mackenzie J., Demartini G.* Natural language processing for the legal domain: a survey of tasks, datasets, models, and challenges // ACM Computing Surveys. 2025. Vol. 58, No. 6. P. 1–37. https://doi.org/10.48550/arXiv.2410.21306

26. *Hosseini M.J., Petrov A., Fabrikant A., Louis A.* A synthetic data approach for domain generalization of NLI models // Proc. of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024). 2024. P. 2212–2226. https://doi.org/10.18653/v1/2024.acl-long.120

ON IMPROVING RAG QUALITY FOR CONTRADICTION DETECTION IN COURT RULINGS

E. A. Kozlova[1] [0009-0002-5884-126X], **A. M. Baiuk**[2] [0009-0002-6721-0005],

A. M. Petrova [3] [0009-0006-0393-1400]

[1, 3]*DM Solutions, Moscow, Russia*

[2, 3]*Institute of Linguistics of the Russian Academy of Sciences, Moscow, Russia*

[1]e.kozlova@dm-solutions.ru, [2]alexandra.m.baiuk@gmail.com,
[3]m.petrova@dm-solutions.ru

Abstract

The paper extends and continues the work on a pilot system for detecting contradictions in court rulings. The tool aims to identify contradictions between court rulings on administrative offense cases and the legislative norms governing them. The previously presented pipeline, which includes rule-based document preprocessing, search over a structured repository of legal norms, and LLM fine-tuning using a LoRA adapter, demonstrated fairly high metrics [1].

However, this implementation had a number of limitations, primarily related to the RAG component. Unlike LLM, the RAG module was not trained on legal texts, which led to a substantial loss of relevant fragments for comparison at the retrieval stage.

The present work therefore focuses on a detailed quality analysis and improving the RAG component of the system. A dataset for evaluating premise retrieval recall was compiled, comprising 30 documents, along with automated tests containing all relevant sentence pairs (more than one hundred pairs in total). Using this data, the recall@n was calculated for the original (baseline) version of the model.

To improve the recall of the retrieved data, the following experiments were conducted: (1) fine-tuning the original embeddings on the system's training dataset; (2) adding a reranking model, both with and without fine-tuning; (3) combining variations of base/fine-tuned embeddings/reranker across several values of the top_k parameter (the number of retrieved matches); and (4) replacing the sentence embedding model with alternative options.

These experiments improved RAG recall from 72% to 90%. However, increasing the number of the retrieved fragments reduced the system's overall precision metrics, indicating a need for further training.

Other notable optimizations include reorganizing the structure of the legal codes database, expanding the test set, and improving the algorithm for filtering out irrelevant sentences from the rulings.

***Keywords**: contradiction detection, legal domain, LoRA (Low-Rank Adaptation) fine-tuning, RTE (Recognizing Textual Entailment), RAG (Retrieval-Augmented Generation).*

Acknowledgments

This research was supported by project FMNE-2025-0002 «Problems of Applied Linguistics and the Humanities in the Light of Modern Digital Technologies», carried out at the Institute of Linguistics, Russian Academy of Sciences.

REFERENCES

1. *Kozlova E., Petrova M., Baiuk A.* Contradiction Detection in Administrative Offense Rulings: A Case Study Analysis // Proc. of Conference on Computational Linguistics and Intellectual Technologies Dialog-2026. 2026. P. 294–310.

2. *Harabagiu S., Hickl A., Lacatusu F.* Negation, contrast and contradiction in text processing // Proc. of the 21st National Conference on Artificial Intelligence (AAAI-06). 2006. P. 755–762.

3. *De Marneffe M.-C., Rafferty A.N., Manning C.D.* Finding contradictions in text // Proc. of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-08: HLT). 2008. P. 1039–1047.

4. *Mehdad Y., Negri M., Federico M.* Towards cross-lingual textual entailment // Proc. of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2010). 2010. P. 321–324.

5. *Bowman S., Angeli G., Potts C., Manning C.D*. A large annotated corpus for learning natural language inference // arXiv:1508.05326. 2015. https://doi.org/10.48550/arXiv.1508.05326

6. *Dragos V.* Detection of contradictions by relation matching and uncertainty assessment // Procedia Computer Science. 2017. Vol. 112. P. 71–80.

7. *Pielka M., et al.* A linguistic investigation of machine learning based contradiction detection models: an empirical analysis and future perspectives // arXiv:2210.10434. 2022. https://doi.org/10.48550/arXiv.2210.10434

8. *Putra I M. S., Siahaan D., Saikhu A*. Recognizing textual entailment: a review of resources, approaches, applications, and challenges // ICT Express. 2024. Vol. 10, No. 1. P. 132–155. https://doi.org/10.1016/j.icte.2023.08.012

9. *Shajalal M., et al.* Textual entailment recognition with semantic features from empirical text representation // arXiv:2210.09723v4. 2022. https://doi.org/10.48550/arXiv.2210.09723

10. *Rahimi Z., ShamsFard M.* A knowledge-based approach for recognizing textual entailments with a focus on causality and contradiction // Research Square preprint. 2024. https://doi.org/10.21203/rs.3.rs-3826973/v1

11. *Asif M., Ahmed Khan T., Song W.-C.* Evaluating large language models for optimized intent translation and contradiction detection using KNN in IBN // IEEE Access. 2025. Vol. 13. P. 20316–20327.

12. *Wu X., Niu X., Rahman R.* Topological analysis of contradictions in text // Proc. of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2022). 2022. P. 2478–2483. https://doi.org/10.1145/3477495.3531881

13. *Williams A., Nangia N., Bowman S.* A broad-coverage challenge corpus for sentence understanding through inference // Proc. of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2018). 2018. P. 1112–1122. https://doi.org/10.18653/v1/N18-1101

14. *Camarda A.D., Ianni G.* A study on contradiction detection using a neurosymbolic approach // Proc. of 40th Italian Conference on Computational Logic (CILC 2025). 2025, paper 08.

15. *Feng X., Hunter A.* Identification of entailment and contradiction relations between natural language sentences: a neurosymbolic approach // arXiv:2405.01259. 2024. https://doi.org/10.48550/arXiv.2405.01259

16. *Walton D., Reed C.* Argumentation schemes and defeasible inferences // Working Notes of the ECAI 2002 Workshop on Computational Models of Natural Argument (CMNA 2002). 2002. P. 45–55.

17. *Ibraim S.S., Mendonça P.C.C.* Book Review: «Argumentation Schemes» // Ensaio Pesquisa em Educação em Ciências. 2013. Vol. 15, No. 3. P. 255–262.

18. *Walton D.* Legal argumentation and evidence // Penn State University Press, 2010. 392 p.

19. *Mantravadi A., et al.* LegalWiz: a multi-agent generation framework for contradiction detection in legal documents // arXiv:2510.03418. 2025. https://doi.org/10.48550/arXiv.2510.03418

20. *Koreeda Y., Manning C.D.* ContractNLI: a dataset for document-level natural language inference for contracts // Findings of the Association for Computational Linguistics: EMNLP 2021. 2021, P. 1907–1919. https://doi.org/10.18653/v1/2021.findings-emnlp.164

21. *Nguyen H.-T., et al.* Transformer-based approaches for legal text processing: JNLP team – COLIEE 2021 // The Review of Socionetwork Strategies. 2022. Vol. 16, No. 1. P. 135–155. https://doi.org/10.1007/s12626-022-00102-2

22. *Nguyen C., et al.* CAPTAIN at COLIEE 2023: efficient methods for legal information retrieval and entailment tasks // arXiv:2401.03551. 2024. https://doi.org/10.48550/arXiv.2401.03551

23. *Aires J.P., Strube de Lima V.L., Meneguzzi F.* Identifying potential conflicts between norms in contracts // 18th International Workshop on Coordination, Organizations, Institutions, and Norms (COIN 2015). 2016. P. 179–188.

24. *Surana S., Dembla S., Bihani P.* Identifying contradictions in the legal proceedings using natural language models // SN Computer Science. 2022. Vol. 3, No. 3. Art. no. 187. https://doi.org/10.1007/s42979-022-01075-3

25. *Ariai F., Mackenzie J., Demartini G.* Natural language processing for the legal domain: a survey of tasks, datasets, models, and challenges // ACM Computing Surveys. 2025. Vol. 58, No. 6. P. 1–37. https://doi.org/10.48550/arXiv.2410.21306

26. *Hosseini M.J., Petrov A., Fabrikant A., Louis A.* A synthetic data approach for domain generalization of NLI models // Proc. of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024). 2024. P. 2212–2226. https://doi.org/10.18653/v1/2024.acl-long.120

СВЕДЕНИЯ ОБ АВТОРАХ



КОЗЛОВА Елена Алексеевна – текстовый аналитик в ООО «Датамайнинг Солюшинс» (DM Solutions, Москва). Область научных интересов: автоматизация юридической деятельности (LegalTech), обработка естественного языка (NLP), большие языковые модели (LLM), логический вывод на естественном языке (NLI), синтетические датасеты для обучения моделей.

Elena Alekseevna KOZLOVA – a text analyst at DM Solutions LLC (Moscow, Russia). Research interests: legal process automation (LegalTech), natural language processing (NLP), large language models (LLMs), natural language inference (NLI), and synthetic datasets for model training.

email: e.kozlova@dm-solutions.ru

ORCID: 0009-0002-5884-126X



БАЮК Александра Михайловна – младший научный сотрудник Института языкознания Российской академии наук. Научные интересы включают машинное обучение, большие языковые модели, извлечение информации, лингвистическую разметку, мультиязычность и автоматическую обработку малоресурсных языков.

Alexandra Mikhailovna BAIUK – junior researcher at the Institute of Linguistics of the Russian Academy of Sciences. Her research interests include machine learning, large language models, information extraction, linguistic annotation, multilingualism, and natural language processing for low-resource languages.

email: alexandra.m.baiuk@gmail.com

ORCID: 0009-0002-6721-0005



ПЕТРОВА Мария Андреевна – кандидат филологических наук, заведующая отделом прикладной лингвистики Института языкознания Российской академии наук, руководитель направления текстовой аналитики в ООО «Датамайнинг Солюшинс» (DM Solutions, Москва). Сфера научных интересов – компьютерная лингвистика, формальные описания, языковые модели, семантика, текстовая аналитика, лингвистическая экспертиза.

Maria Andreevna PETROVA – Candidate of Philological Sciences, head of the Applied Linguistics Department at the Institute of Linguistics of the Russian Academy of Sciences, head of Text Analytics at DM Solutions LLC (Moscow, Russia). Research interests include computational linguistics, formal descriptions, language models, semantics, text analytics, and linguistic expertise.

email: m.petrova@dm-solutions.ru

ORCID: 0009-0006-0393-1400

Материал поступил в редакцию 25 августа 2026 года

УДК 004.934

КОММУНИКАТИВНАЯ ПРЕДВЗЯТОСТЬ ASR ПРИ РАСПОЗНАВАНИИ АТИПИЧНОЙ РЕЧИ НА РУССКОМ ЯЗЫКЕ

А. В. Колмогорова[1] [0000-0002-6425-2050], **Е. В. Явшиц** [2][0000-0001-6344-1565],
М. С. Майорова [3][0009-0009-2106-7689], **С. В. Дмитриева** [4][0009-0001-1949-5360]

[1–4]*Национальный исследовательский университет «Высшая школа экономики», г. Москва, Россия*

[1]akolmogorova@hse.ru, [2]kate_yavshitz@mail.ru, [3]svyatoslavna.m@gmail.com,
[4]sofiadmitrieva82@gmail.com

Аннотация

Статья посвящена проблеме коммуникативной предвзятости систем автоматического распознавания речи при работе с атипичной речью на русском языке. Актуальность исследования обусловлена активным внедрением технологий искусственного интеллекта в социальную сферу и государственное управление в России, что предполагает обеспечение их доступности для различных категорий граждан, включая людей с речевыми нарушениями. Введено понятие коммуникативной предвзятости ASR-моделей как свойства системы в условиях неопределенности (атипичная речь, шум, паузы) принимать решения о заполнении смысловых и акустических лакун на основе собственной обучающей выборки. В эмпирической базы исследования выбран корпус RuAphasiaBank, включающий записи речи пациентов с различными типами афазии (моторной, сенсорной, семантической, сенсомоторной) разной степени тяжести, а также нейротипичных информантов. Для сравнительного анализа были отобраны шесть ASR-моделей, работающих с русским языком: две открытые (GigaAM, T-One) и четыре коммерческие (Yandex SpeechKit, Any2Text, Charla, Speech2Text). Качество распознавания оценено с использованием метрик WER и CER, а также с помощью анализа соотношения вставок, удалений и замен на уровне слов и символов. Результаты показали, что модели реализуют различные стратегии при столкновении с атипичной речью: «паническое укорачивание» (GigaAM, Yandex SpeechKit), «лексическое сохранение» в ущерб точности (T-One) и «стабильная» стратегия сбалансированных ошибок (Any2Text, Speech2Text, Charla). Лингвистический анализ

© А. В. Колмогорова, Е. В. Явшиц, М. С. Майорова, С. В. Дмитриева, 2026.
Данная статья распространяется на условиях международной лицензии Creative Commons License Attribution 4.0 International (CC BY 4.0).

выявил, что каждая модель порождает специфический искаженный коммуникативный стиль: от разговорной диалогической речи до стиля, ассоциируемого с когнитивными нарушениями или употреблением молодежного сленга. На основе полученных данных моделируются потенциальные риски для социального самочувствия и правовой защищенности пациентов с афазией при использовании голосовых помощников в государственных учреждениях.

***Ключевые слова**: речевые технологии, речевая база данных, автоматическое распознавание речи, атипичная речь, афазия, коммуникативная предвзятость, голосовые ассистенты, ассистивные технологии, инклюзивность.*

ВВЕДЕНИЕ

Активное развитие технологий искусственного интеллекта в России и его быстрое проникновение в социальную сферу ставят перед разработчиками и исследователями ряд вопросов, связанных с качеством работы, доступностью современных технологий различным категориям пользователей и их объективно существующими ограничениями. Эти вопросы становятся особенно актуальными в связи со стремлением государства внедрить искусственный интеллект в социальную сферу: «В рамках проекта «Искусственный интеллект» планируется внедрить технологии искусственного интеллекта в экономику, социальную сферу и государственное управление»[1]. При этом согласно Указу Президента «О развитии искусственного интеллекта в Российской Федерации» к основным принципам развития и использования технологий искусственного интеллекта в России относятся, в первую очередь, «обеспечение защиты прав и свобод человека, гарантированных законодательством Российской Федерации, в том числе права на труд, и предоставление гражданам возможности получать знания и

[1] Нацпроект «Экономика данных и цифровая трансформация государства» [Электронный ресурс]. Офиц. сайт национальных проектов РФ. URL: https://xn--80akpjgfht4a0d.xn--p1ai/naczproekt-ekonomika-dannyh-i-czifrovaya-transformacziya-gosudarstva/ (дата обращения: 17.08.2025)

приобретать навыки для успешной адаптации к условиям цифровой экономики»[2], а также безопасность, понимаемая, в том числе, как «предупреждение и минимизация рисков возникновения негативных последствий использования технологий искусственного интеллекта». К таким негативным последствиям следует относить не только риски в сфере информационной и экономической безопасности, но и репутационные потери, снижение уровня социального (субъективного) настроения [1], основанного, в частности, на социальном самочувствии. Последнее есть «определенное состояние переживания индивида по поводу комфортности или дискомфортности своего бытия в социуме» [2]. Иными словами, проблематизирующий контекст настоящего исследования составляет предположение о том, что возможны случаи, когда технологии могут вносить дисбаланс в психологическое равновесие пользующегося ими индивида, вызывать чувство фрустрации, ухудшая тем самым его социальное самочувствие. Эта гипотеза особенно актуальна тогда, когда мы говорим о людях с заболеваниями, накладывающими определенные ограничения на социальное взаимодействие, и при этом предлагаем технологии, предполагающие смягчение этих ограничений. Подобные технологии включены в категорию ассистивных, понимаемых как «любое внешнее изделие (включая оборудование, инструменты, приборы или программное обеспечение), основная цель которого – поддерживать или улучшать функционирование и независимость человека и тем самым способствовать его благополучию» [3].

Рабобта посвящена описанию результатов предварительной оценки качества работы шести моделей автоматического распознавания речи (Automatic Speech Recognition – ASR) с корпусом речи пациентов с системным речевым расстройством – афазией. Ее целью – системно представить наблюдения, свидетельствующие о том, что каждая из моделей имеет собственные стратегии получения результата в ситуации работы с атипичной речью, но в большинстве случаев это приводит к комплексному, в случае каждой из моделей – к специфическому, искажению коммуникативного стиля говорящих, что способно повлечь за

[2] Указ Президента Российской Федерации от 10 октября 2019 г. № 490 «О развитии искусственного интеллекта в Российской Федерации» [Электронный ресурс]. Офиц. интернет-портал правовой информации. URL: https://www.consultant.ru/document/cons_doc_LAW_335184 (дата обращения: 17.08.2025)

собой риски в сфере социального самочувствия пользователей с нарушениями речи.

ПОНЯТИЕ КОММУНИКАТИВНОЙ ПРЕДВЗЯТОСТИ: КАК ЭТО РАБОТАЕТ В ОТНОШЕНИИ МОДЕЛЕЙ ASR

В коммуникативной лингвистике сложилась и существует определенная традиция изучения способов передачи чужой речи – ксенопоказателей. К их числу относят модальные частицы *мол*, *типа*, *якобы* [4], специфическую просодику [5] и мимические движения, не свойственные пересказывающему чужую речь в его обычном коммуникативном поведении. Между таким сугубо лингвистическим подходом и проблематикой моделей ASR достаточно мало общего. Тем не менее вопрос о стратегиях передачи чужой речи важен как в лингвистическом, так и в технологическом контекстах. Так, например, в [5] отмечено, что при передаче содержания чужих мнений или высказываний возможны две стратегии: либо говорящий ориентируется на свои представления, не заботясь о том, согласится ли с такой передачей субъект передаваемого мнения или высказывания, либо говорящий ориентируется на представления субъекта передаваемого мнения или высказывания, даже если сам их не разделяет.

В техническом контексте же обсуждается несколько иное, но близкое по сути [6] понятие «предсказательной предвзятости» моделей: когда качество предсказания моделей ASR зависит от того, насколько подаваемые на вход данные отличаются от тех, на которых модель обучалась. Ставится вопрос о том, что подавляющее большинство моделей обучается и тестируется на данных, репрезентирующих речь некоторой приоритезируемой компаниями-разработчиками социальной группы. Качество моделей оценивается также на ограниченном количестве датасетов, активно циркулирующих внутри индустриальных и исследовательских сообществ. Ошибки и эрративные паттерны, которые допускают модели, работая с нетипичной или атипичной речью не принимаются во внимание, не собираются и не анализируются. Таким образом, по нашему мнению, разработчики ASR моделей фактически становятся создателями определенной языковой политики, оказывая предпочтение одним идиомам и маргинализируя другие. Иными словами, речь идет о том, что модели также могут «выбирать» ту или иную стратегию для передачи распознанного чужого слова.

Таким образом, объединяя лингвистическую и технологическую концепции, мы будем понимать под коммуникативной предвзятостью ASR-моделей такое свойство системы, при котором в условиях неопределенности (атипичная речь, шум, паузы) модель принимает решение о заполнении лакун, опираясь на собственную обучающую выборку, а не на реальное коммуникативное намерение говорящего, что приводит к систематическому искажению речевого портрета пользователя.

ДАННЫЕ, МЕТОДЫ И МОДЕЛИ

Записи атипичной речи производились с использованием микрофонов iPhone 13–16 во время клинических логопедических занятий с пациентами одного из нейрореабилитационных центров. Для обеспечения конфиденциальности доступ в Интернет был отключен. Занятия проходили в тихом помещении, пациент и терапевт сидели за столом, телефон размещался на расстоянии 30–40 см от говорящего, что обеспечивало минимальный уровень фонового шума. Записи речи нейротипичных участников собирались в сопоставимых акустических условиях. Все участники – как нейротипичные лица, так и пациенты с афазией (ПА) – дали информированное согласие на участие в записи. Возрастное распределение ПА и нейротипичных информантов коррелировали, гендерное распределение – нет. В датасете представлены различные типы речи: монолог, диалог, чтение и пересказ. Основные параметры данных представлены в табл. 1, более подробное описание датасета можно найти в [9].

В сравнительный анализ были включены шесть моделей ASR, в документации которых указано, что при их обучении использовались данные на русском языке, а в продвигающем тексте отмечалось, что модель предназначена в том числе для анализа данных на русском языке. Две модели являются открытыми (GigaAM и T-One), еще 4 – коммерческими (Yandex SpeechKit, Any2text, Charla, Speech2Text).

Поскольку для каждой записи уже имелись проверенные экспертами эталонные транскрипты, сделанные вручную, основной методологией стала оценка качества распознавания при помощи стандартных метрик WER (Word Error Rate) [10] и CER (Character Error Rate) [11]. Однако мы решили не просто измерить мет-

рики для каждой из моделей на выборках: а) различных типов афазий, б) различных степеней выраженности нарушений, в) различных типов речи, но и посмотреть в каждом из 18 случаев соотношение вставок, замен и удалений, чтобы проанализировать стратегии, которые выбирают модели при столкновении с атипичными записями.

Табл. 1. Основные параметры датасета.

Параметр	Категория	Объем
Записи	Общая длительность	14 ч
	ПА	9.7 ч
	Нейротипичные	4.3 ч
Информанты	ПА	70 чел.
	Логопеды	8 чел.
	Нейротипичные	20 чел.
Гендер	ПА m/f	54/16 чел.
	Логопеды m/f	0/8 чел.
	Нейротипичные m/f	18/6 чел.
Средний возраст (в годах)	ПА m/f	53.4/61.3
	Логопеды m/f	0/41.2
	Нейротипичные m/f	56.1/70.2

РЕЗУЛЬТАТЫ

Сравнение значений WER при распознавании записей речи пациентов с разными типами афазий демонстрирует (см. рис. 1), что наиболее успешные по этой метрике модели изменяются в зависимости от типа афазии, а аутсайдер неизменен – это Yandex SpeechKit, хотя и его поведение нестабильно: в случае семантической афазии лидируют замены, а при сенсорной – удаления. GigaAM показывает наилучшие значения WER при комплексной моторной, семантической и сенсомоторной афазии, но при этом так же, как и аутсайдер, предпочитает удалять непонятое.

T-One выделяется высоким уровнем замен (до 25%) при относительно низком CER (рис. 2) (15–24%), что говорит о стратегии «лексического сохранения» – модель жертвует точностью на уровне символов, но пытается сохранить слова;

его лучший результат достигается на семантической афазии (WER 29%, CER 19%). Speech2Text, Charla и Any2Text занимают промежуточное положение (WER 27–41%, CER 18–26%) (рис. 1 и 2), демонстрируя сбалансированное соотношение удалений и замен; Speech2Text выделяется аномально высокой долей вставок на семантической афазии (10.1%), а Any2Text показывает лучший результат на эфферентной моторной афазии (WER 32%, CER 18%).

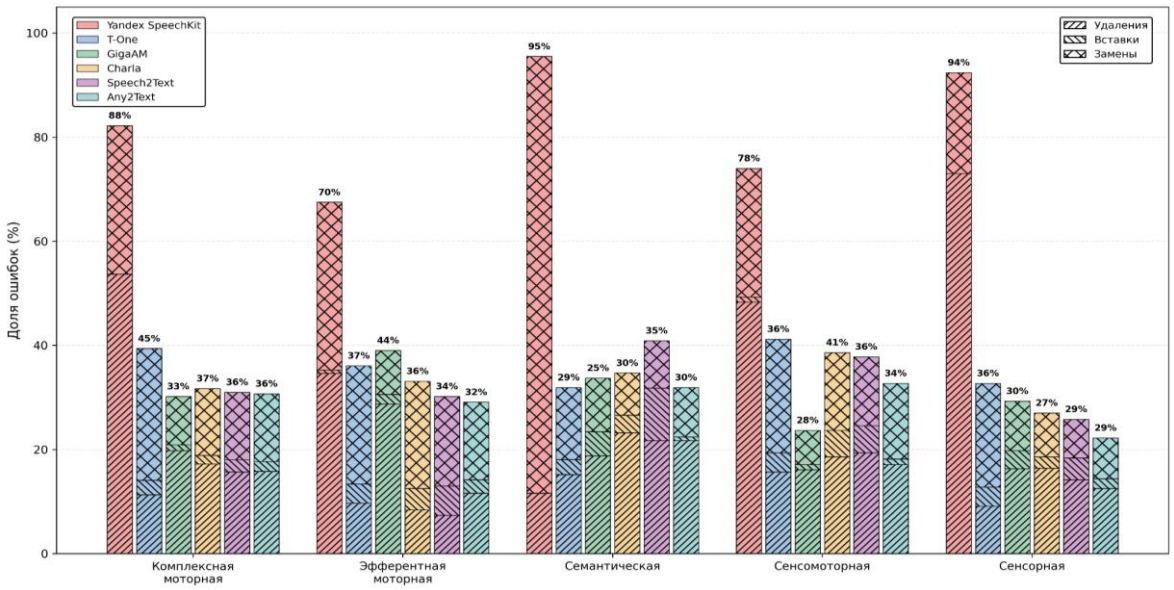


Рис. 1. WER и соотношение вставок/удалений/замен по разным типам афазии.

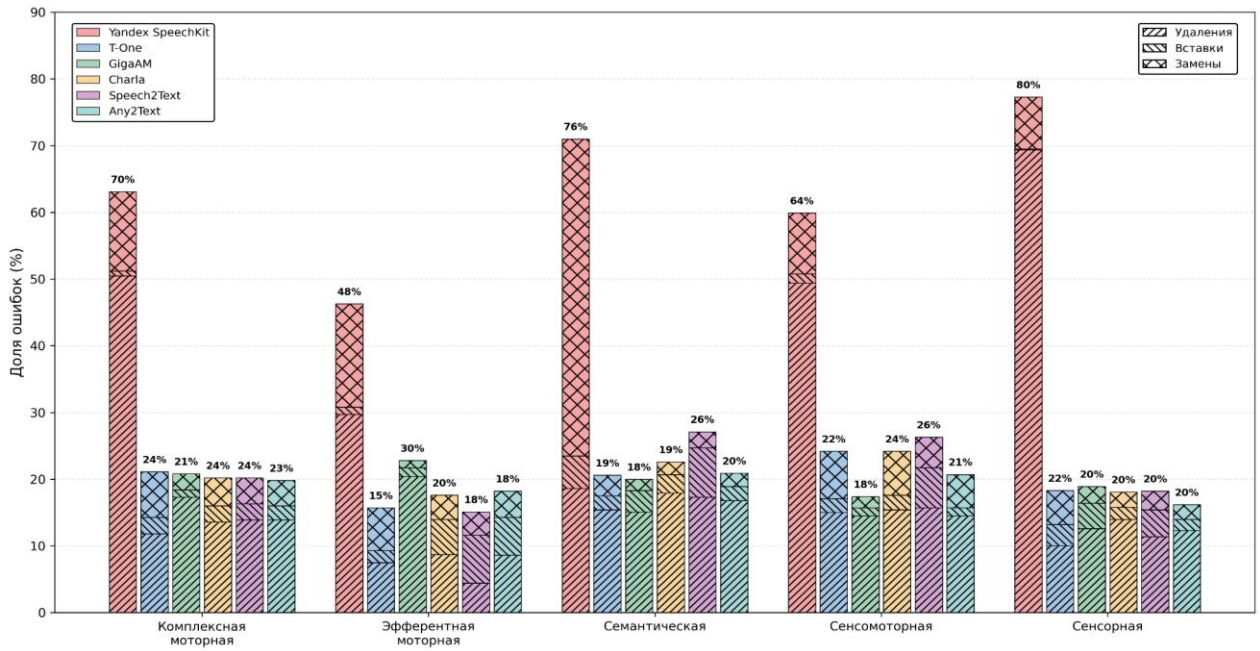


Рис. 2. CER и соотношение вставок/удалений/замен по разным типам афазии.

Поскольку в датасете представлены записи и нейротипичных информантов (монолог и диалог), то при анализе результатов обработки записей с различной степенью выраженности речевого нарушения (рис. 3 и 4) заметен четкий градиент: все модели показывают рост WER от нейротипичных к тяжелой афазии. Исключение составляет Yandex SpeechKit: у этой модели WER и так высокий на нейротипичных (89%), и при афазии он почти не растет, потому что «потолок» достигнут. GigaAM демонстрирует самые низкие значения WER, которые умеренно повышаются вместе с ростом тяжести нарушения: 17% (нейротипичные) → 24% (легкая) → 32% (средняя) → 35% (тяжелая). T-One и Charla показывают наиболее крутой подъем от легкой к тяжелой (40% → 47% и 30% → 41% соответственно).

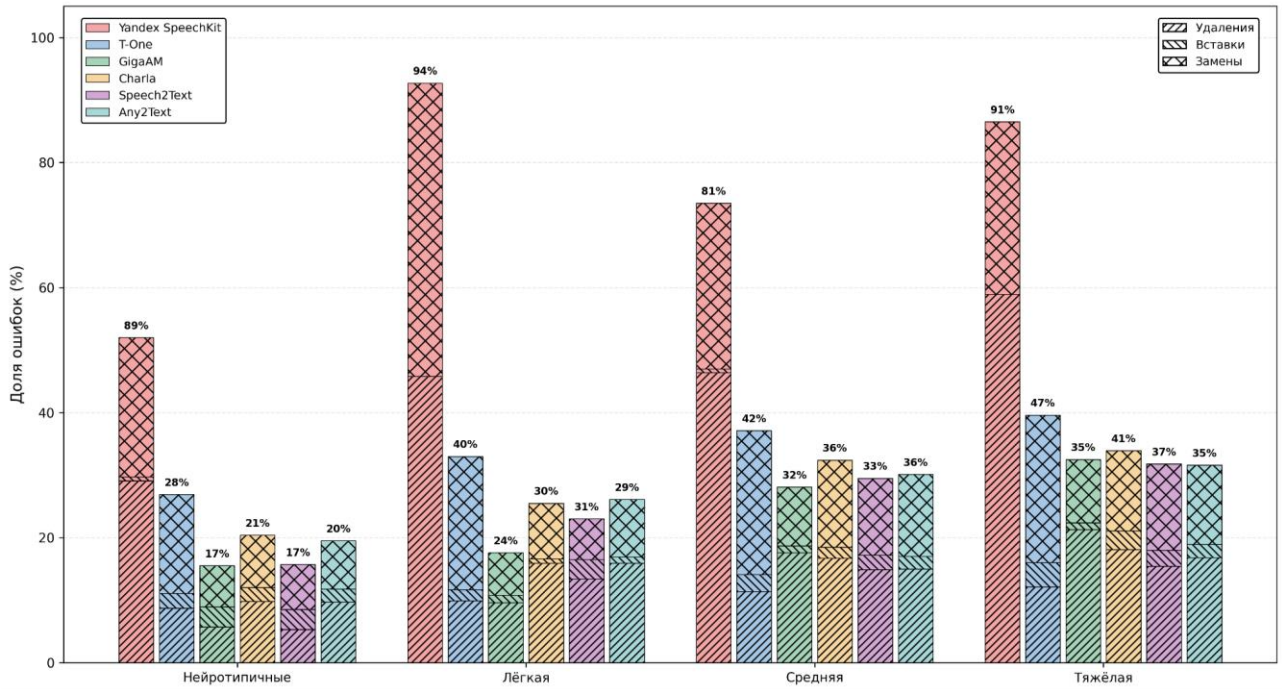


Рис. 3. WER и соотношение вставок/удалений/замен по степени тяжести афазий.

При распознавании речи, имеющей разную степень речевой девиации, на уровне символов (CER) все модели демонстрируют значительно более низкий процент ошибки, чем на уровне слов (WER). Это говорит о том, что большинство ошибок связано с неправильным определением границ слов, а не с искажением отдельных букв. Остальные тенденции коррелируют с WER. Однако, например,

T-One показывает CER 15–25% с преобладанием замен (4.3–7.1%) над удалениями (7.1–13.1%) – модель сохраняет слова, но искажает отдельные символы внутри них. Speech2Text, Charla и Any2Text занимают промежуточное положение (CER 18–26%) со сбалансированным ростом всех типов изменений по сравнению с эталонными транскриптами, при этом у Speech2Text на легкой степени нарушения CER (18%) выше, чем у GigaAM (15%), но ниже, чем у Any2Text (20%).

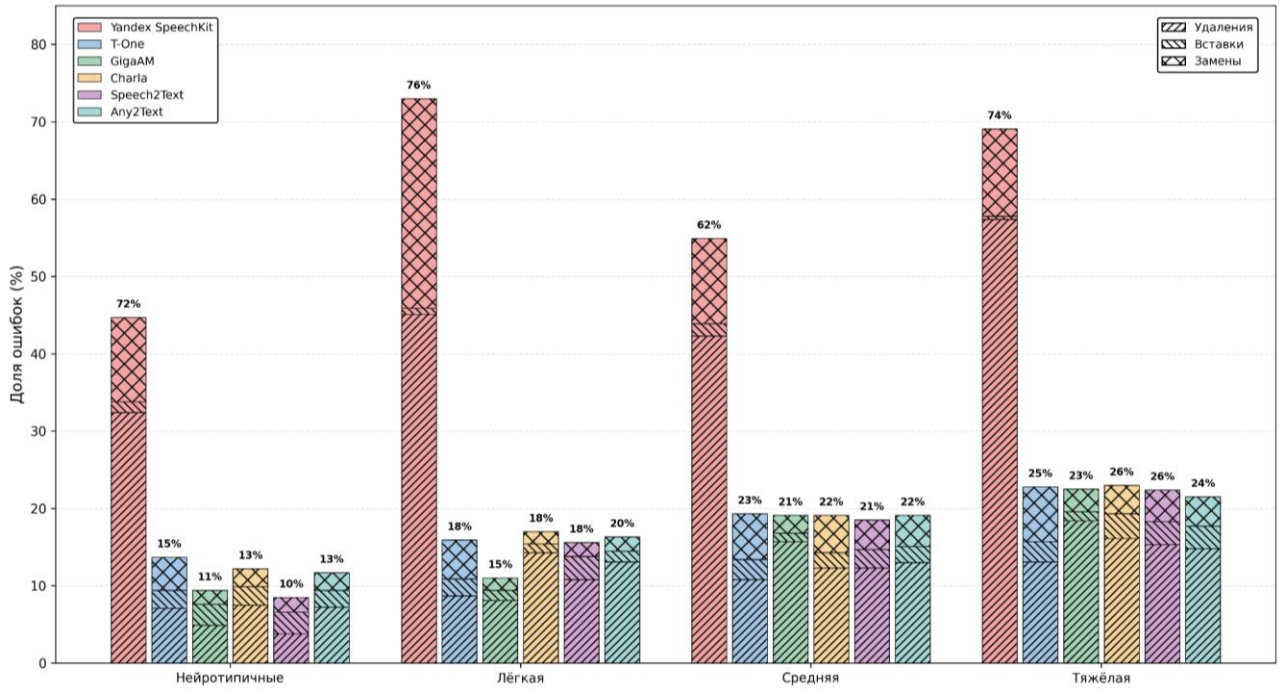


Рис. 4 CER и соотношение вставок/удалений/замен по степени тяжести афазий.

Сравним теперь модели по типам речи (рис. 5 и 6). Медианные значения WER и CER для всех моделей, за исключением Yandex SpeechKit, на уровне слов демонстрируют ожидаемую тенденцию: самые низкие показатели ошибок наблюдаются при распознавании нейротипичных записей, тогда как самые высокие – при обработке пересказов у ПА. Yandex SpeechKit имеет самые высокие показатели ошибок как для нейротипичных, так и для афазических записей диалогов; распознавание же монологов этой моделью значимо лучше для нейротипичных, чем для ПА. GigaAM имеет наиболее сбалансированную и устойчивую производительность на всех типах речи, превосходя другие модели по WER в самом сложном случае – пересказе, хотя при этом использует чаще всего удаления.

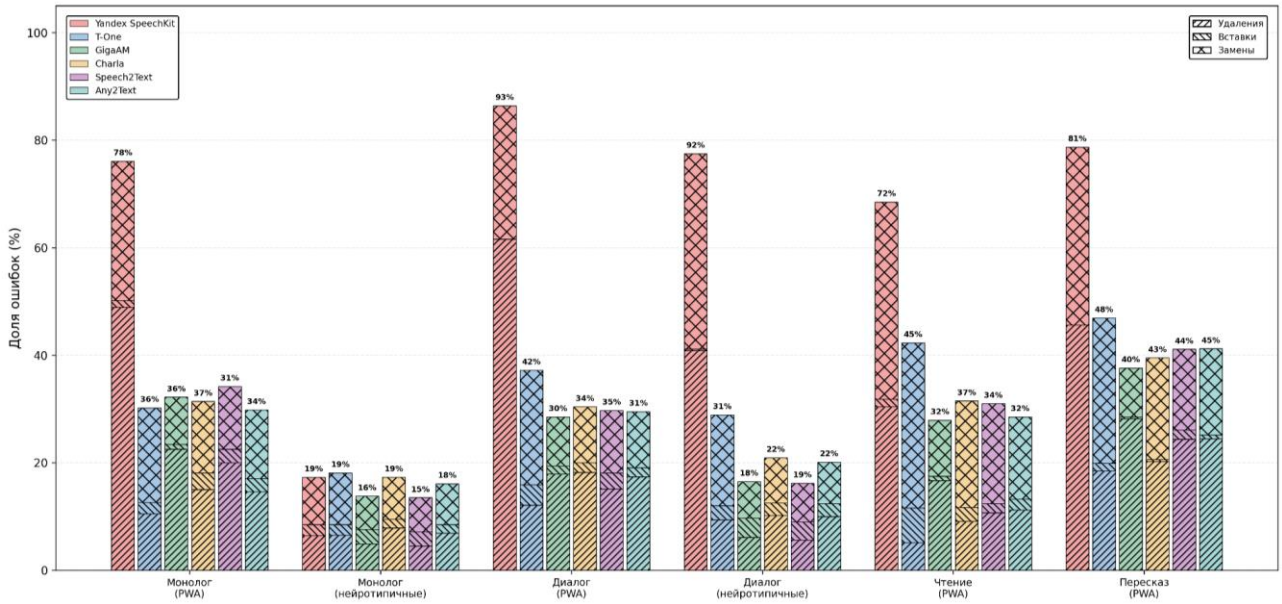


Рис. 5. WER и соотношение вставок/удалений/замен по типу речи у пациентов с афазией.

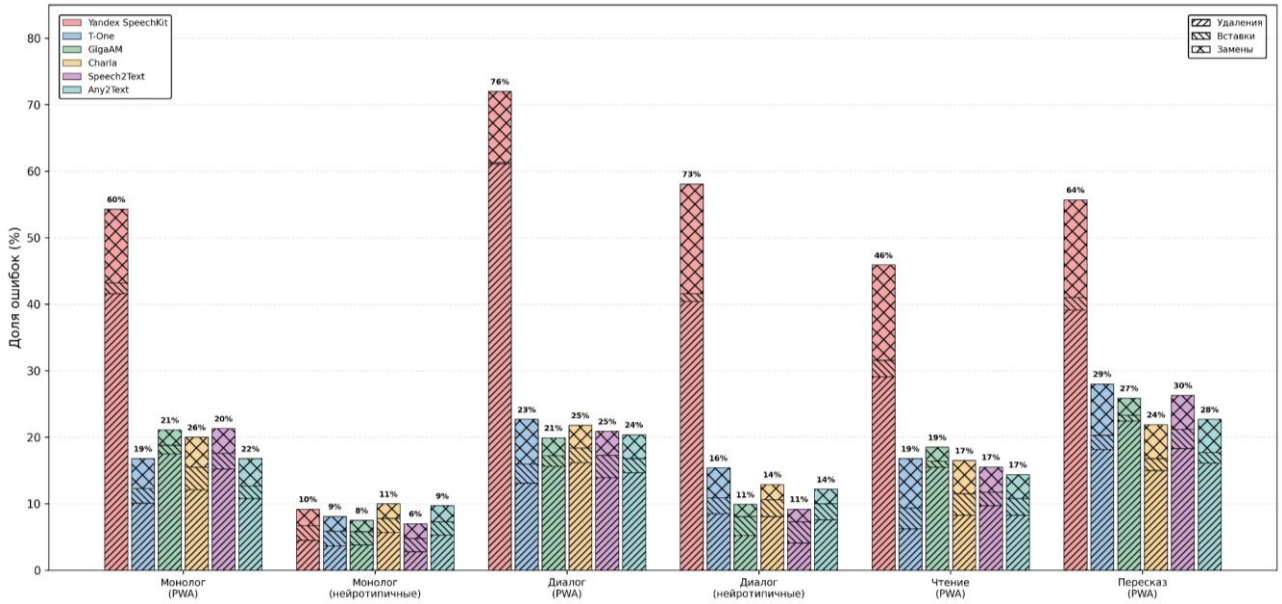


Рис. 6. CER и соотношение вставок/удалений/замен по типу речи.

Speech2Text показывает сравнительно хорошие результаты на монологах (WER – 31%, CER – 20%), но теряет свое преимущество в интерактивных или нарративных контекстах афазической речи. Any2Text и Charla имеют стабильные средние результаты, в то время как T-One отстает, приоритезируя замены.

Таким образом, в результате проведенного анализа установлены три различные стратегии моделей при распознавании речи в условиях неопределенности, в нашем случае – при обработке атипичной речи. GigaAM и Yandex SpeechKit используют «паническое укорачивание» – агрессивное сжатие распознанного текста по сравнению с оригиналом. T-One старается сохранить лексическое наполнение в ущерб точности на уровне символов. Any2Text, Speech2Text и Charla демонстрируют «стабильную» стратегию, при которой длина выходного текста практически не зависит от снижения качества речевых данных (например, при усугублении грубости речевого нарушения у говорящего).

Табл. 2. Результаты сравнения замеров метрик качества распознавания русскоязычных движков автоматического распознавания речи.

ПА						
ASR Model	Yandex SpeechKit	T-One	GigaAM	Charla	Speech2Text	Any2Text
WER	**88%**	43%	**32%**	37%	35%	34%
CER	**69%**	23%	**21%**	23%	23%	23%
Нейротипичные информанты						
ASR Model	Yandex SpeechKit	T-One	GigaAM	Charla	Speech2Text	Any2Text
WER	**89%**	28%	**17%**	21%	**17%**	20%
CER	**72%**	15%	11%	13%	**10%**	13%

Yandex SpeechKit показывает наихудшее значение метрики WER (88%) (табл.2) на фоне чрезвычайно высокого количества удалений и замен. Тем не менее GigaAM имеет самый низкий показатель WER (32%) за счет большого количества удалений на фоне минимального количества вставок и замен. T-One (WER = 43%) выделяется самой высокой долей замен на уровне слов. Большой разрыв между CER и WER (WER = 43% при CER = 23%) для T-One позволяет пред-

положить, что модель правильно распознает большинство символов, но допускает одно-два незначительных искажения на слово – этого как раз достаточно, чтобы все слово было отвергнуто метрикой WER. Такое расхождение указывает на фонетическую близость ошибок: модель испытывает трудности с распознаванием редких или искаженных фонем, типичных для афазической речи, но лексический инвентарь текста в значительной степени сохраняется. Остальные модели – Charla (WER = 37%), Speech2Text (WER = 35%) и Any2Text (WER = 34%) – демонстрируют примерно схожие значения метрик (табл.2). Однако само по себе низкое значение WER или CER не может быть абсолютным показателем качества распознавания. То же самое можно сказать и о соотношении удалений/вставок/замен.

Проанализируем качественно лингвистический характер удалений/вставок/замен и попробуем спрогнозировать, какие черты коммуникативного стиля фиксируются моделями при распознавании и как это может повлиять на характер социального взаимодействия в случае, если клиент с нарушением речи решит воспользоваться услугами голосового помощника в банке или ином учреждении.

Например, при распознавании семантической афазии GigaAM стремится структурировать высказывание, приводя его к некоторым типовым синтаксическим структурам, характерным для разговорной речи. Следует отметить, что семантическая афазия сложна тем, что пациент достаточно случайным образом употребляет служебные слова, ему не удается передача пространственных и причинно-следственных отношений, неправильно употребляются слова и конструкции с непрямым значением. Например, пациент на вопрос логопеда о том, кем он работал, пытается сказать, что он ходил на судне в море, был моряком, но ему удается передать только простую пространственную схему – некоторый сверток смысла, передающий ритм его профессиональной жизни «туда-сюда (ходил в рейсы)»:

(1) [эталонная расшифровка] *мне а я это как его вот туда-сюда туда-сюда*;

(2) [аутпут модели]: *Мне это как его? Вот – Туда-сюда, туда-сюда.*

Речь пациента изобилует повторами (парафазиями) двух первых слогов слова *конечно*: *коне коне коне.* Модель удаляет часть повторов, а то, что было

неразборчиво, заменяет на другие асемантические повторы – *отрадио, радио*, но при этом отчетливо делит транскрипт на короткие ясные диалогические реплики:

(4) [аутпут модели] *Дети живут отдельно отрадио, радио. – Отдельно. – Да, конечно*.

По мере ухудшения качества речи (при движении от слабой степени выраженности речевого нарушения к сильной) GigaAM удаляет хезитации, числительные (например, *тринадцать*), имена-отчества (*Иван Сергеевич*) и многосложные слова (*вернувшись*), заменяя их на более короткие формы: хезитации, предлоги (чаще всего – *в*, *на*), междометия, обсценизмы, местоимения *я* и *ты*, производятся замены литературных форм на более разговорные (*поскользнулся → подскользнулся*). Таким образом, речь говорящего в расшифровках GigaAM предстает как более разговорная, диалогическая, используются короткие, но емкие реплики.

Модель от Yandex лучше всего справляется с речью пациентов с эфферентной моторной афазией. Для такой речи характерны многократные повторы слогов, замены одних звуков на другие (похожие):

(5) [эталонная расшифровка] *доча есть двадцать два года и иже оза занимается оу не знаю она живет ну две нрзб учиться не знаю пор пор званию медик полощ зубы протезирует*.

Yandex SpeechKit правильно расшифровывает ключевое числительное *двадцать два* и глагол *занимается*, но остальные слова заменяет рандомными формами: междометиями (*ага*, *ну*) и формами, частотными для речи молодежи (*ВК, Про образование* (название подкаста)) и игрового сленга (Зерай и Карвез – имена персонажей World of Warcraft):

(6) [аутпут модели] *это Надо двадцать два года и Ага занимается Она жалеет Ну два Катя Вк Про образование Карвез зерай*.

Yandex SpeechKit испытывает наибольшие трудности при распознавании речи пациентов с уже упоминавшейся выше семантической афазией. Например, тот же фрагмент, что и в примере 1, данная модель распознает как:

(7) *Мы живем техникой Что выжили не теряя ни один Благодарю Челябина джэйли диана Нет То О это ж деньги И какими же ладонями была ему музейными какая то феноменальная ночь*.

Рандомно подобранные лексемы отражают, скорее, коммуникативный стиль героев манг, нежели повседневную речь рядового говорящего. По мере огрубления речевого расстройства в записях (от легкой до тяжелой степени) модель Yandex SpeechKit заменяет слова случайным образом на другие существующие, часто выступающие как контекстуальные синонимы (*чиновников → хирургов*, т. е. название профессий, видов деятельности) или синтагматические соседи (*много → суп*, частотная коллокация *много супа*).

Таким образом, Yandex SpeechKit искажает коммуникативный стиль «диктора» (пациента с речевым расстройством) до неузнаваемости, делая смысловую сторону почти полностью бессмысленной, а форму высказывания – ассоциирующейся с молодым человеком, активно пользующимся социальными сетями и сленгом геймеров.

T-One как будто пытается достигнуть наибольшей интерпретабельности говоримого. Например, фраза в примере 3 распознается очень четко благодаря тому, что все асемантические слова просто удаляются:

(8) *дети живут отдельно отдельно*.

По мере ухудшения качества речи модель обнаруживает постепенное усиление следующих тенденций: удалению подвергаются не только короткие слова (*ой*, *там*), несуществующие формы (*пэцхоус*, *чизля*, *пахля*), многосложные слова (например, *захихикал*, *белоствольная*, *революции*), но и термины родства (*жена, внук, дед, дедок)*; вставляются обсценизмы, производятся замены «с потолка» (*а – нагорская*, *ну – звезда*), а также усечение первого слога оригинального многосложного слова, в результате которого появляется другое, тоже существующее, слово с иным значением (*благодарит – дарит*, *оторвалось – рвалось*, *убежал – жал*). Модель стремится быть морфологически изобретательной и легко манипулирует близкими по фонемному составу словами. Особенно это проявляется при распознавании чтения текста пациентом. Сравните:

(9) ориг – *я косточку бросил за окошко;*

t_one – *я кошечку брошу за окошко*.

Таким образом, порождаемое высказывание при поверхностном знакомстве выглядит достаточно правильным и в большинстве случаев отвечает речевому стандарту, однако передаваемое содержание сильно искажено по сравнению с оригиналом, что порождает у слушающего когнитивные диссонансы

между соседними контекстами. Такая речь, в свою очередь, может ассоциироваться с человеком, имеющим когнитивные нарушения.

Speech2text интересна тем, что на месте неразборчивого фрагмента довольно часто вставляет обсценизмы («нрзб» → «п...з...ц»), а также многократно перебирает варианты по одному услышанному слогу («петь», «пять», «почти», «пас»). Склонность к морфологическим экспериментам приводит к тому, что модель часто производит морфологические замены глаголов по лицу, виду, времени, словообразовательной модели, типа предикации (полная/свернутая): *дочитывала → зачитывала*, *накинув → накинула*; *убежали →сбежали*, *сбросился → бросился*. Еще один любопытный паттерн – замена на «угу» («лезняком» → «угу»), т. е. модель просто «сдается» и ставит маркер согласия вместо непонятного слова. Лучше всего модель справляется с эфферентной моторной афазией, достаточно неплохо подбирая контекстуальные варианты, и опираясь на отдельные услышанные слоги. Так, реплику о дочери, работающей зубным техником (пр.5), модель распознает как:

(10) *дочь есть, двадцать два года ее или она занимается, не знаю, она живет, ну две группы в городе учится не знаю по образованию медик волос зубы паразитизирует,* где просторечное *доча* заменяется на литературное *дочь*, *оза* правильно достраивается до *она*, *ну две нрзб* превращаются в *две группы*, а *пор пор званию медик* – в *по образованию медик.* Преимуществом модели является ее способность, поняв общий контекст (например, образование и род занятий), достраивать непонятое самыми частотными коллокациями. Однако специальное слово из медицинского контекста *протезирует*, было заменено на оценочное общеупотребительное *паразитирует*.

Контекст же из примера с семантической афазией (пример 3) модель же передала идеально с точки зрения узуса, удалив все неразборчивые и эрративные формы:

(11) *Дети живут отдельно от вас? Отдельно, да*.

Иными словами, модель порождает такой коммуникативный стиль, который ассоциируется с говорящим, репрезентирующим кооперативный тип коммуникативной личности, склонным соглашаться с собеседником [12], владеющим русской литературной нормой и имеющим достаточно большой словарный запас, но в качестве междометий нередко использующим обсценизмы.

Any2text выделяет вставки, делающие речь максимально разговорной, даже близкой к просторечной – местоимения я и ты, обсценизмы (*бл...ть, них...я*), имена (например, *Владимир*), термины родства (*жена*, *сын*, *внучка*), неправильные формы. При этом заменяет редкие, неправильные и даже частотные правильные формы выдуманными (ивняк → явнях, защищаль → сосисали, удоками → ударками, взгляда → гляда, принадлежностями → пренаслежностями, насили → насылили). В результате фраза о дочери-зубном технике (пример 5) выглядит так:

(12) *Дочь есть. двадцать два года ее. Или... Она занимается... не знаю она живет ну две группы короткий хочется Не знаю. Борзование медика. Уболость зубы паразитизирует.*

Короткие сегменты, многоточия, маркирующие паузы и хезитации, повторы «не знаю», междометия *ну* делают речь очень разговорной, а *борзование* и *уболость* делают ее странной.

Отметим, что модель Charla демонстрирует аналогичные изменения оригинальной расшифровки. Иными словами, порождаемый моделями коммуникативный стиль может быть ассоциирован с таким говорящим, который владеет исключительно разговорной или просторечной речью, возможно, имеет психические или когнитивные нарушения.

ОБСУЖДЕНИЕ

Проведенные эксперименты показали, что при оценке качества распознавания метрики CER и WER дают только общую оценку, которую необходимо дополнять качественно-количественным анализом вставок, удалений и замен. Так, например, среди трех моделей Charla, Any2text и Speech2text при сопоставимых значениях метрик и соотношений вставок/удалений/замен Speech2text генерирует расшифровки, которые могут быть ассоциированы слушающим с коммуникативным стилем носителя языка, владеющего узуальной повседневной разговорной речью, и склонного соглашаться с собеседником и использовать короткие обсценизмы в качестве междометий. Остальные две модели выдают такие распознанные высказывания, которые реализуют коммуникативный стиль, скорее, свойственный малограмотному человеку, владеющему преимущественно

просторечием и имеющему некоторые психические девиации, например, находящемуся в состоянии алкогольного опьянения или под воздействием психотропных веществ. T-One порождает формально правильные высказывания, которые противоречат изначальному смыслу речевого фрагмента, что может создать у слушающего впечатление о том, что собеседник имеет некоторые когнитивные нарушения.

Лучшая по метрикам модель GigaAM продемонстрировала стратегию панического удаления, но при этом ей удается трансформировать усеченные распознанные фразы в короткие диалогические реплики, соответствующие представлению о живой повседневной диалогической речи. Наихудшая же модель от Yandex, демонстрируя ту же стратегию панического удаления, все же делает вставки и замены, хотя бы в малой степени пытаясь компенсировать удаленное, но использует при этом молодежный сленг и сленг геймеров.

Таким образом, можно заключить что наша гипотеза о коммуникативной предвзятости моделей подтвердилась, но в каждом случае в разной степени. Так, в наибольшей степени это верно для Yandex SpeechKit: по-видимому, модель обучалась на записях спонтанной диалогической речи представителей молодежных сообществ, пользователей социальных сетей, которых компания одновременно рассматривает как своих основных потребителей. Скорее всего, GigaAM обучалась на спонтанных повседневных диалогах взрослых (возможно, 30+) жителей мегаполисов (Москва, Санкт-Петербург), имеющих высшее образование, однако не исключено и обучение на некотором сэмпле атипичной речи, поскольку модель хорошо работает с ошибочными и неполными формами. Speech2text опирается на обучающие выборки записей разговорной речи информантов, возможно, имеющих более низкий образовательный ценз, чем информанты, речь которых «слушала» GigaAM. Можно предположить, что остальные три модели не обучались на повседневной разговорной речи, хотя оказались знакомыми с нормированной речью актеров, дикторов, иных, не спонтанных, говорящих.

В дальнейших исследованиях планируется проверить сформулированные на основе лингвистического анализа признаки коммуникативной предвзятости в экспериментальной работе с информантами-носителями русского языка. В про-

водимой в настоящее время работе мы ставим перед информантами задачу проранжировать шестерки транскриптов по степени субъективной понятности смысла сообщаемого, на следующем этапе экспериментальное задание будет предложено сотрудникам госучреждений и сервисов.

Перспективу исследования составляет также дополнение датасета новыми записями речи пациентов и нейротипичных информантов в тех же условиях и их расшифровками. Несбалансированность записей по типам афазии и степени тяжести речевого нарушения составляет серьезное ограничение данных, которое исследовательская команда в полной мере осознает.

ЗАКЛЮЧЕНИЕ

Проведенный анализ шести систем ASR демонстрирует, что предложенное нами определение коммуникативной предвзятости следует считать рабочим и достаточно продуктивным, поскольку нам удалось установить и описать признаки коммуникативной предвзятости для большинства моделей, участвовавших в экспериментах на датасете RuAphasiaBank. То, в какой степени коммуникативная предвзятость может навредить социальному самочувствию пациентов с афазией при пользовании голосовыми сервисами и ассистентами, только предстоит оценить на практике – в настоящей работе мы лишь смоделировали одну из ситуаций гипотетически. Тем не менее уже сама постановка вопроса обращает нас к широкой проблеме технологической инклюзии, которая требует своего обсуждения.

Благодарности

Статья и данные были подготовлены в рамках Программы фундаментальных исследований НИУ ВШЭ (HSEBR-2025-032). Исследование одобрено Комиссией по внутривузовским опросам и этической оценке эмпирических исследовательских проектов Высшей школы экономики. Авторы являются зарегистрированными правообладателями описанной базы данных RuAphasiaBank: база данных фонограмм речи пациентов с афазией зарегистрирована в Федеральной службе по интеллектуальной собственности Российской Федерации (№ 6.0021-2025).

СПИСОК ЛИТЕРАТУРЫ

1. *Харченко С.В.* Феномен и природа социального настроения // Мир науки, культуры, образования. 2011. № 1. С. 168–170.

2. *Михайлова Л.И.* Социальное самочувствие и восприятие будущего россиянами // Социологические исследования. 2010. № 3. С. 45–49.

3. *World Health Organization.* Priority assistive products list: improving access to assistive technology for everyone, everywhere. Geneva, 2016.
URL: https://www.who.int/publications/i/item/priority-assistive-products-list (дата обращения: 20.08.2026).

4. *Левонтина И.Б.* Об арсенале ксенопоказателей в русском языке // Вопросы языкознания. 2020. № 3. С. 52–77.

5. *Шклярук Е.Я.* Маркеры-ксенопоказатели в русской устной речи: дискуссионный статус единиц // Вестник ЮУрГУ. Серия: Лингвистика. 2025. № 2. С. 70–78.

6. *Markl N., McNulty S.J*. Language technology practitioners as language managers: arbitrating data bias and predictive bias in ASR // Proc. of the 13th Conference on Language Resources and Evaluation (LREC 2022). European Language Resources Association (ELRA), 2022. P. 6328–6339.

7. *Вассерман Л.И., Дорофеева С.А., Меерсон Я.А*. Методы нейропсихологической диагностики: практическое руководство. СПб.: Стройлеспечать, 1997. 123 с.

8. *Лурия А.Р.* Высшие корковые функции человека. 3-е изд. СПб.: Питер, 2008. 128 с.

9. *Колмогорова А.В., Явшиц Е.В., Никольская О.Н., Сугян А.Х*. Сервисы автоматического распознавания речи на русском языке: исследование качества и инклюзивности на примере датасета RuAphasiaBank // Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. 2026. Т. 24, № 1. С. 113–127.

10. *Morris A.C., Maier V., Green P*. From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition // Proc. Interspeech. 2004. P. 2765–2768.

11. *Thennal K.D., James Jesin, Gopinath D.P., Ashraf K.M.* Advocating Character Error Rate for Multilingual ASR Evaluation // arXiv.org. 2024.
https://doi.org/10.48550/arXiv.2410.07400

12. *Седов К.Ф.* Речевое поведение и типы языковой личности // Культурно-речевая ситуация в современной России. Екатеринбург, 2000. С. 298–312.

COMMUNICATIVE BIAS IN ASR RECOGNITION OF ATYPICAL RUSSIAN SPEECH

A .V. Kolmogorova[1] [0000-0002-6425-2050], E. V. Yavshits[2] [0000-0001-6344-1565], M. S. Mayorova[3] [0009-0009-2106-7689], S. V. Dmitrieva[4] [0009-0001-1949-5360]

[1–4]*National Research University Higher School of Economics, Moscow, Russia*

[1]akolmogorova@hse.ru, [2]kate_yavshitz@mail.ru, [3]svyatoslavna.m@gmail.com, [4]sofiadmitrieva82@gmail.com

Abstract

The article addresses the problem of communicative bias in automatic speech recognition systems when processing atypical speech in Russian. The relevance of the study is driven by the active integration of artificial intelligence technologies into the social sphere and public administration in Russia, which presupposes ensuring their accessibility for various categories of citizens, including people with speech disorders. The authors introduce the concept of communicative bias of ASR models as a property of a system that, under conditions of uncertainty (atypical speech, noise, pauses), makes decisions about filling semantic and acoustic gaps based on its own training data rather than the speaker's actual communicative intention. The empirical basis of the study is the RuAphasiaBank corpus, which includes speech recordings from patients with various types of aphasia of varying severity, as well as neurotypical informants. For comparative analysis, six ASR models operating with Russian were selected: two open-source (GigaAM, T-One) and four commercial (Yandex SpeechKit, Any2Text, Charla, Speech2Text). Recognition quality was assessed using WER and CER metrics, as well as a qualitative analysis of the ratio of insertions, deletions, and substitutions at the word and character levels. The results show that the models implement different strategies when encountering atypical speech: “panic shortening” (GigaAM, Yandex SpeechKit), “lexical preservation” at the expense of accuracy (T-One), and a “stable” strategy of balanced errors (Any2Text, Speech2Text, Charla). Linguistic analysis

reveals that each model produces a specific distorted communicative style, ranging from colloquial dialogic speech to a style associated with cognitive impairments or the use of youth slang. Based on the obtained data, potential risks to the social well-being and legal security of aphasia patients when using voice assistants in public institutions are modeled.

Keywords*: speech technology, speech database, automatic speech recognition, atypical speech, aphasia, communicative bias, voice assistants, assistive technologies, inclusivity.*

REFERENCES

1. *Kharchenko S.V.* Fenomen i priroda sotsialnogo nastroeniya // Mir nauki, kultury, obrazovaniya. 2011. № 1. P. 168–170.

2. *Mikhaylova L.I.* Sotsialnoe samochuvstvie i vospriyatie budushchego rossiyanami // Sotsiologicheskie issledovaniya. 2010. № 3. P. 45–49.

3. *World Health Organization.* Priority assistive products list: improving access to assistive technology for everyone, everywhere. Geneva, 2016.
URL: https://www.who.int/publications/i/item/priority-assistive-products-list (accessed: 20.08.2026).

4. *Levontina I.B.* Ob arsenale ksenopokazateley v russkom yazyke // Voprosy yazykoznaniya. 2020. № 3. P. 52–77.

5. *Shklyaruk E.Ya.* Markery-ksenopokazateli v russkoy ustnoy rechi: diskussionnyy status edinits // Vestnik YuUrGU. Seriya: Lingvistika. 2025. № 2. P. 70–78.

6. *Markl N., McNulty S.J*. Language technology practitioners as language managers: arbitrating data bias and predictive bias in ASR // Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022). European Language Resources Association (ELRA), 2022. P. 6328–6339.

7. *Vasserman L.I., Dorofeeva S.A., Meerson Ya.A.* Metody neyropsikhologicheskoy diagnostiki: prakticheskoe rukovodstvo. SPb.: Stroylespechat, 1997. 123 p.

8. *Luriya A.R.* Vysshie korkovye funktsii cheloveka. 3-e izd. SPb.: Piter, 2008. 128 p.

9. *Kolmogorova A.V., Yavshits E.V., Nikolskaya O.N., Sugyan A.Kh.* Servisy avtomaticheskogo raspoznavaniya rechi na russkom yazyke: issledovanie kachestva i

inklyuzivnosti na primere dataseta RuAphasiaBank // Vestnik NGU. Seriya: Lingvistika i mezhkulturnaya kommunikatsiya. 2026. Vol. 24, № 1. P. 113–127.

10. *Morris A.C., Maier V., Green P*. From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition // Proc. Interspeech. 2004. P. 2765–2768.

11. *Thennal K.D., James Jesin, Gopinath D.P., Ashraf K.M*. Advocating Character Error Rate for Multilingual ASR Evaluation // arXiv.org. 2024. https://doi.org/10.48550/arXiv.2410.07400

12. *Sedov K.F.* Rechevoe povedenie i tipy yazykovoy lichnosti // Kulturno-rechevaya situatsiya v sovremennoy Rossii. Ekaterinburg, 2000. P. 298–312.

СВЕДЕНИЯ ОБ АВТОРАХ



КОЛМОГОРОВА Анастасия Владимировна – Профессор, доктор филологических наук, заведующая лабораторией языковой конвергенции и профессор департамента филологии НИУ ВШЭ – Санкт-Петербург. Является методологическим руководителем проекта. Под ее руководством сформировалась научная школа, объединяющая более 20 исследователей, 14 из которых – кандидаты наук. Имеет опыт реализации прикладных проектов, включая разработку систем автоматической обработки текстовых и речевых данных по заказам промышленных и коммерческих организаций. Владеет теоретическими основами, так и практическими инструментами компьютерной лингвистики.

Anastasia Vladimirovna KOLMOGOROVA is a Professor, Doctor of Philological Sciences, head of the Laboratory for Language Convergence, and professor at the Department of Philology at HSE University – St. Petersburg. She serves as the project's methodology lead. She heads a scientific school that currently brings together over 20 researchers, 14 of whom have successfully defended their PhD theses under her supervision. Prof. Kolmogorova has extensive experience in managing applied projects, including those involving automated processing of textual and speech data commissioned by industrial and commercial enterprises. She is proficient in both the theoretical foundations and practical methods of computational linguistics.

email: akolmogorova@hse.ru

ORCID: 0000-0002-6425-2050



ЯВШИЦ Екатерина Валерьевна – специалист в области речевых технологий и компьютерной лингвистики, чья профессиональная деятельность сосредоточена на междисциплинарном пересечении автоматического распознавания речи и обработки естественного языка (NLP). Научно-исследовательские интересы охватывают широкий спектр направлений, включая цифровую обработку речевых сигналов, речевую реабилитацию, устойчивое распознавание диктора и речи в условиях акустических шумов, а также речевую аналитику и ее применение для улучшения клиентского опыта (CX).

Ekaterina Valerievna YAVSHITS is a specialist in speech technologies and computational linguistics, whose professional work lies at the interdisciplinary intersection of automatic speech recognition (ASR) and natural language processing (NLP). The author's research interests cover a broad spectrum of areas, including digital speech signal processing, speech rehabilitation, robust speaker and speech recognition in noisy acoustic environments, as well as speech analytics and its application to enhancing customer experience (CX).

email: kate_yavshitz@mail.ru

ORCID: 0000-0001-6344-1565



МАЙОРОВА Милана Святославна – специалист в области компьютерной лингвистики, работающий на стыке анализа данных, машинного обучения и обработки естественного языка. В сферу научных интересов входят прикладная, когнитивная и корпусная лингвистика, а также изучение нарушенной речи.

Milana Svyatoslavna MAYOROVA is a specialist in computational linguistics, working at the intersection of data analysis, machine learning, and natural language processing. Research interests include applied, cognitive, and corpus linguistics, as well as the study of disordered speech.

email: svyatoslavna.m@gmail.com

ORCID: 0009-0009-2106-7689



ДМИТРИЕВА Софья Владимировна – магистрант программы «Языковые технологии в бизнесе и образовании» НИУ ВШЭ – Санкт-Петербург, специалист по разметке данных ООО «КВИНТ». Занимается разметкой текстовых и аудиоданных для голосовых роботов и контролем качества их работы.

Sofia Vladimirovna DMITRIEVA is a master's student in the «Language Technology in Business and Education» program at HSE University – St. Petersburg, and a data labeling specialist at OOO «KVINT». She labels text and audio data for voice bots and monitors the quality of their performance. She has experience preparing, testing, and optimizing dialogue scenarios for voice assistants.

email: sofiadmitrieva82@gmail.com

ORCID: 0009-0001-1949-5360

Материал поступил в редакцию 21 августа 2026 года

УДК 004.822:004.912

ОТ РАЗМЕЧЕННОГО КОРПУСА К ОНТОЛОГИИ: ФОРМАЛЬНАЯ МОДЕЛЬ МНОГОУРОВНЕВОЙ СЕМАНТИЧЕСКОЙ РАЗМЕТКИ ФЕЙКОВЫХ НОВОСТЕЙ В КАЗАХСКО-РУССКОМ МЕДИАПРОСТРАНСТВЕ (KazFakeCorpus/KazFakeOnto)

А. А. Некесова[1] [0009-0001-7642-0648], Е. А. Сидорова[2] [0000-0001-8731-3058], Ж. Б. Ламашева[3] [0000-0002-9535-2636], М. А. Самбетбаева[4] [0000-0001-9358-1614], А. С. Еримбетова[5] [0000-0002-2013-1513], М. Ж. Калдарова[6] [0000-0001-7494-9794], А. А. Назымхан[7] [0009-0007-4366-6709]

[1, 3, 4, 6, 7]*Международный научный комплекс «Астана», г. Астана, Казахстан*

[1, 3, 4, 7]*Евразийский национальный университет имени Л. Н. Гумилева, г. Астана, Казахстан*

[2]*Институт систем информатики им. А. П. Ершова СО РАН, г. Новосибирск, Россия*

[5]*Институт информационных и вычислительных технологий КН МНВО РК, г. Алматы, Казахстан*

[5]*Школа инженерии и информационных технологий, Университет МЕТА, г. Алматы, Казахстан*

[6]*Международный университет Астана, г. Астана, Казахстан*

[1]nekessova_aa_1@enu.kz, [2]lsidorova@iis.nsk.su, [3]lamasheva_zhb@enu.kz, [4]sambetbayeva_ma_1@enu.kz, [5]aigerian8888@gmail.com, [6]mira_kaldarova@aiu.edu.kz, [7]nazymkhan_aa@enu.kz

Аннотация

Рассмотрена проблема формального представления знаний о дезинформации в малоресурсном двуязычном медиапространстве. Автоматическое обнаружение фейковых новостей традиционно сводится к бинарной классификации REAL/FAKE, которая не описывает внутреннюю структуру недостоверного сообщения: тип фейкового контента, примененную технику дезинформации, коммуникативное намерение автора, характеристики источника и доказательной базы.

© А. А. Некесова, Е. А. Сидорова, Ж. Б. Ламашева, М. А. Самбетбаева, А. С. Еримбетова, М. Ж. Калдарова, А. А. Назымхан, 2026.
Данная статья распространяется на условиях международной лицензии Creative Commons License Attribution 4.0 International (CC BY 4.0).

Представлены два взаимосвязанных ресурса. Первый – это двуязычный корпус KazFakeCorpus: 4276 текстов на казахском и русском языках, сбалансированных по классам внутри каждого языка; класс REAL сформирован из официальных материалов портала Gov.kz, класс FAKE – из их контролируемых трансформаций. Разметку по многоуровневой семантической схеме выполнили в среде Label Studio два независимых эксперта. Второй ресурс – это онтология KazFakeOnto, переводящая схему разметки из плоского реляционного представления в формальную модель на языке OWL 2 DL: 38 классов, 26 объектных свойств, 22 свойства данных, 112 индивидов, 1155 утвержденных триплетов. Одиннадцать определяемых классов не утверждаются вручную, а выводятся ризонером из значений свойств. Непротиворечивость проверена с помощью HermiT, запросный слой апробирован на пятнадцати SPARQL- и тринадцати DL-запросах.

Показано, что онтологический слой обеспечивает воспроизводимую спецификацию схемы разметки, автоматический вывод содержательных категорий и единую формальную среду для постановки корпусно-лингвистических вопросов и анализа ошибок моделей машинного обучения. Обсуждены ограничения подхода и направления масштабирования онтологии на полный корпус и на естественно возникающую дезинформацию.

Ключевые слова: *обнаружение фейковых новостей, онтология, OWL 2, SPARQL, логический вывод, двуязычный корпус, казахский язык, многоуровневая аннотация, техники дезинформации, объяснимый NLP, малоресурсные языки, KazFakeCorpus, KazFakeOnto.*

ВВЕДЕНИЕ

Распространение цифровых медиа привело к росту объема недостоверных сообщений, и автоматическое обнаружение фейковых новостей закрепилось как самостоятельная задача обработки естественного языка [1]. Качество таких систем в решающей степени определяется качеством корпусных ресурсов, а подавляющее большинство доступных наборов данных построено для высокоресурсных языков [2, 3]. Для казахского языка проблема стоит особенно остро: при наличии корпусов общего назначения [4] ресурсы, пригодные для анализа дезинформации, практически отсутствуют. Дополнительную сложность создает

двуязычная медиасреда Казахстана, в которой казахский и русский языки используются параллельно.

Существенным ограничением большинства ресурсов является сама постановка задачи. Бинарная схема REAL/FAKE фиксирует итоговый вердикт, но не сохраняет тех сведений, ради которых вердикт и выносится: какого рода искажение допущено, какими риторическими средствами оно оформлено, на что рассчитывал автор, насколько надежен источник. Между тем именно эти сведения нужны и эксперту-фактчекеру, и разработчику объяснимой модели.

В работах [5–7] нами предложены многоуровневая схема семантической разметки фейковых новостей и построенный по ней двуязычный корпус KazFakeCorpus. Схема реализована в среде Label Studio в виде набора связанных сущностей и полей и представляет собой, по существу, ER-модель предметной области. Такая форма удобна для организации ручной разметки, однако семантика полей задается в ней неформально, корректность заполнения не контролируется автоматически, а агрегирование сведений по нескольким уровням требует отдельного программного кода.

Настоящая статья является расширенной версией доклада [7], представленного на конференции «Диалог-2026». По отношению к [5, 6] новыми являются онтологическая формализация схемы (разд. 4), запросный слой и результаты логического вывода (разд. 5). Онтология не заменяет корпус, а надстраивается над ним: она фиксирует смысл каждого уровня разметки в виде логических аксиом, позволяет ризонеру выводить принадлежность материала к содержательным категориям вместо ручного проставления меток и предоставляет общий декларативный язык запросов к разметке.

Вклад работы состоит в следующем: многоуровневая схема разметки формализована средствами OWL 2 DL; построена онтология KazFakeOnto, охватывающая уровни разметки, участников аннотирования, источники, доказательную базу, эвиденциальные маркеры, модели детекции и порождаемые ими объяснения; предложены одиннадцать определяемых классов, выводимых ризонером из значений свойств, а непротиворечивость модели подтверждена ризонером HermiT; апробирован запросный слой из пятнадцати SPARQL- и тринадцати DL-запросов.

1. БЛИЗКИЕ ПО ТЕМАТИКЕ ИССЛЕДОВАНИЯ

Корпусные ресурсы. Состояние области представлено в обзоре [1]: подавляющее большинство ресурсов построено для высокоресурсных языков и ограничено бинарной разметкой достоверности; социальный контекст и метаданные учитываются лишь в отдельных наборах данных – FakeNewsNet [8] (≈23 тыс. материалов) и MCFEND [9] (≈23.8 тыс.). Наиболее глубоко размечен немецкий корпус DeFaktS [10]: ≈ 105 тыс. сообщений, десять признаков дезинформации одновременно со степенью достоверности. Для малоресурсных языков доступные корпуса невелики: 900 документов для урду [2], около 15 тыс. записей для курдского языка [3]. KazFakeCorpus (4276 текстов) уступает крупнейшим ресурсам по объему и отличается происхождением класса FAKE: материалы порождены контролируемой трансформацией верифицированных сообщений, а не собраны в естественной среде. Такой дизайн обеспечивает попарную сопоставимость классов, но ограничивает внешнюю валидность (разд. 6).

Многоуровневые схемы разметки. Дезинформация представляет собой не однородное явление, а совокупность форм и приемов [11], что мотивирует переход от бинарного вердикта к многомерным схемам. Наиболее развитой линией здесь является разметка манипулятивных приемов: в SemEval-2020 Task 11 [12] выделены восемнадцать пропагандистских техник с привязкой к текстовым фрагментам. Таксономия TAXODIS [13] систематизирует лингвистические признаки дезинформационного дискурса и опубликована в виде SKOS-словаря – первый известный нам случай формальной сериализации схемы разметки дезинформации. Существенно, однако, что SKOS фиксирует иерархию понятий и не поддерживает логического вывода: словарь задает допустимые значения меток, но не позволяет вычислять производные категории материала. Именно этот разрыв составляет предмет настоящей работы, которая не расширяет перечень меток [5, 6], а изменяет способ их представления.

Формальные модели и графы знаний. Словарь ClaimReview консорциума schema.org [14] стандартизирует публикацию результатов фактчекинга и служит основой для агрегирующих графов знаний: ClaimsKG [15] собирает проверенные утверждения с нормализованными вердиктами, CimpleKG [16] расширяет этот

подход до непрерывно обновляемого ресурса, объединяющего данные 77 фактчекинговых организаций и свыше 200 тыс. документов на 26 языках. Эти ресурсы отвечают на вопрос, кто, когда и с каким вердиктом проверил утверждение, но не описывают внутреннюю структуру недостоверного сообщения – примененную технику, тип искажения, намерение автора.

Онтология Fandet [17] формализует средствами OWL контекстные признаки публикации в социальных медиа, но описывает преимущественно ее окружение – автора, платформу, распространение, а не семантику искажения, и не содержит конструкций логического вывода. Работа [18], выполненная в казахстанском контексте, близка к настоящей по материалу и инструментам, однако в ней моделируются акторы, целевые аудитории и эмоциональные оценки информационного воздействия, а онтология служит средством структурирования графа; в KazFakeOnto же содержательные категории материала выводятся ризонером из значений элементарных свойств.

Отметим разделение труда между OWL и языком ограничений SHACL: требования вида «поле обязательно заполнено» в OWL не проверяются вследствие допущения открытого мира, поэтому контроль полноты разметки требует SHACL, тогда как OWL 2 DL необходим для вывода производных категорий; в настоящей работе формализуется вторая задача. Архитектурная роль онтологии как слоя интеграции над разнородными ресурсами проработана в проектах цифровых научных экосистем, характерным примером которых служит экосистема OntoMath [19]; настоящая работа переносит эту идею – онтология над корпусом, а не вместо него – в область анализа дезинформации.

Таким образом, лакуна локализуется по трем признакам одновременно. Формальные ресурсы уровня связанных данных [14–16] ориентированы на метаданные проверки и не описывают внутреннюю структуру сообщения. Модели, описывающие эту структуру [17, 18], либо ограничены контекстом публикации, либо не доводят формализацию до логического вывода. Схемы многоуровневой разметки [5, 6, 10, 13] формализованы средствами, не поддерживающими вывода, и ни одна из них не охватывает процесс аннотирования, результаты моделей детекции и порождаемые ими объяснения в едином графе; для казахско-русского материала не существует ресурса ни одного из перечисленных типов.

2. KazFakeCorpus: ПОСТРОЕНИЕ КОРПУСА

Общая схема исследования приведена на рис. 1: верхний ряд соответствует построению корпуса, нижний – его онтологической формализации и запросному слою.

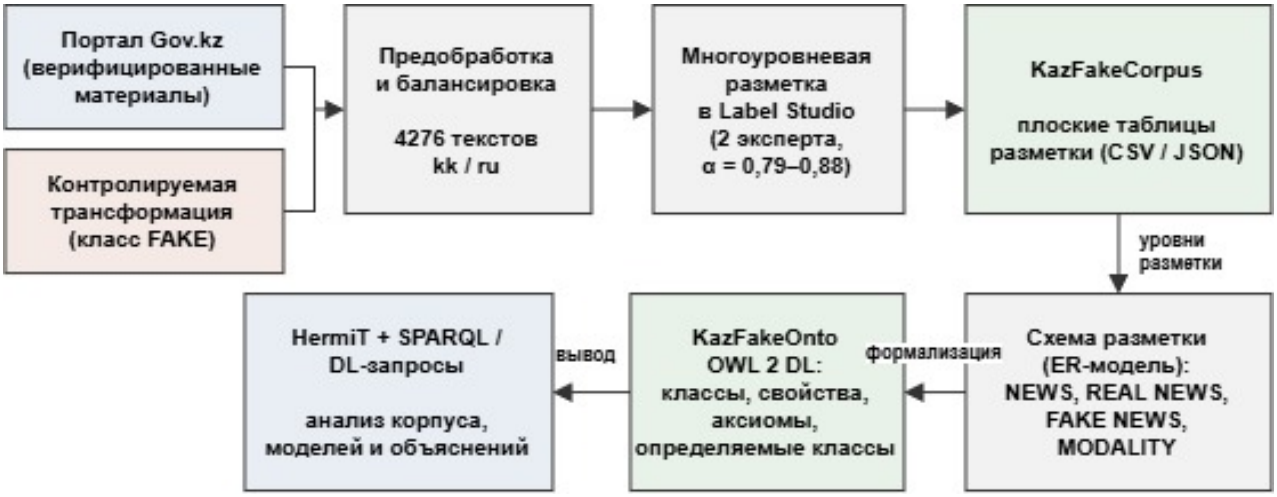


Рис. 1. Схема исследования: от сбора данных к онтологической модели и запросному слою.

Материалы класса REAL. Официальные новостные материалы собирались с портала gov.kz по тематике цифровизации, искусственного интеллекта и государственных инициатив. Ограничение корпуса одной предметной областью снижает тематическую вариативность и обеспечивает контролируемое сопоставление классов. Материалы прошли дополнительную проверку экспертами предметной области до начала разметки.

Материалы класса FAKE. Тексты класса FAKE создавались путем контролируемой трансформации материалов класса REAL с сохранением общей тематической структуры и внесением целенаправленных искажений, соответствующих заранее определенным техникам дезинформации. Для генерации использовалась модель GPT-5 в интерфейсе ChatGPT: на вход подавался исходный текст T и системная инструкция, предписывающая проанализировать его фактологическую структуру, выбрать релевантные значения категорий FAKE_type и DISINFORMATION_TECHNIQUE и выполнить преобразование с сохранением тематической связности; результатом являлся синтетический пример T’. Каждый порожденный текст проходил экспертную проверку; материалы с прямым копированием исходного текста, логическими противоречиями, явными языковыми

ошибками либо не соответствовавшие заданной технике исключались.

Такой дизайн минимизирует вариативность, не связанную с дезинформацией: каждый экземпляр класса FAKE порожден из соответствующего экземпляра класса REAL, поэтому наблюдаемые различия локализованы на уровне семантической манипуляции. Отношение порождения фиксируется свойствами generatedBy и derivedFromSeed (разд. 4), что позволяет ставить вопрос о том, какие признаки – дезинформационные или порождающие – усваивает обучаемая модель.

Предобработка и разбиение. Предобработка включала удаление дубликатов, фильтрацию шума, удаление URL-адресов и нормализацию казахского и русского текстов; дедупликация применялась внутри класса REAL, поскольку пары REAL/FAKE близки по содержанию по построению. Из того же обстоятельства следует требование к разбиению на обучающую и тестовую части: случайное разбиение по документам развело бы порождающий и порожденный тексты по разным подвыборкам и завысило бы оценки качества, поэтому разбиение выполняется по группам, объединяющим исходный материал и все порожденные из него тексты; это требование должно соблюдаться и в последующих работах, использующих корпус. Корпус сбалансирован по классам и языкам (табл. 1).

Табл. 1. Статистика корпуса KazFakeCorpus.

Язык	Класс	Число текстов	Среднее число слов	Доля, %
Казахский	REAL	1085	≈ 278	25.4
Казахский	FAKE	1085	≈ 330	25.4
Русский	REAL	1053	≈ 301	24.6
Русский	FAKE	1053	≈ 448	24.6
Всего	**–**	**4276**	≈ 339	100

Обращает на себя внимание систематическое различие средней длины текстов классов REAL и FAKE, особенно выраженное для русского языка (301 против 448 слов). Это потенциальный поверхностный признак, на который может опереться классификатор, поэтому при оценке моделей на корпусе целесообразно приводить базовое решение, использующее только длину текста; фиксация

этого признака – одно из оснований для включения свойства tokenCount в онтологическую модель.

Протокол аннотирования. Разметка выполнялась двумя независимыми экспертами по предопределенной многоуровневой схеме в среде Label Studio; схема уточнялась на пилотном этапе, охватившем 120 текстов. Помимо категориальных меток каждый аннотатор фиксировал уверенность в решении по пятибалльной шкале. Материалы с расхождением решений и с низкими значениями уверенности передавались на дополнительное рассмотрение; итоговая метка принималась по консенсусу экспертов. Сохраняемая в корпусе оценка уверенности переносится в онтологию как значение свойства annotatorConfidence и используется в определении класса HighConfidenceFake.

Согласованность оценивалась коэффициентом альфа Криппендорфа (Krippendorff's Alpha) [20] по разметке двух аннотаторов до этапа согласования на случайной подвыборке в 15 % корпуса. Получены следующие значения α: достоверность (бинарная номинальная переменная) – 0.88; тип фейкового контента – 0.84; техника дезинформации – 0.81; модальность – 0.79.

Значения α в диапазоне 0.79–0.88 свидетельствуют о высокой согласованности. Наиболее низкие значения получены для техники дезинформации и модальности: эти уровни требуют интерпретации текста и допускают приписывание одному материалу нескольких значений, поэтому коэффициент вычислялся для них в варианте для множественных ответов. Тем же продиктованы ограничения на свойства в онтологии: hasVeracityLabel и hasLanguage объявлены функциональными, тогда как usesTechnique, hasModality и hasFakeNewsType таких ограничений не имеют.

3. МНОГОУРОВНЕВАЯ СХЕМА СЕМАНТИЧЕСКОЙ РАЗМЕТКИ

Центральная сущность. Центральным элементом схемы является сущность NEWS – базовая единица описания новостного сообщения. Она агрегирует метаданные, необходимые для формальной идентификации и контекстуализации текста: уникальный идентификатор, дату разметки, итоговую метку достоверности REAL_or_FAKE, характеристики источника source_type, временные и географические параметры, целевую сущность TARGET_entity и оценку уверенности CONFIDENCE по шкале 1–5. В зависимости от значения REAL_or_FAKE сущность

NEWS связывается с блоком REAL_NEWS либо с блоком FAKE_NEWS.

Блоки REAL_NEWS и FAKE_NEWS. Блок REAL_NEWS содержит параметры, отражающие механизмы верификации: EVIDENCE и SOURCE_CREDIBILITY, что согласуется с подходами автоматизированного фактчекинга, где достоверность определяется проверяемостью утверждения и надежностью источника.

Блок FAKE_NEWS описывает внутреннюю структуру дезинформации и включает fake_type, DISINFORMATION_TECHNIQUE и AUTHOR_INTENT. Компонент fake_type опирается на устоявшиеся типологии недостоверного контента [11] и охватывает тринадцать подтипов, среди которых отметим сфабрикованный контент, вводящий в заблуждение контент и ложный контекст. Компонент DISINFORMATION_TECHNIQUE описывает манипулятивные стратегии – кликбейт, эмоциональное давление, преувеличение, преуменьшение, поиск виновного и ложную атрибуцию; в корпусе преобладают ложная атрибуция, кликбейт и эмоциональное давление, однако это распределение отражает состав инструкций, использованных при контролируемой трансформации, а не наблюдаемую частотность приемов. Включение AUTHOR_INTENT добавляет разметке прагматическое измерение, блок MODALITY фиксирует форму представления содержания (текст, изображение, видео, аудио) и делает схему расширяемой на мультимодальные постановки задачи. Модель включает также явные семантические отношения между сущностями, что предоставляет возможность графового представления корпуса.

Ограничения плоского представления. Реализация схемы в виде связанных таблиц разметки имеет три ограничения. Во-первых, семантика полей задается неформально: утверждение «материал класса FAKE не может иметь заполненное поле EVIDENCE» существует лишь в инструкции для аннотатора. Во-вторых, содержательные категории приходится вводить вручную: чтобы выделить, например, казахскоязычные фейки, порожденные языковой моделью и содержащие эвиденциальный маркер, требуется либо новое поле разметки, либо отдельный программный запрос. В-третьих, эталонная разметка, предсказания моделей, объяснения и метаданные корпуса хранятся в разных форматах, и их совместный анализ требует связующего кода.

Онтологическое представление устраняет второе и третье ограничения полностью: содержательные категории выводятся ризонером, а разнородные

сведения объединяются в едином графе, доступном для декларативных запросов. Первое ограничение снимается частично: аксиомы OWL позволяют обнаружить противоречащие друг другу утверждения, но не выявляют отсутствие обязательного значения.

4. KazFakeOnto: ОНТОЛОГИЧЕСКАЯ ФОРМАЛИЗАЦИЯ СХЕМЫ

Принципы проектирования. Онтология KazFakeOnto построена на языке OWL 2 DL [21] в среде Protégé; идентификатор онтологии – http://www.semanticweb.org/kazfake, префикс пространства имен – `kf:`. Соблюдались четыре принципа. Первый – это соответствие схеме разметки: каждый ее уровень отображается в конструкцию онтологии, что позволяет автоматически преобразовывать выгрузку корпуса в RDF-триплеты (целевая сущность и географические параметры сущности NEWS пока не формализованы). Второй – справочники представлены именованными индивидами классов-справочников, а не классами, что упрощает построение SPARQL-запросов. Третий – содержательные категории не проставляются вручную, а определяются аксиомами эквивалентности и вычисляются ризонером. Четвертый – это проверяемость: каждая конструкция сопровождается запросом, результат которого сопоставляется с ожидаемым.

Количественные характеристики модели: 1155 утвержденных триплетов (без учета выведенных), 38 именованных классов, 26 объектных и 22 свойства-значения данных, 112 индивидов. Наполнение включает демонстрационный фрагмент корпуса из 18 размеченных материалов (9 казахских и 9 русских, 12 FAKE и 6 REAL, из них 3 машинно-порожденные), а также источники, аннотаторов, объекты аннотации, эвиденциальные маркеры, подтверждения, порождающие модели, модели детекции с метриками, предсказания, объяснения и метаданные корпуса. Фрагмент носит иллюстративный характер: он построен так, чтобы задействовать все конструкции модели, поэтому приводимые далее количественные результаты характеризуют состав фрагмента, а не распределения в корпусе.

Структура классов. Таксономия классов представлена на рис. 2. Классы-контейнеры описывают основные типы сущностей предметной области: Corpus, NewsItem, Annotator, Annotation, Source, Evidence, DetectionModel, Prediction, Explanation, EvidentialMarker, Generator. Класс Explanation разделен на подклассы

CommonsenseExplanation, TextualExplanation и EvidenceBasedExplanation. Класс EvidentialMarker разделен на EvidentialAffix, EvidentialParticle и ReportativeConstruction – это разбиение отражает грамматическую типологию средств выражения эвиденциальности [22], существенно различающихся в казахском и русском языках.

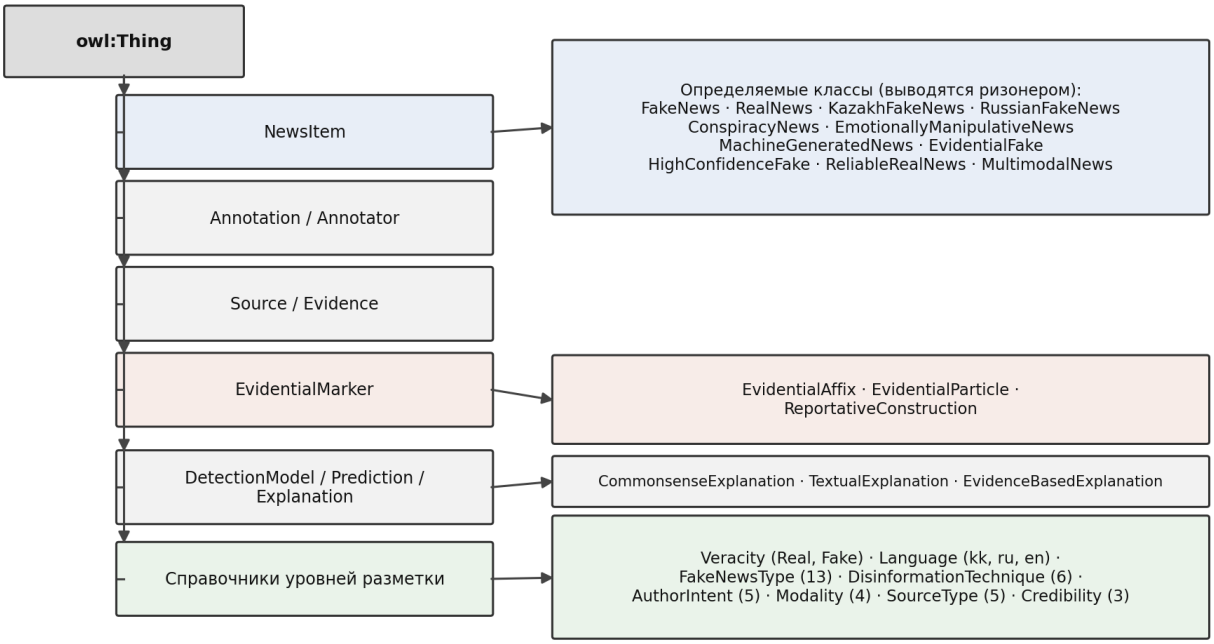


Рис. 2. Фрагмент таксономии классов KazFakeOnto.

Восемь классов-справочников отображают уровни разметки в конструкции онтологии:

- достоверность – Veracity и функциональное свойство hasVeracityLabel (2 значения);
- язык – Language и функциональное hasLanguage (3, одно зарезервировано); тип фейкового контента – FakeNewsType и hasFakeNewsType (13);
- техника дезинформации – DisinformationTechnique и usesTechnique (6); интенция автора – AuthorIntent и hasAuthorIntent (5);
- модальность – Modality и hasModality (4); тип источника – SourceType и hasSourceType (5);
- доверие к источнику – Credibility и функциональное hasCredibility (3).

Первые шесть свойств имеют доменом NewsItem, два последних – Source. Индивиды каждого справочника объявлены попарно различными аксиомой

owl:AllDifferent: в OWL не действует допущение уникальности имен, и без этого объявления ризонер не смог бы ни установить нарушение функциональности свойства, ни вывести принадлежность материала к классу, определенному через ограничение мощности. Значение MixedLanguage предназначено для материалов со смешением кодов; в текущей версии корпуса они отсутствуют, значение включено для расширяемости модели.

Свойства и аксиомы. Свойства сущности NewsItem образуют звездообразную структуру (рис. 3). Дататипные свойства фиксируют идентификатор newsId, заголовок title, дату newsDate, объем tokenCount, уверенность аннотатора annotatorConfidence и примечания notes.

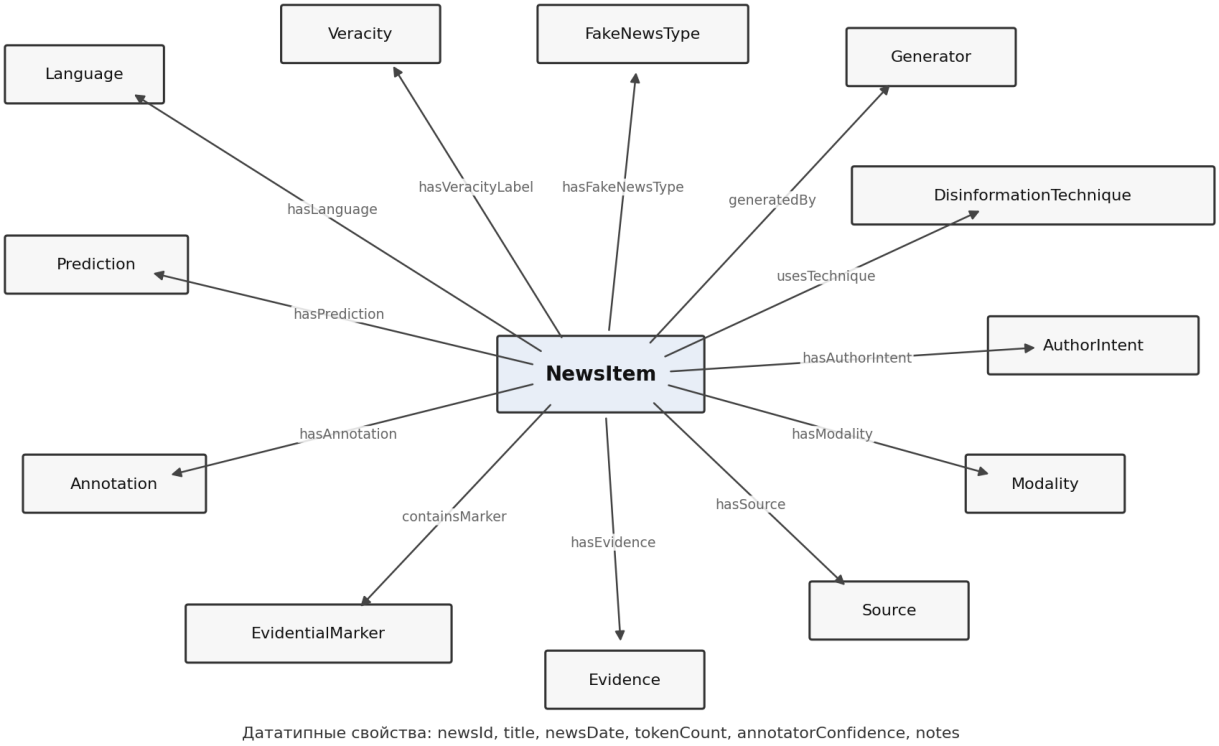


Рис. 3. Граф свойств сущности NewsItem.

Помимо доменов и областей значений заданы четыре типа содержательно значимых аксиом. Свойства hasVeracityLabel, hasLanguage, hasCredibility и madeBy объявлены функциональными: материал не может одновременно относиться к классам REAL и FAKE, иметь два языка разметки, а источник – два уровня доверия. Само по себе это не гарантирует непротиворечивости: не располагая сведениями о различии индивидов, ризонер отождествил бы kf:Fake и

kf:Real. Поэтому функциональность дополнена объявлением индивидов справочников попарно различными и аксиомой дизъюнктности классов FakeNews и RealNews. Пара обратных свойств hasAnnotation/annotates позволяет обходить граф в обоих направлениях без дублирования утверждений.

Аксиома-цепочка SubObjectPropertyOf(ObjectPropertyChain(hasAnnotation madeBy) annotatedBy) связывает материал непосредственно с аннотатором. В OWL 2 цепочка свойств выражается только включением, но не эквивалентностью; вследствие этого свойство annotatedBy становится несоставным и не может использоваться в ограничениях мощности и объявляться функциональным, иначе онтология вышла бы за пределы OWL 2 DL. Свойство annotatedBy нигде не утверждается явно: все его значения выводятся ризонером, что демонстрирует запрос D10.

Свойства происхождения generatedBy и derivedFromSeed фиксируют дизайн контролируемой трансформации (разд. 2) и делают воспроизводимой проверку гипотезы о влиянии порождающей модели на обучаемый детектор.

Отметим ограничение текущей версии, существенное для интерпретации (разд. 5): уровни разметки заданы как свойства сущности NewsItem, а не объекта Annotation. Это соответствует хранению согласованного результата разметки, но не позволяет представить в графе две независимые разметки одного материала, поэтому пересчет коэффициента согласованности по графу невозможен.

Определяемые классы. Ключевым элементом, отличающим онтологию от структурированной таблицы, являются определяемые классы; их перечень приведен в табл. 2 в нотации Manchester Syntax.

Табл. 2. Определяемые классы KazFakeOnto.

Класс	Определение
FakeNews	NewsItem and hasVeracityLabel value Fake
RealNews	NewsItem and hasVeracityLabel value Real
KazakhFakeNews	FakeNews and hasLanguage value Kazakh
RussianFakeNews	FakeNews and hasLanguage value Russian
ConspiracyNews	NewsItem and hasFakeNewsType value ConspiracyTheory

EmotionallyManipulativeNews	NewsItem and usesTechnique value Emotional-Pressure
MachineGeneratedNews	NewsItem and generatedBy some Generator
EvidentialFake	FakeNews and containsMarker some Evidential-Marker
HighConfidenceFake	FakeNews and annotatorConfidence some xsd:integer[>= 4]
ReliableRealNews	RealNews and hasSource some (hasCredibility value HighCredibility)
MultimodalNews	NewsItem and hasModality min 2 Modality

Аннотатор заполняет только элементарные уровни разметки – язык, достоверность, тип, технику, модальность, источник, уверенность; все производные категории вычисляются автоматически. Так, EvidentialFake объединяет фейковые материалы с эвиденциальным маркером; ReliableRealNews требует источника с высоким уровнем доверия, т. е. предполагает обход графа на два шага; HighConfidenceFake использует фасет xsd:integer[>= 4]; MultimodalNews – квалифицированное ограничение мощности. Ни одна из этих категорий не хранится в разметке, и при изменении элементарных значений все производные пересчитываются согласованно.

Приведем фрагмент онтологии в синтаксисе Turtle, описывающий один материал корпуса.

```
kf:N001 a owl:NamedIndividual , kf:NewsItem ;
    kf:newsId "KFC-0001" ; kf:hasLanguage kf:Kazakh ; kf:hasVeracityLabel kf:Fake ;
    kf:hasFakeNewsType kf:ConspiracyTheory ; kf:usesTechnique kf:EmotionalPressure ;
    kf:hasAuthorIntent kf:IdeologicalIntent ; kf:hasModality kf:TextModality ;
    kf:hasSource kf:AnonymousTelegramChannel ; kf:containsMarker kf:Marker_MYS ;
    kf:annotatorConfidence 5 ; kf:tokenCount 148 ; kf:partOfCorpus kf:KazFakeCorpus .
```

Материал kf:N001 не отнесен явно ни к одному определяемому классу; ризонер выводит его принадлежность к FakeNews, KazakhFakeNews, ConspiracyNews, EmotionallyManipulativeNews, EvidentialFake и HighConfidenceFake.

Доступность и проверка непротиворечивости. Онтология распространяется вместе с сопроводительными материалами исследования в двух сериализациях – RDF/XML и Turtle – на условиях лицензии Creative Commons Attribution 4.0 International (CC BY 4.0).

Логический вывод выполнялся ризонером HermiT 1.4.3 [23]. Проверка непротиворечивости не выявила неудовлетворимых классов: ни один именованный класс не отнесен к owl:Nothing. Все одиннадцать определяемых классов получили непустые множества экземпляров. В табл. 3 представлены результаты для девяти классов, имеющих самостоятельное значение для демонстрации возможностей вывода; не включенные в таблицу классы RussianFakeNews и EmotionallyManipulativeNews содержат соответственно 6 и 4 экз. Отметим деталь реализации: даты представлены типом xsd:dateTime, поскольку xsd:date не входит в обязательную карту типов данных OWL 2 и ризонером HermiT в применявшейся конфигурации не обрабатывался.

5. ЗАПРОСНЫЙ СЛОЙ И АПРОБАЦИЯ

Работоспособность модели проверялась SPARQL-запросами [24] к утвержденным триплетам и DL-запросами к выведенной модели: вкладка SPARQL в Protégé работает без учета вывода, поэтому определяемые классы проверяются DL-запросами при запущенном ризонере. Все результаты получены на демонстрационном фрагменте из 18 материалов и характеризуют состав фрагмента, а не корпуса в целом.

Анализ разметки средствами SPARQL. Пятнадцать запросов Q1–Q15 покрывают четыре группы задач: описательную статистику корпуса (распределение по языку и достоверности, частотность техник, нагрузка аннотаторов, метаданные); анализ разметки (профиль материала, мультимодальные фейки, фейки из ненадежных источников, достоверные новости и их подтверждения); лингвистический слой (формы эвиденциальных маркеров и их распределение по классам достоверности); слой моделей детекции (сравнение моделей, машинно-по-

рожденные материалы и их исходные тексты, объяснения, расхождения разметки и предсказаний).

Особый интерес представляет запрос Q5 о распределении эвиденциальных маркеров по классам достоверности:

```
SELECT ?veracity (COUNT(DISTINCT ?n) AS ?withMarker) WHERE {
  ?v a kf:Veracity ; rdfs:label ?veracity . OPTIONAL { ?n kf:hasVe-
racityLabel ?v ;
      kf:containsMarker ?m } FILTER(lang(?veracity) = "ru") }
GROUP BY ?veracity
```

Конструкция OPTIONAL здесь существенна: без нее класс, ни один материал которого не содержит маркера, не попал бы в результат, и нулевое значение было бы неотличимо от отсутствия материалов такого класса. Запрос возвращает две строки – «фейковая: 7» и «достоверная: 0».

Запрос Q4 детализирует картину по формам маркеров: в казахском материале зафиксированы аффикс *-мыс/-міс*, частица *екен*, ренарративные конструкции *деп айтады/дейді* и *көрінеді*; в русском – *якобы*, *по слухам*, *как сообщается*. Различие грамматических средств выражения эвиденциальности делает этот уровень разметки типологически значимым и объясняет введение подклассов класса EvidentialMarker. Наблюдаемая асимметрия является следствием инструкций порождения; воспроизводится ли она на естественно возникающей дезинформации – открытый вопрос.

Логический вывод: DL-запросы. Результаты тринадцати DL-запросов приведены в табл. 3. Ни одна из проверяемых категорий не утверждается в онтологии явно.

Табл. 3. DL-запросы и результаты вывода.

№	Запрос (Manchester Syntax)	Количество экз.
D1	FakeNews	12
D2	RealNews	6
D3	KazakhFakeNews	6
D4	MachineGeneratedNews	3
D5	EvidentialFake	7

D6	HighConfidenceFake	9
D7	ReliableRealNews	5
D8	MultimodalNews	6
D9	FakeNews and (usesTechnique value Clickbait)	3
D10	NewsItem and (annotatedBy value Annotator1)	4
D11	FakeNews and (hasSource some (hasCredibility value LowCredibility))	12
D12	NewsItem and (containsMarker some EvidentialAffix)	1
D13	ConspiracyNews	2

Наиболее показательны три результата. Запрос D7 возвращает пять материалов из шести достоверных: один исключается, поскольку его источник имеет средний уровень доверия, вывод опирается на обход графа через сущность Source. Запрос D10 возвращает материалы, размеченные конкретным аннотатором, при том что свойство annotatedBy нигде не утверждено. Запрос D13 объединяет два материала конспирологического типа, один из которых порожден языковой моделью из второго.

Отдельного комментария требует результат D11: все двенадцать материалов класса FAKE во фрагменте связаны с источниками низкого доверия. Это следствие устройства фрагмента, а не наблюдаемая закономерность; в корпусе в целом уровень доверия к источнику не должен совпадать с меткой достоверности, иначе он превращается в признак, тривиально предсказывающий класс.

Онтология как слой над результатами моделей детекции. Онтология охватывает не только эталонную разметку, но и результаты систем автоматического обнаружения: классы DetectionModel, Prediction и Explanation связывают материал, предсказанную метку, уверенность модели и порожденные объяснения. Метрики моделей, разработанных и оцененных на KazFakeCorpus в рамках проекта AP26195591, представлены свойствами accuracy, macroF1, f1Russian, f1Kazakh и извлекаются запросом Q7 (табл. 4).

Табл. 4. Модели детекции, представленные в онтологии (значения в %).

Модель	Accuracy	Macro-F1	F1 (рус.)	F1 (каз.)
ADVISOR (RAG + двунаправленное кросс-внимание)	87.9	88.0	88.7	86.9
mDeBERTa-v3	84.5	84.5	–	–
multilingual-E5-large	83.3	83.3	–	–
XLM-RoBERTa [25]	82.4	82.4	–	–

Примечание. Приведенные значения получены в рамках проекта AP26195591 на корпусе KazFakeCorpus; описание архитектуры модели ADVISOR и протокола ее оценки готовится к публикации отдельно и здесь не является предметом самостоятельной экспериментальной оценки.

Различие качества по языкам для лучшей модели (86.9 против 88.7) составляет 1.8 п. п. и получено в одном прогоне без оценки доверительных интервалов, поэтому не может служить основанием для содержательных выводов. Как рабочую гипотезу отметим возможную связь с наблюдением по запросу Q4: значительная часть маркеров недостоверности в казахском языке выражается грамматически, что может затруднять их усвоение многоязычными энкодерами.

Запрос Q11 автоматизирует анализ ошибок: он отбирает материалы, для которых предсказанная метка расходится с эталонной, и возвращает обе метки, модель, ее уверенность и технику дезинформации. Во фрагменте единственная ошибка приходится на материал с техникой преуменьшения, ошибочно классифицированный моделью XLM-RoBERTa как достоверный с уверенностью 0.63.

Запрос Q10 извлекает три объяснения, порожденные объяснимой моделью для одного материала: на здравом смысле (E_CS), стилистико-текстовое (E_TX) и на внешних свидетельствах (E_DB). Для материала KFC-0001 при уверенности модели 0.94 объяснение E_TX прямо апеллирует к эвиденциальному аффиксу *-мыс/-міс*, т. е. к признаку, зафиксированному в онтологии на уровне разметки. Возможность сопоставить объяснение модели с независимо размеченным лингвистическим признаком служит практическим инструментом оценки достоверности объяснений.

6. ОБСУЖДЕНИЕ

Онтология как спецификация схемы разметки. Формализация переводит часть инструкции для аннотатора в проверяемую форму: ограничения, ранее существовавшие как соглашения – единственность метки достоверности, единственность уровня доверия источника, несовместимость классов REAL и FAKE, становятся аксиомами, нарушение которых обнаруживается автоматически. Для схем, развиваемых на протяжении нескольких итераций, это снижает риск скрытого рассогласования версий разметки. Требования полноты средствами OWL не проверяются и требуют ограничений SHACL.

Разделение наблюдаемого и выводимого. Принцип «аннотатор размечает только элементарные уровни, содержательные категории вычисляются» ограничивает разметку теми решениями, которые эксперт действительно принимает при чтении текста, и исключает класс ошибок, связанных с ручным дублированием производных меток. Одновременно он делает исследовательские категории пересматриваемыми: изменение определения категории требует правки одной аксиомы, а не переразметки корпуса. Существенно и то, что статистика корпуса, лингвистическая гипотеза об эвиденциальности, анализ ошибок классификатора и оценка объяснений формулируются над одним графом.

Ограничения. Основное ограничение унаследовано от корпуса: материалы класса FAKE синтетические и могут отличаться от естественно возникающей дезинформации стилистически и дискурсивно; контролируемая трансформация обеспечивает сопоставимость классов, но создает риск того, что обучаемая модель усваивает признаки порождающей модели. Свойства generatedBy и derivedFromSeed введены именно для того, чтобы этот риск можно было измерять. Второе ограничение модельное: уровни разметки заданы как свойства материала, а не объекта аннотации, поэтому двойная разметка и процедура арбитража в графе не представлены. Третье: в онтологию перенесен иллюстративный фрагмент, а не весь корпус. Четвертое связано с производительностью: вывод над графом из нескольких десятков тысяч индивидов требует перехода к триплетному хранилищу, причем профили OWL, поддерживаемые правиловыми машинами, не покрывают фасеты типов данных и квалифицированные ограничения мощности; практическим решением представляется предвычисление содержательных категорий запросами CONSTRUCT.

Этические аспекты и условия использования. Ресурс содержит машинно-порожденные тексты, воспроизводящие приемы дезинформации, что создает риск нецелевого использования. Материалы класса REAL – это находящиеся в открытом доступе официальные публикации; персональные данные частных лиц из текстов удалялись. Материалы класса FAKE являются синтетическими обучающими примерами и не предназначены для представления в качестве достоверных сообщений.

Доступ к текстовой части корпуса предоставляется для научных и образовательных целей по обоснованному запросу к авторам проекта и при условии принятия пользовательского соглашения, которое запрещает применение материалов для введения пользователей в заблуждение, имитации достоверных новостей и распространения дезинформации. Пользователь обязуется обеспечивать безопасное хранение данных, не публиковать синтетические тексты без указания на их искусственное происхождение и приводить ссылку на KazFakeCorpus.

Направления развития. Ближайшими задачами являются: перенос полного корпуса в онтологию, перенос уровней разметки на объект аннотации и автоматическая генерация интерфейса аннотатора из справочников онтологии. Содержательные направления – это сопоставление естественной и порожденной дезинформации в терминах определяемых классов и расширение типологии эвиденциальных средств казахского языка, а также согласование KazFakeOnto со словарем ClaimReview [14] и графами ClaimsKG [15] и CimpleKG [16].

ЗАКЛЮЧЕНИЕ

Представлен сбалансированный двуязычный корпус KazFakeCorpus объемом 4276 текстов, размеченный по многоуровневой семантической схеме, которая охватывает достоверность сообщения, тип фейкового контента, технику дезинформации, коммуникативное намерение, модальность и характеристики источника; согласованность разметки по коэффициенту Krippendorff’s Alpha составила 0.79–0.88.

Основной вклад статьи состоит в переводе схемы разметки в формальную онтологическую модель KazFakeOnto на языке OWL 2 DL: 38 классов, 26 объектных и 22 свойства-значения данных, 112 индивидов и одиннадцать определяе-

мых классов, выводимых ризонером из элементарных уровней разметки. Непротиворечивость подтверждена ризонером HermiT, запросный слой апробирован на пятнадцати SPARQL- и тринадцати DL-запросах.

Предложенный подход моделирует фейковую новость как многомерное семантическое явление и обеспечивает воспроизводимую основу для исследований в области анализа дезинформации и кросс-языкового обнаружения фейковых новостей для малоресурсных языков.

Благодарности

Исследование выполнено при поддержке Комитета науки Министерства науки и высшего образования Республики Казахстан (грант № AP26195591).

СПИСОК ЛИТЕРАТУРЫ

1. *Zhou X., Zafarani R.* A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities // ACM Computing Surveys. 2020. Vol. 53, No. 5. P. 1–40. https://doi.org/10.1145/3395046

2. *Amjad M., Sidorov G., Zhila A., Gómez-Adorno H., Voronkov I., Gelbukh A.* "Bend the Truth": Benchmark Dataset for Fake News Detection in Urdu Language and Its Evaluation // Journal of Intelligent & Fuzzy Systems. 2020. Vol. 39, No. 2. P. 2457–2469. https://doi.org/10.3233/JIFS-179905

3. *Azad R., Mohammed B., Mahmud R., Zrar L., Sdiqa S.* Fake News Detection in Low-Resourced Languages "Kurdish Language" Using Machine Learning Algorithms // Turkish Journal of Computer and Mathematics Education. 2021. Vol. 12, No. 6. P. 4219–4225.

4. *Makhambetov O., Makazhanov A., Yessenbayev Z., Matkarimov B., Sabyrgaliyev I., Sharafudinov A.* Assembling the Kazakh Language Corpus // Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2013). 2013. P. 1022–1031. URL: https://aclanthology.org/D13-1104/ (дата обращения 15.08.2026).

5. *Sambetbayeva M., Nekessova A., Yerimbetova A., Bayangali A., Kaldarova M., Telman D., Smailov N.* A Multi-Level Annotation Model for Fake News Detection: Implementing Kazakh-Russian Corpus via Label Studio // Big Data and Cognitive Computing. 2025. Vol. 9, No. 8. Article 215.
https://doi.org/10.3390/bdcc9080215

6. *Lamasheva Zh., Nekessova A., Kantureyeva M., Sambetbayeva M., Kaldarova M., Nazymkhan A.* KazFakeCorpus: A Bilingual Corpus with Multi-Level Semantic Annotation for Fake News Detection // Big Data and Cognitive Computing. 2026. Vol. 10, No. 6. Article 183. https://doi.org/10.3390/bdcc10060183

7. *Nekessova A., Sidorova E., Lamasheva Zh., Sambetbayeva M., Yerimbetova A., Kaldarova M., Nazymkhan A.* KazFakeCorpus: A Bilingual Multi-Level Corpus for Semantic Detection of Fake News // Компьютерная лингвистика и интеллектуальные технологии: по материалам международной конференции «Диалог». Вып. 24. М.: МАКС Пресс, 2026. С. 438–447. https://doi.org/10.29003/2075-7182-2026-24-438-447

8. *Shu K., Mahudeswaran D., Wang S., Lee D., Liu H.* FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media // Big Data. 2020. Vol. 8, No. 3. P. 171–188. https://doi.org/10.1089/big.2020.0062

9. *Li Y., He H., Bai J., Wen D.* MCFEND: A Multi-Source Benchmark Dataset for Chinese Fake News Detection // Proc. of the ACM Web Conference 2024. 2024. P. 4018–4027. https://doi.org/10.1145/3589334.3645385

10. *Ashraf S., Bezzaoui I., Andone I., Markowetz A., Fegert J., Flek L.* DeFaktS: A German Dataset for Fine-Grained Disinformation Detection through Social Media Framing // Proceedings of LREC-COLING 2024. 2024. P. 4580–4591. URL: https://aclanthology.org/2024.lrec-main.409/ (дата обращения 15.08.2026).

11. *Tandoc E.C., Lim Z.W., Ling R.* Defining “Fake News”: A Typology of Scholarly Definitions // Digital Journalism. 2018. Vol. 6, No. 2. P. 137–153. https://doi.org/10.1080/21670811.2017.1360143

12. *Da San Martino G., Barrón-Cedeño A., Wachsmuth H., Petrov R., Nakov P.* SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles // Proceedings of the Fourteenth Workshop on Semantic Evaluation (SemEval-2020). 2020. P. 1377–1414. https://doi.org/10.18653/v1/2020.semeval-1.186

13. *Bezzaoui I., Fafalios P., Fegert J., Rettinger A., Todorov K.* Decoding Deception with TAXODIS – a Taxonomy of Disinformation Cues for Fine-Grained Text Labeling // Semantic Web Journal. URL: https://zenodo.org/records/14264593 (дата обращения 15.08.2026).

14. ClaimReview // Schema.org. URL: https://schema.org/ClaimReview (дата

обращения 15.08.2026).

15. *Tchechmedjiev A., Fafalios P., Boland K., Gasquet M., Zloch M., Zapilko B., Dietze S., Todorov K.* ClaimsKG: A Knowledge Graph of Fact-Checked Claims // The Semantic Web – ISWC 2019. Lecture Notes in Computer Science. Vol. 11779. Cham: Springer, 2019. P. 309–324. https://doi.org/10.1007/978-3-030-30796-7_20

16. *Burel G., Mensio M., Peskine Y., Troncy R., Papotti P., Alani H.* CimpleKG: A Continuously Updated Knowledge Graph on Misinformation, Factors and Fact-Checks // The Semantic Web – ISWC 2024. Lecture Notes in Computer Science. Vol. 15233. Cham: Springer, 2025. P. 97–114. https://doi.org/10.1007/978-3-031-77847-6_6

17. *Bani-Hani A., Adedugbe O., Benkhelifa E., Majdalawieh M.* Fandet Semantic Model: An OWL Ontology for Context-Based Fake News Detection on Social Media // Combating Fake News with Computational Intelligence Techniques. Studies in Computational Intelligence. Vol. 1001. Cham: Springer, 2022. P. 91–125. https://doi.org/10.1007/978-3-030-90087-8_5

18. *Abdygalym B.Kh., Adali E., Sambetbayeva M.A., Sadirmekova Zh.B., Nazymkhan A.A.* A Conceptual Model for Ontology-Based Detection of Information Operations in Digital Media // Bulletin of Shakarim University. Technical Sciences. 2025. No. 4(20). P. 36–44. https://doi.org/10.53360/2788-7995-2025-4(20)-5

19. *Елизаров А.М., Кириллович А.В., Липачев Е.К., Невзорова О.А.* Цифровая экосистема OntoMath как подход к построению пространства математических знаний // Электронные библиотеки. 2023. Т. 26, № 2. С. 154–202. https://doi.org/10.26907/1562-5419-2023-26-2-154-202

20. *Krippendorff K.* Content Analysis: An Introduction to Its Methodology. 4th ed. Thousand Oaks: SAGE Publications, 2018.

21. OWL 2 Web Ontology Language Document Overview (Second Edition). W3C Recommendation, 11 December 2012. URL: https://www.w3.org/TR/owl2-overview/ (дата обращения 15.08.2026).

22. *Aikhenvald A.Y.* Evidentiality. Oxford: Oxford University Press, 2004.

23. *Glimm B., Horrocks I., Motik B., Stoilos G., Wang Z.* HermiT: An OWL 2 Reasoner // Journal of Automated Reasoning. 2014. Vol. 53, No. 3. P. 245–269. https://doi.org/10.1007/s10817-014-9305-1

24. SPARQL 1.1 Query Language. W3C Recommendation, 21 March 2013. URL: https://www.w3.org/TR/sparql11-query/ (дата обращения 15.08.2026).

25. *Conneau A., Khandelwal K., Goyal N., Chaudhary V., Wenzek G., Guzmán F., Grave E., Ott M., Zettlemoyer L., Stoyanov V.* Unsupervised Cross-Lingual Representation Learning at Scale // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020. P. 8440–8451.
https://doi.org/10.18653/v1/2020.acl-main.747

FROM AN ANNOTATED CORPUS TO AN ONTOLOGY: A FORMAL MODEL OF MULTI-LEVEL SEMANTIC ANNOTATION OF FAKE NEWS IN THE KAZAKH–RUSSIAN MEDIA SPACE (KazFakeCorpus / KazFakeOnto)

A. A. Nekessova[1] [0009-0001-7642-0648], E. A. Sidorova[2] [0000-0001-8731-3058], Zh. B. Lamasheva[3] [0000-0002-9535-2636], M. A. Sambetbayeva[4] [0000-0001-9358-1614], A. S. Yerimbetova[5] [0000-0002-2013-1513], M. Zh. Kaldarova[6] [0000-0001-7494-9794], A. A. Nazymkhan[7] [0009-0007-4366-6709]

[1, 3, 4, 6, 7]*International Science Complex Astana, Astana, Kazakhstan*
[1, 3, 4, 7]*L.N. Gumilyov Eurasian National University, Astana, Kazakhstan*
[2]*A.P. Ershov Institute of Informatics Systems, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia*
[5]*Institute of Information and Computational Technologies CS MSHE RK, Almaty, Kazakhstan*
[5]*School of Engineering and Information Technology, META University, Almaty, Kazakhstan*
[6]*Astana International University, Astana, Kazakhstan*
[1]nekessova_aa_1@enu.kz, [2]lsidorova@iis.nsk.su, [3]lamasheva_zhb@enu.kz, [4]sambetbayeva_ma_1@enu.kz, [5]aigerian8888@gmail.com, [6]mira_kaldarova@aiu.edu.kz, [7]nazymkhan_aa@enu.kz

Abstract

The paper addresses the formal representation of knowledge about disinformation in a low-resource bilingual media environment. Automatic fake news detection is traditionally formulated as a binary REAL/FAKE classification task, which does not capture the internal structure of an unreliable message: the type of fake content, the disinformation technique employed, the author's communicative intent, and the

characteristics of the source and evidence base.

Two interrelated resources are presented. The first is KazFakeCorpus, a bilingual corpus of 4276 texts in Kazakh and Russian, balanced by class within each language; the REAL class was drawn from official materials of the Gov.kz portal and the FAKE class from their controlled transformations. The texts were annotated in Label Studio by two independent experts following a multi-level semantic scheme; Krippendorff's Alpha for the principal levels ranged from 0.79 to 0.88. The second resource is KazFakeOnto, an ontology that recasts the annotation scheme from a flat relational representation into a formal OWL 2 DL model: 38 classes, 26 object properties, 22 data properties, 112 individuals and 1155 asserted triples. Eleven defined classes are not asserted manually but inferred by a reasoner from property values; a property chain links a news item to its annotator through an annotation object. Consistency was verified with HermiT, and the query layer was evaluated on fifteen SPARQL and thirteen DL queries.

The ontological layer is shown to provide a reproducible specification of the annotation scheme, the automatic inference of semantic categories, and a unified formal environment for formulating corpus-linguistic questions and analysing machine-learning model errors. The limitations of the approach and directions for scaling the ontology to the full corpus and to naturally occurring disinformation are discussed.

Keywords: *fake news detection, ontology, OWL 2, SPARQL, logical inference, bilingual corpus, Kazakh language, multi-level annotation, disinformation techniques, explainable NLP, low-resource languages, KazFakeCorpus, KazFakeOnto.*

Acknowledgements. This research was supported by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (grant No. AP26195591).

REFERENCES

1. *Zhou X., Zafarani R.* A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities // ACM Computing Surveys. 2020. Vol. 53, No. 5. P. 1–40. https://doi.org/10.1145/3395046

2. *Amjad M., Sidorov G., Zhila A., Gómez-Adorno H., Voronkov I., Gelbukh A.* "Bend the Truth": Benchmark Dataset for Fake News Detection in Urdu Language and Its Evaluation // Journal of Intelligent & Fuzzy Systems. 2020. Vol. 39, No. 2. P. 2457–

2469. https://doi.org/10.3233/JIFS-179905

3. *Azad R., Mohammed B., Mahmud R., Zrar L., Sdiqa S.* Fake News Detection in Low-Resourced Languages “Kurdish Language” Using Machine Learning Algorithms // Turkish Journal of Computer and Mathematics Education. 2021. Vol. 12, No. 6. P. 4219–4225.

4. *Makhambetov O., Makazhanov A., Yessenbayev Z., Matkarimov B., Sabyrgaliyev I., Sharafudinov A.* Assembling the Kazakh Language Corpus // Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2013). 2013. P. 1022–1031. URL: https://aclanthology.org/D13-1104/ (дата обращения 15.08.2026).

5. *Sambetbayeva M., Nekessova A., Yerimbetova A., Bayangali A., Kaldarova M., Telman D., Smailov N.* A Multi-Level Annotation Model for Fake News Detection: Implementing Kazakh-Russian Corpus via Label Studio // Big Data and Cognitive Computing. 2025. Vol. 9, No. 8. Article 215.
https://doi.org/10.3390/bdcc9080215

6. *Lamasheva Zh., Nekessova A., Kantureyeva M., Sambetbayeva M., Kaldarova M., Nazymkhan A.* KazFakeCorpus: A Bilingual Corpus with Multi-Level Semantic Annotation for Fake News Detection // Big Data and Cognitive Computing. 2026. Vol. 10, No. 6. Article 183. https://doi.org/10.3390/bdcc10060183

7. *Nekessova A., Sidorova E., Lamasheva Zh., Sambetbayeva M., Yerimbetova A., Kaldarova M., Nazymkhan A.* KazFakeCorpus: A Bilingual Multi-Level Corpus for Semantic Detection of Fake News // Компьютерная лингвистика и интеллектуальные технологии: по материалам международной конференции «Диалог». Вып. 24. М.: МАКС Пресс, 2026. С. 438–447.
https://doi.org/10.29003/2075-7182-2026-24-438-447

8. *Shu K., Mahudeswaran D., Wang S., Lee D., Liu H.* FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media // Big Data. 2020. Vol. 8, No. 3. P. 171–188. https://doi.org/10.1089/big.2020.0062

9. *Li Y., He H., Bai J., Wen D.* MCFEND: A Multi-Source Benchmark Dataset for Chinese Fake News Detection // Proc. of the ACM Web Conference 2024. 2024. P. 4018–4027. https://doi.org/10.1145/3589334.3645385

10. *Ashraf S., Bezzaoui I., Andone I., Markowetz A., Fegert J., Flek L.* DeFaktS: A

German Dataset for Fine-Grained Disinformation Detection through Social Media Framing // Proceedings of LREC-COLING 2024. 2024. P. 4580–4591.
URL: https://aclanthology.org/2024.lrec-main.409/ (дата обращения 15.08.2026).

11. *Tandoc E.C., Lim Z.W., Ling R.* Defining “Fake News”: A Typology of Scholarly Definitions // Digital Journalism. 2018. Vol. 6, No. 2. P. 137–153.
https://doi.org/10.1080/21670811.2017.1360143

12. *Da San Martino G., Barrón-Cedeño A., Wachsmuth H., Petrov R., Nakov P.* SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles // Proceedings of the Fourteenth Workshop on Semantic Evaluation (SemEval-2020). 2020. P. 1377–1414. https://doi.org/10.18653/v1/2020.semeval-1.186

13. *Bezzaoui I., Fafalios P., Fegert J., Rettinger A., Todorov K.* Decoding Deception with TAXODIS – a Taxonomy of Disinformation Cues for Fine-Grained Text Labeling // Semantic Web Journal. URL: https://zenodo.org/records/14264593 (дата обращения 15.08.2026).

14. ClaimReview // Schema.org. URL: https://schema.org/ClaimReview (дата обращения 15.08.2026).

15. *Tchechmedjiev A., Fafalios P., Boland K., Gasquet M., Zloch M., Zapilko B., Dietze S., Todorov K.* ClaimsKG: A Knowledge Graph of Fact-Checked Claims // The Semantic Web – ISWC 2019. Lecture Notes in Computer Science. Vol. 11779. Cham: Springer, 2019. P. 309–324. https://doi.org/10.1007/978-3-030-30796-7_20

16. *Burel G., Mensio M., Peskine Y., Troncy R., Papotti P., Alani H.* CimpleKG: A Continuously Updated Knowledge Graph on Misinformation, Factors and Fact-Checks // The Semantic Web – ISWC 2024. Lecture Notes in Computer Science. Vol. 15233. Cham: Springer, 2025. P. 97–114. https://doi.org/10.1007/978-3-031-77847-6_6

17. *Bani-Hani A., Adedugbe O., Benkhelifa E., Majdalawieh M.* Fandet Semantic Model: An OWL Ontology for Context-Based Fake News Detection on Social Media // Combating Fake News with Computational Intelligence Techniques. Studies in Computational Intelligence. Vol. 1001. Cham: Springer, 2022. P. 91–125.
https://doi.org/10.1007/978-3-030-90087-8_5

18. *Abdygalym B.Kh., Adali E., Sambetbayeva M.A., Sadirmekova Zh.B., Nazymkhan A.A.* A Conceptual Model for Ontology-Based Detection of Information Operations in Digital Media // Bulletin of Shakarim University. Technical Sciences. 2025. No. 4(20). P. 36–44. https://doi.org/10.53360/2788-7995-2025-4(20)-5

19. *Elizarov A.M., Kirillovich A.V., Lipachev E.K., Nevzorova O.A.* Tsifrovaya ekosistema OntoMath kak podkhod k postroeniyu prostranstva matematicheskikh znanii [Digital Ecosystem OntoMath as an Approach to Building the Space of Mathematical Knowledge] // Russian Digital Libraries Journal. 2023. Vol. 26, No. 2. P. 154–202. https://doi.org/10.26907/1562-5419-2023-26-2-154-202 (In Russ.)

20. *Krippendorff K.* Content Analysis: An Introduction to Its Methodology. 4th ed. Thousand Oaks: SAGE Publications, 2018.

21. OWL 2 Web Ontology Language Document Overview (Second Edition). W3C Recommendation, 11 December 2012. URL: https://www.w3.org/TR/owl2-overview/ (дата обращения 15.08.2026).

22. *Aikhenvald A.Y.* Evidentiality. Oxford: Oxford University Press, 2004.

23. *Glimm B., Horrocks I., Motik B., Stoilos G., Wang Z.* HermiT: An OWL 2 Reasoner // Journal of Automated Reasoning. 2014. Vol. 53, No. 3. P. 245–269. https://doi.org/10.1007/s10817-014-9305-1

24. SPARQL 1.1 Query Language. W3C Recommendation, 21 March 2013. URL: https://www.w3.org/TR/sparql11-query/ (дата обращения 15.08.2026).

25. *Conneau A., Khandelwal K., Goyal N., Chaudhary V., Wenzek G., Guzmán F., Grave E., Ott M., Zettlemoyer L., Stoyanov V.* Unsupervised Cross-Lingual Representation Learning at Scale // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020. P. 8440–8451.

СВЕДЕНИЯ ОБ АВТОРАХ



НЕКЕСОВА Анаргуль Аймуратовна – докторант Евразийского национального университета имени Л.Н. Гумилева и младший научный сотрудник Международного научного комплекса «Астана». Область научных интересов: обработка естественного языка, обнаружение дезинформации, корпусная лингвистика для малоресурсных языков.

Anargul A. NEKESSOVA – PhD student, L.N. Gumilyov Eurasian National University and junior researcher at International Science Complex Astana. Research interests: natural language processing, disinformation detection, corpus linguistics for low-resource languages.

email: nekessova_aa_1@enu.kz

ORCID: 0009-0001-7642-0648

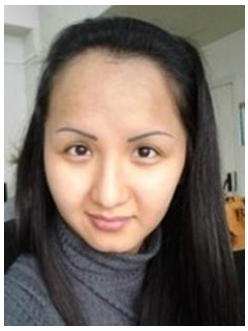


СИДОРОВА Елена Анатольевна – кандидат физико-математических наук, старший научный сотрудник Института систем информатики имени А.П. Ершова СО РАН. Область научных интересов: обработка естественного языка, онтологии и представление знаний, автоматический анализ текстов.

Elena A. SIDOROVA – PhD, senior researcher at the A.P. Ershov Institute of Informatics Systems, Siberian Branch of the Russian Academy of Sciences. Research interests: natural language processing, ontologies and knowledge representation, automatic text analysis.

email: lsidorova@iis.nsk.su

ORCID: 0000-0001-8731-3058



ЛАМАШЕВА Жанар Бейбутовна – PhD, ассоциированный профессор Евразийского национального университета имени Л. Н. Гумилева и старший научный сотрудник Международного научного комплекса «Астана». Область научных интересов: компьютерная лингвистика, анализ медиатекстов, многоязычные языковые ресурсы.

Zhanar B. LAMASHEVA – PhD, associate professor at L. N. Gumilyov Eurasian National University and senior researcher at International Science Complex Astana. Research interests: computational linguistics, media text analysis, multilingual language resources.

email: lamasheva_zhb@enu.kz

ORCID: 0000-0002-9535-2636



САМБЕТБАЕВА Мадина Аралбаевна – PhD, ассоциированный профессор Евразийского национального университета имени Л. Н. Гумилева, ведущий научный сотрудник Международного научного комплекса «Астана». Область научных интересов: информационные системы, графы знаний, многоязычные системы обработки естественного языка.

Madina A. SAMBETBAYEVA – PhD, associate professor at L. N. Gumilyov Eurasian National University, leading researcher at International Science Complex Astana. Research interests: information systems, knowledge graphs, multilingual natural language processing systems.

email: sambetbayeva_ma_1@enu.kz

ORCID: 0000-0001-9358-1614



ЕРИМБЕТОВА Айгерим Сембековна – PhD, ведущий научный сотрудник Института информационных и вычислительных технологий КН МНВО РК, ассоциированный профессор Школы инженерии и информационных технологий Университета МЕТА. Область научных интересов: машинное обучение, обработка естественного языка, интеллектуальные информационные системы.

Aigerim S. YERIMBETOVA – PhD, leading researcher at the Institute of Information and Computational Technologies, associate professor at the School of Engineering and Information Technology, META University. Research interests: machine learning, natural language processing, intelligent information systems.

email: aigerian8888@gmail.com

ORCID: 0000-0002-2013-1513



КАЛДАРОВА Мира Жорабековна – PhD Международного университета Астана и старший научный сотрудник Международного научного комплекса «Астана». Область научных интересов: анализ данных, интеллектуальные системы обработки текстовой информации.

Mira Zh. KALDAROVA – PhD at Astana International University and senior researcher at International Science Complex Astana. Research interests: data analysis, intelligent text processing systems.

email: mira_kaldarova@aiu.edu.kz

ORCID: 0000-0001-7494-9794



НАЗЫМХАН Ақсәуле Абзалқызы – инженер-программист Международного научного комплекса «Астана». Область научных интересов: обнаружение фейковых новостей, машинное обучение, аннотирование языковых данных.

Aksaule A. NAZYMKHAN – software engineer at International Science Complex Astana. Research interests: fake news detection, machine learning, language data annotation.

email: nazymkhan_aa@enu.kz

ORCID: 0009-0007-4366-6709

Материал поступил в редакцию 22 августа 2026 года

УДК 004

RusHallu-RAG: КОМПЛЕКСНЫЙ ПОДХОД К ОБНАРУЖЕНИЮ ГАЛЛЮЦИНАЦИЙ В РУССКОЯЗЫЧНЫХ RAG-СИСТЕМАХ

Ф. А. Садковский[1] [0009-0006-4502-9466], **Р. Р. Насырова**[2] [0009-0008-6280-0667], **А. А. Сорокин**[3] [0000-0003-3877-4223]

[1–3]*Московский государственный университет имени М. В. Ломоносова, Исследовательский Центр в сфере ИИ, г. Москва, Россия*
[3]*ООО «Яндекс», г. Москва, Россия*

[1]sadkovskiyfa@my.msu.ru, [2]NasyrovaRR_01@my.msu.ru, [3]sorokinaa@my.msu.ru

Аннотация

Представлен RusHallu-RAG – первый комплексный бенчмарк для обнаружения галлюцинаций в русскоязычных RAG-системах. Набор данных из 1000 пар «запрос – ответ» охватывает области общего знания и научных текстов с контролируемой балансировкой релевантности контекста. Предложена детализированная таксономия из шести типов галлюцинаций, использованная для разметки на уровне всего ответа и отдельных его фрагментов с привлечением экспертов и больших языковых моделей.

Для обучения детектора собран сбалансированный тренировочный датасет объемом 3000 примеров с применением синтетических методов генерации. На его основе методом SFT обучен детектор на базе YandexGPT-5-Lite-8B. Проведена оценка 15 моделей с открытыми и закрытыми весами, а также обученного детектора. Результаты показали, что число параметров не гарантирует высокой производительности: модели среднего размера (20–33 млрд) иногда превосходят более крупные. Обученный детектор достигает средней точности 63.7%, занимая второе место после закрытой Gemini-2.5-Pro, с приростом 47.2 п.п. относительно исходной модели. Разрыв между закрытыми и открытыми моделями сокращается с 20–35 до 8.4 п.п., что подтверждает эффективность подхода. Бенчмарк, код и обученная модель опубликованы в открытом доступе.

Ключевые слова*: обнаружение галлюцинаций, дополненная поиском генерация, большие языковые модели, бенчмаркинг, обработка естественного*

© Ф. А. Садковский, Р. Р. Насырова, А. А. Сорокин, 2026.
Данная статья распространяется на условиях международной лицензии Creative Commons License Attribution 4.0 International (CC BY 4.0).

языка, вопросно-ответные системы, оценка моделей, дообучение с учителем, таксономия галлюцинаций.

ВВЕДЕНИЕ

Обнаружение галлюцинаций является центральной проблемой при разработке надежных RAG-систем на основе больших языковых моделей (Large Language Models, LLM). В таких системах галлюцинации проявляются как фрагменты ответа, не соответствующие извлеченному контексту, – противоречащие ему, содержащие неподтвержденную информацию или отклоняющиеся от него.

Для английского языка существует множество бенчмарков для обнаружения галлюцинаций [1–7], в том числе специализированных для RAG [1, 5]. Для русского языка подобные ресурсы отсутствуют. Существующие русскоязычные RAG-бенчмарки [8] опираются на поверхностные метрики (ROUGE) или LLM-as-a-judge, но не содержат детальной разметки галлюцинаций. Этот пробел препятствует разработке и систематической оценке детекторов галлюцинаций для RAG-систем на русском языке.

В настоящей работе представлен RusHallu-RAG – первый комплексный бенчмарк для обнаружения галлюцинаций в русскоязычных RAG-системах. Бенчмарк включает набор данных из 1000 пар «запрос – ответ» из областей общего знания и научных текстов с контролируемой балансировкой релевантности контекста. Предложена новая таксономия из шести классов галлюцинаций, специфических для RAG, выполнена разметка на уровне ответов и отдельных фрагментов.

На основе бенчмарка проведена оценка 15 LLM с открытыми и закрытыми весами в роли детекторов галлюцинаций. Результаты показали, что рост числа параметров не гарантирует повышения качества, а между закрытыми и открытыми моделями для русского языка сохраняется значительный разрыв.

Кроме того, для решения прикладной задачи обучения детектора галлюцинаций собран тренировочный набор из 3000 примеров, сбалансированный с помощью синтетических методов. На этих данных методом SFT обучен детектор на базе YandexGPT-5-Lite-8B[1]. Обученная модель достигает средней точности

[1] https://huggingface.co/yandex/YandexGPT-5-Lite-8B-instruct

63.7%, занимая второе место после Gemini-2.5-Pro, с приростом 47.2 п.п. относительно исходной модели. Разрыв между закрытыми и открытыми моделями сокращается с 20–35 до 8.4 п.п., что подтверждает эффективность подхода. Бенчмарк, код и модель публикуются в открытом доступе[2].

ОБЗОР ЛИТЕРАТУРЫ

Существующие англоязычные ресурсы для оценки галлюцинаций в LLM можно разделить на три категории. Первая группа – общие бенчмарки для оценки RAG-систем [9–12], которые не фокусируются специально на обнаружении галлюцинаций, рассматривая любой неверный ответ как галлюцинацию без детальной разметки. Вторая, более широко представленная группа, интегрирует метрики, специфичные для галлюцинаций, в оценочные фреймворки, такие как RAGAS [13] и аналогичные [14–16], которые измеряют точность следования фактам (faithfulness) путем декомпозиции ответов на атомарные утверждения. Однако эти методы не идентифицируют конкретные фрагменты текста, содержащие галлюцинации.

Третья группа включает бенчмарки, разработанные специально для обнаружения галлюцинаций: HaluEval [4] с синтетическими примерами, HaluEval-Wild [6], включающий примеры из реальных диалогов, RAGTruth [5] с детальной разметкой на уровне ответов и фрагментов, а также FaithBench [17] и FaithJudge [18] с таксономиями по степени критичности искажений. Все перечисленные ресурсы являются исключительно англоязычными.

Для русского языка подобные ресурсы остаются малочисленными. Бенчмарк DRAGON [8] представляет собой динамически пополняющийся RAG-бенчмарк, построенный на новостных корпусах, однако его оценка опирается исключительно на метрики ROUGE и элементарные сравнения с привлечением LLM в роли судьи. Российский нормативно-корпоративный RAG-бенчмарк (RRNCB)[3] ориентирован на обработку корпоративных документов и характеризуется сходными ограничениями. Ни один из этих ресурсов не предоставляет детальной разметки галлюцинаций.

[2] https://github.com/feudor2/RusHallu-RAG

[3] https://fractalagents.ai/rrncb-rag-benchmark/

ДАННЫЕ

Существующие русскоязычные RAG-бенчмарки ориентированы на новостные и корпоративные области. Настоящая работа восполняет этот пробел, сосредоточившись на документах из области общего знания и научных текстов.

Контрольная выборки была сформирована из двух источников. Из тестовой части SberQuAD [19] – датасета оценки понимания прочитанного на русском языке, взято 655 запросов из области общего знания. Для научной сферы из бенчмарка ruSciBench [20], представляющего собой набор научных аннотаций с запросами, сгенерированными LLM по методологии AIR-Bench [21], отобраны 345 запросов. Итоговый тестовый набор составил 1000 пар «запрос – ответ».

Для обучения детектора галлюцинаций сформирован тренировочный датасет объемом 3000 примеров. Как и для контрольной выборки, использованы два источника – ruSciBench и SberQuAD. В первом случае основу составили примеры из тестового набора AIR-Bench (1374 примера), для которых эталонные релевантные документы определены с помощью пайплайна на основе модели BGE-M3 [22] (порог реранкера 0.85). Оставшиеся 1626 примеров взяты из тестовой подвыборки SberQuAD, не использованной при формировании контрольной выборки.

ПАЙПЛАЙН RAG

Разработанный пайплайн RAG состоит из двух этапов: поиска документов и генерации ответа.

Поиск и формирование контекста

RusHalluRAG интегрирован с фреймворком RusBeIR [23] – платформой для оценки качества информационного поиска на русском языке. Модели BGE-M3 [22] используются как в качестве ретривера, так и в качестве реранкера. После ранжирования для формирования контекста при каждом запросе извлекаются пять наиболее релевантных фрагментов.

Для повышения разнообразия контекста и введения контролируемого уровня сложности применяется этап балансировки. Набор данных разбивается на шесть групп в зависимости от позиции первого релевантного документа (от 1 до 5 или отсутствие). Меньшие группы дополняются путем переноса примеров

из группы с позицией 1 на целевую позицию *k* и замены прочих релевантных документов до *k* на наиболее похожие нерелевантные фрагменты, подобранные с помощью BM-25. Это повышает сложность обработки контекста, поскольку релевантные документы заменяются на похожие по форме, но не совпадающие по содержанию. Описанная процедура применялась как для контрольной, так и для тренировочной выборок.

Генерация ответа

В качестве генератора ответа выбрана модель YandexGPT-5-Lite-8B-instruct с 8 млрд параметров. Вывод модели выполняется с использованием библиотеки vllm [24]. Наборы данных, включающие пять наиболее релевантных документов и соответствующие им сгенерированные ответы, публикуются в открытом доступе[4, 5, 6].

РАЗМЕТКА

Классы галлюцинаций

Предлагается таксономия из шести классов, ориентированная на ответы, не соответствующие контексту:

1) **противоречие** (Contradiction): ответ прямо противоречит информации в релевантном контексте (ср. Conflict в [5]);

2) **неподтвержденная деталь** (Unconfirmed): ответ содержит информацию, не имеющую подкрепления в релевантном контексте (ср. Baseless в [5]);

3) **выдуманный ответ** (Excess): ответ целиком построен на нерелевантных документах;

4) **подмена вопроса** (Missing): ответ не отвечает на поставленный вопрос, что приводит к бессвязности (ср. [2]);

5) **неполный ответ** (Partial): ответ содержит неполную информацию без указания на ее неполноту;

6) **ложный отказ** (Oversight): модель ошибочно отказывается отвечать при наличии релевантного контекста.

[4] https://huggingface.co/datasets/feudor2/sberquad-rag
[5] https://huggingface.co/datasets/feudor2/ruscibench-rag
[6] https://huggingface.co/datasets/feudor2/RuHallu-RAG-train

Для иллюстрации рассмотрим характерные примеры из контрольной выборки.

При запросе о факторах, подталкивающих врачей к применению нового лекарства, модель включает устную коммуникацию в число движущих факторов, тогда как контекст прямо указывает, что она оказывала тормозящее воздействие – это противоречие.

При запросе о педагогических взглядах Б. Рассела модель утверждает, что он выступал против религиозного образования, однако контекст упоминает лишь сходство его взглядов со взглядами Р. Оуэна и Ш. Фурье, которые выступали против, но не самого Рассела – это неподтвержденная деталь.

При запросе о преимуществах модифицированной системы остаточных классов для параллельной арифметики модель дает развернутый ответ на основе документа, который вообще не упоминает эту систему, а другие документы нерелевантны – это выдуманный ответ.

При запросе о крупнейших транснациональных корпорациях модель перечисляет только Pepsi и Coca-Cola, хотя контекст содержит еще несколько компаний (Нестле, Новартис, АББ) – это неполный ответ.

При запросе о вовлечении студентов в физический практикум модель выдает развернутый ответ о занятиях физической культурой и спортом, ошибочно связав запрос с этой темой вместо физики – это подмена вопроса.

При запросе определения геоэкономики модель утверждает, что в документах нет прямого определения, хотя контекст описывает геоэкономику как формирование единой мировой сетевой рыночной экономики – это ложный отказ.

Процесс разметки

В разметке контрольной выборки участвовали два эксперта и модель Google Gemini-2.5-Pro. Каждый пример независимо размечен человеком и LLM; разногласия разрешались при участии второго эксперта. Разметчики определяли наличие галлюцинации, классифицировали ее тип(ы) и выделяли соответствующие фрагменты текста, следуя блок-схеме на рис. 1. Разметка выполнялась в Label Studio [25].

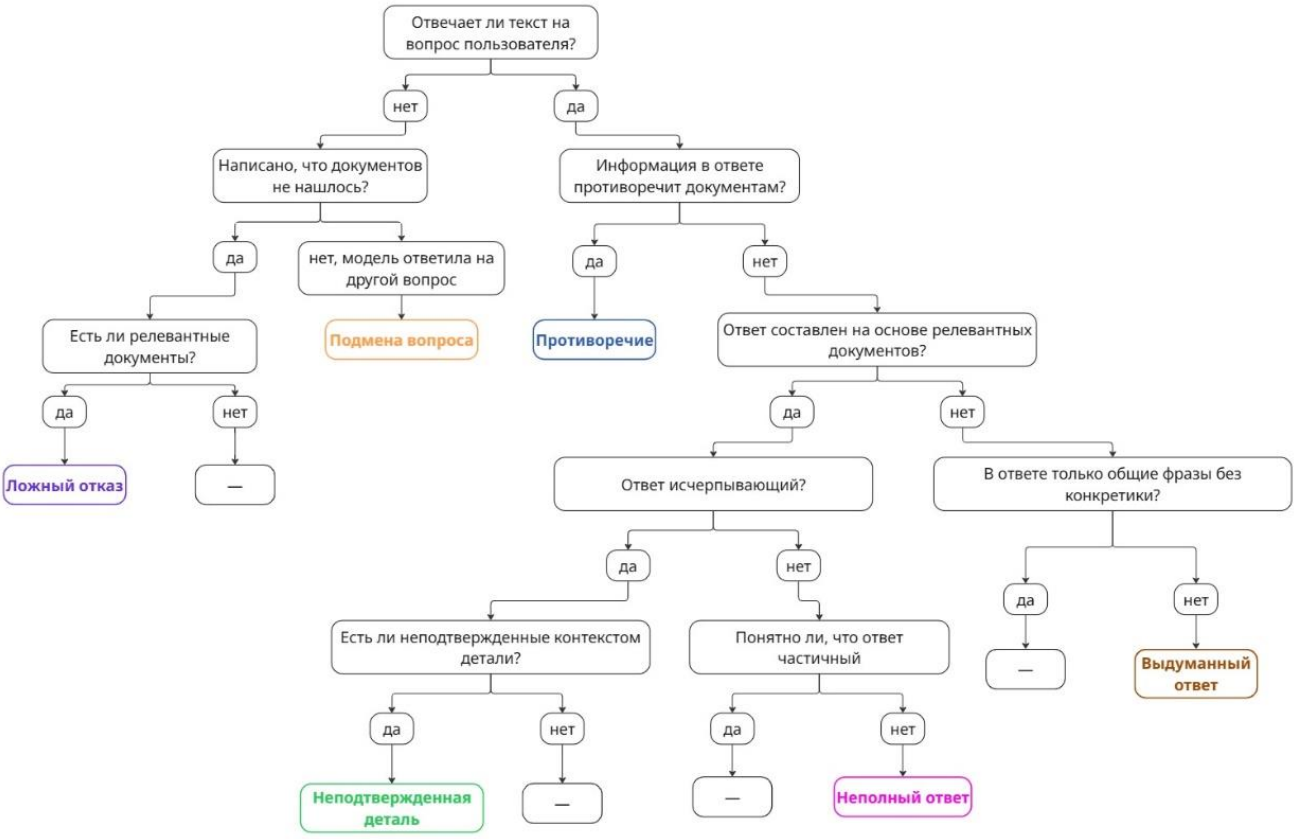


Рис. 1. Блок-схема с алгоритмом классификации для разметчиков.

Статистика

Согласованность между разметчиками составила 0.69–0.76 по коэффициенту альфа Криппендорфа. Распределение ответов представлено на рис. 2 и 3. Доля ответов без галлюцинаций (≈66%) отражает эффективность пайплайна RAG. ruSciBench существенно сложнее SberQuAD: доля ответов с галлюцинациями составляет 51.6% против 23.7%.

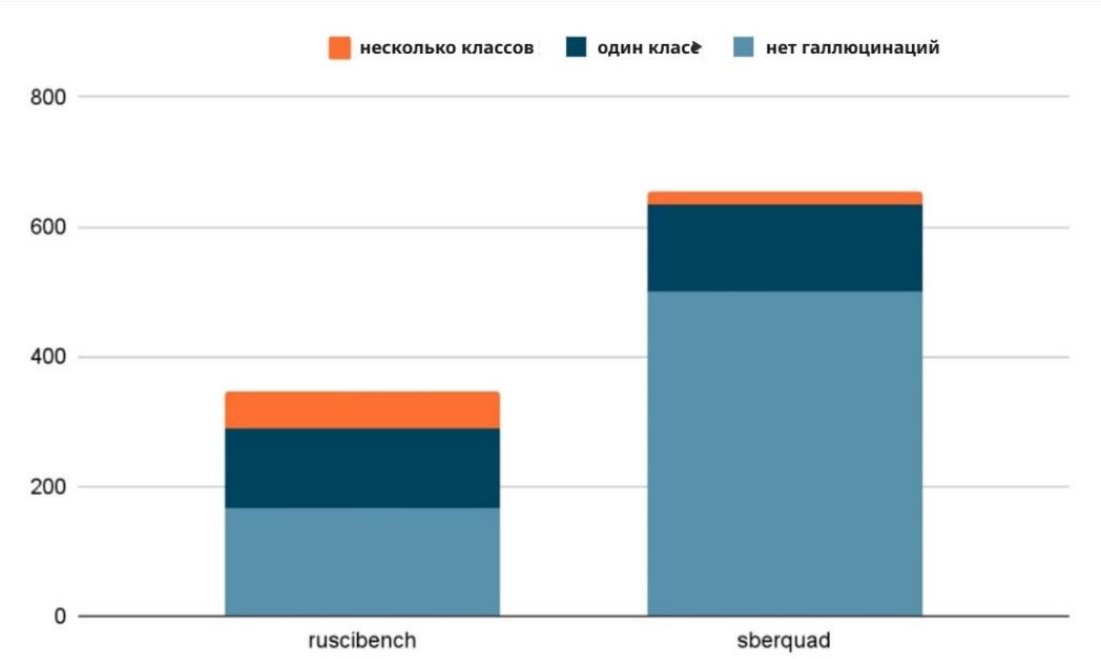


Рис. 2. Распределение ответов без галлюцинаций, ответов с одной галлюцинацией и ответов с несколькими галлюцинациями в контрольной подвыборке.

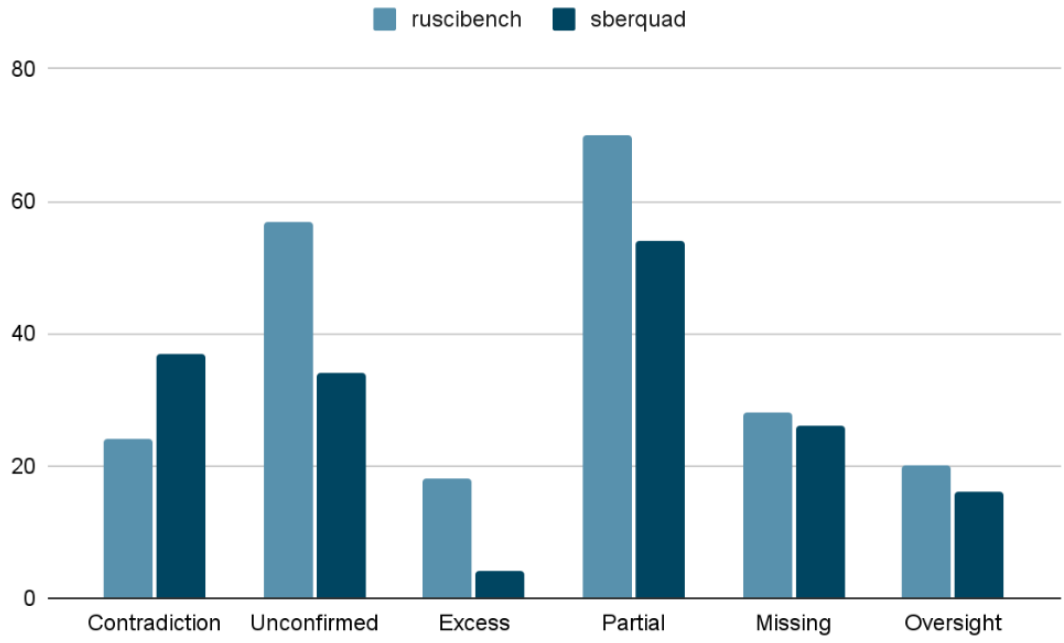


Рис. 3. Распределение галлюцинаций по каждому типу в контрольной подвыборке.

Для тренировочного набора (3000 примеров) разметка выполнена автоматически с использованием Gemini-2.5-Pro. Исходное распределение было несбалансированно: 71% примеров без галлюцинаций, среди галлюцинаций преобладают «Неподтвержденная деталь» (19.4%) и «Противоречие» (4.6%), остальные классы – менее 2.2%. Доля примеров с несколькими галлюцинациями составила 5.6%. Для устранения дисбаланса была разработана процедура синтетической балансировки, которая позволила поднять число примеров всех классов галлюцинаций до 200 за счет более представленных классов.

МЕТРИКИ

Качество обнаружения галлюцинаций оценивается на трех уровнях детализации, соответствующих схеме разметки: на уровнях бинарной классификации (наличие или отсутствие галлюцинации), многоклассовой классификации и выделения фрагментов, содержащих галлюцинации.

Для бинарной классификации использованы *точность* (Accuracy), *макроусредненные точность* (Precision), *полнота* (Recall) и F1-*мера*. Для многоклассовой классификации между наборами меток вычислялись *точность совпадения* (Exact Match Accuracy), *расстояние Хэмминга* (Hamming Loss) и *коэффициент Жаккара* (Jaccard Similarity). Для оценки выделения фрагментов применялись *ROUGE-L* [26] для оценки пересечения на уровне токенов, а также *точность выделения фрагментов* (Span Accuracy), *коэффициент Жаккара* и *расстояние Хэмминга* для множеств фрагментов. Вычислялся также коэффициент Success Rate – доля ответов с корректным форматированием JSON.

МОДЕЛИ

Выбор базовых моделей

Набор базовых моделей включает модели, перечисленные в табл. 1.

Табл. 1. Базовые модели и их параметры.

Модель	Число параметров	Доступность
YandexGPT-5-Lite-8B-instruct[7]	8 млрд	открытые веса
Qwen3-8B[8]	8 млрд	открытые веса

[7] https://huggingface.co/yandex/YandexGPT-5-Lite-8B-instruct
[8] https://huggingface.co/Qwen/Qwen3-8B

T-lite-it-2.1[9]	8 млрд	открытые веса
gpt-oss-20b[10]	20 млрд	открытые веса
gemma-3-27b-it[11]	27 млрд	открытые веса
T-pro-it-2.0[12]	33 млрд	открытые веса
Llama-3.3-70B-Instruct[13]	70 млрд	открытые веса
glm-4.6v[14]	106 млрд	открытые веса
Qwen3-235B-A22B-Instruct-2507[15]	235 млрд (22 млрд активных)	открытые веса
DeepSeek-v3-0324[16]	685 млрд	открытые веса
kimi-k2-thinking[17]	1.1 трлн	открытые веса
gemini-2.5-flash[18]	неизвестно	по API
GPT5.1-pro[19]	неизвестно	по API
YandexGPT-5.1[20]	неизвестно	по API
gemini-2.5-pro[21]	неизвестно	по API

Такой выбор позволил провести систематический анализ взаимосвязи между размером модели и качеством обнаружения галлюцинаций, а также учесть архитектурное разнообразие в сопоставимых диапазонах размеров.

Настройка вывода

Вывод для моделей с числом параметров до 33 млрд выполнялся локально на графических процессорах NVIDIA A100 с использованием библиотеки vllm [24] и тензорного параллелизма для распределенного вывода. Более крупные модели с открытым весом и закрытые модели использовались через API

[9] https://huggingface.co/t-tech/T-lite-it-2.1
[10] https://huggingface.co/openai/gpt-oss-20b
[11] https://huggingface.co/google/gemma-3-27b-it
[12] https://huggingface.co/t-tech/T-pro-it-2.0
[13] https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct
[14] https://huggingface.co/deepseek-ai/DeepSeek-V3-0324
[15] https://huggingface.co/Qwen/Qwen3-235B-A22B-Instruct-2507
[16] https://huggingface.co/deepseek-ai/DeepSeek-V3-0324
[17] https://huggingface.co/moonshotai/Kimi-K2-Thinking
[18] https://ai.google.dev/gemini-api/docs/models/gemini-2.5-flash
[19] https://openai.com/index/gpt-5-1/
[20] https://ya.ru/ai/gpt
[21] https://ai.google.dev/gemini-api/docs/models/gemini-2.5-pro

провайдера VseLLM[22]. Температура семплирования была установлена на 0 для всех моделей для обеспечения детерминированности результатов.

Для автоматизированной оценки модели были ограничены выводом структурированного JSON-массива. Каждый массив содержит объекты с двумя полями:

1) **span**: точная текстовая подстрока, содержащая галлюцинацию (null при отсутствии);

2) **type**: один из классов: Contradiction, Unconfirmed, Excess, Partial, Missing, Oversight (null при отсутствии галлюцинации).

Такой формат обеспечивает единообразную обработку ответов и вычисление метрик для всех базовых моделей.

Параметры обучения детектора

Для обучения детектора на сбалансированном тренировочном датасете применялась методика SFT на базе YandexGPT-5-Lite-8B-instruct. Обучение выполнялось на одном GPU NVIDIA H100 в течение трех эпох с использованием оптимизатора PagedAdam8bit, learning rate 5×10^{-5}, weight decay 0.05, точности bf16 и параметров LoRA ($\alpha = 32$, $r = 32$). Размер валидационной подвыборки составлял 10%. Модель обучалась предсказывать разметку в формате JSON-массива с полями type и span.

РЕЗУЛЬТАТЫ

Общий обзор

В табл. 2–4 представлены результаты для трех подзадач. Метрики даны с макроусреднением[23], результаты отсортированы по Accuracy.

Табл. 2. Общие метрики бинарной классификации (P = Precision, R = Recall, SR = Success Rate).

Модель	Accuracy	F1	P	R	SR
gemini-2.5-pro	**81.3**	**78.5**	**79.2**	**78.0**	**100**
gemini-2.5-flash	75.2	72.2	72.1	72.3	**100**

[22] https://vsellm.ru/

[23] Следует отметить, что в некоторых случаях F1-мера может быть ниже как точности, так и полноты. Это не является ошибкой: макроусредненная F1-мера вычисляется не из макроточности и макрополноты, а путем усреднения F1-мер каждого класса.

glm-4.6v	70.1	45.0	50.2	42.0	89.5
gpt5.1-pro	69.5	45.2	45.1	46.0	99.4
T-pro-it-2.0	65.0	42.3	51.1	38.1	81.9
gpt-oss-20b	63.0	38.8	49.6	35.5	81.9
kimi-k2-thinking	56.4	38.3	57.6	31.8	63.9
YandexGPT-5.1	54.3	33.5	41.9	29.9	73.5
llama-3.3-70b-instruct	41.6	39.0	59.4	54.6	**100**
deepseek-chat-v3-0324	37.0	25.9	39.6	32.7	79.7
Qwen3-235B-A22B-Instruct-2507	35.5	29.5	55.0	50.9	**100**
Qwen3-8B	30.2	16.0	10.9	30.2	92.7
YandexGPT-5-8B-Lite-instruct	28.6	15.9	30.7	28.5	88.3
T-lite-it-2.1	28.0	15.5	30.5	27.9	88.4
gemma-3-27b-it	26.8	17.9	46.6	26.7	67.0
YandexGPT-5-Lite-SFT	66.7	29.1	39.0	34.0	99.9

Табл. 3. Общие метрики многоклассовой классификации (P = Precision, R = Recall, HL = Hamming Loss,).

Модель	Accuracy	HL ↓	Jaccard	F1	P	R
gemini-2.5-pro	**71.6**	**7.0**	**34.1**	**47.0**	**49.8**	**46.0**
gemini-2.5-flash	63.1	9.5	21.9	31.7	33.2	34.0
glm-4.6v	61.7	10.0	23.7	34.6	40.7	32.0
T-pro-it-2.0	58.1	11.1	19.9	29.3	38.8	25.4
gpt-oss-20b	57.4	11.3	15.8	23.1	41.9	20.8
gpt5.1-pro	55.0	11.4	17.1	25.9	33.0	29.1
YandexGPT-5.1	52.0	12.6	12.6	18.2	30.3	17.4
kimi-k2-thinking	51.9	12.6	18.6	27.6	48.7	21.2
deepseek-chat-v3-0324	17.3	21.3	12.9	22.2	25.1	33.7
llama-3.3-70b-instruct	16.1	22.1	8.3	14.7	19.6	24.7
Qwen3-235B-A22B-Instruct-2507	11.4	22.9	10.5	18.1	24.7	34.8
gemma-3-27b-it	6.5	24.2	7.8	13.9	25.0	22.0
YandexGPT-5-8B-Lite-instruct	5.3	24.0	5.2	9.6	16.8	17.1
Qwen3-8B	4.7	26.1	7.0	12.6	23.8	25.9
T-lite-it-2.1	3.2	27.6	3.8	7.1	12.9	17.9
YandexGPT-5-Lite-SFT	65.8	9.2	9.1	11.5	17.1	13.1

Табл. 4. Общие метрики выделения фрагментов (HL = Hamming Loss).

Модель	Accuracy	HL ↓	Jaccard	ROUGE-L
gemini-2.5-pro	**71.7**	**25.7**	**73.6**	**79.4**
gpt-oss-20b	67.7	30.7	68.7	70.6
T-pro-it-2.0	66.1	32.6	67.0	71.1
glm-4.6v	65.8	32.5	66.9	72.2
kimi-k2-thinking	65.2	34.2	65.5	69.4
YandexGPT-5.1	63.8	36.1	63.9	65.2
gemini-2.5-flash	61.9	35.5	63.6	70.7
gpt5.1-pro	56.7	41.9	58.0	65.5
deepseek-chat-v3-0324	35.5	60.7	37.8	47.9
gemma-3-27b-it	28.7	68.4	30.1	42.0
llama-3.3-70b-instruct	13.5	83.0	15.1	30.5
YandexGPT-5-8B-Lite-instruct	12.7	86.2	13.5	26.3
T-lite-it-2.1	10.4	86.4	11.9	22.8
Qwen3-8B	8.9	87.9	10.4	23.2
Qwen3-235B-A22B-Instruct-2507	8.3	87.5	10.3	25.3
YandexGPT-5-Lite-SFT	65.7	34.3	65.7	66.3

Бинарная классификация. Наивысшее качество демонстрируют закрытые модели: Gemini-2.5-Pro достигает 81.3%, Gemini-2.5-Flash – 75.2%, GPT5.1-Pro – 69.5%, YandexGPT-5.1 – 54.3%. Среди моделей с открытым весом GLM-4.6v показывает конкурентоспособный результат 70.1%, T-pro-it-2.0 и gpt-oss-20b достигают 65% и 63% соответственно, однако их соблюдение формата составляет лишь 81.9%. Крупные модели Llama-3.3-70B и Qwen3-235B демонстрируют низкую точность (41.6% и 35.5%) при идеальном соблюдении формата. Модели с числом параметров до 8 млрд не превышают 31%. Дообученная YandexGPT-5-Lite-8B достигает точности 66.7% (прирост +38.1 п.п. относительно zero-shot-версии с 28.6%), занимая четвертое место и решая проблему несоблюдения формата (99.9% SR).

Многоклассовая классификация. Наилучший результат принадлежит Gemini-2.5-Pro с точностью 71.6%. Gemini-2.5-Flash следует с заметным отставанием (63.1%), а GLM-4.6v приближается к нему (61.7%). Модели T-pro-it-2.0, gpt-oss-20b, GPT5.1-Pro, YandexGPT-5.1 и Kimi-k2-thinking группируются в диапазоне

51.9–58.1% с различными компромиссами между точностью и полнотой. Значительный разрыв наблюдается ниже 20%: DeepSeek-v3, Llama-3.3-70B и Qwen3-235B показывают 17.3%, 16.1% и 11.4% соответственно, тогда как модели с числом параметров ≤ 8 млрд опускаются ниже 7%. Обученная модель достигает 65.8% – второй результат после Gemini-2.5-Pro, с приростом +60.5 п.п. относительно исходной модели.

Выделение фрагментов. Gemini-2.5-Pro уверенно лидирует с точностью 71.7%. Следующий уровень, включающий gpt-oss-20b, T-pro-it-2.0, GLM-4.6v, Kimi-k2-thinking и YandexGPT-5.1, показывает конкурентоспособные результаты 63.8–67.7% с ROUGE-L 65.2–72.2. Gemini-2.5-Flash опускается на седьмое место (61.9%), что позволяет предположить, что выделение фрагментов в меньшей степени выигрывает от оптимизации Flash. Остальные модели резко деградируют: DeepSeek достигает лишь 35.5%, Gemma-3-27b-it – 28,7%, Llama-3.3-70B и модели с числом параметров ≤ 8 млрд опускаются ниже 15%. Обученная модель показывает 65.7% (четвертое место), уступая только Gemini-2.5-Pro, gpt-oss-20b и T-pro-it-2.0, с приростом +53 п.п. относительно исходной модели.

Анализ по датасетам. Значения Accuracy по датасетам представлены в табл. 5. Все модели демонстрируют более низкие результаты на ruSciBench по сравнению с SberQuAD, что подтверждает бóльшую сложность научного домена. Для бинарной классификации точность Gemini-2.5-Pro снижается с 84.6% до 75.1% (падение 9.5 п.п.), для Gemini-2.5-Flash – с 79.8% до 66.4% (13.4 п.п.), для GLM-4.6v – с 76.8% до 57.4% (19.4 п.п.). Разрыв еще более выражен для многоклассовой классификации: Gemini-2.5-Pro теряет 21.2 п.п., Gemini-2.5-Flash – 28.7 п.п. Выделение фрагментов демонстрирует наибольшее падение: Gemini-2.5-Pro – 22.3 п.п., Gemini-2.5-Flash – 29.9 п.п., gpt-oss-20b – 27.2 п.п., что подчеркивает дополнительные сложности научной терминологии и точного определения границ. Дообученная модель показывает устойчивую производительность: на SberQuAD точность составляет 76.2% (бинарная), 75.3% (многоклассовая) и 75.3% (фрагменты); на ruSciBench – 48.5%, 47.8% и 47.5% соответственно. Падение при переходе к научному домену (27.7–27.8 п.п.) сопоставимо с показателями ведущих моделей. Средняя точность обученной модели 63.7% выводит ее на второе

место после Gemini-2.5-Pro, превосходя исходную модель (16.5%) на 47.2 п.п., что подтверждает эффективность предложенного подхода.

Табл. 5. Метрики точности (бинарная, многоклассовая и по фрагментам) на датасетах SberQuAD и ruSciBench (B = BINARY, ML = MULTI-LABEL, SQ = SberQuAD, RB = ruSciBench, Сред. = среднее по ряду)

Model	B SQ	B RB	ML SQ	ML RB	span SQ	span RB	Сред.
gemini-2.5-pro	**84.6**	**75.1**	**78.9**	**57.7**	**79.4**	**57.1**	**72.1**
gemini-2.5-flash	79.8	66.4	73.0	44.3	72.2	42.3	63.0
glm-4.6v	76.8	57.4	71.9	42.3	74.5	49.3	62.0
gpt-oss-20b	72.5	44.9	69.0	35.4	77.1	49.9	58.1
T-pro-it-2.0	75.1	45.8	70.4	34.8	76.9	45.5	58.1
gpt5.1-pro	73.4	62.0	64.6	36.8	66.3	38.6	57.0
YandexGPT-5.1	64.3	35.4	62.6	31.9	72.8	46.7	52.3
kimi-k2-thinking	76.0	19.1	71.6	14.5	74.8	47.0	50.5
deepseek-chat-v3-0324	29.9	50.4	17.3	17.4	44.7	18.0	29.6
llama-3.3-70b-instruct	36.2	51.9	18.6	11.3	17.9	5.2	23.5
gemma-3-27b-it	16.8	45.8	5.3	8.7	39.2	8.7	20.8
Qwen3-235B-A22B…	26.9	51.9	10.4	13.3	9.6	5.8	19.7
YandexGPT-5-8B…	20.9	43.2	4.4	7.0	14.7	9.0	16.5
Qwen3-8B	22.3	45.2	3.8	6.4	9.5	7.8	15.8
T-lite-it-2.1	21.8	39.7	3.2	3.2	10.2	10.7	14.8
YandexGPT-5-Lite-SFT	76.2	48.5	75,3	47.8	75.3	47.5	63.7

ОБСУЖДЕНИЕ

Оценка 15 моделей выявила несколько ключевых закономерностей в способностях LLM к обнаружению галлюцинаций.

Размер модели не гарантирует высокого качества. Модели среднего размера (20–33 млрд параметров) могут превосходить более крупные: T-pro-it-2.0 и gpt-oss-20b (средняя точность 58.1%) опережают Llama-3.3-70B (23.5%) и Qwen3-235B (19.7%), несмотря на в 2–11 раз меньшее число параметров. Крупные модели демонстрируют катастрофический провал в многоклассовой классификации (ниже 19%) и выделении фрагментов на ruSciBench (5–6%), что указывает на недостаточность одного лишь масштаба для семантически точного анализа галлюцинаций. Обученная модель YandexGPT-5-Lite-8B (8 млрд) достигает средней

точности 63.7%, подтверждая, что целенаправленная донастройка на сбалансированных синтетических данных может быть эффективнее простого увеличения числа параметров.

Соблюдение формата необходимо, но недостаточно. Модели с идеальным соблюдением формата демонстрируют среднюю точность в диапазоне от 19.7% до 72.1% – разрыв в 52.4 п.п. Llama-3.3-70B и Qwen3-235B безупречно следуют ограничениям на вывод, но достигают лишь 23.5% и 19.7% средней точности с практически нулевой производительностью в задачах детальной разметки. Среди открытых моделей конкурентоспособные варианты имеют лишь 64–82% соблюдения формата, что создает операционные риски. Обученная модель с почти идеальным соблюдением формата (99.9% SR) полностью устраняет эту проблему.

Закрытые модели сохраняют значительное преимущество. Gemini-2.5-Pro уверенно лидирует (средняя точность 72.1%, на 8.4 п.п. выше лучшей открытой модели) с превосходной устойчивостью на ruSciBench: 75.1% (бинарная), 57.7% (многоклассовая) и 57.1% (фрагменты). Разрыв между закрытыми и открытыми моделями достигает 20–35 п.п. на сложных примерах. Обученная модель сокращает этот разрыв до 8.4 п.п. по средней точности.

Сложность задачи создает разрывы в возможностях. Бинарное обнаружение решается наиболее эффективно (наилучший результат 84.6%), но показывает разрывы между датасетами в 27–57 п.п. Многоклассовая классификация демонстрирует самое большое расхождение: ведущие модели достигают 70–79% на SberQuAD, но только 35–58% на ruSciBench. Выделение фрагментов показывает схожие разрывы ниже первых восьми моделей (5–18% точности).

Только четыре модели превышают 60% средней точности, причем лишь две (Gemini-2.5-Pro и обученная YandexGPT-5-Lite-8B) находятся выше 63%. Наиболее остро этот разрыв проявляется для научных областей и многоклассовой классификации, что подчеркивает критическую потребность в надежных, локально развертываемых русскоязычных детекторах и определяет приоритетные направления для будущих разработок, ориентированных на научный и технический контент.

ОГРАНИЧЕНИЯ

Настоящее исследование имеет несколько важных ограничений, которые следует признать. Во-первых, и это наиболее существенно, все ответы в RAG-системе были сгенерированы одной моделью. Это означает, что RusHallu-RAG не является классическим RAG-бенчмарком, оценивающим множество генерационных моделей; скорее, он фокусируется исключительно на способностях к обнаружению галлюцинаций в рамках фиксированной конфигурации. Данное проектное решение ограничивает обобщаемость полученных результатов на различные языковые модели и может не охватывать полный спектр паттернов галлюцинаций, которые могли бы возникать при использовании различных архитектур.

Во-вторых, бенчмарк ограничен двумя конкретными областями: общим знанием и научными документами. Хотя эти области восполняют недостаточно изученный пробел в оценке русскоязычных RAG-систем, они не охватывают другие важные области, такие как новости, корпоративные документы или специализированные технические сферы, которые могут демонстрировать иные характеристики галлюцинаций.

В-третьих, объем тестового набора данных из 1000 пар «запрос – ответ», хотя и достаточен для первоначальных целей исследования, может быть ограничен для обучения устойчивых моделей обнаружения галлюцинаций или для сбора редких типов галлюцинаций в достаточном количестве.

Наконец, автоматическая разметка тренировочного датасета с помощью Gemini-2.5-Pro потенциально вносит смещение, поскольку модель обучается имитировать решения этой конкретной LLM, что может не всегда быть оптимальным с точки зрения согласованности с экспертной оценкой человека. Кроме того, синтетические методы балансировки классов, хотя и эффективны для устранения дисбаланса, могут не полностью воспроизводить распределение и структуру галлюцинаций в реальных данных.

ЗАКЛЮЧЕНИЕ

Впервые представлен RusHallu-RAG – первый бенчмарк для обнаружения галлюцинаций в русскоязычных RAG-системах. Бенчмарк объединяет наборы данных из областей общего знания и науки с контролируемой балансировкой

контекста. В основу разметки была заложена расширенная таксономию галлюцинаций из шести классов, адаптированная для RAG. Разметка примеров была проведена как на уровне ответов, так и на уровне фрагментов текста, что обеспечивает большую детализацию по сравнению с множеством альтернативных решений.

Помимо контрольного множества, был сформирован сбалансированный тренировочный датасет объемом 3000 примеров с использованием синтетических методов генерации для выравнивания распределения классов. На его основе с применением дообучения с учителем был обучен детектор галлюцинаций на базе YandexGPT-5-Lite-8B.

Оценка шестнадцати LLM, включая обученную модель, позволяет сделать следующие выводы.

Научные тексты являются более сложными для обнаружения галлюцинаций, чем тексты из области общего знания.

Качество обнаружения не масштабируется монотонно с ростом числа параметров: модели среднего размера иногда превосходят более крупные. Обученная модель с 8 млрд параметров достигает средней точности 63.7 п.п., обходя значительно более крупные модели с открытыми весами.

Дообучение с учителем на сбалансированных синтетических данных является эффективной стратегией: прирост средней точности по сравнению с исходной моделью без дообучения составил 47.2 п.п., так что обученный детектор занимает второе место после ведущей закрытой модели.

Значительный разрыв сохраняется между закрытыми моделями и моделями с открытым весом. Этот разрыв подчеркивает потребность в точных, локально развертываемых русскоязычных детекторах галлюцинаций. Представленный подход к синтетической балансировке и дообучению демонстрирует многообещающий путь к созданию таких детекторов, особенно для научно-технического контента, где потребность в надежных решениях наиболее остра.

Благодарности

Авторы выражают благодарность Дарье Парамоновой за ценный вклад в процесс разметки данных. Исследование выполнено с использованием суперкомпьютера МГУ-270 Московского государственного университета имени

М. В. Ломоносова. Часть работы по обучению и тестированию моделей подготовлена при поддержке Центра социальных технологий Yandex.Cloud и реализована на платформе Yandex Cloud.

СПИСОК ЛИТЕРАТУРЫ

1. *Ridder F., Schilling M.* The HalluRAG Dataset: Detecting Closed-Domain Hallucinations in RAG Applications Using an LLM's Internal States // Proc. of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2025. https://doi.org/10.48550/arXiv.2412.17056

2. *Agarwal V., et al.* MedHalu: Hallucinations in Responses to Healthcare Queries by Large Language Models // Findings of the Association for Computational Linguistics: ACL 2025. 2025. https://doi.org/10.1609/icwsm.v20i1.42623

3. *Liu T., Zhang Y., Brockett C., Mao Y., Sui Z., Chen W., Dolan B.* A Token-Level Reference-Free Hallucination Detection Benchmark for Free-Form Text Generation // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers). Dublin, Ireland, May 2022. P. 6723–6737. https://doi.org/10.18653/v1/2022.acl-long.464

4. *Li J., Cheng X., Zhao X., Nie J.-Y., Wen J.-R*. HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models // Proc. of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore, December 2023. P. 6449–6464. https://doi.org/10.18653/v1/2023.emnlp-main.397

5. *Niu C., Wu Y., Zhu J., Xu S., Shum K., Zhong R., Song J., Zhang T.* RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models // Proc. of the 62nd Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers). Bangkok, Thailand, August 2024. P. 10862–10878. https://doi.org/10.18653/v1/2024.acl-long.585

6. *Zhu Z., Yang Y., Sun Z.* HaluEval-Wild: Evaluating Hallucinations of Language Models in the Wild // Advances in Neural Information Processing Systems. 2024. Vol. 37. https://doi.org/10.48550/arXiv.2403.04307

7. *Vazquez R., et al.* SemEval-2025 Task 3: Mu-SHROOM, the Multilingual Shared-Task on Hallucinations and Related Observable Overgeneration Mistakes //

Proc. of the 19th International Workshop on Semantic Evaluation (SemEval-2025). Vienna, Austria, July 2025. P. 2472–2497. URL: https://aclanthology.org/2025.semeval-1.322/ (дата обращения 20.08.2026).

8. *Chernogorskii F., et al.* DRAGOn: Designing RAG on periodically updated corpus // Proc. of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Vol. 4: Student Research Workshop). 2026. P. 622–638. https://doi.org/10.18653/v1/2026.eacl-srw.48

9. *Xiong G., Jin Q., Lu Z., Zhang A.* Benchmarking Retrieval-Augmented Generation for Medicine // Findings of the Association for Computational Linguistics: ACL 2024. Bangkok, Thailand, August 2024. C. 6233–6251. https://doi.org/10.18653/v1/2024.findings-acl.372

10. *Tang Y., Yang Y.* Multihop-RAG*:* Benchmarking Retrieval-Augmented Generation for Multi-Hop Queries // arXiv:2401.15391. 2024. https://doi.org/10.48550/arXiv.2401.15391

11. *Yang X. et al.* CRAG – Comprehensive RAG Benchmark // Advances in Neural Information Processing Systems. 2024. Vol. 37. P. 10470–10490. https://doi.org/10.52202/079017-0335

12. *Pipitone N., Alami G.H.* LegalBench-RAG: A Benchmark for Retrieval-Augmented Generation in the Legal Domain // arXiv:2408.10343. 2024. https://doi.org/10.48550/arXiv.2408.10343

13. *Es S., James J., Espinosa Anke L., Schockaert S.* RAGAs: Automated Evaluation of Retrieval Augmented Generation // Proc. of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations. St. Julians, Malta, March 2024. P. 150–158. https://doi.org/10.18653/v1/2024.eacl-demo.16

14. *Saad-Falcon J., Khattab O., Potts C., Zaharia M.* ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems // Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Vol. 1: Long Papers). Mexico City, Mexico, June 2024. P. 338–354. https://doi.org/10.18653/v1/2024.naacl-long.20

15. *Friel R., Belyi M., Sanyal A.* RAGBench: Explainable Benchmark for Retrieval-Augmented Generation Systems // arXiv:2407.11005. 2024. https://doi.org/10.48550/arXiv.2407.11005

16. *Wang Z., et al.* RAGRouter-Bench: A Dataset and Benchmark for Adaptive RAG Routing // arXiv:2602.00296. 2026. https://doi.org/10.48550/arXiv.2602.00296

17. *Bao F.S., et al.* FaithBench: A Diverse Hallucination Benchmark for Summarization by Modern LLMs // Proc. of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Vol. 2: Short Papers). Albuquerque, New Mexico, April 2025. P. 448–461. https://doi.org/10.18653/v1/2025.naacl-short.38

18. *Tamber M.S., et al.* Benchmarking LLM Faithfulness in RAG with Evolving Leaderboards // Proc. of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track. Suzhou (China), November 2025. P. 799–811. https://doi.org/10.18653/v1/2025.emnlp-industry.54

19. *Efimov P., Chertok A., Boytsov L., Braslavski P.* SberQuAD–Russian Reading Comprehension Dataset: Description and Analysis // 11th Conference and Labs of the Evaluation Forum, CLEF 2020. Springer Science and Business Media Deutschland GmbH, 2020. P. 3–15. https://doi.org/10.1007/978-3-030-58219-7_1

20. *Vatolin A., Gerasimenko N., Ianina A., Vorontsov K.* RuSciBench: Open Benchmark for Russian and English Scientific Document Representations // Doklady Mathematics. 2024. Vol. 110. P. S251–S260.
https://doi.org/10.1134/S1064562424602191

21. *Chen J., et al.* AIR-Bench: Automated Heterogeneous Information Retrieval Benchmark // Proc. of the 63rd Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers). Vienna, Austria, July 2025. P. 19991–20022. https://doi.org/10.18653/v1/2025.acl-long.982.

22. *Chen J., et al.* M3-Embedding: Multilinguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation // Findings of the Association for Computational Linguistics: ACL 2024. Bangkok, Thailand, August 2024. P. 2318–2335. https://doi.org/10.18653/v1/2024.findings-acl.137

23. *Kovalev G., Tikhomirov M., Kozhevnikov E., Kornilov M., Loukachevitch N.* Building Russian Benchmark for Evaluation of Information Retrieval Models // arXiv:2504.12879. 2025. https://doi.org/10.48550/arXiv.2504.12879

24. *Kwon W., et al.* Efficient Memory Management for Large Language Model Serving with PagedAttention // Proc. of the 29th Symposium on Operating Systems

Principles, SOSP '23. New York, NY, USA: Association for Computing Machinery, 2023. P. 611–626. https://doi.org/10.1145/3600006.3613165

25. *Tkachenko M., Malyuk M., Holmanyuk A., Liubimov N.* Label Studio: Data Labeling Software. 2020–2025. [Электронный ресурс].
URL: https://github.com/HumanSignal/label-studio (дата обращения: 20.08.2026).

26. *Lin C.-Y.* ROUGE: A Package for Automatic Evaluation of Summaries // Text Summarization Branches Out. Barcelona, Spain, July 2004. P. 74–81.
URL: https://aclanthology.org/W04-1013/ (дата обращения: 20.08.2026).

RusHallu-RAG: A COMPREHENSIVE APPROACH TO HALLUCINATION DETECTION IN RUSSIAN-LANGUAGE RAG SYSTEMS

F. A. Sadkovskii [1] [0009-0006-4502-9466], R. R. Nasyrova [2] [0009-0008-6280-0667], A. A. Sorokin [3, 4] [0000-0003-3877-4223]

[1–3]*Lomonosov Moscow State University AI Center, Moscow, Russia*

[4]*Yandex, Moscow, Russia*

[1]sadkovskiyfa@my.msu.ru, [2]NasyrovaRR_01@my.msu.ru, [3]sorokinaa@my.msu.ru

Abstract

This paper introduces RusHallu-RAG, the first comprehensive benchmark for detecting hallucinations in Russian-language RAG systems. The dataset of 1,000 query-answer pairs covers general knowledge and scientific domains with controlled context relevance balancing. A fine-grained taxonomy of six hallucination types is proposed and used for annotation at both response and span levels with human experts and LLMs.

For training a hallucination detector, a balanced dataset of 3,000 examples is collected using synthetic generation methods. A detector based on YandexGPT-5-Lite-8B is trained via SFT. We evaluate 15 open-weight and proprietary models alongside the trained detector. Results show that model size does not guarantee performance: medium-sized models (20–33B) sometimes outperform larger ones. The trained detector achieves an average accuracy of 63.7%, ranking second after proprietary Gemini-2.5-Pro, with a gain of 47.2 p.p. over the baseline. The gap between proprietary

and open-weight models is reduced from 20–35 to 8.4 p.p., confirming the effectiveness of the approach. The benchmark, code, and trained model are publicly released.

***Keywords**: hallucination detection, retrieval-augmented generation (RAG), large language models (LLMs), benchmarking, russian language, natural language processing (NLP), question answering (QA), model evaluation, supervised fine-tuning (SFT), hallucination taxonomy.*

REFERENCES

1. *Ridder F., Schilling M.* The HalluRAG Dataset: Detecting Closed-Domain Hallucinations in RAG Applications Using an LLM's Internal States // Proc. of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2025. https://doi.org/10.48550/arXiv.2412.17056

2. *Agarwal V., et al.* MedHalu: Hallucinations in Responses to Healthcare Queries by Large Language Models // Findings of the Association for Computational Linguistics: ACL 2025. 2025. https://doi.org/10.1609/icwsm.v20i1.42623

3. *Liu T., Zhang Y., Brockett C., Mao Y., Sui Z., Chen W., Dolan B.* A Token-Level Reference-Free Hallucination Detection Benchmark for Free-Form Text Generation // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers). Dublin, Ireland, May 2022. P. 6723–6737. https://doi.org/10.18653/v1/2022.acl-long.464

4. *Li J., Cheng X., Zhao X., Nie J.-Y., Wen J.-R*. HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models // Proc. of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore, December 2023. P. 6449–6464. https://doi.org/10.18653/v1/2023.emnlp-main.397

5. *Niu C., Wu Y., Zhu J., Xu S., Shum K., Zhong R., Song J., Zhang T.* RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models // Proc. of the 62nd Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers). Bangkok, Thailand, August 2024. P. 10862–10878. https://doi.org/10.18653/v1/2024.acl-long.585

6. *Zhu Z., Yang Y., Sun Z.* HaluEval-Wild: Evaluating Hallucinations of Language Models in the Wild // Advances in Neural Information Processing Systems. 2024. Vol. 37. https://doi.org/10.48550/arXiv.2403.04307

7. *Vazquez R., et al.* SemEval-2025 Task 3: Mu-SHROOM, the Multilingual Shared-Task on Hallucinations and Related Observable Overgeneration Mistakes // Proc. of the 19th International Workshop on Semantic Evaluation (SemEval-2025). Vienna, Austria, July 2025. P. 2472–2497.
URL: https://aclanthology.org/2025.semeval-1.322/.

8. *Chernogorskii F., et al.* DRAGOn: Designing RAG on periodically updated corpus // Proc. of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Vol. 4: Student Research Workshop). 2026. P. 622–638. https://doi.org/10.18653/v1/2026.eacl-srw.48

9. *Xiong G., Jin Q., Lu Z., Zhang A.* Benchmarking Retrieval-Augmented Generation for Medicine // Findings of the Association for Computational Linguistics: ACL 2024. Bangkok, Thailand, August 2024. C. 6233–6251.
https://doi.org/10.18653/v1/2024.findings-acl.372

10. *Tang Y., Yang Y.* Multihop-RAG*:* Benchmarking Retrieval-Augmented Generation for Multi-Hop Queries // arXiv:2401.15391. 2024.
https://doi.org/10.48550/arXiv.2401.15391

11. *Yang X. et al.* CRAG – Comprehensive RAG Benchmark // Advances in Neural Information Processing Systems. 2024. Vol. 37. P. 10470–10490.
https://doi.org/10.52202/079017-0335

12. *Pipitone N., Alami G.H.* LegalBench-RAG: A Benchmark for Retrieval-Augmented Generation in the Legal Domain // arXiv:2408.10343. 2024.
https://doi.org/10.48550/arXiv.2408.10343

13. *Es S., James J., Espinosa Anke L., Schockaert S.* RAGAs: Automated Evaluation of Retrieval Augmented Generation // Proc. of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations. St. Julians, Malta, March 2024. P. 150–158.
https://doi.org/10.18653/v1/2024.eacl-demo.16

14. *Saad-Falcon J., Khattab O., Potts C., Zaharia M.* ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems // Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Vol. 1: Long Papers). Mexico City, Mexico, June 2024. P. 338–354. https://doi.org/10.18653/v1/2024.naacl-long.20

15. *Friel R., Belyi M., Sanyal A.* RAGBench: Explainable Benchmark for Retrieval-Augmented Generation Systems // arXiv:2407.11005. 2024. https://doi.org/10.48550/arXiv.2407.11005

16. *Wang Z., et al.* RAGRouter-Bench: A Dataset and Benchmark for Adaptive RAG Routing // arXiv:2602.00296. 2026. https://doi.org/10.48550/arXiv.2602.00296

17. *Bao F.S., et al.* FaithBench: A Diverse Hallucination Benchmark for Summarization by Modern LLMs // Proc. of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Vol. 2: Short Papers). Albuquerque, New Mexico, April 2025. P. 448–461. https://doi.org/10.18653/v1/2025.naacl-short.38

18. *Tamber M.S., et al.* Benchmarking LLM Faithfulness in RAG with Evolving Leaderboards // Proc. of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track. Suzhou (China), November 2025. P. 799–811. https://doi.org/10.18653/v1/2025.emnlp-industry.54

19. *Efimov P., Chertok A., Boytsov L., Braslavski P.* SberQuAD–Russian Reading Comprehension Dataset: Description and Analysis // 11th Conference and Labs of the Evaluation Forum, CLEF 2020. Springer Science and Business Media Deutschland GmbH, 2020. P. 3–15. https://doi.org/10.1007/978-3-030-58219-7_1

20. *Vatolin A., Gerasimenko N., Ianina A., Vorontsov K.* RuSciBench: Open Benchmark for Russian and English Scientific Document Representations // Doklady Mathematics. 2024. Vol. 110. P. S251–S260. https://doi.org/10.1134/S1064562424602191

21. *Chen J., et al.* AIR-Bench: Automated Heterogeneous Information Retrieval Benchmark // Proc. of the 63rd Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers). Vienna, Austria, July 2025. P. 19991–20022. https://doi.org/10.18653/v1/2025.acl-long.982.

22. *Chen J., et al.* M3-Embedding: Multilinguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation // Findings of the Association for Computational Linguistics: ACL 2024. Bangkok, Thailand, August 2024. P. 2318–2335. https://doi.org/10.18653/v1/2024.findings-acl.137

23. *Kovalev G., Tikhomirov M., Kozhevnikov E., Kornilov M., Loukachevitch N.* Building Russian Benchmark for Evaluation of Information Retrieval Models // arXiv:2504.12879. 2025. https://doi.org/10.48550/arXiv.2504.12879

24. *Kwon W., et al.* Efficient Memory Management for Large Language Model Serving with PagedAttention // Proc. of the 29th Symposium on Operating Systems Principles, SOSP '23. New York, NY, USA: Association for Computing Machinery, 2023. P. 611–626. https://doi.org/10.1145/3600006.3613165

25. *Tkachenko M., Malyuk M., Holmanyuk A., Liubimov N.* Label Studio: Data Labeling Software. 2020–2025. URL: https://github.com/HumanSignal/label-studio (Accessed 20.08.2026).

26. *Lin C.-Y.* ROUGE: A Package for Automatic Evaluation of Summaries // Text Summarization Branches Out. Barcelona, Spain, July 2004. P. 74–81. URL: https://aclanthology.org/W04-1013/ (Accessed: 20.08.2026).

СВЕДЕНИЯ ОБ АВТОРАХ



САДКОВСКИЙ Федор Алексеевич – младший научный сотрудник ИЦИИ МГУ имени М. В. Ломоносова, г. Москва, Россия.

Fedor Alekseevich SADKOVSKII – junior researcher at AI Center, Lomonosov Moscow State University, Moscow, Russia.

email: sadkovskiyfa@my.msu.ru

ORCID: 0009-0006-4502-9466



НАСЫРОВА Регина Руслановна – младший научный сотрудник ИЦИИ МГУ имени М. В. Ломоносова, г. Москва, Россия.

Regina Ruslanovna NASYROVA – junior researcher at AI Center, Lomonosov Moscow State University, Moscow, Russia.

email: nasyrovarr_01@my.msu.ru

ORCID: 0009-0008-6280-0667



СОРОКИН Алексей Андреевич – старший научный сотрудник ИЦИИ МГУ имени М. В. Ломоносова, г. Москва, Россия.

Alexey Andreevich SOROKIN – senior researcher at AI Center, Lomonosov Moscow State University, Moscow, Russia.

Email: sorokinaa@my.msu.ru

ORCID: 0000-0003-3877-4223

Материал поступил в редакцию 27 августа 2026 года

УДК 004.934

АВТОМАТИЧЕСКИЙ МОРФОЛОГИЧЕСКИЙ АНАЛИЗ ДЛЯ ОСЕТИНСКОГО ЯЗЫКА НА МАТЕРИАЛЕ УСТНОГО КОРПУСА

А. С. Шатских[1] [0009-0006-9025-7783], А. А. Сорокин[2] [0000-0003-3877-4223]

[1]*Институт языкознания РАН, г. Москва, Россия*
[2]*Московский государственный университет имени М. В. Ломоносова, г. Москва, Россия*
[2]*Яндекс, г. Москва, Россия*

[1]anapriselec@gmail.com, [2]alexey.sorokin@list.ru

Аннотация

Работа посвящена созданию первого корпуса для осетинского языка с морфологической разметкой в формате Universal Dependencies. Корпус состоит из размеченных вручную текстов Устного корпуса осетинского языка, содержит 5454 предложения и имеет суммарный объем 73042 токена. Представлен морфологический анализатор на основе архитектуры BERT, достигающий пословной корректности 95.60%. Приведены результаты экспериментов по улучшению качества классификации. Сделан вывод, что фильтрация выдачи модели с помощью бесконтекстного анализатора и многозадачный подход не приводят к значительному улучшению. Представлены также результаты тестирования анализатора на внедоменных данных и продемонстрировано что на них модель достигает пословной корректности 91.45%.

Ключевые слова*: осетинский язык, иронский диалект, стандарт морфологической аннотации, морфологический анализ.*

ВВЕДЕНИЕ

Осетинский язык является представителем индоевропейской семьи (иранская группа) индоевропейской языковой семьи. Язык распространен преимущественно на Северном Кавказе. Согласно Всероссийской переписи населения 2020 г., число его носителей на территории России составляет 344907 человек. На сегодняшний день существуют два основных корпуса осетинского

© А. С. Шатских, А. А. Сорокин, 2026.
Данная статья распространяется на условиях международной лицензии Creative Commons License Attribution 4.0 International (CC BY 4.0).

языка: Осетинский национальный корпус (ОНК)[1] и Устный корпус осетинского (УКО)[2].

Осетинский национальный корпус содержит литературные тексты преимущественно современных авторов (1999–2004 гг.), а также произведения наиболее известных авторов XX в. Кроме того, он включает тексты литературных журналов «Мах дуг» («Наша эпоха») и «Ногдзау» («Пионер») и материалы интернет-газеты «Спутник», а также наиболее полное собрание Нартовского эпоса. Объем корпуса составляет приблизительно 12 млн словоупотреблений. В корпусе автоматически размечены части речи, леммы и морфологические признаки; аннотация выполнена с помощью анализатора Uniparser-Morph [1]. Поскольку анализатор применяется бесконтекстно, в корпусе не снята омонимия, и для каждого словоупотребления предлагается множество всех возможных разборов.

Устный корпус осетинского содержит устные тексты, собранные в ходе лингвистических экспедиций 2007–2013 гг. в различных районах Республики Северная Осетия – Алания. Корпус также содержит выпуски радиопередачи «Одноклассники». Объем корпуса составляет приблизительно 60 тыс. словоупотреблений. Он снабжен поморфемными разборами (глоссами), которые выполнены вручную в программе SIL FieldWorks Language Explorer.

Несмотря на то что из формата FieldWorks возможно извлечь частеречные метки и морфологические признаки, аннотация УКО является неполной и недостаточно подробной. С одной стороны, это затрудняет автоматическое приведение данных к международному стандарту морфологической аннотации Universal Dependencies [2], с другой – это препятствует использованию корпуса в образовательных и исследовательских целях. Таким образом, существует необходимость в создании корпуса осетинского языка с морфологической разметкой со снятой омонимией, который бы соответствовал стандартам Universal Dependencies версии 2 (UD v2). Настоящее исследование направлено на решение этой задачи. Настоящая статья является расширенной версией работы, представленной на конференции «Диалог-2026» [3].

[1]http://corpus.ossetic-studies.org (дата обращения 14.08.2026)
[2]https://www.ossetic-studies.org/ru/texts/iron (дата обращения 14.08.2026).

ЛИНГВИСТИЧЕСКИЕ СВЕДЕНИЯ

Осетинский язык подразделяется на два диалекта: иронский и дигорский. В настоящей работе термины «осетинский язык» и «иронский диалект осетинского языка» используются как синонимы.

Осетинский язык обладает богатой морфологией. Именная морфология является агглютинативной, однако демонстрирует регулярные чередования гласных и согласных при образовании форм множественного числа [4]. Глагольная морфология имеет ряд синтетических черт, таких как наличие словоизменительных классов, супплетивное или нерегулярное образование основы прошедшего времени, а также кумулятивное выражение лица и числа. Продуктивным способом образования новых глаголов является создание сложных глаголов. Сложный глагол состоит из именной части и легкого глагола, который выбирается из закрытого класса.

Порядок слов в клаузе свободный, немаркированным порядком является SOV. Как правило, вершина составляющей в осетинском языке расположена справа. Порядок слов в именной группе фиксирован и следует иерархии Poss > Det > Num > Adj, Gen > N [5], где Poss – генитив посессора, Gen – низкий нереферентный генитив.

ИСТОЧНИК ДАННЫХ

Цель настоящего исследования – создание корпуса с морфологической аннотацией для осетинского языка. В качестве источника данных выбран корпус УКО, поскольку он уже содержит ручную морфологическую аннотацию в нестандартном формате со снятой омонимией. Исходные морфологические метки приводятся к формату CoNLL-U согласно правилам разметки UD v2 [2] и их расширениям, которые описаны в следующем разделе.

Приведение данных УКО к формату CoNLL-U включает следующие этапы:

1) фильтрация данных не на целевом языке, в частности удаление всех полностью русскоязычных предложений;

2) ручная сегментация предложений по многоточию и символу верхнего регистра, а также сегментация многословных токенов;

3) ручное приведение меток частей речи и морфологических признаков к универсальному перечню;

4) ручное исправление аннотации корпуса в соответствии с универсальными правилами разметки и их расширениями. Этот этап требует знания осетинской грамматики и является наиболее трудоемким;

5) автоматическое приведение данных к формату CoNLL-U, включая заполнение по умолчанию тех значений признаков, которые отсутствуют в исходных поморфемных разборах (например, именительный падеж существительных).

Итоговый датасет состоит из 5454 предложений и имеет суммарный объем 73 042 токена. Датасет разделен на обучающую, валидационную и тестовую выборки в отношении 8:1:1. Распределение частот поджанров в каждой из подвыборок приведено в Приложении А, распределение частот морфологических меток – в Приложении Б.

ПРАВИЛА АННОТАЦИИ

Ряд особенностей осетинской грамматики требует расширения стандартного набора признаков UD v2 и их значений. В этом разделе описаны расширения, принятые в исходной системе аннотации корпуса и введенные нами в настоящем исследовании.

Многие исследователи (например, В. И. Абаев [4]) отмечают, что в осетинском языке не существует четких границ между существительными, прилагательными, наречиями и послелогами. Одна и та же последовательность символов может выполнять функции различных частей речи; так, например, слово «хорз» может быть переведено как «хороший» или как «хорошо». В УКО существительным, прилагательным, послелогам и некоторым наречиям присваивается одна и та же метка части речи. Система UD v2, в свою очередь, задает универсальный инвентарь частей речи, а часть речи словоформы в рамках этой системы определяется на основании морфосинтаксических критериев. Для различения послелогов, наречий, прилагательных и существительных приняты следующие эвристики:

1) имена без падежно-числовой морфологии, модифицирующие глагол и управляющие генитивом или посессивной проклитикой, считаются послелогами (ADP) (например, «бын» в примере 1);

2) имена без падежно-числовой морфологии, модифицирующие глагол, считаются наречиями (ADV) (например, «рæсугъд» в примере 2);

3) имена без падежно-числовой морфологии, модифицирующие другие имена (кроме закрытого класса интенсификаторов типа тынг ‘очень’), считаются прилагательными (ADJ) (например, «ирон» в примере 3);

4) имена с падежно-числовой морфологией и имена без падежно-числовой морфологии, не покрытые в предыдущих пунктах, считаются существительными (NOUN).

Приведём примеры.

(1) *Æмæ зæй-ы бын фе-сты адæм…*
и лавина-gen под pv-быть.3pl человек.pl
‘И под лавиной оказались люди…’

(2) *Ды тынг рæсугъд дзур-ыс.*
2sg очень красиво говорить-prs.2sg
‘Ты очень красиво говоришь.’

(3) *Фæлæ ирон æгъдæу-тт-æ…*
но осетинский обычай-pl-nom
‘Но осетинские обычаи…’

В именных предикациях и сложных глаголах синтаксические критерии, описанные выше, неприменимы. В таких случаях имя по умолчанию считается существительным; в случае, если имя имеет показатель сравнительной степени или зависимый интенсификатор, оно относится к прилагательным (ADJ). Так, прилагательным считается «кадджын» в примере 4.

(4) *Æмæ гъе уыдон ны-ххаст-а, æмæ уый тынг кадджын уыд-и.*
и ну DemDist.pl pv-брать-pst.3sg и DemDist очень почетный быть-pst.3sg
‘И он взял их, и это было очень почетно.’

В осетинском языке существует закрытый класс глаголов, способных функционировать в качестве вспомогательных; однако все представители этого класса также могут выступать в роли лексических глаголов. Хотя в исходной разметке УКО глаголы в лексической и вспомогательной функциях аннотируются одинаково, мы утверждаем, что их различение не только методологически необходимо, но и осуществимо на основе синтаксических критериев. Глагол считается вспомогательным (AUX) в следующих конфигурациях:

1) легкий глагол в сложных глаголах (например, «кодтат» в примере 5);

2) глагол-связка в именных предикациях (например, «уыди» в примере 6);

3) собственно вспомогательный глагол, т. е. финитный глагол в аналитических глагольных конструкциях (например, «нæй» в примере 7).

Приведём примеры.

(5) *Куыд ахуыр код-т-ат скъола-йы?*
как учеба делать-tr-pst.2pl школа-in
‘Как вы учились в школе?’

(6) *Алагирский район-ы хицау уыд-и.*
Алагирский район-gen хозяин быть-pst.3sg
‘Он был главой Алагирского района.’

(7) *Уым дæр нæй фыст.*
DemDist.in add быть.neg.prs.3sg писать.ptcp.pst
‘Этого там тоже не написано.’

Для классификации местоименных единиц на местоимения-существительные (PRON) и местоимения-прилагательные (DET) применяются те же морфосинтаксические критерии, которые используются для различения существительных (NOUN) и прилагательных (ADJ). Решение обусловлено тем, что лексемы, обычно выступающие как именные модификаторы, могут подвергаться субстантивации, т. е. возглавлять именную группу, выступать в роли аргумента глагола или выражать грамматические категории существительных (например, «ахæмтæ» в примере 8). В таких случаях им присваиваются метка PRON и признаки падежа и числа. Те же принципы применяются к лексеме-плейсхолдеру «йед» (как в примере 9). Следует отдельно отметить, что у лексемы «йед» отсутствует признак местоименного разряда (PronType), так как в инвентаре значений этого признака отсутствует подходящее.

Рассмотрим примеры.

(8) *Уыд-и дзы ахæм-т-æ, æртын фондз*
быть-PST.3SG 3SG.IN такой-PL-NOM тридцать пять
аз-ы кæ-уыл цыд-и.
год-NUM кто-SUPER идти-PST.3SG
‘Были там и те, кому было 35 лет.’

(9) *Æмæ дын йед-мæ куы ба-хæццæ дæн…*

и 2SG.DAT HES-ALL когда PV-смешанный быть.PRS.1SG
бын-мæ…
низ-ALL
‘И когда я пришла туда… вниз…’

Субстантивация местоименных наречий в первую очередь происходит путем присоединения падежной морфологии. При этом получающиеся словоформы продолжают функционировать как модификаторы имен или глаголов. Чтобы подчеркнуть отличие субстантивированных наречий от полноценных случаев субстантивации, описанных выше, мы не присваиваем признаку числа значение по умолчанию (Number = Sing) и аннотируем этот признак только тогда, когда число выражено ненулевым показателем (например, «афтæтæ» в примере 10).

(10) *Ам* *дæр* *афтæ-т-æ нæ* *уыд-ысты.*
DemProx.IN ADD так-PL-NOM NEG быть-3PL
‘И здесь так не было.’

В именной парадигме можно выделить от девяти до двенадцати падежей [6]. Директив и рецессив, которые были грамматикализованы позднее других, отсутствуют в классических грамматиках; однако мы включаем их в парадигму в соответствии с отмеченным исследованием [6]. Для директива и рецессива вводятся дополнительные значения признака Case с названиями Dir и Rcs.

Статус нумератива как отдельного падежа для существительных в конструкциях с числительными остается предметом дискуссий. В первую очередь это связано с тем, что показатель счетной формы (*-ы*) омонимичен показателю генитива. Однако, можно привести как минимум два убедительных аргумента в пользу его включения в падежную парадигму.

Во-первых, в дигорском диалекте осетинского языка показатель счетной формы может сочетаться с любым падежом. Это указывает на то, что нумератив представляет собой отдельную грамматическую функцию, а не одну из функций генитива. С другой стороны, в то время, как дигорский диалект предоставляет достаточные основания для отнесения нумератива к категории числа, в иронском диалекте таких оснований недостаточно, при этом необходимость различать нумератив и генитив существует.

Во-вторых, в иронском диалекте у существительных *pluralia tantum* нумератив омонимичен номинативу, а не генитиву. Согласно двум приведенным соображениям, Num вводится в качестве дополнительного значения признака падежа.

Когда конструкция с числительным выступает в роли прямого дополнения, для определения падежа существительного применяются правила дифференцированного маркирования объекта (DOM) [7]. Согласно этим правилам одушевленные прямые дополнения, обозначающие людей, маркируются генитивом, неодушевленные прямые дополнения – номинативом. Соответственно, в конструкциях с числительными генитив присваивается существительным, обозначающим людей (например, одушевленное *лæппуйы* в примере 11), нумератив – неодушевленным существительным (например, неодушевленное *рæнхъы* в примере 12). Поскольку правила DOM для одушевленных объектов, которые не обозначают людей, недостаточно описаны, такие имена обрабатываются аналогично неодушевленным существительным согласно эвристике (например, *карчы* в примере 13).

Приведём примеры.

(11) *Уыцы* *дыууæ* *лæппу-йы* *с-хъомыл* *код-т-а.*
DemDist два мальчик-GEN PV-взрослый делать-TR-PST.3SG
‘Она вырастила этих двух мальчиков.’

(12) *…аст* *рæнхъ-ы* *ны-ффыст-он.*
восемь строка-NUM PV-писать-PST.1SG
‘...я написала восемь строк.’

(13) *…æмæ* *иу-дæс* *карч-ы* *а-ргæвст-он,*
и один-десять курица-NUM PV-забивать-PST.1SG
иуæндæс *карч-ы.*
одиннадцать курица-NUM
‘...и забила около десяти куриц, одиннадцать куриц.’

Широкий перечень нефинитных глагольных форм в осетинском языке требует введения новых значений для признака VerbForm. Введены значения Inf2, Part2, Dest и Red. Соответствие нефинитных форм значениям признака VerbForm приведено в табл. 1 на примере глагола *дзурын* ‘говорить’.

Табл. 1. Нефинитные глагольные формы.

Форма	Значение VerbForm	Функция
дзур-ын	Inf	Инфинитив
дзур-æн	Inf2	II инфинитив
дзырд	Part	Причастие прошедшего времени (страдательное)
дзур-инаг	Part	Причастие будущего времени
дзур-æг	Part	Причастие настоящего времени (действительное)
дзур-аг	Part2	Причастие настоящего времени (действительное, хабитуальное)
дзур-гæ	Conv	Конверб
дзур-æггаг	Dest	Дестинативное причастие
ра-дзур-ба-дзур	Red	Номинализация в сложных глаголах

Использование для осетинских глагольных форм признака вида (Aspect) привело бы к формированию неоднородных категорий, не соответствующих традиции описания языка. Вместо этого предлагается использовать признак Preverb = Yes для словоформ, содержащих глагольные приставки (превербы), поскольку именно они выражают видовые значения (например, *æрбахонон* в примере 14).

(14) *Æрба-хон-он* *дын* *æй?*
PV-звать-SBJV.1SG 2SG.DAT 3SG.GEN
‘Мне позвать ее?’

Кроме того, в языке существуют единицы *æд* и *æнæ,* которые могут быть отражены на письме как именные приставки или предлоги (см. *æнæ* в примерах 15 и 16). Нами вводится признак Prenoun = Yes и используется в первом случае для указания на наличие именной приставки в словоформе.

Представим примеры.

(15) *Адæм* *æнæ-хæлд* *гъе* *уæд* *уыд-ысты.*

Человек.PL NEG-портить.PTCP.PST ну тогда быть-PST.3PL
‘Люди тогда не были испорченными.’

(16) *Æз та ænæ очки-т-æй næ уыд-т-он бинаг фыст.*
1SG PTCL без очки-PL-ABL NEG видеть-TR-PST.1SG нижний писать.PTCP.PST
‘Без очков я не увидела нижнюю надпись.’

СОЗДАНИЕ МОРФОЛОГИЧЕСКОГО АНАЛИЗАТОРА

Помимо создания корпуса с морфологической аннотацией, целью нашего исследования было создание морфологического анализатора, обученного на данных этого корпуса. Учитывая успех подхода трансферного обучения и предобученных энкодерных языковых моделей в задаче морфологического анализа (см., в частности, [8, 9]), мы предложили анализатор на основе архитетуры BERT [10]. Поскольку языковые модели для осетинского языка отсутствуют, модель архитектуры BERT обучается стандартной задаче маскированного языкового моделирования, а затем дообучается задаче морфологического анализа. Предобученная модель архитектуры BERT также может быть использована для других задач автоматической обработки текста на осетинском языке.

Для предобучения используются неразмеченные данные ОНК. Данные проходят автоматическую предобработку, которая состоит из следующих этапов:

1) NFC-нормализация;

2) локальная дедупликация путем удаления повторяющихся строк в пределах одного документа;

3) исключение фрагментов HTML и JavaScript с помощью регулярных выражений;

4) глобальная дедупликация строк, встречающихся более 50 раз во всем корпусе;

5) фильтрация документов по порогу минимальной длины в 200 символов;

6) применение алгоритма LSH (Locality Sensitive Hashing) для обнаружения и удаления приближенных дубликатов.

Предбработанные данные разделяются на предложения с помощью регулярного выражения, и в них нормализуется регистр. Кроме того, удаляются предложения, соответствующие по крайней мере одному из следующих критериев:

- предложение содержит только издательскую информацию или метаданные;
- в предложении отсутствуют кириллические символы или символы нижнего регистра;
- в предложении содержатся кириллические и латинские символы в пределах одного токена (за исключением латинского символа *æ*).

В качестве кандидатов на оптимальную отправную точку используются следующие модели:

1) mBERT: мультиязычная модель архитектуры BERT от Google среднего размера, учитывающая регистр[3];

2) осетинская модель BERT: версия mBERT с замененным слоем эмбеддингов, который адаптирован к осетинскому языку согласно методу [11].

Для совместимости мультиязычной модели с символьным составом обучающей выборки стандартная нормализация токенизатора mBERT дополняется следующими шагами:

- кириллический символ *æ* приводится к его латинскому аналогу, поскольку только последний присутствует в словаре модели mBERT;
- различные варианты написания кавычек и апострофа (U+2018, U+2019, U+02BC, U+0027, U+00B4, U+0060) приводятся к символу U+02BC. Единообразие представлений различных вариантов критически важно, поскольку символ апострофа или кавычки в осетинском языке выполняют функцию показателя сандхи на границе слов.

Инициализация весов осетинской модели BERT производится следующим образом. На данных ОНК обучается токенизатор системы WordPiece с размером словаря в 25000 токенов. Для сохранения знаний исходной модели каждое вхождение из словаря нового токенизатора разбивается с помощью токенизатора исходной модели и эмбеддинги полученных токенов усредняются. После

[3]https://huggingface.co/google-bert/bert-base-multilingual-cased (дата обращения 14.08.2026)

этого слой эмбеддингов исходной модели заменяется на конкатенацию эмбеддингов вхождений осетинского словаря.

Обе модели дообучаются задаче маскированного языкового моделирования в течение пяти эпох с размером батча 64 и темпом обучения 5e-5. Значения функции потерь моделей после обучения представлены в табл. 2. Отметим, что эти значения нельзя сравнивать напрямую, поскольку словари моделей несопоставимы. Это отражено в среднем числе токенов на слово, которое измерено на тестовой выборке датасета Ossetic-COT и также приведено в табл. 2.

Табл. 2. Результаты предобучения моделей.

Модель	Значение функции потерь	Среднее число токенов на слово
Мультиязычная	0.6086	3.29
Осетинская	3.4311	1.21

В табл. 3 представлены результаты дообучения в течение 10 эпох для мультиязычной модели и в течение 13 эпох для осетинской модели с размером батча 8 и темпом обучения 5e-5. Значения усреднены по результатам пяти независимых запусков обучения. Для каждой модели оптимальные гиперпараметры подобраны с помощью валидационной выборки, и результаты приведены только для наилучшего набора гиперпараметров.

Табл. 3. Результаты дообученияmоделей.

Модель	Корректность	Корректность по предложениям
Мультиязычная	**95.60**	**61.54**
Осетинская	95.47	60.11

Поскольку различие между моделями в корректности классификации незначительно, в качестве итоговой модели выбран классификатор на основе мультиязычной модели BERT (mBERT). Значения точности, полноты и F-меры для этой модели приведены в табл. 4. Таблица содержит значения, усредненные по

всем классам с помощью макроусреднения и взвешенного усреднения, а также значения, усредненные по всем классам, соответствующим каждой части речи, с помощью макроусреднения и взвешенного усреднения. Все значения также усреднены по результатам пяти независимых запусков обучения.

Табл. 4. Расширенные метрики мультиязычной модели.

	Точность		Полнота		F-мера	
Часть речи	Макро	Вес	Макро	Вес	Макро	Вес
ADJ	0.78	0.91	0.81	0.92	0.79	0.91
ADP	0.99	0.99	0.96	0.96	0.97	0.97
ADV	0.84	0.95	0.88	0.97	0.86	0.96
AUX	0.64	0.86	0.66	0.88	0.64	0.86
CCONJ	0.99	0.99	0.99	0.99	0.99	0.99
DET	0.89	0.96	0.87	0.93	0.88	0.94
INTJ	0.97	0.98	0.97	0.98	0.97	0.98
NOUN	0.78	0.93	0.78	0.94	0.78	0.93
NUM	0.85	0.99	0.85	0.98	0.85	0.98
PART	0.97	0.98	0.99	0.99	0.98	0.99
PRON	0.81	0.92	0.82	0.92	0.80	0.92
PROPN	0.44	0.95	0.50	0.96	0.46	0.96
PUNCT	1.00	1.00	1.00	1.00	1.00	1.00
SCONJ	0.93	0.86	0.91	0.82	0.92	0.84
VERB	0.77	0.92	0.79	0.92	0.77	0.92
X	0.87	0.87	0.66	0.66	0.75	0.75
Макро	0.77		0.78		0.77	
Вес	0.95		0.96		0.95	

Для фильтрации выдачи модели на этапе получения предсказаний создается бесконтекстный морфологический анализатор на основе конечного преобразователя в программе Foma [12], а также список основ из закрытых классов и словоформ из закрытых классов для единиц с нерегулярными парадигмами. Алгоритм фильтрации для одного токена состоит из следующих этапов.

1. Поиск словоформы в списке словоформ из закрытых классов.

2. Порождение множества разборов с помощью бесконтекстного анализатора.

3. Поиск основы из каждого из разборов в списке основ из закрытых классов.

4. Получение распределения вероятностей по классам из нейронной сети для классификации.

5. Приведение к нулю вероятностей всех классов, не входящих в множество бесконтекстных разборов для данного токена.

6. Выбор наиболее вероятного класса из распределения в качестве предсказания для данного токена.

Из токенов тестовой выборки лишь для меньше 1% наиболее вероятный класс в распределении нейросетевого классификатора исключается бесконтекстным анализатором. Таким образом, фильтрация представляется избыточной и не используется в итоговой модели.

Для качественной оценки работы модели использован набор псевдоминимальных пар предложений для основных случаев морфологической омонимии в осетинском языке. Каждая пара состоит из предложений, где целевое слово имеет различный верный разбор, при этом окружающий контекст отличается минимально. Задачей классификатора является предсказание правильного разбора в обоих случаях. Тестовый набор представлен в Приложении В, целевые слова выделены полужирным. В табл. 5 приведены предсказания модели для каждой пары в тестовом наборе.

Табл. 5. Разрешение омонимии мультиязычной моделью.

Часть речи	Контекст 1	Контекст 2
Adj/Noun	Adj	Noun
Adj/Adv	Adj	Adv
Noun/Adp	Noun	Noun
1SG/INF	1SG	INF
GEN/IN	GEN	IN
GEN/NUM (неодуш.)	GEN	NUM
GEN/NUM (одуш.)	NUM	NUM

IN/PRS.3SG	GEN	GEN
DemDist.NOM/DemDist.GEN	DemDist.NOM	DemDist.GEN
1PL.NOM/1PL.GEN	1PL.NOM	1PL.GEN
1SG.POSS/1SG.GEN (энкл.)	1SG.POSS	1SG.GEN (энкл.)
1SG.POSS/союз ‘и’	1SG.POSS	союз ‘и’
2SG.POSS/2SG.GEN (энкл.)	2SG.POSS	2SG.GEN (энкл.)
2SG.GEN (энкл.)/быть.2SG	2SG.GEN (энкл.)	быть.2SG
1PL.POSS/1PL.GEN (энкл.)	1PL.POSS	NEG
1PL.POSS/NEG	1PL.POSS	NEG
EMPH/NEG	EMPH	NEG
ADD/NEG	ADD	NEG
ITER/‘один’	ITER	‘один’

Анализ результатов показывает, что наиболее сложными являются следующие случаи омонимии: Noun/Adp (существительное и послелог), GEN/NUM (генитив и нумератив) для существительных, обозначающих людей, 1PL.GEN/NEG (генитив местоимения 1-го лица единственного числа и показатель отрицания) и IN/PRS.3SG (инессив и показатель 3-го лица единственного числа настоящего времени глагола).

Отметим, что для некоторых случаев омонимии предсказанная метка отличается для разных запусков обучения. Кроме того, незначительные лексические изменения в тестовых примерах могут приводить к предсказанию неверного разбора для целевых словоформ. Эти явления могут быть обусловлены количественными и качественными ограничениями датасета Ossetic-COT. Отмеченные ограничения можно преодолеть с помощью обогащения обучающей выборки более объемными и лексически разнообразными языковыми данными.

Детальный анализ ошибок модели на тестовой выборке позволяет выделить следующие основные классы ошибок:

- неразрешимая неоднозначность, или случаи, в которых контекст не позволяет определить верный разбор либо существует более одного верного разбора;
- ошибки из-за несбалансированности данных, или случаи, когда для словоформы существует более одного верного разбора, при этом один из них значительно более частотен в обучающей выборке, чем другие, и ошибочно предсказывается вместо других;
- орфографические ошибки, или случаи, когда моделью предсказан верный разбор для словоформы, похожей на целевую, но не для целевой;
- дистантные зависимости, или случаи, когда фактор, определяющий верный разбор, находится в предложении на значительном расстоянии от целевой словоформы.

Стоит отметить языковые категории, представляющие наибольшую сложность. Ошибки возникают при определении нефинитных глагольных форм, а также при разграничении вспомогательных и лексических употреблений глаголов.

ОГРАНИЧЕНИЯ

Примечательно, что как валидационная, так и тестовая выборки содержат комбинации признаков, отсутствующие в обучающей выборке. Более того, подмножество теоретически возможных комбинаций признаков не встречается ни в одной из подвыборок. Проблема распознавания таких незнакомых классов может быть решена с помощью подхода многозадачного обучения (см., например, [13, 9]). При таком подходе значение каждого морфологического признака предсказывается моделью отдельно. В табл. 6 представлены результаты обучения многозадачных морфологических анализаторов на основе мультиязычной и осетинской моделей BERT. Использованы размер батча 8 и темп обучения 5e-5, обучение длится 25 эпох для мультиязычной модели и 20 эпох для осетинской модели. Можно видеть, что использование многозадачного подхода не позволяет существенно превзойти показатели классификатора, предсказывающего цельные морфологические метки. Таким образом, приоритет отдается последнему в силу его простоты и вычислительной эффективности.

Табл. 6. Результаты дообучения многозадачных моделей.

Модель	Корректность	Корректность по предложениям
Мультиязычная	95.54	62.39
Осетинская	95.45	60.92

Как было отмечено при анализе ошибок, корпус обладает ограниченным лексическим разнообразием. Более того, синтаксические структуры в большинстве предложений либо просты, либо характерны для устной речи. Эти особенности данных приводят к более низкому качеству при применении модели к данным из других источников и других жанров. Так, при измерении качества модели на выборке из 200 предложений из осетинской Википедии корректность по словам составляет 91.45, корректность по предложениям – 41.00. Наиболее часто ошибки совершаются при распознавании сокращений, собственных имен и числительных, записанных цифрами. Адаптация анализатора к данным различных жанров остается направлением для будущих исследований.

ЗАКЛЮЧЕНИЕ

Представлен первый корпус осетинского языка с морфологической аннотацией, выполненной в соответствии со стандартом Universal Dependencies версии 2. Корпус подходит как для обучения нейросетевых морфологических анализаторов, так и для исследовательских и образовательных целей, поскольку его разметка отражает большинство морфологических признаков и различий, присущих языку. Кроме того, представлена базовая модель, которая может служить для оценки сложности датасета и является первым нейросетевым морфологическим анализатором для осетинского языка. Корпус и модель находятся в свободном доступе[4][5].

Основным направлением будущих исследований является преодоление текущих количественных и качественных ограничений корпуса путем его обогащения разнообразными литературными текстами из ОНК.

[4]https://github.com/ania3000/Ossetic-COT (дата обращения 14.08.2026)
[5]https://huggingface.co/ossetic-encoders/ossbert-morph (дата обращения 14.08.2026)

Благодарности

Авторы выражают благодарность Олегу Игоревичу Беляеву за идею исследования и за помощь в разработке специфичных для языка правил аннотации, а также Олегу Игоревичу Беляеву и Арсению Павловичу Выдрину за предоставление неразмеченных данных ОНК.

Исследование поддержано проектом FMNE-2025-0002, выполняемом в Институте языкознания РАН.

Список сокращений

1 – 1 лицо, 2 – 2 лицо, 3 – 3 лицо, ABL – аблатив, ADD – аддитивная частица, ALL – аллатив, DAT – датив, DemDist – дальний дейктик, DemProx – ближний дейктик, GEN – генитив, HES – плейсхолдер, IN – инессив, INF – инфинитив, ITER– итеративная частица, NEG – показатель отрицания, NOM – номинатив, NUM – нумератив, PL – множественное число, POSS – притяжательное местоимение, PRS – настоящее время, PST – прошедшее время, PTCL – частица, PTCP – причастие, PV – преверб, SBJV – конъюнктив, SG – единственное число, SUPER – суперэссив, TR – показатель переходности.

СПИСОК ЛИТЕРАТУРЫ

1. *Arkhangelskiy T., Belyaev O., Vydrin A.* The Creation of Large-Scale Annotated Corpora of Minority Languages using UniParser and the EANC platform. // Proc. of COLING 2012: Posters, Mumbai, India. The COLING 2012 Organizing Committee: 2012. P. 83–92.

2. *Nivre J., de Marneffe M., Ginter F., Hajic J., Manning C., Pyysalo S., Schuster S., Tyers F., Zeman D.* Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. // Proc. of the Twelfth Language Resources and Evaluation Conference, Marseille, France. European Language Resources Association: 2020. P. 4034–4043.

3. *Shatskikh A., Sorokin A.* Ossetic-COT: Designing a morphologically annotated corpus and morphological analyzer for Ossetic. // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference “Dialogue” (2026). 2026. P. 544-555.

4. *Abaev V.* A grammatical sketch of Ossetic. The Hague: Mouton, 1964.

5. *Беляев О.И.* Коррелятивная конструкция в осетинском языке (диссертация на соискание степени кандидата филологических наук). 2014.

6. *Belyaev O.* Evolution of Case in Ossetic // Iran and the Caucasus. 2010. No. 14 (2). P. 287–322.

7. *Serdobolskaya N., Tuzhik O.* Differential Object Marking in Modern Ossetic: Referential Properties and Animacy // Tomsk Journal of Linguistics and Anthropology. 2025. No. 48 (2). P. 91–107.

8. *Kondratyuk D., Straka M.* 75 Languages, 1 Model: Parsing Universal Dependencies Universally // Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China. Association for Computational Linguistics: 2019. P. 2779–2795.

9. *Straka M., Straková J., Hajic, J.* UDPipe at SIGMORPHON 2019: Contextualized Embeddings, Regularization with Morphological Categories, Corpora Merging. // Proc. of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology, Florence, Italy. Association for Computational Linguistics: 2019. P. 95–103.

10. *Devlin J., Chang M., Lee K., Toutanova K.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1, Minneapolis, Minnesota. Association for Computational Linguistics: 2020. P. 4171–4186.

11. *Kuratov Y., Arkhipov M.* Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. 2019.

12. *Hulden M.* Foma: a Finite-State Compiler and Library // Proc. of the Demonstrations Session at EACL 2009, Athens, Greece. Association for Computational Linguistics: 2009. P. 29–32.

13. *Inoue G., Shindo H., Matsumoto Y.* Joint Prediction of Morphosyntactic Categories for Fine-Grained Arabic Part-of-Speech Tagging Exploiting Tag Dictionary Information // Proc. of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), Vancouver, Canada. Association for Computational Linguistics: 2017. P. 421–431.

ПРИЛОЖЕНИЕ А. РАСПРЕДЕЛЕНИЕ ПОДЖАНРОВ

	Обучающая	Валидационная	Тестовая
Радио	253 (6.15%)	37 (6.79%)	29 (5.32%)
Экспедиции	4111 (93.85%)	508 (93.21%)	516 (94.68%)

ПРИЛОЖЕНИЕ Б. РАСПРЕДЕЛЕНИЕ МОРФОЛОГИЧЕСКИХ МЕТОК

	Обучающая	Валидационная	Тестовая
PUNCT	14913 (25.58%)	1865 (24.92%)	1824 (25.16%)
NOUN, Case = Nom|Number = Sing	3269 (5.61%)	450 (6.01%)	444 (6.12%)
PART	3236 (5.55%)	419 (5.60%)	394 (5.43%)
CCONJ	2794 (4.79%)	351 (4.69%)	355 (4.90%)
PROPN, Case = Nom	1888 (3.24%)	253 (3.38%)	277 (3.82%)
ADV	1027 (1.76%)	109 (1.46%)	127 (1.75%)
ADV, PronType = Int	1009 (1.73%)	112 (1.50%)	116 (1.60%)
NOUN, Case = Nom|Number = Plur	1003 (1.72%)	126 (1.68%)	140 (1.93%)
ADV, PronType = Dem	999 (1.71%)	123 (1.64%)	139 (1.92%)
PART, Polarity = Neg	988 (1.69%)	111 (1.48%)	124 (1.71%)
ADJ	882 (1.51%)	120 (1.60%)	104 (1.43%)
ADV, Deixis = Remt| PronType = Dem	851 (1.46%)	132 (1.76%)	111 (1.53%)
VERB, Mood = Ind| Number = Sing|Person = 3| Tense = Pres| VerbForm = Fin	830 (1.42%)	100 (1.34%)	85 (1.17%)
VERB, Mood = Ind| Number = Sing|Person = 3| Preverb = Yes|Tense = Past| VerbForm = Fin	749 (1.28%)	95 (1.27%)	87 (1.20%)

INTJ	714 (1.22%)	113 (1.51%)	132 (1.82%)
ADV, Deixis = Prox| PronType = Dem	701 (1.20%)	77 (1.03%)	92 (1.27%)
VERB, Mood = Ind| Number = Sing|Person = 3| Tense = Past| VerbForm = Fin	673 (1.15%)	96 (1.28%)	98 (1.35%)
PRON, Case = Gen| Number = Sing| Person = 3| PronType = Prs	643 (1.10%)	82 (1.10%)	88 (1.21%)
NUM, Case = Nom| NumType = Card	610 (1.05%)	63 (0.84%)	81 (1.12%)
NOUN, Case = Gen|Number = Sing	590 (1.01%)	95 (1.27%)	66 (0.91%)

ПРИЛОЖЕНИЕ В. ТЕСТОВЫЙ НАБОР ДЛЯ РАЗРЕШЕНИЯ ОМОНИМИИ

	Контекст 1	Контекст 2
Adj/Noun	**Зæронд** ус æрбацыди.	**Зæронд** æрбацыди.
Adj/Adv	Уый **рæсугъд** чызгимæ кафы.	Уый **рæсугъд** кафы.
Noun/Adp	Лæппу хæдзары **сæр** уыны.	Лæппу хæдзары **сæр** бады.
1SG/INF	Æз **кусын**.	Æз **кусын** байдыдтон.
GEN/INE	Залинæ **хæдзары** мидæг уыди.	Залинæ **хæдзары** уыди.
GEN/NUM (неодуш.)	Залинæ **хæдзары** хицауы федта.	Залинæ дыууæ **хæздары** федта.
GEN/NUM (одуш.)	Дыууæ **лæджы** федтон.	Дыууæ **лæджы** æрбацыдысты.
INE/PRS.3SG	Мах **хъæуы** барæгбон кæнæм.	Мах **хъæуы** барæгбон.
DemDist.NOM/DemDist.GEN	**Уый** æрбацыдис.	**Уый** фæстæ æрбацыдтæн.
1PL.NOM/1PL.GEN	**Мах** уым цæрæм.	**Мах** бинонтæ уым цæрынц.

1SG.POSS/1SG.GEN (энкл.)	Уый **мæ** мады федта.	Уый **мæ** федта.
1SG.POSS/союз ‘и’	**Мæ** чызг уым ахуыр кодта.	Лæппу **мæ** чызг уым ахуыр кодтой.
2SG.POSS/2SG.GEN (энкл.)	Уый **дæ** фыды федта.	Уый **дæ** федта.
2SG.GEN (энкл.)/быть.2SG	Ирон лæг **дæ** федта.	Ирон лæг **дæ**.
1PL.POSS/1PL.GEN (энкл.)	Уый **нæ** чызджы федта.	Уый **нæ** федта.
1PL.POSS/NEG	Уый **нæ** хæдзар федта.	Уый хæдзар **нæ** федта.
EMPH/NEG	Радзур-**ма** исты.	**Ма** дзур.
ADD/NEG	Хъæумæ **ма** цыдтæн.	Хъæумæ **ма** цæу.
ITER/‘один’	Æмæ-**иу** лæг æрбацыдис.	Æмæ **иу** лæг æрбацыдис.

AUTOMATIC MORPHOLOGICAL TAGGING FOR OSSETIC BASED ON DATA FROM THE CORPUS OF ORAL TEXTS

A. S. Shatskikh[1] [0009-0006-9025-7783], A. A. Sorokin[2] [0000-0003-3877-4223]

[1]*Institute of Linguistics of the Russian Academy of Sciences, Moscow, Russia*
[2]*Lomonosov Moscow State University, Moscow, Russia*
[2]*Yandex, Moscow, Russia*

[1]anapriselec@gmail.com, [2]alexey.sorokin@list.ru

Abstract

In this work we present the first morphologically annotated corpus for Iron Ossetic that conforms to the Universal Dependencies schema. The corpus includes 5454 manually annotated sentences from the Iron Ossetic Corpus of Oral Texts, containing 74032 tokens. We use this corpus to train a BERT-based morphological analyzer. The analyzer achieves tag accuracy of 95.60%. Furthermore, the paper presents the results of experiments aimed at improving classification quality. It is shown that neither filtering the model's output using a context- free analyzer nor multi- task approach lead to a significant improvement. Finally, the paper presents the results of testing the analyzer on out- of- domain data and shows that the model achieves tag accuracy of 91.45% on them.

Keywords*: Iron Ossetic, Universal Dependencies, Morphological Tagging, Morphological Analysis*

REFERENCES

1. *Arkhangelskiy T., Belyaev O., Vydrin A.* The Creation of Large-Scale Annotated Corpora of Minority Languages using UniParser and the EANC platform // Proc. of COLING 2012: Posters, Mumbai, India. 2012. P. 83–92.

2. *Nivre J., de Marneffe M., Ginter F., Hajic J., Manning C., Pyysalo S., Schuster S., Tyers F., Zeman D.* Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection // Proc. of the Twelfth Language Resources and Evaluation Conference, Marseille, France. European Language Resources Association: 2020. P. 4034–4043.

3. *Shatskikh A., Sorokin A.* Ossetic-COT: Designing a morphologically annotated corpus and morphological analyzer for Ossetic. // Computational

Linguistics and Intellectual Technologies: Papers from the Annual International Conference “Dialogue” (2026). 2026. P. 544-555.

4. *Abaev V.* A grammatical sketch of Ossetic. The Hague: Mouton, 1964.

5. *Belyaev O.* Korreljativnaja konstrukcija v osetinskom jazyke v tipologicheskom osveschenii (PhD thesis). 2014.

6. *Belyaev O.* Evolution of Case in Ossetic. // Iran and the Caucasus. 2010. No. 14 (2). P. 287–322.

7. *Serdobolskaya N., Tuzhik O.* Differential Object Marking in Modern Ossetic: Referential Properties and Animacy // Tomsk Journal of Linguistics and Anthropology. 2025. No. 48 (2). P. 91–107.

8. *Kondratyuk D., Straka M.* 75 Languages, 1 Model: Parsing Universal Dependencies Universally // Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th International Joint Conf. on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China. 2019. P. 2779–2795.

9. *Straka M., Straková J., Hajic J.* UDPipe at SIGMORPHON 2019: Contextualized Embeddings, Regularization with Morphological Categories, Corpora Merging // Proc. of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology, Florence, Italy. 2019. P. 95–103.

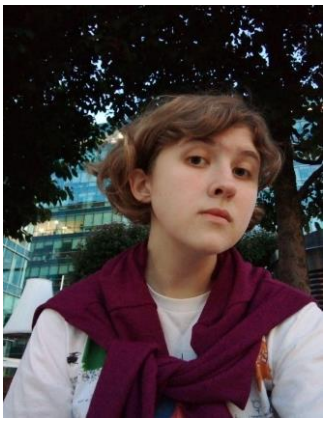
10. *Devlin J., Chang M., Lee K., Toutanova K.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. // Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1, Minneapolis, Minnesota. Association for Computational Linguistics: 2020. P. 4171–4186.

11. *Kuratov Y., Arkhipov M.* Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. 2019.

12. *Hulden M.* Foma: a Finite-State Compiler and Library // Proceedings of the Demonstrations Session at EACL 2009, Athens, Greece. 2009. P. 29–32.

13. *Inoue G., Shindo H., Matsumoto Y.* Joint Prediction of Morphosyntactic Categories for Fine-Grained Arabic Part-of-Speech Tagging Exploiting Tag Dictionary Information // Proc. of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), Vancouver, Canada. 2017. P. 421–431.

СВЕДЕНИЯ ОБ АВТОРАХ



ШАТСКИХ Анна Сергеевна – лаборант-исследователь в секторе прикладной лингвистики Института языкознания РАН. Область научных интересов: обработка естественного языка, языковые модели, компьютерная морфология, малоресурсные языки, иранские языки, осетинский язык.

Anna S. SHATSKIKH is a research assistant at the Department of Applied Linguistics, Institute of Linguistics of the Russian Academy of Sciences. Research interests: natural language processing, large language models, computational morphology, low-resource languages, Iranian languages, Ossetic language.

email: anapriselec@gmail.com

ORCID: 0009-0006-9025-7783



СОРОКИН Алексей Андреевич – старший научный сотрудник Центра ИИ МГУ, доцент кафедры математической логики и теории алгоритмов механико-математического факультета МГУ. Область научных интересов: обработка естественного языка, языковые модели, компьютерная морфология, обработка малоресурсных языков, автоматическое исправление грамматических ошибок.

Alexey A. SOROKIN is a senior researcher at the MSU AI Centre. Research interests: natural language processing, large language models, computational morphology, NLP for low-resource languages, grammatical error correction.

email: alexey.sorokin@list.ru

ORCID: 0000-0003-3877-4223

Материал поступил в редакцию 25 августа 2026 года

УДК 004.8

МОРФОЛОГИЧЕСКИЕ ФОРМЫ ГЛАГОЛА В ЗАДАЧЕ РАСПОЗНАВАНИЯ ЭМОЦИЙ: ХАРАКТЕРИСТИКИ ДАННЫХ И РЕЗУЛЬТАТЫ КЛАССИФИКАЦИИ

П. В. Ярошенко[1] [0000-0002-9425-5842], **А. А. Петрова**[2] [0009-0005-0176-9836], **Н. В. Лукашевич**[3] [0000-0002-1883-4121]

[1–3]*Московский государственный университет имени М. В. Ломоносова, г. Москва, Россия*

[1]polina.iaroshenko@yandex.ru, [2]petrova.anastasi999@gmail.com, [3]louk_nat@mail.ru

Аннотация

Распознавание эмоций в тексте является одной из актуальных задач обработки естественного языка. В работе исследовано влияние морфологической формы глагола на качество эмоциональной классификации моделями-энкодерами и выявлены факторы, определяющие морфологическую чувствительность языковых моделей. Для эксперимента использованы три модели (ruBERT-tiny2, ruBERT-base, ruRoBERTa-large), дообученные на объединенном корпусе из четырех русскоязычных датасетов (22852 примера). Каждая модель классифицировала 5676 морфологических форм 473 глаголов из Russian Emotion Lexicon. В результате зафиксировано систематическое влияние морфологии: различия в качестве классификации между формами по метрике F1 достигают 20–25% для всех моделей. Для объяснения установленных закономерностей был проведен анализ характеристик обучающих данных. Показано, что частотность морфологических форм в корпусе частично коррелирует с качеством классификации, однако не является единственным определяющим фактором. При анализе распределения эмоций по морфологическим формам выявлено, что в обучающем корпусе императив преимущественно встречается в контекстах с эмоцией «Радость». Однако глаголы класса «Грусть» в форме повелительного наклонения систематически классифицируются энкодерами как «Злость». Это расхождение между характеристиками обучающих данных и результатами классификации

© П. В. Ярошенко, А. А. Петрова, Н. В. Лукашевич, 2026.
Данная статья распространяется на условиях международной лицензии Creative Commons License Attribution 4.0 International (CC BY 4.0).

указывает на то, что императив в сочетании с семантикой грусти интерпретируется моделями как выражение агрессии, что свидетельствует о взаимодействии грамматической формы с семантикой глагола при определении эмоционального класса.

***Ключевые слова**: классификация эмоций, BERT, обработка естественного языка, русский язык, морфология.*

ВВЕДЕНИЕ

Распознавание эмоций в тексте (textual emotion recognition) является одной из актуальных задач обработки естественного языка. Эмоциональные лексиконы, состоящие из списков слов и выражений, ассоциированных с различными эмоциональными категориями [1], остаются важным инструментом в данной области [2]. Однако существующие лексиконы, как и традиционные словари, как правило, не учитывают морфологическую вариативность: представление слов в виде лемм. Однако грамматическая форма может оказывать влияние на эмоциональный компонент лексической единицы. Например, глагол *мучиться* в ряде источников отнесен к категории «Грусть» [3, 4], при этом форма императива (*мучься*) может быть интерпретирована как более агрессивная, выражающая скорее злость, нежели печаль.

Проблема морфологической вариативности актуальна также в свете специфики текстовых данных, анализируемых при распознавании эмоций, нередко это комментарии из социальных сетей, характеризующиеся краткостью и фрагментарностью. Как отмечено в работе [5], короткие высказывания часто не содержат достаточного контекста для определения эмоции, что потенциально увеличивает значимость морфологических характеристик отдельных лексических единиц.

Таким образом, остается открытым вопрос о роли морфологических характеристик глаголов в условиях минимального контекста, типичного для задач автоматического распознавания эмоций. Настоящее исследование направлено на восполнение этого пробела. Были сформулированы следующие исследовательские вопросы.

1. Влияет ли морфологическая форма на предсказание класса эмоции языковыми моделями?

2. Существуют ли закономерности классификации, связанные с конкретными морфологическими формами?

Для ответа на эти вопросы было проведено экспериментальное исследование. На основе ресурса Russian Emotion Lexicon (RusEmoLex) [6] был составлен список глаголов, размеченных по эмоциям. Для каждого глагола были сгенерированы различные морфологические формы. Три модели-энкодера (ruBERT-tiny2, ruBERT-base, ruRoBERTa-large), дообученные на объединенном корпусе из четырех русскоязычных датасетов (общий объем корпуса: 22852 размеченных примера), классифицировали каждую форму. Качество оценивалось с использованием метрики F1.

Настоящая статья является расширенной версией работы, представленной на конференции «Диалог-2026» [7]. В исходной публикации были представлены результаты экспериментов по классификации морфологических форм глаголов из RusEmoLex моделями-энкодерами и генеративными моделями. Было установлено, что морфологическая форма глагола систематически влияет на предсказания моделей. Вклад настоящей статьи заключается в расширении исходного исследования за счет анализа характеристик обучающих данных – это корпус размеченных по эмоциям текстов, составленных из доступных русскоязычных датасетов для классификации эмоций. В работе представлены следующие дополнительные результаты: количественный анализ частотности морфологических форм глаголов из RusEmoLex в объединенном обучающем корпусе; анализ распределения эмоциональных классов по морфологическим формам в обучающих данных; интерпретация наблюдаемых паттернов предсказаний моделей через характеристики обучающего корпуса.

1. РЕСУРСЫ ДЛЯ РАСПОЗНАВАНИЯ ЭМОЦИЙ В ТЕКСТЕ

Область распознавания эмоций в тексте получила значительное развитие с появлением методов глубокого обучения [8, 9]. Несмотря на значительное количество доступных датасетов для этой задачи, в обзорных статьях [5, 10] нехватка размеченных данных обозначена как одна из главных проблем в данной области.

Недостаточность данных объясняется двумя факторами. Во-первых, это доминирование англоязычных ресурсов: подавляющее большинство доступных

датасетов являются англоязычными (см., например, описание различных ресурсов в обзоре [5]). Таким образом, нехватка данных на языках помимо английского ощущается особенно остро [11]. Во-вторых, существует значительное количество теоретических подходов к описанию эмоций, соответственно, и подходы к разметке данных различаются в зависимости от того, какая теоретическая модель лежит в основе [5]. Согласно статистике, представленной в [12], чаще всего в современных исследованиях, посвященных анализу эмоций, используются подходы Пола Экмана [13] и Роберта Плутчика [14]. Результатом разнообразия теоретических подходов является также проблема несбалансированности эмоциональных классов – одни эмоции оказываются избыточно представленными, другие – недостаточно.

Для русского языка, помимо датасетов для машинного обучения (описанных в разд. 3), существуют специализированные лексиконы и базы данных, куда внесены результаты опросов носителей языка.

Лексикон RuSentiLex [3] содержит более 12000 ед. оценочной лексики, из которых 1760 относятся к эмоциональной лексике, сгруппированной под общим тегом feeling. NRC Emotion Lexicon [15] – это англоязычный краудсорсинговый ресурс, автоматически переведенный на 108 языков, включая русский. Лексикон содержит более 14000 ед., размеченных по восьми базовым эмоциям из модели Плутчика [14]: злость, страх, предвкушение, доверие, грусть, радость, отвращение и удивление. Разметка проведена по бинарной шкале 0/1. Оба лексикона включают различные части речи. База данных ENRuN (Emotional Norms for Russian Nouns) [16] содержит эмоциональные рейтинги для 1800 существительных по пяти классам (радость, грусть, злость, страх, отвращение) по шкале от 0 до 5. RusEmoLex [6] объединяет данные из лексикографических источников (словарей и тезаурусов), семантически размеченной части Национального корпуса русского языка [17], результаты опросов об ассоциациях между словами и эмоциями, а также датасетов, используемых в задачах распознавания эмоций. RusEmoLex включает 1024 лексические единицы, для которых указаны частеречная принадлежность, а также метка эмоционального класса (радость, грусть, злость, удивление, страх).

Однако всем перечисленным ресурсам присуще общее ограничение: лексические единицы представлены в виде лемм, что не позволяет оценить влияние грамматической формы на эмоциональное содержание.

2. МАТЕРИАЛ ИССЛЕДОВАНИЯ

Общая идея эксперимента заключается в том, чтобы определить, в какой степени предсказания энкодерной языковой модели зависят от формы глагола. Необходимый для проведения эксперимента список глаголов, размеченных по эмоциям, был взят из лексикона RusEmoLex[1].

RusEmoLex содержит 473 глагола, размеченных по пяти классам эмоций, каждому глаголу присвоена только одна метка: грусть (181 глагол), радость (107), злость (81), страх (78) и удивление (26). Для каждого глагола генерировались морфологические формы с учетом видовых характеристик. Итоговый набор включал 12 форм: инфинитив – базовая словарная форма (infinitive); настоящее время (для глаголов несовершенного вида) или будущее время (для глаголов совершенного вида): формы 1-го, 2-го, 3-го лица единственного и множественного чисел (6 форм: pres_sing_1per, pres_sing_2per, pres_sing_3per, pres_plur_1per, pres_plur_2per, pres_plur_3per); прошедшее время: формы мужского и женского родов единственного числа, множественное число (3 формы: past_masc, past_femn, past_plur); повелительное наклонение 2-е лицо единственного и множественного чисел (impr_sing, impr_plur)

Общее количество проанализированных форм составило 5676 (473 глагола, для каждого глагола по 12 форм). Каждая форма подавалась в модель изолированно, без дополнительного контекста.

3. МОДЕЛИ И ОБУЧАЮЩИЕ ДАННЫЕ

Для экспериментов были использованы три модели-энкодера различного размера: ruBERT-tiny2[2], ruBERT-base[3] и ruRoBERTa-large[4]. Все модели дообуча-

[1] https://github.com/nl-pi/rusemolex
[2] https://huggingface.co/cointegrated/rubert-tiny2
[3] https://huggingface.co/DeepPavlov/rubert-base-cased
[4] https://huggingface.co/ai-forever/ruRoberta-large

лись на задачу классификации по эмоциям на объединенном корпусе, составленном из доступных русскоязычных датасетов. В настоящем разделе описан процесс формирования обучающего корпуса (3.1) и дообучения (fine-tuning) энкодеров (3.2).

3.1. Наборы данных для дообучения энкодеров

Для дообучения энкодеров был создан объединенный корпус на основе доступных русскоязычных датасетов, размеченных по эмоциям. Объединение имеющихся датасетов позволило обеспечить более широкий охват эмоциональных лексических единиц для эффективного последующего дообучения энкодеров. Датасеты имеют схожую структуру: они включают предложения или короткие тексты, каждому из которых присвоена категория эмоции. Краткое описание использованных ресурсов приведено ниже, основные характеристики каждого датасета представлены в табл. 1.

CEDR [18] – это набор данных, где каждому предложению присвоены одна или несколько меток эмоциональных классов; присутствуют также предложения без метки класса, т. е., по мнению разметчиков, не содержащие эмоционального компонента. Источниками текстов послужили комментарии в социальных сетях, блогах и новостных ресурсах[5].

AIST [19] – для дообучения моделей была использована часть набора данных, находящаяся в открытом доступе. Источниками для формирования датасета послужили комментарии в социальных сетях (VK, Telegram)[6].

BRIGHTER [20] – коллекция размеченных вручную датасетов для распознавания эмоций на 28 языках, использовавшихся в соревновании SemEval-2025 по многоязычному распознаванию эмоций. Русскоязычная часть датасета составлена на основе комментариев из социальных сетей[7].

GoEmotions [21] – датасет, состоящий из размеченных по эмоциям комментариев из Reddit. Для дообучения энкодеров в рамках эксперимента использовалась автоматически переведенная на русский язык версия датасета.

[5] https://huggingface.co/datasets/sagteam/cedr_v1
[6] https://github.com/asbabiy/AIST
[7] https://brighter-dataset.github.io/

Табл. 1. Датасеты для классификации эмоций на русском языке.

Датасет	Метки классов	Количество размеченных текстов
CEDR	радость, грусть, удивление, страх, злость	9410
AIST	радость, грусть, злость, неуверенность, нейтральный класс	9447
BRIGHTER	злость, отвращение, страх, радость, грусть, удивление	2679 (тренировочная выборка для русского языка)
GoEmotions	27 эмоций и нейтральный класс	54263

Для дообучения энкодеров в объединенный корпус были отобраны только тексты с метками грусть, радость, злость, страх и удивление, соответствующие целевым классам эмоций из RusEmoLex. В табл. 2 представлено распределение примеров по эмоциям в объединенном датасете с учетом мультиклассовых примеров.

Табл. 2. Объединенный корпус на основе доступных датасетов для дообучения энкодеров.

Метка эмоционального класса	Количество размеченных текстов
Радость	9884
Грусть	4993
Злость	3583
Удивление	2257
Страх	1703
Удивление, радость	107
Злость, грусть	79
Страх, грусть	68

Злость, удивление	48
Грусть, радость	37
Удивление, грусть	31
Страх, радость	19
Страх, удивление	19
Злость, страх	13
Злость, радость	10
Злость, грусть, радость	1
Всего	22852

Анализ длины текстов в объединенном корпусе показал следующее распределение: медианная длина составляет 8 слов, при этом около четверти текстов корпуса (25.8%) содержат менее 5 слов, чуть больше половины корпуса (56.9%) – менее 10 слов, 76.7% – менее 15 слов. Таким образом, подавляющее большинство текстов корпуса представляет собой очень короткие высказывания, которые, как было сказано ранее, характерны для такого типа данных. В табл. 3 приведены примеры коротких высказываний из корпуса.

Табл. 3. Примеры коротких высказываний из объединенного корпуса.

Текст	Эмоция
Не скучай)	Радость
а вы боялись)	Радость
уже не чем рыдать!!	Грусть
грусть печаль точка обида :(	Грусть
#Школа #ненавижу	Злость
бесит все уже	Злость
боюсь завтра.	Страх
Вечером удивились.	Удивление

В приведенных примерах глагол часто является единственным словом, содержащим эмоциональный компонент. Другие маркеры эмоции в контексте либо отсутствуют, либо представлены минимально (эмотиконы, хештеги). В этой связи морфологическая форма глагола может быть более значима для интерпретации эмоционального содержания высказывания.

3.2. Дообучение энкодеров

Объединенный корпус, составленный на основе русскоязычных датасетов для классифкации эмоций, использовался для дообучения всех трех моделей-энкодеров: ruBERT-tiny2, ruBERT-base и ruRoBERTa-large. Все слои моделей дообучались на задачу классификации эмоций. Датасет был разделен в соотношении 80/20 на обучающую и валидационную выборки. Тексты токенизировались с максимальной длиной 100 токенов с применением усечения (truncation) и дополнения (padding). Использовалась следующая конфигурация: оптимизатор AdamW, размер батча 64, линейный планировщик с обновлениями после каждого батча и функция потерь BCEWithLogitsLoss. Скорость обучения была установлена на уровне 1×10^{-5} (ruBERT-tiny2) и 5×10^{-6} (ruBERT-base, ruRoBERTa-large). Обучение проводилось до достижения оптимального качества на валидационной выборке: 19 эпох (ruBERT-tiny2), 7 эпох (ruBERT-base) и 3 эпохи (ruRoBERTa-large). На валидационной выборке модели достигли следующих значений F1: ruBERT-tiny2 – 0.78, ruBERT-base – 0.79, ruRoBERTa-large – 0.9.

По результатам дообучения энкодеров были получены метрики по классификации глаголов в форме инфинитива из RusEmoLex (табл. 4). Форма инфинитива рассматривается в настоящей работе в качестве базового результата (baseline).

Табл. 4. Базовые результаты классификации глаголов по эмоциям.

Модель	Точность	Полнота	F1
ruRoBERTa-large	0.77	0.76	0.77
rubert-base	0.53	0.51	0.52
ruBERT-tiny2	0.56	0.54	0.55

Таким образом, можно заключить, что определение класса эмоции изолированной глагольной формы представляет сложность для моделей-энкодеров, обученных на контекстуальных представлениях. Модель ruRoBERTa-large с F1 = 0.77 показывает приемлемое качество, что может свидетельствовать о способности энкодеров большего размера распознавать эмоциональный компонент на уровне отдельных лексических единиц. Значительное падение качества для ruBERT-base (F1 = 0.52) и ruBERT-tiny2 (F1 = 0.55) подчеркивает более выраженную зависимость меньших моделей от контекстуальной информации для определения класса эмоции.

4. ПРЕДСКАЗАНИЕ ЭМОЦИОНАЛЬНОГО КЛАССА ДЛЯ ГЛАГОЛЬНЫХ ФОРМ

Энкодеры, дообученные на объединенном датасете, затем были применены к формам глаголов из RusEmoLex. В табл. 5 представлены лучшие и худшие результаты классификации эмоций для каждой модели, а также соответствующие глагольные формы.

Табл. 5. Влияние глагольной формы на качество эмоциональной классификации моделями-энкодерами.

Модель	Baseline (инфинитив)	F1 best	F1 worst
ruRoBERTa-large	0.77	0.83 past_femn	0.64 impr_sing
ruBERT-base	0.52	0.61 past_femn	0.49 impr_sing
ruBERT-tiny2	0.55	0.64 past_femn	0.48 pres_sing_2per

Результаты демонстрируют систематическое влияние глагольной формы на определение эмоционального класса лексических единиц. Для всех трех моделей форма женского рода прошедшего времени обеспечивает максимальные значения F1 (0.83, 0.61 и 0.64 соответственно), превосходя базовый уровень инфинитива на 0.07–0.09 п.п. Относительный прирост составляет 9% для

ruRoBERTa-large, 17% для ruBERT-base и 16% для ruBERT-tiny2. Минимальные значения F1 характерны для формы повелительного наклонения в единственном числе (ruRoBERTa-large, ruBERT-base) и формы настоящего времени 2-го лица в единственном числе (ruBERT-tiny2).

Таким образом, наблюдаются существенные колебания метрик в зависимости от формы глагола. Морфологическая чувствительность, рассчитанная как относительное снижение качества между наилучшей и наихудшей формами, составляет 23%, 20% и 25% для ruRoBERTa-large, ruBERT-base и ruBERT-tiny2 соответственно.

На примере модели с наиболее высоким показателем F1 рассмотрим основные тенденции предсказаний эмоционального класса. На рис. 1 представлено сравнение матриц ошибок для ruRoBERTa-large для наилучшей (форма женского рода прошедшего времени) и наихудшей (форма императива единственного числа) форм. Поскольку классы несбалансированы (например, глаголов с меткой «Удивление» значительно меньше, чем глаголов с меткой «Грусть»), в матрицах для наглядности представлены процентные соотношения, а не абсолютные значения: каждое значение показывает, какой процент глаголов данного истинного класса был классифицирован как соответствующий предсказанный класс.

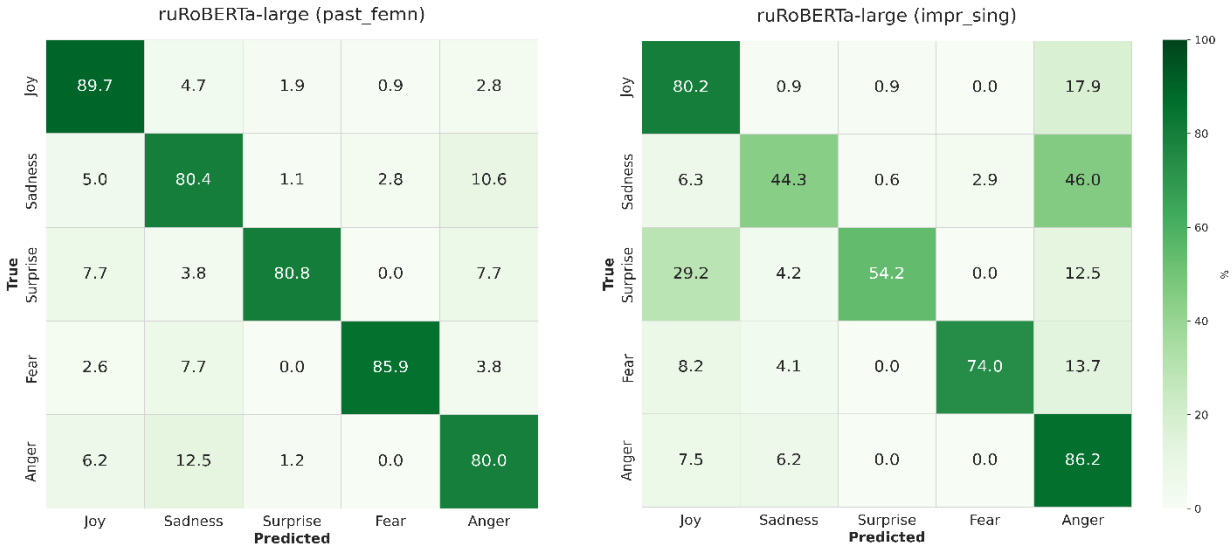


Рис. 1. Матрицы ошибок модели ruRoBERTa-large для оптимальной (past_femn, слева) и наименее эффективной (impr_sing, справа) глагольных форм.

Если для формы прошедшего времени женского рода модель демонстрирует относительно высокую точность предсказаний, то для формы императива точность резко падает для двух классов: «Грусть» и «Удивление». Наиболее явно прослеживается ошибка «Грусть – Злость»: 46% случаев грусти классифицируются как злость. Класс «Удивление» демонстрирует смещение в сторону класса «Радость» (29.2%), хотя и в меньшей степени. В табл. 6 представлены наиболее частотные смещения предсказаний для каждой модели.

Табл. 6. Паттерны смещения предсказаний эмоционального класса для каждой модели.

Модель	Глагольная форма	Истинная метка класса	Предсказанная метка класса	Количество случаев
ruRoBERTa-large	impr_sing	Грусть	Злость	80
	impr_plur	Грусть	Злость	76
	pres_plur_2per	Грусть	Злость	61
ruBERT-base	pres_sing_2per	Грусть	Злость	81
	impr_plur	Грусть	Злость	79
	impr_sing	Грусть	Злость	77
ruBERT-tiny2	pres_sing_2per	Грусть	Злость	113
	impr_plur	Грусть	Злость	102
	infinitive	Грусть	Злость	100

Для всех моделей наиболее частым паттерном является смещение «Грусть – Злость». Эта тенденция особенно характерна для форм императива, а также настоящего времени 2-го лица единственного числа. Если рассмотреть лексические единицы класса «Грусть», которые систематически классифицируются всеми тремя моделями как «Злость», можно сделать следующие наблюдения: как правило, это глаголы с высокой эмоциональной интенсивностью (*убиваться, затравить, изнурять*). Значительное число составляют глаголы с се-

мантикой подавления или насилия (*сломить, сокрушить, раздавить, подавить, давить*).

Таким образом, тенденции предсказаний, по-видимому, обусловлены двумя факторами: глагольными формами, подразумевающими ориентацию на другое лицо (повелительное наклонение, настоящее время второго лица), и семантикой глагола.

5. ХАРАКТЕРИСТИКИ ОБУЧАЮЩИХ ДАННЫХ И ИХ СВЯЗЬ С РЕЗУЛЬТАТАМИ КЛАССИФИКАЦИИ

Эксперименты показали систематическое влияние морфологической формы глагола на классификацию по эмоциональным категориям. Однако необходимо также оценить, в какой степени результаты классификации обусловлены характеристиками обучающих данных. В настоящем разделе проведен анализ обучающего корпуса в двух аспектах: частотность морфологических форм и их распределение по эмоциональным классам.

5.1. Частотность морфологических форм глаголов из RusEmoLex в объединенном корпусе

Для оценки влияния характеристик обучающих данных на результаты классификации были проанализированы частотности глаголов-лемм из RusEmoLex и их морфологических форм в корпусе, составленном из русскоязычных датасетов.

В обучающем корпусе встречаются 230 из 473 глаголов RusEmoLex (48.6%). Для каждого вхождения определялась морфологическая форма глагола. Результаты представлены в табл. 7.

Табл. 7. Частотность морфологических форм глаголов из RusEmoLex в обучающем корпусе.

Форма	Количество вхождений	% от общего числа
pres_sing_1per	1136	35.37
infinitive	470	14.63
pres_sing_3per	426	13.26

past_masc	304	9.46
past_femn	253	7.87
pres_plur_3per	166	5.17
past_plur	153	4.76
impr_sing	108	3.36
pres_sing_2per	76	2.37
impr_plur	60	1.87
pres_plur_1per	47	1.46
pres_plur_2per	13	0.40
Всего вхождений	**3212**	100

Наиболее частотной оказалась форма настоящего времени 1-го лица единственного числа (35.37% всех вхождений), такой результат ожидаем, принимая во внимание специфику материала: комментарии из социальных сетей зачастую содержат много высказываний от 1-го лица. Инфинитив, являющийся базовой словарной формой глагола, составляет 14.63% вхождений. Формы повелительного наклонения представлены значительно слабее (суммарно 5.2%). Наименее частотной является форма настоящего времени 2-го лица множественного числа (0.40%).

Для количественной оценки связи между частотностью морфологических форм в обучающем корпусе и качеством их классификации был рассчитан коэффициент корреляции Спирмена. Анализ для модели ruRoBERTa-large показал умеренную статистически значимую корреляцию ($\rho = 0.58$; $p = 0.048$), что подтверждает влияние частотности на качество классификации, однако указывает на существенный вклад и других факторов. Для моделей меньшего размера (ruBERT-base и ruBERT-tiny2) корреляция оказалась слабее и статистически незначимой.

Отметим также, что распределение случаев употребления глаголов из RusEmoLex в корпусе характеризуется неравномерностью: 10 наиболее частотных глаголов составляют 49% всех вхождений (табл. 8).

Табл. 8. Наиболее частотные глаголы в объединенном корпусе.

Глагол	Эмоциональный класс	Частотность в корпусе
бояться	Страх	613
любить	Радость	309
ненавидеть	Злость	291
скучать	Грусть	218
плакать	Грусть	179
надеяться	Радость	172
удивляться	Удивление	139
пугать	Страх	121
злиться	Злость	101
убить	Грусть	90

Анализ распределения морфологических форм у наиболее частотных глаголов в корпусе выявил усиление смещения в сторону 1-го лица единственного числа: у 10 наиболее частотных глаголов эта форма составляет более половины от всех случаев употребления (56.9%), в то время как в среднем по корпусу процент ниже, всего 35.4%. При этом доля повелительного наклонения (2.7%) почти вдвое ниже среднего показателя, что усугубляет проблему недостаточной представленности этих форм в обучающих данных.

Таким образом, малое количество форм повелительного наклонения и форм второго лица в обучающем корпусе может рассматриваться как одна из причин систематического снижения качества их эмоциональной классификации моделями-энкодерами.

5.2. Распределение эмоций по морфологическим формам глагола в объединенном корпусе

Для более глубокого понимания факторов, влияющих на классификацию глагольных форм, было проанализировано распределение различных морфоло-

гических форм в обучающем корпусе по классам эмоций. Для каждого вхождения глагола в корпусе фиксировались морфологическая форма глагола и эмоциональная метка всего текста. Результаты анализа для ключевых форм представлены в табл. 9.

Табл. 9. Распределение меток эмоциональных классов по морфологическим формам в обучающем корпусе (%).

Форма	Радость	Грусть	Удивление	Страх	Злость
infinitive	32.24	29.77	10.27	11.50	16.22
past_femn	30.74	30.74	8.89	21.48	8.15
impr_sing	46.85	30.63	1.80	3.60	17.12
pres_sing_2per	26.25	17.50	12.50	7.50	36.25

Инфинитив и форма прошедшего времени женского рода демонстрируют относительно равномерное распределение между основными эмоциональными категориями.

Формы повелительного наклонения единственного числа показывают смещение в сторону эмоции «Радость»: в 46.85% случаев глаголы в форме императива встречаются в текстах, размеченных как выражающие радость (примеры: «*не серчай ;)*»; «*Радуй всех своей улыбкой и больше не болей!=))*». Это превышает соответствующий показатель для инфинитива (32.24%) и формы прошедшего времени женского рода (30.74%). Одновременно для императива характерна низкая представленность эмоций «Удивление» (1.80%) и «Страх» (3.60%).

Форма настоящего времени 2-го лица единственного числа демонстрирует противоположную тенденцию: для нее характерно смещение к эмоции «Злость» (36.25% против 16.22% для инфинитива).

Сопоставление полученных данных о распределении эмоций в корпусе с результатами классификации выявляет следующую закономерность. Несмотря на то что в обучающих данных формы императива преимущественно встречаются в контексте эмоции «Радость» (46.85%), модель ruRoBERTa-large систематически классифицирует формы повелительного наклонения глаголов истинного класса «Грусть» как «Злость»: 46% глаголов, истинной меткой которых является

«Грусть», получают предсказанную метку «Злость» (см. рис. 1). При этом важно отметить, что для других эмоциональных классов форма императива не приводит к столь выраженным искажениям, что может указывать на взаимодействие между эмоциональным семантическим компонентом и грамматической формой: императив в сочетании с семантикой грусти интерпретируется моделью как выражение злости.

Таким образом, можно заключить, что морфологическая чувствительность энкодеров при классификации эмоций определяется комплексом факторов: не только частотностью форм в обучающих данных, но и характеристиками грамматических форм.

ЗАКЛЮЧЕНИЕ

Настоящее исследование позволило установить, что морфологическая форма глагола оказывает систематическое влияние на предсказания языковых моделей-энкодеров при классификации эмоций. Проведенные эксперименты показали, что морфологическая чувствительность моделей составляет 20–25%: различие в качестве классификации между лучшей и худшей глагольными формами достигает этого значения для всех трех исследованных энкодеров (ruBERT-tiny2, ruBERT-base, ruRoBERTa-large). Наиболее высокие результаты классификации демонстрирует форма прошедшего времени женского рода, превосходящая базовый уровень инфинитива. Напротив, формы повелительного наклонения и настоящего времени 2-го лица единственного числа показывают систематическое снижение качества на 20–23%. Анализ результатов также выявил устойчивый паттерн смещения предсказаний: глаголы класса «Грусть» в форме императива систематически классифицируются как «Злость». Глаголы, которым свойственна эта тенденция, как правило, характеризуются с высокой эмоциональной интенсивностью, зачастую глаголы включают семантические компоненты деструктивного воздействия.

Исследование характеристик обучающих данных показало, что чувствительность моделей-энкодеров к морфологии определяется рядом факторов. Частотность морфологических форм в корпусе частично коррелирует с качеством классификации: мало представленные в обучающем корпусе формы (императив, настоящее время 2-го лица множественного числа) демонстрируют худшие

результаты. Однако частотность не является единственным определяющим фактором: модели систематически ассоциируют императив в сочетании с семантикой грусти с выражением агрессии, злости, хотя формы императива в обучающем корпусе преимущественно встречаются в контексте эмоции «Радость». Это может указывать на то, что модели-энкодеры распознают не только статистические корреляции в обучающих данных, но и характеристики грамматических форм. Таким образом, в условиях минимального контекста морфологические характеристики глаголов могут быть значимы для выявления и классификации эмоционального компонента.

Благодарности

Исследование выполнено за счет гранта Российского научного фонда № 25-11-00191, https://rscf.ru/project/25-11-00191/.

СПИСОК ЛИТЕРАТУРЫ

1. *Cavicchio F.* Emotion Detection in Natural Language Processing. Cham: Springer, 2025.

2. *Mohammad S.* Best Practices in the Creation and Use of Emotion Lexicons // Findings of the Association for Computational Linguistics: EACL 2023. Dubrovnik, Croatia: Association for Computational Linguistics, 2023. P. 1825–1836.

3. *Loukachevitch N., Levchik A.* Creating a General Russian Sentiment Lexicon // Proc. of the Tenth International Conference on Language Resources and Evaluation (LREC'16). Portorož, Slovenia: European Language Resources Association (ELRA), 2016. P. 1171–1176.

4. *Бабенко Л.Г.* Алфавит эмоций: словарь-тезаурус эмотивной лексики, 2-е издание, стереотипное. Екатеринбург, М.: Кабинетный ученый, 2022.

5. *Deng J., Ren F.* A Survey of Textual Emotion Recognition and Its Challenges // IEEE Transactions on Affective Computing. 2023. Vol. 14, No. 1. P. 49–67. https://doi.org/10.1109/TAFFC.2021.3053275

6. *Ярошенко П.В., Лукашевич Н.В.* RusEmoLex: эмоциональный лексикон для русского языка // Russian Journal of Linguistics. 2025. Vol. 29, No. 3. P. 659–687. https://doi.org/10.22363/2687-0088-44439

7. *Iaroshenko P.V., Petrova A.A., Loukachevitch N.V.* Morphology and emotion recognition: How verb form affects language model predictions // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference Dialogue 2026. 2026. P. 134–143.

8. *Acheampong F.A., Nunoo-Mensah H., Chen W.* Transformer models for text-based emotion detection: a review of BERT-based approaches // Artif Intell Rev. 2021. Vol. 54. P. 5789–5829. https://doi.org/10.1007/s10462-021-09958-2

9. *Kumar R.P., Ramya P., Teja G.S., Krishna V.R.* BERTing Emotions: Exploring Textual Emotion Recognition using NLP // International Conference on Intelligent Computing and Control Systems (ICICCS). Erode, India, 2025. P. 888–895. https://doi.org/10.1109/ICICCS65191.2025.10984670

10. *Kusal S., Patil S., Kotecha K., Aluvalu R., Varadarajan V.* AI Based Emotion Detection for Textual Big Data: Techniques and Contribution // Big Data and Cognitive Computing. 2021. Vol. 5, No. 3. Article 43. https://doi.org/10.3390/bdcc5030043

11. *De Bruyne L.* The paradox of multilingual emotion detection // Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis. Toronto, Canada: Association for Computational Linguistics, 2023. P. 458–466.

12. *Plaza-del-Arco F.M., Cercas Curry A.A., Cercas Curry A., Hovy D.* Emotion analysis in NLP: Trends, gaps and roadmap for future directions // Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation. 2024. P. 5696–5710.

13. *Ekman P.* Are there basic emotions? // Psychological Review. 1992. Vol. 99, No. 3. P. 550–553.

14. *Plutchik R.* A general psycho-evolutionary theory of emotion // Theories of emotion. Cambridge: Academic Press, 1980. P. 3–33.

15. *Mohammad S., Turney P.* Crowdsourcing a word-emotion association lexicon // Computational Intelligence. 2013. Vol. 29, No. 3. P. 436–465. http://doi.org/10.1111/j.1467-8640.2012.00460.x

16. *Сысоева Т.А., Люсин Д.В.* Разработка расширенной базы данных с эмоциональными оценками существительных ENRuN-2: успехи, проблемы и перспективы // Психология познания: материалы Всероссийской научной конференции. ЯрГУ, 6–8 декабря 2024 г. / под ред. И.Ю. Владимирова, С.Ю. Коровкина.

Ярославль: Филигрань, 2024. С. 316–320.

17. *Савчук С.О., Архангельский Т.А., Бонч-Осмоловская А.А., Донина О.В., Кузнецова Ю.Н., Ляшевская О.Н., Орехов Б.В., Подрядчикова М.В.* Национальный корпус русского языка 2.0: новые возможности и перспективы развития // Вопросы языкознания. 2024. № 2. С. 7–34.

18. *Sboev A., Naumov A., Rybka R.* Data-driven model for emotion detection in Russian texts // Procedia Computer Science. 2021. Vol. 190. P. 637–642.

19. *Kazyulina M., Babii A., Malafeev A.* Emotion Classification in Russian: Feature Engineering and Analysis // Analysis of Images, Social Networks and Texts. AIST 2020. Lecture Notes in Computer Science. Vol. 12602. Cham: Springer, 2021. https://doi.org/10.1007/978-3-030-72610-2_10

20. *Muhammad S.H., Ousidhoum N., Abdulmumin I. et al.* BRIGHTER: BRIdging the Gap in Human-Annotated Textual Emotion Recognition Datasets for 28 Languages // Proc. of the 63rd Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers). Vienna, Austria: Association for Computational Linguistics, 2025. P. 8895–8916. https://doi.org/10.18653/v1/2025.acl-long.436

21. *Demszky D., Movshovitz-Attias D., Ko J., Cowen A., Nemade G., Ravi S.* GoEmotions: A Dataset of Fine-Grained Emotions // Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2020. P. 4040–4054.

VERB FORMS IN EMOTION RECOGNITION TASKS: ANALYSIS OF DATA CHARACTERISTICS AND CLASSIFICATION RESULTS

P. V. Iaroshenko[1][0000-0002-9425-5842], A. A. Petrova[2][0009-0005-0176-9836], N. V. Loukachevitch[3][0000-0002-1883-4121]

[1–3]*Lomonosov Moscow State University, Moscow, Russia*

[1]polina.iaroshenko@yandex.ru, [2]petrova.anastasi999@gmail.com, [3]louk_nat@mail.ru

Abstract

Emotion recognition in text represents an important task in natural language processing. This study investigates the influence of verb morphological form on emotion classification performance by encoder models and identifies factors that may determine morphological sensitivity in language models. Three models (ruBERT-tiny2, ruBERT-base, ruRoBERTa-large), fine-tuned on a combined corpus of four Russian-language datasets (22,852 examples), were employed in the experiment. Each model classified 5,676 morphological forms of 473 verbs from the Russian Emotion Lexicon. Results revealed systematic morphological influence: differences in classification performance between forms, as measured by F1-score, reached 20–25% across all models. To account for the observed patterns, training data characteristics were analysed. Frequency of morphological forms in the corpus was shown to partially correlate with classification performance, though it does not appear to be the sole determining factor. Analysis of emotion distribution across morphological forms revealed that in the training corpus, imperative forms occur predominantly in contexts associated with Joy. However, Sadness-labelled verbs in imperative mood are systematically classified by encoders as Anger. This discrepancy between training data characteristics and classification outcomes suggests that imperatives combined with sadness semantics are interpreted by models as expressions of aggression, indicating an interaction between grammatical form and verb semantics in determining emotion class.

Keywords*: emotion classification, BERT, natural language processing, Russian language, morphology.*

Acknowledgements

The research was supported by the Russian Science Foundation, project No. 25-11-00191, https://rscf.ru/project/25-11-00191/.

REFERENCES

1. *Cavicchio F.* Emotion Detection in Natural Language Processing. Cham: Springer, 2025.

2. *Mohammad S.* Best Practices in the Creation and Use of Emotion Lexicons // Findings of the Association for Computational Linguistics: EACL 2023. Dubrovnik, Croatia: Association for Computational Linguistics, 2023. P. 1825–1836.

3. *Loukachevitch N., Levchik A.* Creating a General Russian Sentiment Lexicon // Proc. of the Tenth International Conference on Language Resources and Evaluation (LREC'16). Portorož, Slovenia: European Language Resources Association (ELRA), 2016. P. 1171–1176.

4. *Babenko L.G.* Alphabet of Emotions: Thesaurus of Emotive Lexis, 2nd edition. Yekaterinburg; Moscow: Armchair Scientist, 2022.

5. *Deng J., Ren F.* A Survey of Textual Emotion Recognition and Its Challenges // IEEE Transactions on Affective Computing. 2023. Vol. 14, No. 1. P. 49–67. https://doi.org/10.1109/TAFFC.2021.3053275

6. *Iaroshenko P.V., Loukachevitch N.V.* RusEmoLex: Russian Emotion Lexicon // Russian Journal of Linguistics. 2025. Vol. 29, No. 3. P. 659–687.
URL: https://doi.org/10.22363/2687-0088-44439

7. *Iaroshenko P.V., Petrova A.A., Loukachevitch N.V.* Morphology and emotion recognition: How verb form affects language model predictions // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference Dialogue 2026. 2026. P. 134–143.

8. *Acheampong F.A., Nunoo-Mensah H., Chen W.* Transformer models for text-based emotion detection: a review of BERT-based approaches // Artif Intell Rev. 2021. Vol. 54. P. 5789–5829. https://doi.org/10.1007/s10462-021-09958-2

9. *Kumar R.P., Ramya P., Teja G.S., Krishna V.R.* BERTing Emotions: Exploring Textual Emotion Recognition using NLP // International Conference on Intelligent Computing and Control Systems (ICICCS). Erode, India, 2025. P. 888–895.
https://doi.org/10.1109/ICICCS65191.2025.10984670

10. *Kusal S., Patil S., Kotecha K., Aluvalu R., Varadarajan V.* AI Based Emotion Detection for Textual Big Data: Techniques and Contribution // Big Data and Cognitive Computing. 2021. Vol. 5, No. 3. Article 43. URL: https://doi.org/10.3390/bdcc5030043

11. *De Bruyne L.* The paradox of multilingual emotion detection // Proc. of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis. Toronto, Canada: Association for Computational Linguistics, 2023. P. 458–466.

12. *Plaza-del-Arco F.M., Cercas Curry A.A., Cercas Curry A., Hovy D.* Emotion analysis in NLP: Trends, gaps and roadmap for future directions // Proc. of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation. 2024. P. 5696–5710.

13. *Ekman P.* Are there basic emotions? // Psychological Review. 1992. Vol. 99, No. 3. P. 550–553.

14. *Plutchik R.* A general psycho-evolutionary theory of emotion // Theories of emotion. Cambridge: Academic Press, 1980. P. 3–33.

15. *Mohammad S., Turney P.* Crowdsourcing a word-emotion association lexicon // Computational Intelligence. 2013. Vol. 29, No. 3. P. 436–465.
URL: http://doi.org/10.1111/j.1467-8640.2012.00460.x

16. *Sysoeva T.A., Lyusin D.V.* Development of an extended database with emotional ratings of nouns ENRuN-2: Successes, problems and prospects // Psychology of cognition: Proc.of the All-Russian Scientific Conference, YARSU, December 6-8, 2024 / I.Yu. Vladimirov, S.Yu. Korovkin (eds.). Yaroslavl: Filigran, 2024. P. 316–320.

17. *Savchuk S.O., Arkhangelskiy T., Bonch-Osmolovskaya A.A., Donina O.V., Kuznetsova Yu.N., Lyashevskaya O.N., Orekhov B.V., Podryadchikova M.V.* Russian National Corpus 2.0: New opportunities and development prospects // Voprosy Jazykoznanija. 2024. No. 2. P. 7–34.

18. *Sboev A., Naumov A., Rybka R.* Data-driven model for emotion detection in Russian texts // Procedia Computer Science. 2021. Vol. 190. P. 637–642.

19. *Kazyulina M., Babii A., Malafeev A.* Emotion Classification in Russian: Feature Engineering and Analysis // Analysis of Images, Social Networks and Texts. AIST 2020. Lecture Notes in Computer Science. Vol. 12602. Cham: Springer, 2021. https://doi.org/10.1007/978-3-030-72610-2_10

20. *Muhammad S.H., Ousidhoum N., Abdulmumin I. et al.* BRIGHTER: BRIdging the Gap in Human-Annotated Textual Emotion Recognition Datasets for 28 Languages // Proc. of the 63rd Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers). Vienna, Austria: Association for Computational Linguistics, 2025. P. 8895–8916. https://doi.org/10.18653/v1/2025.acl-long.436

21. *Demszky D., Movshovitz-Attias D., Ko J., Cowen A., Nemade G., Ravi S.* GoEmotions: A Dataset of Fine-Grained Emotions // Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2020. P. 4040–4054.

СВЕДЕНИЯ ОБ АВТОРАХ



ЯРОШЕНКО Полина Владимировна – кандидат филологических наук, научный сотрудник Лаборатории анализа информационных ресурсов Научно-исследовательского вычислительного центра Московского государственного университета имени М. В. Ломоносова. Сфера научных интересов: компьютерная семантика, обработка естественного языка, искусственный интеллект, анализ эмоций.

Polina V. IAROSHENKO, PhD, is a research fellow in the Laboratory for Information Resources Analysis at the Research Computing Center, Lomonosov Moscow State University. Research interests: computational semantics, natural language processing, artificial intelligence, emotion analysis.

email: polina.iaroshenko@yandex.ru

ORCID: 0000-0002-9425-5842



ПЕТРОВА Анастасия Алексеевна — аспирант факультета вычислительной математики и кибернетики Московского государственного университета имени М. В. Ломоносова. Сфера научных интересов: анализ эмоций, обработка естественного языка.

Anastasiia A. PETROVA is a PhD student at the Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University. Research interests: emotion analysis, natural language processing.

email: petrova.anastasi999@gmail.com

ORCID: 0009-0005-0176-9836



ЛУКАШЕВИЧ Наталья Валентиновна – доктор технических наук, ведущий научный сотрудник Лаборатории анализа информационных ресурсов Научно-исследовательского вычислительного центра Московского государственного университета имени М. В. Ломоносова. Сфера научных интересов: компьютерная семантика, обработка естественного языка, искусственный интеллект, анализ тональности.

Natalia V. LOUKACHEVITCH, PhD (Advanced Doctorate), is a leading research fellow in the Laboratory for Information Resources Analysis at the Research Computing Center, Lomonosov Moscow State University. Research interests: computational semantics, natural language processing, artificial intelligence.

email: louk_nat@mail.ru

ORCID: 0000-0002-1883-4121

Материал поступил в редакцию26 августа 2026 года

УДК 004.934+159.9.019.4

PROMPTALIGN: АВТОМАТИЧЕСКОЕ ИТЕРАТИВНОЕ ВЫРАВНИВАНИЕ БОЛЬШОЙ ЯЗЫКОВОЙ МОДЕЛИ ПОД ЭКСПЕРТНУЮ РАЗМЕТКУ БЕЗ ДООБУЧЕНИЯ

А. А. Агафонов[1] [0000-0001-6924-6701], **А. Р. Хусаинов**[2] [0009-0002-1916-2109], **Н. А. Прокопьев**[3] [0000-0003-0066-7465]

[1–3]*Казанский (Приволжский) федеральный университет, г. Казань, Россия*

[1]a.a.agathonov@gmail.com, [2]khusainov.akhat@mail.ru, [3]nikolai.prokopyev@gmail.com

Аннотация

Большие языковые модели все активнее применяются в социальных и психологических науках, однако их предсказания систематически смещены относительно экспертной разметки (эффект смещения выравнивания, alignment shift), а попытки улучшить инструкции вручную или с помощью примеров приводят к так называемому эффекту «перетягивания одеяла» – деградации миноритарных классов.

В работе предложен PromptAlign – автоматический конвейер оптимизации инструкций, не требующий ни доступа к весам модели, ни участия человека в итеративной коррекции, так как таксономия и пространство признаков задаются экспертом априори. Конвейер извлекает гибридные лингвистические и семантические признаки с помощью большой языковой модели, строит на их основе интерпретируемые L1-логистические регрессии для локализации расхождений с экспертом, выполняет метаанализ ошибок и компилирует новые инструкции, управляя балансом через метрику индекса дисбаланса классов (Blanket Pulling Index, BPI). Эксперимент на корпусе, включающем 541 текст на русском языке, с тремя психологическими классами показал рост macro-F1 на отложенной тестовой выборке с 0.655 до 0.749 и минимальной классовой F1 с 0.449 до 0.656 ($p > 0.99$ по парному бутстрэп-тесту). Метод значимо превосходит базовые уровни — без примеров (zero-shot) и с подачей примеров (few-shot). Анализ вклада компонентов подтвердил критическую роль семантической интерпретации признаков и периодической перекомпоновки инструкций. Итого-

© А. А. Агафонов, А. Р. Хусаинов, Н. А. Прокопьев, 2026.
Данная статья распространяется на условиях международной лицензии Creative Commons License Attribution 4.0 International (CC BY 4.0).

вая инструкция содержит пять правил, отражающих латентные критерии экспертов; в отличие от неинтерпретируемых моделей автоматических оптимизаторов, эти правила доступны для проверки и корректировки специалистом.

***Ключевые слова**: большая языковая модель, оптимизация инструкций, выравнивание, интерпретируемость, L1-регуляризация, бутстрэп-тест, классификация текстов, психологическая разметка, суррогатная модель, индекс дисбаланса классов, русскоязычный корпус.*

ВВЕДЕНИЕ

Большие языковые модели (LLM) активно внедряются в исследовательскую практику социальных и поведенческих наук, автоматически выявляя личностные черты, токсичность и агрессию, эмоциональную окраску и коммуникативные стратегии – задачи, еще недавно требовавшие трудоемкой ручной разметки. Однако быстрое распространение этих моделей обнажило методологическую проблему: их предсказания систематически расходятся с эталонными экспертными оценками.

Этот феномен – смещение выравнивания (alignment shift) – порождается неявными предпочтениями, заложенными на стадиях предобучения и обучения с подкреплением на основе отзывов человека (RLHF), и в контекстах высокой содержательной достоверности (психодиагностика, социологический контент-анализ) подрывает надежность автоматической классификации.

Ситуация усложняется, когда дообучение невозможно (например, закрытые коммерческие LLM через API), и единственным инструментом воздействия остается текстовая инструкция (prompt). Ручная итеративная рефлексия эксперта повышает точность, но, как показало предшествующее исследование [1], провоцирует эффект «перетягивания одеяла» (blanket pulling): улучшение целевого миноритарного класса достигается за счет деградации мажоритарных, а сама оптимизация трудоемка, плохо масштабируется и зависит от субъективности аналитика.

Альтернативой служат методы автоматической оптимизации инструкций – Automatic Prompt Engineer (APE) [2], OPRO [3], программный каркас DSPy [4]. Они успешно повышают целевую метрику, но для социальных наук обладают двумя

критическими недостатками: генерируемые инструкции – непрозрачные «черные ящики», не поддающиеся экспертной верификации, а ни один из методов не контролирует баланс точности между классами, что делает многоклассовую оптимизацию рискованной.

Настоящая работа базируется на русскоязычном материале с трехклассовой таксономией «информативное – деструктивное – конструктивное» из [1], где также показано, что предобученные LLM без адаптации смещены в пользу информативного и деструктивного классов за счет конструктивного: граница между информативным текстом (факты, аналитика) и конструктивным (факты плюс эмпатия или решение) остается для модели размытой. Решаемая задача – преодолеть это через автоматизацию цикла коррекции, контроль баланса и экспликацию латентных экспертных критериев в виде правил на естественном языке.

Указанные ограничения ставят следующие исследовательские вопросы.

- Можно ли автоматически итеративно корректировать инструкцию LLM, приближая ее ответы к экспертной разметке, не имея доступа к весам и не привлекая человека к циклу коррекции?
- Можно ли формализовать и автоматически удерживать дисбаланс классов, предотвращая эффект «перетягивания одеяла»?
- Можно ли извлечь из процесса оптимизации интерпретируемые правила на естественном языке, отражающие латентные критерии экспертов?

Ответы на эти вопросы получены с помощью PromptAlign – метода автоматического выравнивания LLM-классификатора под экспертную разметку. Метод реализован в виде конвейера на основе суррогатных интерпретируемых моделей (L1-логистических регрессий), которые локализуют расхождения между экспертом и LLM в пространстве гибридных признаков и транслируют их в правила на естественном языке. Таксономия и признаковое пространство задаются экспертом априори, но цикл коррекции, анализ ошибок и компиляция инструкций автономны; для защиты от деградации миноритарных классов введена метрика BPI (Blanket Pulling Index).

БЛИЗКИЕ ПО ТЕМАТИКЕ РАБОТЫ

Работа опирается на три направления исследований: автоматическую оптимизацию инструкций, выравнивание языковых моделей без доступа к весам и

автоматическую классификацию психологических характеристик текста. Ниже рассмотрены их ключевые результаты и пробелы, которые восполняет PromptAlign.

Автоматическая оптимизация инструкций и контекстное обучение

Ручное конструирование инструкций все чаще уступает автоматическим методам, где инструкция является объектом оптимизации: в APE [2] LLM генерирует и оценивает набор инструкций-кандидатов на размеченном множестве; OPRO [3] формулирует это как численную оптимизацию, где внешняя LLM порождает новые инструкции по значениям целевой функции с предыдущих итераций; DSPy [4] абстрагирует вызовы LLM до декларативных модулей, автоматически подбираемых стохастической оптимизацией.

Альтернативой служит контекстное обучение с подачей примеров (few-shot prompting) [5, 6], где в инструкцию вручную включают репрезентативные примеры каждого класса.

Все перечисленные подходы обладают для социальных наук двумя ограничениями: порождают инструкции-«черные ящики» и оптимизируют одну скалярную метрику без контроля поклассового баланса. Как показано далее, простое добавление примеров не спасает от деградации миноритарных классов. PromptAlign наследует идею автоматической генерации, но выстраивает инструкцию целенаправленно, на основе явно извлеченных признаков и правил из систематического анализа расхождений с эталоном.

Выравнивание языковых моделей без дообучения

Задача выравнивания (alignment) ответов модели с человеческими предпочтениями обычно решается изменением параметров: классическая схема RLHF [7, 8] обучает модель наград и затем настраивает LLM с помощью обучения с подкреплением, а более поздние подходы вроде Direct Preference Optimization (DPO) [9] устраняют модель наград, но по-прежнему требуют доступа к весам, т. е. неприменимы, когда LLM доступна только через закрытый API (GPT, Claude, DeepSeek).

В парадигме «выравнивание через инструкцию» тщательно сконструированные инструкции заметно смещают поведение модели: цепочечные рассуждения (chain-of-thought) [10] вызывают сложное логическое поведение за счет

формата инструкции, а ручная итеративная рефлексия эксперта также повышает точность, но, как показано в [1], неизбежно вызывает эффект «перетягивания одеяла».

До сих пор не был реализован систематический метод, который автоматически корректировал бы инструкцию по сравнению с эталонной разметкой и одновременно контролировал бы баланс классов. PromptAlign закрывает этот пробел замкнутым циклом анализа расхождений, генерации правил и компиляции инструкции в чисто инструктивной парадигме с опорой на статистический критерий приемки и метрику BPI.

Классификация психологических и поведенческих характеристик текстов

Значительная часть работ по автоматическому определению психологических и поведенческих особенностей авторов по текстам посвящена выявлению агрессии, токсичности и оскорбительной речи. В работе [11] проведен анализ автоматического обнаружения языка вражды; показано, что модели склонны смешивать оскорбительную лексику с объективно агрессивными высказываниями. В [12] предложена более детальная таксономия, различающая тип и цель оскорбительных сообщений. В области конструктивного общения в [13] представлен корпус комментариев к новостным статьям и продемонстрировано, что отделение конструктивных аргументов от деструктивных требует глубокого семантического анализа. Более широкие личностные модели также извлекаются из текстов: в [14] показано, что лингвистические признаки предсказывают личностные черты с точностью, сравнимой с оценками знакомых людей.

МЕТОДОЛОГИЯ

Метод PromptAlign реализует замкнутый автоматический цикл, приближающий предсказания LLM к экспертной разметке без доступа к ее весам: модель используется для извлечения структурированных лингвистических и семантических характеристик текста, выявления систематических расхождений с экспертом в этом пространстве характеристик и преобразования найденных закономерностей в правила на естественном языке.

Формальная постановка задачи и общая архитектура

Пусть $D = \{(x_i, y_i^*)\}_{i=1}^N$ – выборка текстов с экспертными метками, где $x_i \in$ X – текст, а $y_i^* \in Y = \{1, \dots, K\}$ – один из K классов. Здесь X – пространство (множество) всех возможных текстов на естественном языке, а *K* – число классов таксономии (в настоящей работе *K* = 3); *Y* = {1, …, *K*} – множество меток классов. Пусть $f_\theta(x, p)$ – большая языковая модель с замороженными параметрами θ, принимающая текст x и текстовую инструкцию $p \in P$. Модель возвращает предсказание $\hat{y} = \arg\max_k P_\theta(k \mid x, p)$.

Целевая метрика – macro-усредненная F1 – имеет вид:

$$\text{macro} - \text{F1}(p) = \frac{1}{K}\sum_{k=1}^{K} \frac{2\,TP_k(p)}{2\,TP_k(p)+FP_k(p)+FN_k(p)}.$$

Задача – найти инструкцию p^*, максимизирующую macro-F1 без превышения допустимой деградации наихудшего класса:

$$p^* = \arg\max_{p \in P} \text{macro} - F1(p) \;\; s.t. \;\; \min_k F1_k(p) \geq \min_k F1_k(p_0) - \epsilon.$$

Здесь p_0 – начальная базовая инструкция, $\epsilon \geq 0$ – разрешенное снижение минимальной классовой F1. Это ограничение противодействует эффекту «перетягивания одеяла» – улучшению одних классов ценой резкого падения других.

Пространство P дискретно и экспоненциально велико, поэтому PromptAlign строит приближение p^* итеративно (общая архитектура конвейера показана на рис. 1), как следующую последовательность шагов.

1. *Классификация.* Текущая инструкция p_t применяется к обучающему и валидационному подмножествам, формируются предсказания $\hat{y}_i^{(t)} = f_\theta(x_i, p_t)$.

2. *Локализация ошибок.* Выделяется множество $E_t = \{i \mid \hat{y}_i^{(t)} \neq y_i^*\}$.

3. *Извлечение признаков*. Вычисляется гибридный признаковый вектор $\phi(x_i)$.

4. *Обучение суррогатных моделей.* Строятся две интерпретируемые L1-логистические регрессии (экспертная и LLM-версия); по разности их коэффициентов определяются признаки, по которым модель расходится с экспертом (формально – ниже).

5. *Метаанализ ошибок.* Стратифицированная выборка из E_t подается LLM с инструкцией сформулировать системные причины ошибок и породить обобщенные правила.

6. *Компиляция инструкции.* Правила и семантическая интерпретация расходящихся признаков встраиваются в инструкцию инкрементально; при накоплении правил делается полная перекомпоновка «с нуля».

7. *Оценка и приемка.* Новая инструкция p_{t+1} оценивается на валидации по парному бутстрэп-тесту и метрике BPI (индекс дисбаланса классов, формализован ниже).

Далее каждый шаг рассмотрен подробно.

Начальная инструкция p_0 содержит только краткие определения трех классов без примеров (см. Приложение А). Классификация выполняется с температурой 0.0 для максимальной воспроизводимости.

Признаковое описание текста $\phi(x)$:

- детерминированные признаки (всего 13) – частотные и статистические показатели без LLM (см. Приложение Б); стабилизирующий «якорь», не зависящий от стохастичности модели;
- LLM-признаки (всего 10) охватывают недоступные правилам семантические аспекты: текст подается в ту же LLM (t = 0.1) с запросом JSON-объекта с бинарными, непрерывными и категориальным полями (см. Приложение Б).

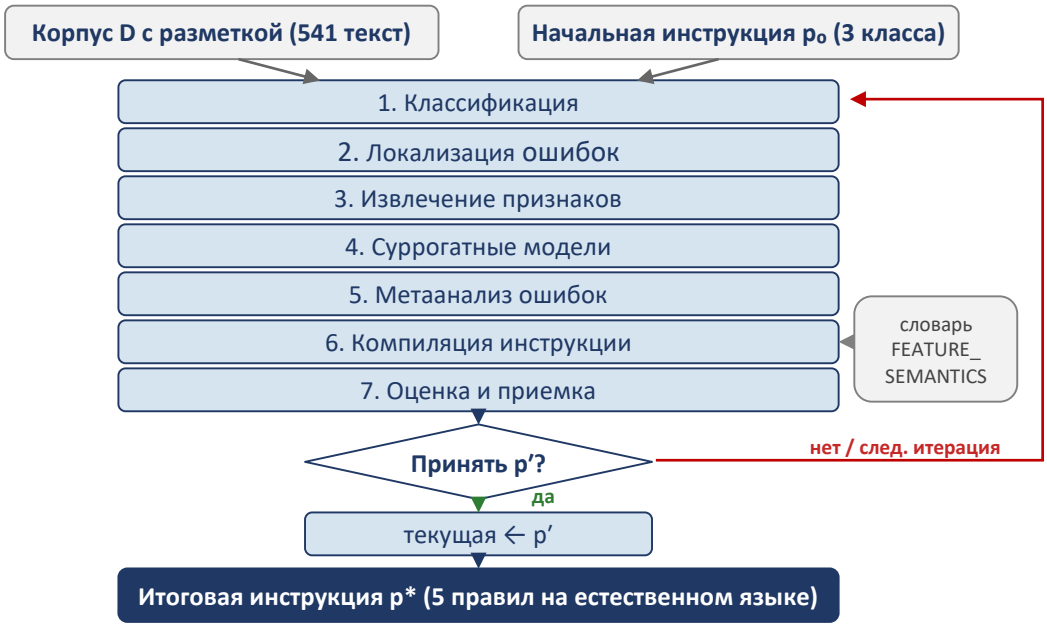


Рис. 1. Архитектура конвейера PromptAlign.

Для повышения устойчивости каждый текст обрабатывается экстрактором дважды ($n_{\text{votes}} = 2$). Бинарные поля усредняются мажоритарным голосованием, непрерывные – средним, категориальный speech_act кодируется унитарно (one-hot). Итоговый вектор $\phi(x)$ образуется конкатенацией 13 детерминированных характеристик и преобразованных LLM-признаков.

На матрице признаков *X* (все числовые поля лежат в отрезке [0, 1]) обучаются две мультиномиальные логистические регрессии:

$$g_{\text{exp}}(x) = \operatorname{argmax}_k P_{\text{exp}}(k \mid x), \qquad g_{\text{LLM}}(x) = \arg\max_k P_{\text{LLM}}(k \mid x).$$

Первая модель предсказывает экспертную метку y^*, вторая – метку, которую присвоила LLM ($\hat{y}^{\text{LLM}}$). Обе обучаются с L1-штрафом (lasso) [15], что автоматически обращает в нуль коэффициенты нерелевантных переменных и дает разреженное, легко интерпретируемое решение; при небольшой обучающей выборке такая регуляризация особенно важна, так как подавляет переобучение и выделяет действительно значимые признаки.

Для каждого *j*-го признака вычисляется величина

$$\delta_j = \frac{1}{K}\sum_{k=1}^{K} |w_{kj}^{\text{exp}}| - \frac{1}{K}\sum_{k=1}^{K} |w_{kj}^{\text{LLM}}|,$$

где w_{kj}^{exp}, w_{kj}^{LLM} – веса *j*-го признака в *k*-м классе для экспертной и LLM-моделей соответственно. Если $\delta_j > 0$, эксперт использует этот признак интенсивнее, чем LLM; если $\delta_j < 0$, признак избыточно эксплуатируется моделью.

Восемь признаков с наибольшими абсолютными значениями δ_j передаются генератору инструкции через фиксированный словарь FEATURE_SEMANTICS, переводящий числовые расхождения в содержательные указания (например, для has_facts: «Наличие фактов, цифр и статистики – главный признак Информативного класса»).

Из множества ошибок E_t формируется стратифицированная подвыборка (не более двух примеров на каждый из шести переходов «истинный класс → предсказанный класс», не более 15 текстов всего) – баланс представительности ошибок и затрат на API.

Каждый отобранный текст снабжается метаинструкцией, которая требует от LLM проанализировать причину ошибки и сформулировать обобщенное правило в формате JSON (поля reasons и rule). Пример: «Если текст содержит критику, аргументацию и предложение решения (даже резкое), то класс Конструктивный, а не Деструктивный». Правила накапливаются в наборе до 20 записей, дубликаты отбрасываются.

Каждая итерация включает инкрементальную компиляцию: текущая инструкция p_t дополняется не более чем четырьмя свежими правилами из метаанализа и семантическими подсказками о признаках с наибольшими δ_j. Генератор (та же LLM) сохраняет исходную структуру, лишь добавляя переданные правила.

Когда получено не менее пяти правил и пройдено две итерации, запускается полная перекомпоновка. Генератор инструкций получает базовый шаблон (p_0), весь набор правил и подсказки, синтезируя новую инструкцию с обобщением правил и устранением противоречий и избыточности. Постобработка гарантирует маркер {text} и фразу «Ответь ОДНИМ СЛОВОМ (Информативный/Деструктивный/Конструктивный)»; при превышении лимита в 1500 символов следует запрос на сокращение.

Статистический критерий принятия инструкции и контроль баланса

Решение о замене текущей лучшей инструкции p_{best} на кандидата p_{t+1} опирается на комбинацию бутстрэп-теста и анализа поклассовых метрик.

Парный бутстрэп-тест. На валидационной выборке вычисляется разность macro-F1 между новой и лучшей инструкциями. Генерируется 1000 бутстрэп-выборок ресэмплингом с возвращением [16] по парам «текст – истинная метка» с пересчетом той же разности; по эмпирическому распределению оцениваются:

- p_{better} – доля бутстрэп-выборок, где новая инструкция показала более высокую macro-F1;
- 95%-й доверительный интервал (CI) для разности.

Новая инструкция принимается, если выполняется хотя бы одно из двух условий:

1) $p_{\text{better}} \geq 0.75$;

2) прирост macro-F1 превышает 0.015 при снижении минимальной классовой F1 не более чем на 0.05 от p_{best}.

Такой составной критерий выработан эмпирически: он обеспечивает чувствительность к реальным улучшениям, не пропуская при этом ложные срабатывания на малых выборках.

Контроль «перетягивания одеяла». Параллельно вычисляется индекс дисбаланса классов (Blanket Pulling Index, BPI):

$$\text{BPI}_t = \max_k \Delta F_{1,k} - \min_k \Delta F_{1,k}, \quad \Delta F_{1,k} = F_{1,k}(p_{t+1}) - F_{1,k}(p_{\text{best}}).$$

BPI измеряет размах изменений классовых F1: большое значение говорит о том, что рост одних классов получен за счет падения других. Если худший класс деградирует более чем на 0.03, а BPI превышает 0.15, то кандидат отвергается даже при пройденном бутстрэп-тесте.

Все пороговые параметры (0.75; 0.015; 0.05; 0.03; 0.15) зафиксированы для всех экспериментов и подобраны предварительными тестами для консервативной приемки улучшений при ограниченном объеме валидационных данных.

ОРГАНИЗАЦИЯ ЭКСПЕРИМЕНТА

Исходные данные

Был использован тот же корпус из 541 текстов на русском языке, что и в [1], собранный с форумов, из социальных сетей и блогов. Эксперты-психологи разметили каждый текст по трехклассовой таксономии: информативный (нейтральная передача фактов, аналитика, статистика без выраженной авторской позиции и призывов к действию), деструктивный (агрессия, провокации, токсичность, выражение безысходности и эмоционального страдания без попыток решения) и конструктивный (эмпатия, сотрудничество, предложение решений, критика с аргументацией и рекомендациями).

Выборка стратифицированно была разделена на три части (random_state = 42). Тестовый набор (20% от корпуса, $N_{\text{test}} = 109$) зарезервирован исключительно для итоговой оценки. Оставшиеся 432 текста разделены на обучающую ($N_{\text{train}} = 302$, что составляет 70% от оставшихся 432 текстов) и валидационную ($N_{\text{val}} = 130$, что составляет 30% от тех же 432 текстов) выборки. Распределение классов представлено в табл. 1.

Табл. 1. Распределение классов по выборкам.

Выборка	Информативный	Деструктивный	Конструктивный	Всего
Обучающая	101	101	100	302
Валидационная	43	43	44	130
Тестовая	36	36	37	109
Итого	**180**	**180**	**181**	**541**

Языковая модель и протокол взаимодействия

Все запросы отправлялись через публичный API deepseek-chat (DeepSeek-V3) с температурой 0.0 для классификации и метаанализа (воспроизводимость), 0.1 для признаков (разнообразие при усреднении) и 0.2 для генерации инструкции (баланс креативности и контроля).

Для извлечения признаков каждый текст обрабатывали дважды ($n_{\text{votes}} = 2$) с усреднением результатов (см. Методологию). Длинные тексты (свыше 4000 символов) обрезали до последнего полного предложения. Вызовы соблюдали ограничения API (28 запросов/мин, 12 000 токенов/мин) с автоматическим ожиданием при превышении лимитов.

Параметры PromptAlign

Конфигурация конвейера зафиксирована заранее и не изменялась в ходе итераций (табл. 2); значения подобраны предварительными тестами на подмножестве данных без оптимизации под финальный результат, чтобы избежать переобучения.

Табл. 2. Параметры конвейера.

Параметр	Значение
Максимальное число итераций	8
Бутстрэп-повторов	1000
Порог p_{better}	≥ 0.75
Альтернативный порог прироста macro-F1	> 0.015
Допустимое снижение min-F1	≤ 0.05
Порог BPI для отклонения	> 0.15
Периодичность перекомпоновки	Каждые 2 итерации

Максимальное число правил в наборе	20
Максимальная длина инструкции	1500 символов
C для логистической регрессии	0.5
Решатель логистической регрессии	liblinear
Автоматическое расширение признаков	Выключено

Вычислительные ресурсы и воспроизводимость

Конвейер реализован на Python 3.10 (scikit-learn 1.3, pandas 2.0, numpy 1.24, openai 1.0) в среде Google Colab (Intel Xeon, 12 ГБ ОЗУ). Для полной воспроизводимости все ответы LLM кэшировались в файлах JSON и восстанавливались при повторном запуске, что устраняло влияние случайности и позволяло продолжать прерванные вычисления без потерь.

Полный цикл из восьми итераций занял около 20 ч (почти все время – ожидание ответов API) и около 2600 вызовов API; обучение регрессий и бутстрэп-тесты заняли менее минуты. Стоимость по тарифам DeepSeek (июнь 2025 г.) не превысила 1–2 долларов США, т. е. метод доступен даже при минимальном бюджете (для сравнения: few-shot – 240 вызовов, без семантического перевода – 130 USD). Код, данные и результаты доступны в открытом репозитории (https://github.com/AlexanderAgathonov/PromptAlign).

РЕЗУЛЬТАТЫ

Основной результат

Полный цикл из восьми итераций дал два статистически значимых улучшения; дальнейшие модификации конвейер устойчиво отклонял.

Динамика на валидационной выборке показана в табл. 3. Базовая инструкция p_0 (лишь краткие определения классов) показала macro-F1 = 0.373 и min-F1 = 0.219 – наихудшим был конструктивный класс. Первая итерация (инкрементальная генерация с семантическими подсказками) подняла macro-F1 сразу на +0.265. На второй итерации сработала полная перекомпоновка на основе 22 накопленных правил, добавив еще +0.139 по macro-F1 и доведя min-F1 до 0.723. Итерации 3–8 не проходили пороги бутстрэп-теста и BPI, поэтому были отвергнуты.

Разрыв между базовыми значениями на валидации (0.373) и тесте (0.655) объясняется тем, что валидационная выборка при стратифицированном разбиении случайно оказалась сложнее для исходной инструкции; поскольку сравнения ведутся на одних и тех же разбиениях, относительный прирост остается корректным.

На отложенной тестовой выборке (табл. 4) PromptAlign статистически значимо превзошел базовый уровень (zero-shot) по всем ключевым метрикам: macro-F1 вырос с 0.655 до 0.749, min-F1 – с 0.449 до 0.656. Парный бутстрэп-тест (1000 повторов) дает 95%-ный доверительный интервал для разности macro-F1 [+0.014; +0.174] и $p_{\text{better}} = 0.992$.

Табл. 3. Динамика качества на валидационной выборке (130 текстов).

Итерация	Действие	macro-F1	min-F1	BPI	Статус
0 (базовый)	–	0.373	0.219	–	–
1	Инкрементальная генерация	0.638	0.608	0.195	Принят
2	Полная перекомпоновка	0.777	0.723	0.051	Принят
3–8	Инкрементальная перекомпоновка	≤ 0.703	≤ 0.609	–	Отклонен

Примечание: в строке «3–8» приведены максимальные значения метрик среди отклоненных итераций.

Табл. 4. Сравнение на тестовой выборке (109 текстов).

Метрика	Базовый уровень (zero-shot)	PromptAlign
macro-F1	0.655	**0.749**
min-F1	0.449	**0.656**
F1 Информативный	0.706	0.756
F1 Деструктивный	0.810	0.833
F1 Конструктивный	0.449	**0.656**
Коэффициент каппа Коэна (к)	0.531	0.628
BPI (относительно базового уровня)	–	0.183

Основная неопределенность сохраняется между конструктивным и информативным классами, но ее масштаб существенно уменьшился. Базовая инструкция ошибочно относила 17 из 36 конструктивных текстов к информативным; PromptAlign сократил число таких ошибок до 11. Интегральный показатель дисбаланса BPI = 0.183 подтверждает, что прирост min-F1 на +0.207 достигнут без сильной деградации других классов: величина F1 информативного класса выросла на 0.050, деструктивного – на 0.023.

Сравнение с базовым уровнем «несколько примеров» (few-shot)

PromptAlign сравнивался с базовым уровнем few-shot prompting – инструкцией с вручную добавленными тремя репрезентативными примерами каждого класса из обучающей выборки. Примеры были подобраны так, чтобы выразительно проиллюстрировать маркеры классов и не превысить разумные ограничения контекста (см. Приложение В; полные тексты – в репозитории).

Результаты на тестовой выборке представлены в табл. 5. Метод с подачей примеров дает лишь незначительный прирост над методом без примеров (macro-F1 на +0.017, min-F1 на +0.042) и одновременно ухудшает распознавание деструктивного класса (F1 с 0.810 до 0.782). Это прямое проявление «перетягивания одеяла»: акцент на конструктивных и информативных примерах смещает вероятности модели в ущерб деструктивному классу.

PromptAlign значимо превосходит метод с подачей примеров (few-shot) по всем метрикам (macro-F1 выше на 0.077, min-F1 – на 0.165), причем наибольший прирост приходится на конструктивный класс (+0.165) благодаря явному контролю баланса через BPI и суррогатные модели.

Табл. 5. Сравнение методов на тестовой выборке (109 текстов).

Метод	macro-F1	min-F1	F1 Информ.	F1 Дестр.	F1 Констр.
Метод без примеров (базовый)	0.655	0.449	0.706	0.810	0.449
Метод с подачей примеров (3 примера/класс)	0.672	0.491	0.744	0.782	0.491
PromptAlign	**0.749**	**0.656**	**0.756**	**0.833**	**0.656**

Анализ вклада компонентов

Чтобы оценить роль ключевых архитектурных решений, был проведен анализ вклада компонентов (ablation study). Полный перезапуск конвейера в различных конфигурациях крайне ресурсоемок, поэтому результаты основного эксперимента комбинировались с контролируемыми локальными тестами. Конфигурация «Без семантического перевода» получена работой генератора инструкций на данных первой итерации, но с передачей ему сырых чисел δ_j вместо словаря FEATURE_SEMANTICS. Конфигурация «Без перекомпоновки» соответствует состоянию после первой (инкрементальной) итерации. Конфигурация «Жесткий критерий» моделирует применение жесткого порога $p_{\text{better}} \geq 0.85$ вместо составного мягкого критерия.

Результаты на валидационной выборке представлены в табл. 6.

Табл. 6. Анализ вклада компонентов (валидационная выборка, 130 текстов).

Конфигурация	macro-F1	min-F1	Δ macro-F1
Полный метод (итерация 2)	**0.777**	**0.723**	**–**
Без семантического перевода	0.612	0.559	−0.165
Без перекомпоновки (только итерация 1)	0.638	0.608	−0.139
Жесткий критерий ($p \geq 0.85$)	0.777	–	0.000

Анализ вклада компонентов (табл. 6) позволяет сформулировать три ключевых наблюдения.

1. *Семантический перевод признаков.* Без словаря FEATURE_SEMANTICS генератор инструкций не получает из сырых числовых расхождений весов достаточного контекста для синтеза осмысленных правил (табл. 6) – лингвистическая трансляция статистики оказывается критичной.

2. *Периодическая перекомпоновка.* Инкрементальная компиляция дает прирост, но без обобщения накопленных правил качество стабилизируется (табл. 6); полная перекомпоновка устраняет противоречия и избыточность в разросшейся инструкции.

3. *Мягкий критерий принятия.* В данном эксперименте более жесткий порог не изменил бы итогового результата, но критерий служит защитой от ложных срабатываний на более шумных или малых выборках.

ОБСУЖДЕНИЕ

Интерпретируемость и экспликация экспертных критериев

Одним из главных требований к автоматическим системам в социальных и поведенческих науках является прозрачность решений [17, 18]. В отличие от APE или OPRO, порождающих «черные ящики», итоговая инструкция PromptAlign содержит пять обобщенных правил (см. Приложение А), которые напрямую отражают критерии экспертов-психологов, не сформулированные ими явно до эксперимента, и доступны для независимой проверки специалистом.

Например, правило, предписывающее относить тексты с фактами и аналитикой без предложения решений к информативному классу, устраняет проблему из [1]: модель склонна была принимать любую выраженную авторскую позицию за признак конструктивного поведения. Таким образом, конвейер не просто повышает метрики, но и служит инструментом экспликации неявного экспертного знания и уточнения таксономий.

Анализ оставшихся ошибок: семантические границы и ирония

Несмотря на значительный прогресс, смешение конструктивных и информативных классов остается основным источником ошибок (11 из 36 конструктивных текстов отнесены к информативным): такие тексты содержат развернутый анализ, аргументацию и выводы, но лишены явного призыва к действию или предложения решения – граница между «глубокой аналитикой с авторской позицией» и «конструктивным разбором причин и следствий» оказывается тонкой и для эксперта, и для модели.

Более сложную проблему создает культурная специфика русскоязычной иронии: в примере из Приложения В текст формально изобилует позитивной лексикой («толковая», «верно», «грамотно», «красиво»), но его реальный смысл состоит в едкой насмешке и скрытой агрессии. Итоговая инструкция указывает, что сарказм может маскировать деструктив, но не дает формальных маркеров

для его распознавания. Эта проблема хорошо известна в компьютерной лингвистике [19] и особенно остра для русского языка.

Сравнение с ручной рефлексией и обучением с подачей примеров

Сопоставление с ручной итеративной рефлексией [1] демонстрирует преимущество автоматического контроля баланса: ручная оптимизация в [1] достигла пиковых F1 (0.972 информативный, 0.967 деструктивный), но min-F1 не превысила 0.619 из-за «перетягивания одеяла», тогда как PromptAlign обеспечил min-F1 = 0.656 при одновременном росте F1 информативного (0.756) и деструктивного (0.833) классов, т. е. уступает в пиковых значениях, но дает более сбалансированное и устойчивое решение, что для прикладных социальных исследований часто предпочтительнее, причем весь цикл коррекции выполняется без участия эксперта, в отличие от ручной рефлексии.

Сравнение с методом с подачей примеров (табл. 5) подтверждает, что простое добавление примеров не устраняет смещение выравнивания и не спасает от деградации деструктивного класса – явное извлечение признаков, локализация расхождений и генерация правил заметно эффективнее неявного обучения на примерах.

Ограничения метода и эксперимента

Перечислим основные ограничения.

1. Объем выборки, статистическая мощность и локальная сходимость. При выборке из 541 текста доверительные интервалы бутстрэп-теста остаются широкими, а правила метаанализа могут быть переобучены на формулировки обучающей выборки и не обобщаться на иные жанры. Кроме того, конвейер гарантирует улучшение лишь в окрестности стартовой инструкции p_0; сходимость к глобальному оптимуму в пространстве всех текстовых инструкций не гарантируется.

2. Граница применимости: экспертное проектирование таксономии и словаря. Метод не автоматизирует проектирование таксономии и признакового пространства, это требует участия эксперта в каждой новой области (определить классы, признаки, словарь FEATURE_SEMANTICS), хотя сама архитектура от предметной области не зависит (см. «Перспективы»). Автоматизирован только цикл поиска правил после того, как классы и признаки заданы.

3. Ограничение длины инструкции и зависимость от качества LLM-признаков. Лимит в 1500 символов (экономия, компактность) ограничивает сложность и число представимых правил; отдельно, плохое распознавание иронии, эмпатии или скрытой агрессии зашумляет суррогатные признаки, а пороги BPI (0.15 и 0.03), подобранные эмпирически, при смещении могут изменить набор принимаемых инструкций, хотя общая логика защиты от дисбаланса устойчива.

Перспективы применения и дальнейшие исследования

Архитектура PromptAlign универсальна по отношению к предметной области, однако ее перенос в новую область по-прежнему требует априорного участия эксперта (см. разд. «Ограничения метода и эксперимента»): замены описаний классов, адаптации набора признаков и предоставления размеченной выборки. Потенциальные сферы применения включают медицинскую классификацию по срочности (триаж), модерацию пользовательского контента, анализ юридических документов и выявление маркеров психологического неблагополучия.

Среди направлений дальнейших исследований – автоматическая ранняя остановка по динамике прироста метрик вместо фиксированного лимита итераций; ансамблирование нескольких лучших инструкций для устойчивости предсказаний; интеграция контекстных векторных представлений для распознавания иронии и скрытой агрессии; масштабирование и валидация метода на более крупных мультиязычных наборах данных.

ЗАКЛЮЧЕНИЕ

Предложен и экспериментально исследован PromptAlign – автоматический конвейер, выравнивающий предсказания LLM под экспертную разметку без дообучения весов и решающий две ключевые проблемы: систематическое смещение и «перетягивание одеяла» – деградацию миноритарных классов при ручной или контекстной оптимизации инструкций.

PromptAlign опирается на гибридные признаки, интерпретируемые L1-логистические регрессии для выявления расхождений с экспертом и метаанализ ошибок, порождающий правила на естественном языке. Эксперимент на корпусе из 541 текста на русском языке (три класса) показал: уже за две итерации конвейер формирует инструкцию, статистически значимо превосходящую и

zero-shot-, и few-shot-подходы, а также ручную итеративную рефлексию эксперта. Величина Macro-F1 на отложенной тестовой выборке выросла с 0.655 до 0.749, минимальная классовая F1 – с 0.449 до 0.656 ($p > 0.99$ по парному бутстрэп-тесту). Метрика BPI удержала «перетягивание» на низком уровне (0.183), позволив усилить конструктивный класс без потерь для остальных.

Итоговая инструкция содержит пять прозрачных правил: в отличие от «черных ящиков», порождаемых многими оптимизаторами, они не просто повышают метрики, но эксплицируют латентные критерии экспертов и доступны специалисту для независимой проверки и коррекции.

Предложенный конвейер универсален по отношению к предметной области и применим в любых задачах автоматической классификации текстов, где доступна экспертная разметка и важна прозрачность принимаемых решений.

Благодарности

Исследование выполнено за счет гранта Российского научного фонда № 24-18-01125 «Психологическая модель поведения субъекта в цифровой среде», https://rscf.ru/project/24-18-01125/.

СПИСОК ЛИТЕРАТУРЫ

1. *Агафонов А.А., Хусаинов А.Р., Устин П.Н, Попов Л.М.* Большие языковые модели как инструмент прогнозирования типов поведения личности // Российский психологический журнал. 2026. № 23(2) (в печати).

2. *Zhou Y., Muresanu A.I., Han Z., et al.* Large Language Models Are Human-Level Prompt Engineers. arXiv preprint arXiv:2211.01910. 2023. https://doi.org/10.48550/arXiv.2211.01910 (дата обращения: 15.05.2026).

3. *Yang C., Wang X., Lu Y., et al.* Large Language Models as Optimizers. arXiv preprint arXiv:2309.03409. 2024. https://doi.org/10.48550/arXiv.2309.03409 (дата обращения: 15.05.2026).

4. *Khattab O., Singhvi A., Maheshwari P., et al.* DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines. arXiv preprint arXiv:2310.03714. 2023. https://doi.org/10.48550/arXiv.2310.03714 (дата обращения: 15.05.2026).

5. *Brown T.B., Mann B., Ryder N., et al*. Language Models are Few-Shot Learners // Advances in Neural Information Processing Systems. 2020. Vol. 33. P. 1877–1901.

6. *Reynolds L., McDonell K.* Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm // Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems. 2021. P. 1–7. https://doi.org/10.1145/3411763.3451760

7. *Christiano P.F., Leike J., Brown T.B., et al.* Deep Reinforcement Learning from Human Preferences // Advances in Neural Information Processing Systems. 2017. Vol. 30. P. 4299–4307.

8. *Ouyang L., Wu J., Jiang X., et al.* Training Language Models to Follow Instructions with Human Feedback // Advances in Neural Information Processing Systems. 2022. Vol. 35. P. 27730–27744.

9. *Rafailov R., Sharma A., Mitchell E., et al.* Direct Preference Optimization: Your Language Model is Secretly a Reward Model // Advances in Neural Information Processing Systems. 2023. Vol. 36. P. 53728–53741.

10. *Wei J., Wang X., Schuurmans D., et al.* Chain-of-Thought Prompting Elicits Reasoning in Large Language Models // Advances in Neural Information Processing Systems. 2022. Vol. 35. P. 24824–24837.

11. *Davidson T., Warmsley D., Macy M., Weber I.* Automated Hate Speech Detection and the Problem of Offensive Language // Proceedings of the 11th International AAAI Conference on Web and Social Media. 2017. P. 512–515. https://doi.org/10.1609/icwsm.v11i1.14955

12. *Zampieri M., Malmasi S., Nakov P., et al.* Predicting the Type and Target of Offensive Posts in Social Media // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019. P. 1415–1420. https://doi.org/10.18653/v1/N19-1144

13. *Kolhatkar V., Wu H., Cavasso L., et al.* The SFU Opinion and Comments Corpus: A Corpus for the Analysis of Online News Comments // Corpus Pragmatics. 2020. Vol. 4, No. 2. P. 155–190. https://doi.org/10.1007/s41701-019-00065-w

14. *Park G., Schwartz H.A., Eichstaedt J.C., et al.* Automatic Personality Assessment through Social Media Language // Journal of Personality and Social Psychology. 2015. Vol. 108, No. 6. P. 934–952. https://doi.org/10.1037/pspp0000020

15. *Tibshirani R.* Regression Shrinkage and Selection via the Lasso // Journal of the Royal Statistical Society. Series B (Methodological). 1996. Vol. 58, No. 1. P. 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

16. *Efron B.* Bootstrap Methods: Another Look at the Jackknife // The Annals of Statistics. 1979. Vol. 7, No. 1. P. 1–26. https://doi.org/10.1214/aos/1176344552

17. *Doshi-Velez F., Kim B.* Towards a Rigorous Science of Interpretable Machine Learning. arXiv preprint arXiv:1702.08608. 2017. https://doi.org/10.48550/arXiv.1702.08608 (дата обращения: 15.05.2026).

18. *Lipton Z.C.* The Mythos of Model Interpretability // Communications of the ACM. 2018. Vol. 61, No. 10. P. 36–43. https://doi.org/10.1145/3233231

19. *Van Hee C., Lefever E., Hoste V.* We Usually Don't Like Going to the Dentist: Using Common Sense to Detect Irony on Twitter // Computational Linguistics. 2018. Vol. 44, No. 4. P. 793–832. https://doi.org/10.1162/coli_a_00337

ПРИЛОЖЕНИЕ А

Итоговая инструкция, сгенерированная PromptAlign. Полный текст инструкции (после второй итерации), использованной для финального тестирования (тег <no_think> – служебная команда отключения режима расширенных рассуждений модели):

```
<no_think>
Определи тип поведения автора текста. Есть три класса:
• Деструктивный – прямые оскорбления, угрозы, унижения, призывы к разрушению, паника, агрессия, а также выражение безысходности, эмоционального страдания или беспомощности без попыток решения.
• Информативный – аналитика, факты, статистика, цифры, описание событий или проектов; НЕТ личной позиции автора и призывов к действию. Даже при наличии эмоциональной окраски, сарказма или критики, если нет предложений решений.
• Конструктивный – предложения решений, эмпатия, поддержка, призывы к созиданию, спокойный тон. Сюда же относится критика с аргументацией и предложением улучшений, анализ причин и следствий с выводами, личные наблюдения с призывом к диалогу.

Обобщенные правила:
1. Если текст содержит факты, цифры, статистику или аналитические выводы
```

```
без явных предложений по решению проблемы – это Информативный, даже при
наличии эмоций или критики.
2. Если текст содержит критику, аргументацию и предложение решения (даже
резкое) – это Конструктивный, а не Деструктивный.
3. Если текст содержит эмоциональное страдание, безысходность, жалобы на
безвыходность без конкретных шагов – это Деструктивный, даже при наличии
просьбы о совете.
4. Если текст содержит развернутое сравнение, анализ причин и следствий,
выводы или позитивную оценку развития – это Конструктивный, даже при эле-
ментах пересказа.
5. Важно: Конструктивный класс часто путают с Информативным. Критика +
предложение решения = Конструктивный. Просто факты = Информативный.

Ответь ОДНИМ СЛОВОМ (Информативный / Деструктивный / Конструктивный).

Текст: {text}
```

ПРИЛОЖЕНИЕ Б

Список извлекаемых признаков. В табл. Б.1 перечислены детерминированные признаки, вычисляемые без обращения к LLM; в табл. Б.2 – LLM-признаки, извлекаемые отдельным запросом к модели. Признак speech_act кодируется унитарно (one-hot), непрерывные значения масштабируются в отрезок [0, 1].

Табл. Б.1. Детерминированные признаки.

Номер	Признак	Описание
1	rb_exclamation	доля восклицательных знаков/предложение
2	rb_question	доля вопросительных знаков/предложение
3	rb_caps_ratio	доля заглавных букв
4	rb_negative_words	частота отрицательных частиц
5	rb_first_person	доля местоимений 1-го лица
6	rb_we_form	доля форм «мы», «нам», «наш»
7	rb_imperative	доля императивных конструкций
8	rb_avg_sent_len	средняя длина предложения (слов)
9	rb_text_len_log	логарифм длины текста
10	rb_ellipsis	количество многоточий

11	rb_quotes	число кавычек
12	rb_numbers	доля числовых выражений
13	rb_url	наличие URL (бинарный)

Табл. Б.2. LLM-признаки.

Номер	Признак	Тип	Описание
1	has_aggression	бинарный	прямая агрессия
2	has_indirect_aggression	бинарный	косвенная/пассивная агрессия
3	has_empathy	бинарный	эмпатия, поддержка
4	has_solution	бинарный	наличие предлагаемого решения
5	has_facts	бинарный	наличие фактов/статистики
6	has_author_position	бинарный	выраженная авторская позиция
7	has_call_to_action	бинарный	призыв к действию
8	speech_act	категориальный	тип речевого акта (6 категорий)
9	irony_probability	непрерывный [0,1]	вероятность иронии
10	emotional_intensity	непрерывный [0,1]	эмоциональная интенсивность

ПРИЛОЖЕНИЕ В

Примеры текстов для few-shot базового уровня. Ниже приведен показательный пример деструктивного класса (на него ссылается разд. «Обсуждение»); полный набор из 9 примеров (по 3 на класс), использованный при формировании few-shot-инструкции, доступен в репозитории проекта.

Пример (деструктивный класс)

Летом прошлого года Ученый совет Российского государственного гуманитарного университета утвердил создание Высшей политической школы имени Ивана Ильина. Разобраться в назначении данной конторы непросто. Для этого достаточно прочитать цель из положения о школе: «Разработка и реализация нового подхода… с опорой на российскую цивилизационную идентичность и традиционные российские духовно-нравственные ценности». Ну толковая же контора? Кто поспорит? По крайней мере, на бумаге все очень верно и грамотно построено. В общем, все красиво… Не то слово. Красивее не бывает.

[Приложение В. Примеры текстов для few-shot базового уровня] (https://github.com/AlexanderAgathonov/PromptAlign-Code/blob/main/supplementary_materials/appendix_b_few_shot_examples.md).

ПРИЛОЖЕНИЕ Г

Правила, сгенерированные метаанализом. Метаанализ ошибок на разных итерациях сформировал набор до 20 правил на естественном языке (формат JSON с полями reasons и rule); в обобщенном виде они вошли в итоговую инструкцию (Приложение А). Полный список правил доступен в репозитории.

[Приложение Г. Правила, сгенерированные метаанализом] (https://github.com/AlexanderAgathonov/PromptAlign-Code/blob/main/supplementary_materials/appendix_c_generated_rules.md).

PROMPTALIGN: AUTOMATIC ITERATIVE ALIGNMENT OF A LARGE LANGUAGE MODEL TO EXPERT ANNOTATIONS WITHOUT FINE-TUNING

A. A. Agathonov[1] [0000-0001-6924-6701], A. R. Khusainov[2] [0009-0002-1916-2109], N. A. Prokopyev[3] [0000-0003-0066-7465]

[1-3]*Kazan (Volga region) Federal University, Kazan, Russia*

[1]a.a.agathonov@gmail.com, [2]ahat2182@gmail.com, [3]nikolai.prokopyev@gmail.com

Abstract

Large language models are increasingly used in social and psychological sciences; however, their predictions are systematically biased relative to expert annotations (alignment shift), and attempts to manually refine instructions or to supply examples lead to a “blanket-pulling” effect – degradation of minority classes.

We propose PromptAlign, an automatic prompt-optimization pipeline that requires neither access to model weights nor human involvement in iterative correction, because the taxonomy and the feature space are specified by an expert a priori. The pipeline extracts hybrid linguistic and semantic features using the large language model itself, builds interpretable L1-regularized logistic regressions to localize disagreements with the expert, performs a meta-analysis of errors, and compiles new prompts while controlling the class balance through a class-imbalance index (Blanket Pulling Index, BPI). An experiment on a corpus of 541 Russian-language texts with three psychological classes showed an increase in macro-F1 on the held-out test set from 0.655 to 0.749 and in the minimum per-class F1 from 0.449 to 0.656 ($p>0.99$ by a paired bootstrap test). The method significantly outperforms both the zero-shot

baseline and few-shot prompting. A component-contribution analysis confirms the critical role of semantic interpretation of features and of periodic prompt rebuilding. The final prompt contains five rules that reflect latent expert criteria; unlike the “black boxes” produced by automatic optimizers, these rules are available for inspection and correction by a specialist.

Keywords*: large language model, prompt optimization, alignment, interpretability, L1 regularization, bootstrap test, text classification, psychological annotation, surrogate model, class imbalance index, Russian-language corpus.*

REFERENCES

1. *Agathonov, A.A., Khusainov, A.R., Ustin, P.N., & Popov, L.M.* Large language models as a tool for predicting types of personal behavior // Russian Psychological Journal. 2026. № 23(2) (in print).

2. *Zhou Y., Muresanu A.I., Han Z., et al.* Large Language Models Are Human-Level Prompt Engineers. arXiv preprint arXiv:2211.01910. 2023. https://doi.org/10.48550/arXiv.2211.01910 (дата обращения: 15.05.2025).

3. *Yang C., Wang X., Lu Y., et al.* Large Language Models as Optimizers. arXiv preprint arXiv:2309.03409. 2024. https://doi.org/10.48550/arXiv.2309.03409 (accessed May 15, 2026).

4. *Khattab O., Singhvi A., Maheshwari P., et al.* DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines. arXiv preprint arXiv:2310.03714. 2023. https://doi.org/10.48550/arXiv.2310.03714 (accessed May 15, 2026).

5. *Brown T.B., Mann B., Ryder N., et al*. Language Models are Few-Shot Learners // Advances in Neural Information Processing Systems. 2020. Vol. 33. P. 1877–1901.

6. *Reynolds L., McDonell K.* Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm // Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems. 2021. P. 1–7. https://doi.org/10.1145/3411763.3451760

7. *Christiano P.F., Leike J., Brown T.B., et al.* Deep Reinforcement Learning from Human Preferences // Advances in Neural Information Processing Systems. 2017. Vol. 30. P. 4299–4307.

8. *Ouyang L., Wu J., Jiang X., et al.* Training Language Models to Follow Instructions with Human Feedback // Advances in Neural Information Processing Systems. 2022. Vol. 35. P. 27730–27744.

9. *Rafailov R., Sharma A., Mitchell E., et al.* Direct Preference Optimization: Your Language Model is Secretly a Reward Model // Advances in Neural Information Processing Systems. 2023. Vol. 36. P. 53728–53741.

10. *Wei J., Wang X., Schuurmans D., et al.* Chain-of-Thought Prompting Elicits Reasoning in Large Language Models // Advances in Neural Information Processing Systems. 2022. Vol. 35. P. 24824–24837.

11. *Davidson T., Warmsley D., Macy M., Weber I.* Automated Hate Speech Detection and the Problem of Offensive Language // Proceedings of the 11th International AAAI Conference on Web and Social Media. 2017. P. 512–515. https://doi.org/10.1609/icwsm.v11i1.14955

12. *Zampieri M., Malmasi S., Nakov P., et al.* Predicting the Type and Target of Offensive Posts in Social Media // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019. P. 1415–1420. https://doi.org/10.18653/v1/N19-1144

13. *Kolhatkar V., Wu H., Cavasso L., et al.* The SFU Opinion and Comments Corpus: A Corpus for the Analysis of Online News Comments // Corpus Pragmatics. 2020. Vol. 4, No. 2. P. 155–190. https://doi.org/10.1007/s41701-019-00065-w

14. *Park G., Schwartz H.A., Eichstaedt J.C., et al.* Automatic Personality Assessment through Social Media Language // Journal of Personality and Social Psychology. 2015. Vol. 108, No. 6. P. 934–952. https://doi.org/10.1037/pspp0000020

15. *Tibshirani R.* Regression Shrinkage and Selection via the Lasso // Journal of the Royal Statistical Society. Series B (Methodological). 1996. Vol. 58, No. 1. P. 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

16. *Efron B.* Bootstrap Methods: Another Look at the Jackknife // The Annals of Statistics. 1979. Vol. 7, No. 1. P. 1–26. https://doi.org/10.1214/aos/1176344552

17. *Doshi-Velez F., Kim B.* Towards a Rigorous Science of Interpretable Machine Learning. arXiv preprint arXiv:1702.08608. 2017. https://doi.org/10.48550/arXiv.1702.08608 (accessed May 15, 2026).

18. *Lipton Z.C.* The Mythos of Model Interpretability // Communications of the ACM. 2018. Vol. 61, No. 10. P. 36–43. https://doi.org/10.1145/3233231

19. *Van Hee C., Lefever E., Hoste V.* We Usually Don't Like Going to the Dentist: Using Common Sense to Detect Irony on Twitter // Computational Linguistics. 2018. Vol. 44, No. 4. P. 793–832. https://doi.org/10.1162/coli_a_00337

СВЕДЕНИЯ ОБ АВТОРАХ



АГАФОНОВ Александр Алексеевич – к. ф.-м. н., доцент, доцент кафедры высшей математики и математического моделирования Института математики и механики им. Н.И. Лобачевского, ФГАОУ ВО «Казанский (Приволжский) федеральный университет». Сфера научных интересов: обработка естественного языка, большие языковые модели, машинное обучение, искусственный интеллект в медицине, анализ поведения в цифровой среде, электронное образование.

Alexander Alekseevich AGATHONOV – Cand. Sc. (Physics and Mathematics), associate professor at the Department of Higher Mathematics and Mathematical Modeling, N.I. Lobachevsky Institute of Mathematics and Mechanics, Kazan Federal University. Scientific interests: natural language processing, large language models, machine learning, artificial intelligence in medicine, behavior analysis in digital environment, e-Learning.

email: a.a.agathonov@gmail.com

ORCID: 0000-0001-6924-6701



ХУСАИНОВ Ахат Рустемович – лаборант, ФГАОУ ВО «Казанский (Приволжский) федеральный университет». Сфера научных интересов: обработка естественного языка, машинное обучение, искусственный интеллект.

Akhat Rustemovich KHUSAINOV – laboratory assistant, Kazan Federal University. Scientific interests: natural language processing, machine learning, artificial intelligence.

email: khusainov.akhat@mail.ru

ORCID 0009-0002-1916-2109



ПРОКОПЬЕВ Николай Аркадиевич – к. т. н., доцент кафедры информационных систем Института вычислительной математики и информационных технологий, ФГАОУ ВО «Казанский (Приволжский) федеральный университет». Научный сотрудник Института прикладной семиотики Академии наук РТ. Сфера научных интересов: обработка естественного языка, искусственный интеллект, электронное образование.

Nikolai Arkadievich PROKOPYEV – Cand. Sc. (Technology), associate professor at the Department of Information Systems, Institute of Computational Mathematics and Information Technologies, Kazan Federal University. Researcher at the Institute of Applied Semiotics of the Academy of Sciences of the Republic of Tatarstan. Scientific interests: natural language processing, artificial intelligence, e-Learning.

email: nikolai.prokopyev@gmail.com

ORCID: 0000-0003-0066-7465

Материал поступил в редакцию 24 июня 2026 года

УДК 004.42+004.65

ВОЗМОЖНОСТИ РАЗРАБОТАННОЙ ИНФОРМАЦИОННОЙ СИСТЕМЫ СОПРОВОЖДЕНИЯ ЛИЦЕНЗИЙ

Е. И. Александров[1][0000-0003-2057-5955], И. Н. Александров[2][0000-0001-7381-6762], Д. В. Беляков[3][0000-0001-6516-1479], Н. А. Давыдова[4][0009-0002-3740-508X], П. В. Зрелов[5][0000-0002-0642-3042], Л. А. Калмыкова[6][0009-0006-0307-1279], М. А. Любимова[7][0000-0002-7146-7164], Т. Ф. Сапожникова[8][0009-0006-9327-1716], Т. С. Сыресина[9][0009-0001-3160-921X], А. В. Яковлев[10][0000-0002-9534-9357]

[1–10]*Объединенный институт ядерных исследований, г. Дубна, Россия*

[1, 2]*Государственный университет «Дубна», г. Дубна, Россия*

[1]aleksand@jinr.ru, [2]alexand@jinr.ru, [3]dmitry@jinr.ru, [4]luna@jinr.ru, [5]zrelov@jinr.ru, [6]lidak@jinr.ru, [7]lma@jinr.ru, [8]tsap@jinr.ru, [9]syresina@jinr.ru, [10]yakovleva@jinr.ru

Аннотация

Основная цель создания информационной системы сопровождения лицензий (ИССЛ) состояла в автоматизации управления, приобретения, обслуживания и использования лицензионных программных продуктов. Представлены результаты развития этой системы за последние два года. Реализован механизм согласования заявок, и на его основе созданы различные типы заявок: заявки пользователя на новую лицензию, добавление новых программных продуктов в каталог поддерживаемого программного обеспечения, заявки аудитора на закупку дополнительных лицензий. Выполнены работы по наполнению базы данных ИССЛ, удобному представлению информации о лицензиях и различной статистической информации, а также интеграции с другими сервисами в рамках Цифровой экосистемы Объединенного института ядерных исследований.

***Ключевые слова**: информационная система сопровождения лицензий, база данных, веб-интерфейс.*

ВВЕДЕНИЕ

В Лаборатории информационных технологий (ЛИТ) имени М. Г. Мещерякова Объединенного института ядерных исследований (ОИЯИ) активно работают над оптимизацией использования лицензионного программного обеспечения.

© Е. И. Александров, И. Н. Александров, Д. В. Беляков, Н. А. Давыдова, П. В. Зрелов, Л. А. Калмыкова, М. А. Любимова, Т. Ф. Сапожникова, Т. С. Сыресина, А. В. Яковлев, 2026.
Данная статья распространяется на условиях международной лицензии Creative Commons License Attribution 4.0 International (CC BY 4.0).

Существуют различные аспекты работы, связанные с оптимизацией использования лицензий. Разнообразие типов приобретаемых лицензий создает потенциальную сложность при рассмотрении и выборе оптимальных условий лицензирования под конкретные нужды лаборатории. Дополнительные трудности возникают при решении задачи сбора динамической информации об использовании сетевых лицензий. Более подробно эти аспекты рассмотрены в статье [1].

Основной целью создания Информационной системы сопровождения лицензий (ИССЛ) является автоматизация управления, приобретения, сопровождения и использования лицензионных программных продуктов, вызванная необходимостью как планирования и оптимизации закупок лицензионного программного обеспечения, так и необходимостью контроля соблюдения требований законодательства Российской Федерации, касающихся правил лицензионной политики.

В настоящей работе представлен опыт разработки, внедрения и эксплуатации ИССЛ в Лаборатории информационных технологий ОИЯИ, а также развития этой системы за последние два года с целью дальнейшего ее использования во всех лабораториях Института.

1. АРХИТЕКТУРА И ЛОГИЧЕСКОЕ ПРЕДСТАВЛЕНИЕ ИССЛ

ИССЛ состоит из трех частей: системы распределения сетевых лицензий (Network Licensing System NLS), баз данных и веб-сервисов. Общая архитектура ИССЛ представлена на рис. 1. Техническая реализация каждой из компонент подробно рассмотрена в [1].

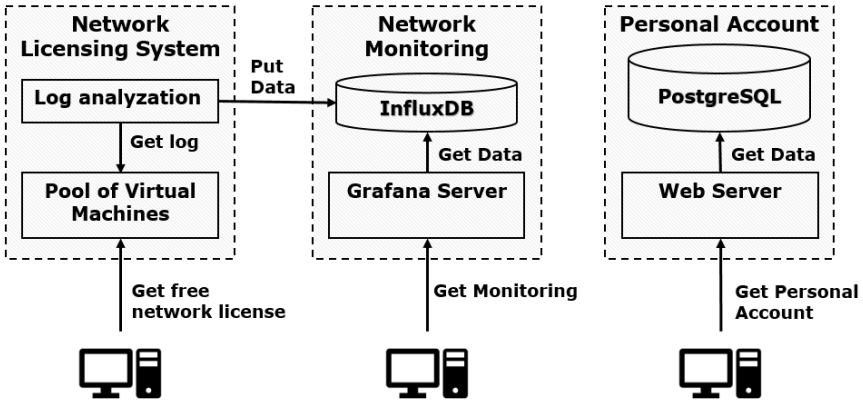


Рис. 1. Общая архитектура ИССЛ.

С точки зрения пользователей ИССЛ состоит из системы распределения сетевых лицензий как отдельной подсистемы и набора различных веб-сервисов. Логическое представление ИССЛ приведено на рис. 2.

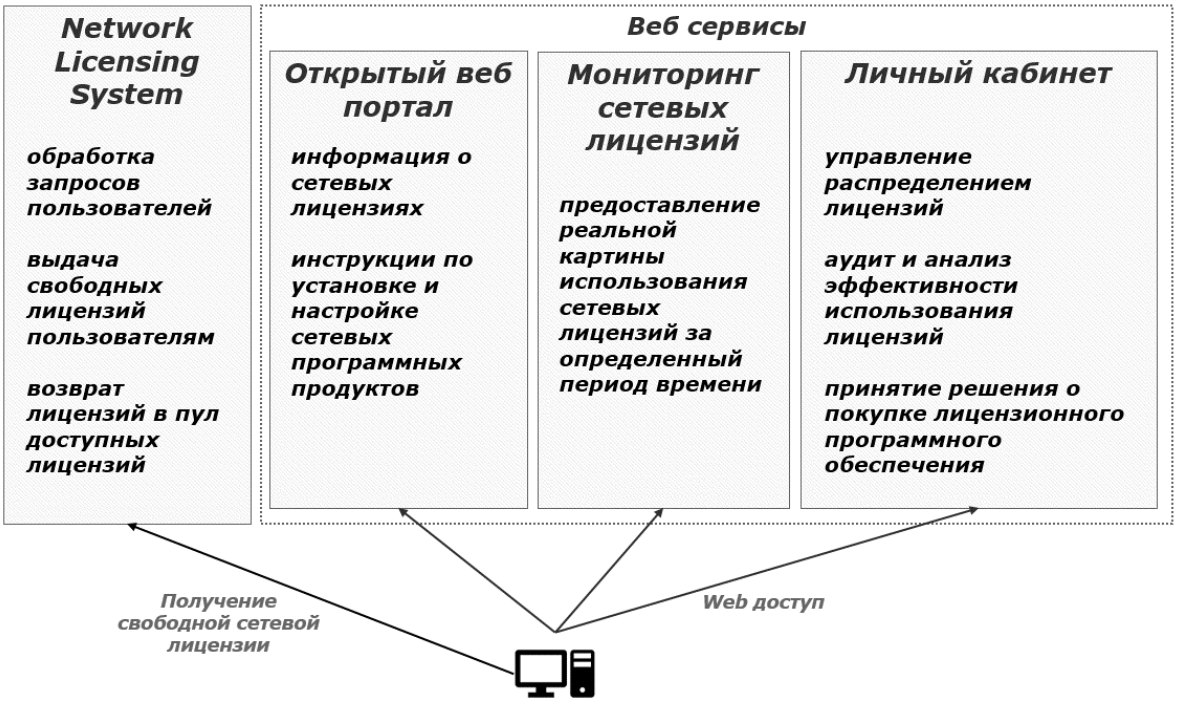


Рис. 2. Логическое представление ИССЛ.

Система распределения сетевых лицензий является отдельной подсистемой ИССЛ. Она состоит из отдельного сегмента сети с частными сетевыми адресами и системы доступа к ним, основанной на общем входе в пул виртуальных машин и передаче запроса на использование лицензии соответствующей виртуальной машине через указанные в запросе порты. Каждая виртуальная машина содержит менеджер соответствующего лицензионного программного продукта (ЛПП) и анализатор журналов менеджера для получения информации о контроле использования ЛПП. При запуске программного пакета на компьютере пользователя происходит обращение к серверу лицензий. Сервер перенаправляет запрос на соответствующую виртуальную машину с необходимым менеджером лицензий и при наличии свободной лицензии предоставляет ее пользователю. Соответственно, по окончании использования программного пакета лицензия возвращается в пул доступных лицензий. Менеджер записывает в файл журнала все действия, связанные с предоставлением и освобождением лицен-

зий, включая время занятия и освобождения лицензии, а также имя пользователя, запросившего лицензию. Эта информация обрабатывается и вносится в базу данных Influx (InfluxDB).

Веб-портал [2] содержит актуальные данные об имеющихся в лаборатории лицензиях на программное обеспечение и условиях их использования.

Мониторинг информации об использовании сетевой лицензии основан на данных, хранящихся в InfluxDB. Эти данные используются непосредственно сервером Grafana [3] и его плагином, реализованным в ИССЛ. Мониторинг использования сетевых лицензий дает реальную картину использования сетевых лицензий за определенный период времени: количество занятых и доступных лицензий, конкретных пользователей, использующих те или иные лицензии, и т. д. Вид страницы мониторинга представлен на рис. 3. График на рисунке показывает количество сетевых лицензий, которые пользователи используют для семейства Maple. В данном случае Maple2021 и Maple2022 – это отдельные программные продукты, с отдельными лицензиями для каждого из них. Страницу мониторинга можно использовать как отдельную веб-страницу или как содержимое вкладки личного кабинета любого пользователя.

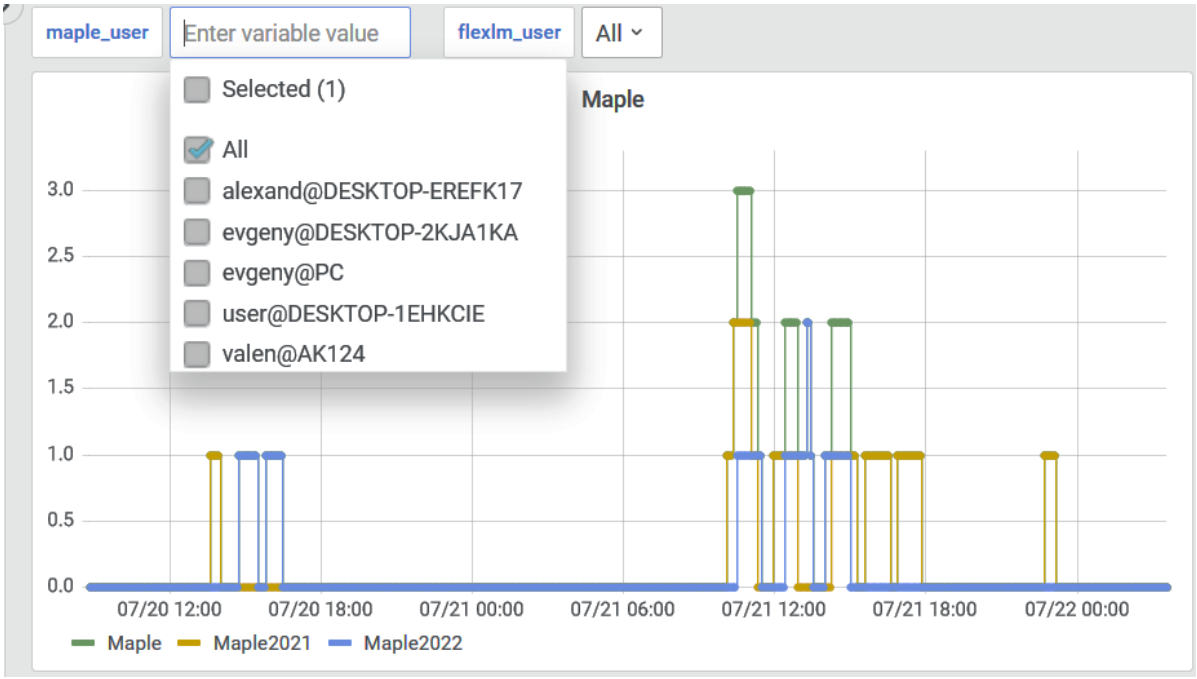


Рис. 3. Страница мониторинга сетевых лицензий.

Приватная часть веб-сервисов представляет собой личный кабинет пользователя, позволяющий различным категориям пользователей управлять распространением и покупкой лицензий, а также вносить изменения в базу данных.

2. ЛИЧНЫЙ КАБИНЕТ ПОЛЬЗОВАТЕЛЯ

Основное назначение личного кабинета пользователя заключается в следующем:

- управление распределением лицензий;
- аудит и анализ эффективности использования лицензий;
- на основе анализа информации принятие решения о покупке лицензионного программного обеспечения.

Важными частями системы являются веб-службы и связанная с ними база данных, реализованная в PostgreSQL. Эта информация используется для формирования личного кабинета пользователя.

Вся информация в базе данных структурирована и представлена соответствующими связанными сущностями (объектами): это данные о лицензиях и их использовании; программных продуктах и компаниях, их производящих; поставщиках, предоставляющих лицензии на программное обеспечение (провайдеры); компьютерах, на которых установлено соответствующее программное обеспечение, и их видах, данных, содержащие информацию о пользователях, их ролях и регистрации в системе; данные для функционирования подсистемы «Справочники»; данные, необходимые для интеграции с другими информационными сервисами ОИЯИ, в частности с системой электронного документооборота (СЭД) «Дубна» [4]. В базе данных ядро ИССЛ реализовано в виде 19 таблиц (подробнее см. в [1]).

Веб-интерфейс ИССЛ реализован в среде разработки СЭД «Дубна» с использованием всех возможностей инструментария Web Application Lego Toolkit (WALT) [5].

Веб-интерфейс выполнен в виде набора личных кабинетов пользователей, которые различаются в зависимости от ролей пользователей. Система единого входа (SSO) ОИЯИ используется для аутентификации пользователя и его роли при входе в личный кабинет ИССЛ. В нем имеется набор вкладок вверху веб-страницы, и любая вкладка может иметь набор подвкладок соответствующей

вкладки. Вид личного кабинета зависит от роли пользователя (см. рис. 4). Личный кабинет для роли «пользователь» показан на рисунке вверху, а для роли «администратор» – внизу.

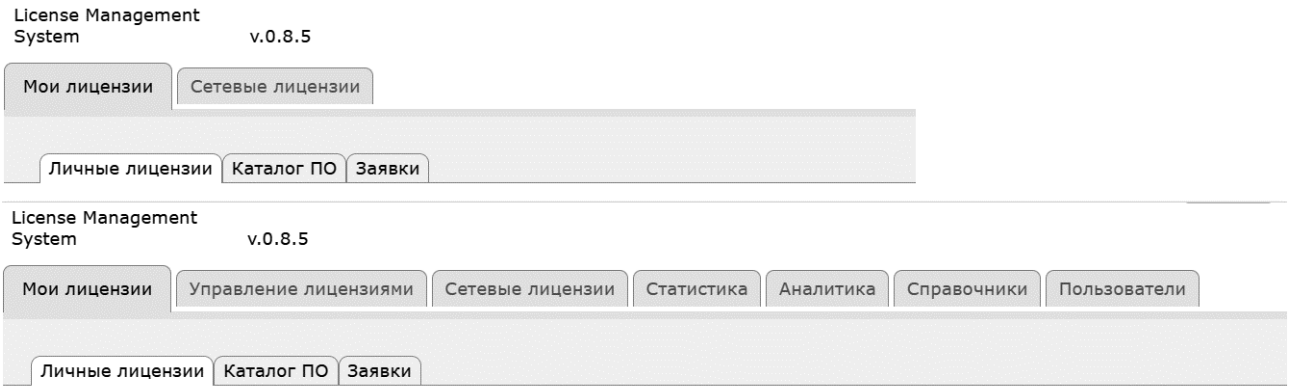


Рис. 4. Вид личного кабинета для пользователей с различными ролями.

ИССЛ определяет различные роли в зависимости от функциональных требований. Роль «пользователь» по умолчанию назначается всем пользователям ИССЛ. Другие роли должны быть назначены явно. Пользователь с ролью «пользователь» может видеть информацию обо всех публичных лицензиях, любую информацию о собственных лицензиях и сроке их действия. Пользователь может запросить новую лицензию из списка существующих лицензий или продлить старую. Может также запросить добавление новой лицензии в каталог доступного программного обеспечения для приобретения лицензий. Для внесения изменений в базу данных необходима одна из нескольких ролей: оператор, аудитор, специалист, администратор и пр. Так, например, оператор (и некоторые другие роли) имеет все права пользователя и может дополнительно внести в БД все данные, касающиеся лицензий, их владельцев, а также другую информацию, которая должна быть представлена в базе данных ИССЛ.

3. ЭКСПЛУАТАЦИЯ ИССЛ И ДАЛЬНЕЙШАЯ МОДЕРНИЗАЦИЯ

После ввода ИССЛ в эксплуатацию все данные по лицензионным пакетам ЛИТ, которые были приобретены и назначены пользователям начиная с 2018 г., внесены в базу данных и верифицированы. На текущий момент в системе содержится следующая информация:

- 114 пользователей ЛИТ, каждому из которых была назначена хотя бы одна лицензия;
- 54 позиции в каталоге программного обеспечения;
- 128 закупленных лицензий (разных типов, различного ПО, для разного количества пользователей и т. п.).

При распространении ИССЛ на другие подразделения ОИЯИ общий объем информации за эти годы оценочно возрастет в 12–17 раз.

ИССЛ предоставляет статистическую информацию об эффективности использования купленных лицензий. Реализована первая версия статистики. Интерфейс позволяет фильтровать выдаваемую информацию по различным параметрам (по типу ПО, сроку действия лицензии и т. п.). Он также обеспечивает различную степень детализации информации.

Из СЭД «Дубна» импортирован механизм согласования запросов (workflow). Этот модуль был адаптирован под архитектуру ИССЛ. В данной реализации представлены все ключевые возможности согласования заявок из СЭД «Дубна» (подробнее см. в [4]). На основе этого механизма были созданы различные запросы: запрос пользователя на новые лицензии и добавление новых программных продуктов в каталог поддерживаемого ПО. Пользователи формируют эти запросы, которые затем должны быть одобрены в нескольких инстанциях (см. рис. 5). Кроме того, был реализован запрос на покупку дополнительных лицензий. Этот запрос могут создавать пользователи с ролью «специалист», «аудитор» и некоторые другие. Далее на основе данных этого запроса (в ИССЛ) создается заявка на покупку в СЭД «Дубна».

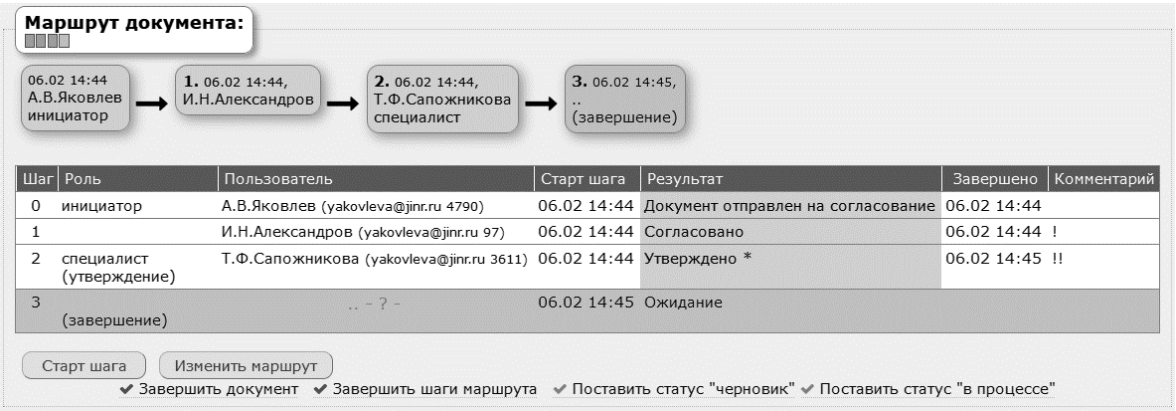


Рис. 5. Маршрут согласования заявки пользователя на новую лицензию.

При постоянном взаимодействии с операторами производятся регулярное тестирование функциональности подсистем ИССЛ и, при необходимости, доработка базы данных и интерфейса. Кроме того, ИССЛ была включена в сервисы Цифровой экосистемы ОИЯИ [6].

4. ЭФФЕКТ ОТ ВНЕДРЕНИЯ СИСТЕМЫ

Анализ информации, предоставляемой ИССЛ, позволяет повысить эффективность использования лицензий и оптимизировать затраты на их покупку.

Существует большое количество типов лицензий и различные способы их использования. Например, часть лицензий предоставляется для конкретных компьютеров. Последние могут устареть или выйти из строя, могут смениться пользователи, которым эти лицензии будут уже не нужны. Информация, предоставляемая ИССЛ, помогает отслеживать такие случаи и перераспределять лицензии внутри лаборатории.

Кроме того, в рамках статистики можно отслеживать различные, в том числе наиболее дорогостоящие лицензии, что, в свою очередь, помогает выявлять степень их полезности и интенсивность использования, а также правильно количественно компоновать заказы на новые лицензии. Например, поскольку ценообразование является достаточно гибким процессом, одномоментная закупка большего количества лицензий может быть выгоднее в пересчете на одного пользователя.

Не меньший эффект дает анализ использования сетевых лицензий, которые также относятся к наиболее дорогостоящим. Мониторинг сетевых лицензий показывает реальные данные использования лицензий: количество занятых и свободных лицензий, кто конкретно из пользователей использует определенное ПО и т. п. Это показывает избыточность или недостаточность текущего набора сетевых лицензий и позволяет корректировать их количество при последующих новых закупках или продлении срока действующих лицензий.

ЗАКЛЮЧЕНИЕ

Разработаны архитектура и объектная модель базы данных ИССЛ. Завершен дизайн веб-интерфейса системы. После ввода ИССЛ в эксплуатацию постоянно проводятся работы по ее совершенствованию и модернизации. Проведено

наполнение базы данных ИССЛ. Система предоставляет различную статистическую информацию. Реализован механизм согласования заявок. Проведены работы по интеграции ИССЛ с другими сервисами Цифровой экосистемы ОИЯИ.

ИССЛ разработана и используется в настоящее время в ЛИТ ОИЯИ. Есть запросы на использование ИССЛ в других лабораториях. При распространении ИССЛ на все подразделения Института объем информации возрастет более чем на порядок. Вся необходимая модернизация для этого в разработанной системе выполнена.

СПИСОК ЛИТЕРАТУРЫ

1. *Alexandrov E.I., Alexandrov I.N., Belyakov D.V., Davydova N.A., Kalmykova L.A., Lyubimova M.A., Sapozhnikova T.F., Syresina T.S., Yakovlev A.V., Zrelov P.V.* Development of an Information-analytical System for the Support and Maintenance of Licenses at MLIT JINR // Proceedings of the Institute for Systems Analysis Russian Academy of Sciences (ISA RAS). 2023. Vol. 73, No. 1. P. 17–23.
URL: http://www.isa.ru/proceedings/images/documents/2023-73-1/17-23.pdf

2. Сайт лицензионного программного обеспечения ЛИТ.
URL: https://soft-lit.jinr.ru/

3. Grafana портал. URL: https://grafana.com/

4. *Александров И.Н., Белякова О.В., Кореньков В.В., Куняев С.В., Печникова Л.Н., Пляшкевич М.С., Семашко С.В., Трубников Г.В., Устенко П.В., Чихалина С.Н., Яковлев А.В.* Разработка и внедрение системы электронного документооборота «СЭД Дубна» в ОИЯИ // CEUR Workshop Proceedings. 2016. Vol. 1787. P. 97–102. URL: https://ceur-ws.org/Vol-1787/97-102-paper-15.pdf

5. *Korenkov V.V., Kuniaev S.V., Semashko S.V., Sokolov I.A.* WALT platform for Web Application development // CEUR Workshop Proceedings. 2021. Vol. 3041. P. 387–392. URL: https://ceur-ws.org/Vol-3041/387-392-paper-71.pdf

6. Цифровая ЭкоСистема ОИЯИ. URL: https://digital.jinr.ru/.

CAPABILITIES OF THE DEVELOPED LICENSE SUPPORT AND MAINTENANCE INFORMATION SYSTEM

E. I. Alexandrov[1] [0000-0003-2057-5955], I. N. Alexandrov[2] [0000-0001-7381-6762], D. V. Belyakov[3] [0000-0001-6516-1479], N. A. Davydova[4] [0009-0002-3740-508X], P. V. Zrelov[5] [0000-0002-0642-3042], L. A. Kalmykova[6] [0009-0006-0307-1279], M. A. Lyubimova[7] [0000-0002-7146-7164], T. F. Sapozhnikova[8] [0009-0006-9327-1716], T. S. Syresina[9] [0009-0001-3160-921X], A. V. Yakovlev[10] [0000-0002-9534-9357]

[1–10]*Joint Institute for Nuclear Research (JINR), Dubna, Russia*

[1, 2]*Dubna State University, Dubna, Russia*

[1] aleksand@jinr.ru, [2]alexand@jinr.ru, [3]dmitry@jinr.ru, [4]luna@jinr.ru, [5]zrelov@jinr.ru, [6]lidak@jinr.ru, [7]lma@jinr.ru, [8]tsap@jinr.ru, [9]syresina@jinr.ru, [10]yakovleva@jinr.ru

Abstract

The purpose of creating the License Support and Maintenance Information System is to automate the management, acquisition, maintenance and use of licensed software products. The results of the system development over the last two years are presented. A request approval mechanism has been implemented, and various requests have been created on its basis: user requests for a new license, for adding new software products to the catalog of supported software, auditor requests for purchasing additional licenses. Work has been completed to fill the database, conveniently present information on licenses and various statistical information, as well as integrate with other services within the JINR Digital Ecosystem.

Keywords*: license management system, database, web interface.*

REFERENCES

1. *Alexandrov E.I., Alexandrov I.N., Belyakov D.V., Davydova N.A., Kalmykova L.A., Lyubimova M.A., Sapozhnikova T.F., Syresina T.S., Yakovlev A.V., Zrelov P.V.* Development of an Information-analytical System for the Support and Maintenance of Licenses at MLIT JINR // Proceedings of the Institute for Systems Analysis Russian Academy of Sciences (ISA RAS). 2023. Vol. 73, No. 1. P. 17–23.
URL: http://www.isa.ru/proceedings/images/documents/2023-73-1/17-23.pdf

2. MLIT licensed software site. URL: https://soft-lit.jinr.ru/

3. Grafana website. URL: https://grafana.com/

4. *Alexandrov I.N., Belyakova O.V., Korenkov V.V., Kuniaev S.V., Pechnikova L.N., Plyashkevich M.S., Semashko S.V., Trubnikov G.V., Ustenko P.V., Chikhalina S.N., Yakovlev A.V.* Development and implementation of electronic document management system “EDMS Dubna” at JINR // CEUR Workshop Proceedings. 2016. Vol. 1787. P. 97–102. URL: https://ceur-ws.org/Vol-1787/97-102-paper-15.pdf

5. *Korenkov V.V., Kuniaev S.V., Semashko S.V., Sokolov I.A.* WALT platform for Web Application development // Proceedings, 9th International Conference “Distributed Computing and Grid Technologies in Science and Education” (GRID’2021). 2021. Vol. 3041. P. 387–392. URL: https://ceur-ws.org/Vol-3041/387-392-paper-71.pdf

6. JINR Digital Ecosystem. URL: https://digital.jinr.ru/

СВЕДЕНИЯ ОБ АВТОРАХ



АЛЕКСАНДРОВ Евгений Игоревич – кандидат технических наук, научный сотрудник Научно-технического отдела программного и информационного обеспечения, Лаборатория информационных технологий, Объединенный институт ядерных исследований, доцент кафедры Распределенных информационно-вычислительных систем, институт системного анализа и управления (Институт САУ), университета «Дубна».

Evgeny Igorevich ALEXANDROV – Doctor of Philosophy (PhD) in Technical Sciences, researcher, department of Software and Information Support, Meshcheryakov Laboratory of Information Technologies (MLIT), Joint Institute for Nuclear Research (JINR), associate professor, Department of Distributed Information and Computing Systems, Institute of Systems Analysis and Management, Dubna University.

email: aleksand@jinr.ru

ORCID: 0000-0003-2057-5955



АЛЕКСАНДРОВ Игорь Николаевич – кандидат физико-математических наук, начальник сектора Научно-технического отдела программного и информационного обеспечения, Лаборатория информационных технологий, Объединенный институт ядерных исследований, доцент кафедры Распределенных информационно-вычислительных систем, институт системного анализа и управления (Институт САУ), университета «Дубна».

Igor Nikolaevich ALEXANDROV – Doctor of Philosophy (PhD) in Physics and Mathematics, head of sector, department of Software and Information Support, Meshcheryakov Laboratory of Information Technologies (MLIT), Joint Institute for Nuclear Research (JINR), associate professor, department of Distributed Information and Computing Systems, Institute of Systems Analysis and Management, Dubna University.

email: alexand@jinr.ru

ORCID: 0000-0001-7381-6762



БЕЛЯКОВ Дмитрий Викторович – ведущий программист Научно-технического отдела внешних коммуникаций и распределенных информационных систем, Лаборатория информационных технологий, Объединенный институт ядерных исследований.

Dmitry Viktorovich BELYAKOV – lead programmer, department of External Communications and Distributed Information Systems, Meshcheryakov Laboratory of Information Technologies (MLIT), Joint Institute for Nuclear Research (JINR).

email: dmitry@jinr.ru

ORCID: 0000-0001-6516-1479



ДАВЫДОВА Наталия Анатольевна – ведущий программист Научно-технического отдела программного и информационного обеспечения, Лаборатория информационных технологий, Объединенный институт ядерных исследований.

Natalia Anatolyevna DAVYDOVA – lead programmer, Department of Software and Information Support, Meshcheryakov Laboratory of Information Technologies (MLIT), Joint Institute for Nuclear Research (JINR).

email: luna@jinr.ru

ORCID: 0009-0002-3740-508X



ЗРЕЛОВ Петр Валентинович – кандидат физико-математических наук, начальник Научно-технического отдела программного и информационного обеспечения, Лаборатория информационных технологий, Объединенный институт ядерных исследований.

Petr Valentinovich ZRELOV – Doctor of Philosophy (PhD) in Physics and Mathematics, head of the Department of Software and Information Support, Meshcheryakov Laboratory of Information Technologies (MLIT), Joint Institute for Nuclear Research (JINR).

email: zrelov@jinr.ru

ORCID: 0000-0002-0642-3042



КАЛМЫКОВА Лидия Анатольевна – заместитель начальника Научно-технического отдела программного и информационного обеспечения, Лаборатория информационных технологий, Объединенный институт ядерных исследований.

Lidiya Anatolyevna KALMYKOVA – deputy head of the Department of Software and Information Support, Meshcheryakov Laboratory of Information Technologies (MLIT), Joint Institute for Nuclear Research (JINR).

email: lidak@jinr.ru

ORCID: 0009-0006-0307-1279



ЛЮБИМОВА Мария Александровна – инженер-программист Научно-технического отдела внешних коммуникаций и распределенных информационных систем, Лаборатория информационных технологий, Объединенный институт ядерных исследований.

Maria Alexandrovna LYUBIMOVA – software engineer, department of External Communications and Distributed Information Systems, Meshcheryakov Laboratory of Information Technologies (MLIT), Joint Institute for Nuclear Research (JINR).

email: lma@jinr.ru

ORCID: 0000-0002-7146-7164



САПОЖНИКОВА Татьяна Федоровна – кандидат физико-математических наук, начальник сектора Научно-технического отдела программного и информационного обеспечения, Лаборатория информационных технологий, Объединенный институт ядерных исследований.

Tatiana Fedorovna SAPOZHNIKOVA – Doctor of Philosophy (PhD) in Physics and Mathematics, head of sector at the Department of Software and Information Support, Meshcheryakov Laboratory of Information Technologies (MLIT), Joint Institute for Nuclear Research (JINR).

email: tsap@jinr.ru

ORCID: 0009-0006-9327-1716



СЫРЕСИНА Татьяна Серафимовна – ведущий программист Научно-технического отдела программного и информационного обеспечения, Лаборатория информационных технологий, Объединенный институт ядерных исследований.

Tatiana Serafimovna SYRESINA – lead programmer at the Department of Software and Information Support, Meshcheryakov Laboratory of Information Technologies (MLIT), Joint Institute for Nuclear Research (JINR).

email: syresina@jinr.ru

ORCID: 0009-0001-3160-921X



ЯКОВЛЕВ Александр Владимирович – научный сотрудник Научно-технического отдела программного и информационного обеспечения, Лаборатория информационных технологий, Объединенный институт ядерных исследований.

Alexander Vladimirovich YAKOVLEV – researcher at the department of Software and Information Support, Meshcheryakov Laboratory of Information Technologies (MLIT), Joint Institute for Nuclear Research (JINR).

email: yakovleva@jinr.ru

ORCID: 0000-0002-9534-9357

Материал поступил в редакцию 2 апреля 2026 года

УДК 004.65

ТРЕБОВАНИЯ К АРХИТЕКТУРЕ СИСТЕМЫ УПРАВЛЕНИЯ НАУЧНЫМИ ДАННЫМИ И ВЫБОР РЕШЕНИЙ ДЛЯ РЕАЛИЗАЦИИ ЕЕ КОМПОНЕНТОВ

К. В. Гарев[0009-0001-1961-3195]

НИЦ «Курчатовский институт» – НИИСИ, г. Москва, Россия

kv@garev.ru

Аннотация

Рассмотрена задача выбора технических решений для компонентов системы управления научными данными. Последние разделены на три категории: научные датасеты, научно-техническая и наукометрическая информации – различающиеся по структуре, объему, характеру обновления, профилю доступа и требованиям к метаданным, что усложняет эффективное управление ими в рамках единого универсального хранилища. На основе типологизации данных и сценариев использования для ключевых ролей (исследователь, аналитик, инженер данных, администратор) сформированы требования к компонентам системы. Был проведен сравнительный анализ альтернативных решений по каждому компоненту и предложена модульная архитектура, разграничивающая исходные данные и их производные представления. Определены согласованный набор технологий, совместимый с локальным развертыванием, а также сквозные механизмы идентификации, авторизации и управления ресурсами.

Ключевые слова*: управление научными данными, научные датасеты, научно-техническая информация, наукометрическая информация, программная архитектура, требования к системе, распределенное хранение, семантический поиск.*

ВВЕДЕНИЕ

Современная научно-технологическая деятельность генерирует разнородные информационные объекты, которые необходимо собирать, хранить, обрабатывать и представлять пользователям. Научные данные трактуются как широкий класс информации, формируемой в ходе исследований. В рамках настоящей

© К. В Гарев, 2026.
Данная статья распространяется на условиях международной лицензии Creative Commons License Attribution 4.0 International (CC BY 4.0).

работы они подразделены на три категории: научные датасеты, научно-техническая информация (НТИ) и наукометрическая информация (НМИ). Каждая категория отличается форматами и сценариями применения [1], а также требованиями к метаданным, идентификации и условиям повторного использования [2]. НТИ объединяет отчеты, патенты, технические описания и требует средств текстового поиска и анализа; НМИ включает показатели публикационной активности и цитирования, необходимые для оценки результативности. Управление научными данными целесообразно рассматривать через единый жизненный цикл, однако на уровне архитектуры этот цикл реализуется различными технологическими средствами.

В российском контексте проблема состоит не только в отсутствии репозитория научных датасетов, но и в недостатке комплексных решений для всей совокупности научных данных. Существующие системы охватывают лишь отдельные аспекты и не имеют широкого междисциплинарного применения. Это обусловливает потребность в интегрированном решении, объединяющем работу с научными данными в едином информационном пространстве, что согласуется с принципами открытой науки [3]. Интеграция дает эффект сквозной связности. Возможность сопоставить датасеты с публикациями, патентами и показателями публикационной активности упрощает поиск и анализ данных и повышает точность мониторинга. Ранее сформулированные принципы проектирования системы обмена научными датасетами выявили ключевые недостатки экосистемы: отсутствие единых регламентов описания, разрозненность инфраструктуры, недостаточную межсистемную совместимость и сложность обеспечения воспроизводимости [4].

Цель настоящей работы – на основании исследования сценариев использования научных данных выработать требования к компонентам системы, провести сравнительный анализ альтернативных решений, предложить архитектуру и набор технологий, обеспечивающих масштабирование, контейнеризацию вычислений для исследователей, ключевые каналы доставки данных и инструментальную поддержку анализа с привлечением высокопроизводительных вычислительных систем [5].

1. ТИПЫ НАУЧНЫХ ДАННЫХ И ИХ ХАРАКТЕРИСТИКИ

Исходной предпосылкой проектирования системы является то, что научные данные не образуют однородного класса. Данные, как правило, различаются по структуре, объему и динамике формирования, требованиям к воспроизводимости и метаданным, а также по правовому режиму доступа и сценариям использования. Все эти различия трансформируются в архитектурные требования к хранению и обработке данных.

Приведенная ниже типологизация является необходимым шагом, связывающим свойства данных с требованиями к компонентам системы.

Научные датасеты – это структурированные и полуструктурированные массивы результатов экспериментов, наблюдений и моделирования. Характеризуются большим объемом и гетерогенностью, требуют стандартизации форматов, присвоения устойчивых идентификаторов (DOI) и долгосрочного распределенного хранения для обеспечения доступности и воспроизводимости экспериментов [2, 6]; хранение крупных табличных массивов требует специализированных методов представления и сжатия [7].

Научно-техническая информация – это корпус неструктурированных текстовых данных, например отчеты НИР/ОКР, статьи, патенты, документация. Их обработка требует методов автоматического извлечения сущностей и связей из научных текстов [8], а хранение – поддержки поиска по смысловой близости наряду с полнотекстовым поиском [9].

Наукометрическая информация – это структурированные данные о научной деятельности персоны или организационной сущности: связи, цитирования, индексы, патентная активность. Для поддержки управленческих решений требуется хранить эти данные в структурированном виде для последующего агрегирования показателей и визуализации.

Различия между категориями проявляются на уровне инженерных требований и зависят от степени структурированности, объема и темпа обновления, профиля доступа, требований к метаданным и контролю качества. Следовательно, система должна проектироваться как совокупность специализированных подсистем хранения и обработки, объединенных общими механизмами идентификации, интероперабельности и управления доступом. Сопоставление

категорий научных данных позволяет определить требования к компонентам и обосновать выбор технологий для их реализации.

В табл. 1 представлены различия типов данных между собой в разрезе их характеристик, что влияет на функциональные требования к системе.

Табл. 1. Сравнительные характеристики категорий научных данных.

Характеристика	Научные датасеты	Научно-техническая информация (НТИ)	Наукометрическая информация (НМИ)
Объект управления	Набор данных исследования	Документ или корпус документов (статья, отчет, патент) и их версии	Запись о научной активности и производные показатели
Степень структурированности	Структурированные и полуструктурированные массивы	Преимущественно неструктурированный полный текст и структурированные метаданные	Структурированные таблицы, отношения и графы взаимосвязей
Типичные масштабы объема	От МБ до ТБ и выше; возможны потоковые поступления	От КБ до десятков МБ на документ; до миллионов документов	От десятков МБ до ТБ на совокупность записей и связей
Темп обновления	Пакетный или потоковый	Нерегулярный (по факту публикации)	Регулярный (пересчет метрик)
Идентификация и цитирование	Устойчивые идентификаторы (DOI), цитирование набора данных по схеме DataCite	DOI и иные устойчивые идентификаторы документов	Идентификаторы сущностей как ключи связывания
Требования к метаданным	Машиночитаемые, пригодные для поиска и повторного использования	Метаданные и полный текст; классификация и рубрикация	Метаданные сущностей и правила расчета показателей

Доминирующий профиль доступа	Преимущественно чтение и работа с большими фрагментами; вычисления рядом с данными	Поиск: частые запросы к индексу, чтение фрагментов и документов	Агрегации и отчетность: группировки, срезы, детализация
Требования к хранению	Распределенное хранение, репликация, контроль целостности, версионирование	Полные тексты и структурированный слой извлечений	Реляционное аналитическое хранилище
Требования к поиску	По метаданным и по структуре и подмножествам	Полнотекстовый и семантический поиск	По атрибутам и связям, фильтрация и иерархии
Требования к вычислениям	Контейнеризованные вычисления рядом с данными; воспроизводимые окружения	Конвейеры обработки естественного языка (извлечение, классификация)	Конвейеры агрегации и нормализации, пересчет метрик
Контроль качества	Валидация формата и целостности, контроль версий, репликация	Дедупликация, контроль версий, качество извлечений	Валидация источников, воспроизводимость расчетов

2. ПОЛЬЗОВАТЕЛЬСКИЕ РОЛИ И ТРЕБОВАНИЯ

Система обслуживает несколько категорий пользователей, для каждой из которых определены типовые задачи, определяющие функциональные и технологические требования. В табл. 2 представлены ключевые роли и характерные сценарии взаимодействия пользователей с системой.

Табл. 2. Пользовательские роли и требования к системе управления научными данными.

Роль	Сценарии взаимодействия	Функциональные требования
Исследователь	Поиск, получение и публикация научных датасетов; вычисления над данными; доступ к связанной НТИ	Поиск датасетов в каталоге по метаданным и работа с их карточками; скачивание и программный доступ к данным; запуск изолированных вычислительных сессий и расчеты «рядом с данными»; сохранение и воспроизведение результатов; поиск и доступ к связанным документам НТИ, в том числе семантический
Аналитик	Анализ агрегированных наукометрических данных и содержимого коллекций документов; выявление трендов	SQL-доступ к витринам НМИ, построение визуальных представлений данных и их анализ; выгрузка данных в среды анализа; полнотекстовый и семантический поиск по НТИ; получение структурированных извлечений и сравнение тематик; формирование подборок документов со ссылками на источники
Инженер данных	Подключение и эксплуатация контуров поступления данных; автоматизированное обновление витрин и индексов	Интеграция внешних источников (API, файлы, веб-ресурсы) и инкрементальная загрузка; нормализация и контроль качества данных; сопровождение оркестрации ETL/ELT-пайплайнов; формирование витрин НМИ и структурированных извлечений НТИ, индексация поискового контура; обеспечение воспроизводимости и наблюдаемости процессов
Администратор	Поддержка инфраструктуры и данных; управление доступом, безопасностью и устойчивостью системы	Управление учетными записями, ролями и политиками доступа, аудит действий; конфигурирование федеративной аутентификации; мониторинг сервисов и управление ресурсами и квотами; резервное копирование и восстановление; управление жизненным циклом данных и обновлениями компонентов

Исследователь работает с датасетами и связанными материалами, ведет поиск в каталоге по метаданным, просматривает карточки датасетов, описание датасета, версия, состав файлов, условия доступа, связанные публикации и документы НТИ. Исследователь получает доступ к данным согласно политикам разграничения прав и загружает данные целиком или выборочно, в том числе с привлечением программных средств для их обработки. Ему необходимы персонализированные вычислительные среды с минимальной инфраструктурной задержкой доступа к данным («рядом с данными») и возможностью воспроизвести вычисления в дальнейшем, а также необходимо осуществлять поиск релевантной НТИ, включая семантический.

Аналитик проводит анализ агрегированной НМИ и содержимого коллекций документов. Он обращается к витринам наукометрических данных через SQL, строит дашборды и отчеты в BI[1]-инструментах (агрегации по времени, организациям, направлениям), выполняет детализирующий анализ (drill-down) от агрегированных показателей к отдельным сущностям, а также выгружает подмножества данных в интерактивную среду для расширенной обработки с фиксацией даты среза и методики расчета. К задачам аналитика относится полнотекстовый и семантический поиск по НТИ с фильтрацией по метаданным и формирование структурированных извлечений для последующего анализа.

Инженер данных отвечает за процессы поступления данных и воспроизводимое обновление витрин и индексов: интеграцию внешних источников (API, выгрузки в JSON/CSV, веб-ресурсы и реестры), настройку и сопровождение конвейеров ETL/ELT, нормализацию и очистку, контроль качества загрузки, оркестрацию (автоматизированную обработку данных по расписанию и событию), формирование структурированных представлений НТИ и витрин НМИ, индексацию в поисковом контуре и поддержку наблюдаемости процессов с фиксацией версий кода и параметров расчета.

Администратор обеспечивает эксплуатацию инфраструктуры и данных: управляет учетными записями, ролями и политиками доступа, настраивает федеративную аутентификацию и интеграцию с институциональными провайде-

[1] Business intelligence (BI) — это комплекс методов и компьютерных инструментов для преобразования данных в наглядные отчёты и дашборды, помогающие принимать управленческие решения.

рами идентификации, ведет аудит действий, обеспечивает мониторинг сервисов, резервное копирование и восстановление (базы данных, объектное хранилище, поисковые индексы), управление вычислительными ресурсами и жизненным циклом данных, а также обновление компонентов без нарушения доступности.

3. КОМПОНЕНТЫ СИСТЕМЫ, ИХ ФУНКЦИИ И РЕШЕНИЯ

Для реализации разнообразных запросов на основе рассмотренных выше пользовательских ролей и сценариев взаимодействия с системой предлагается модульная архитектура системы, объединяющая следующие контуры: доступа и управления, хранения и каталогизации, обработки, вычислений и пользовательской аналитики, эксплуатации и устойчивости. Такое разделение отражает различия в природе управляемых объектов и позволяет масштабировать и развивать подсистемы независимо при сохранении единой модели идентификации и управления доступом. Для каждого компонента далее приводятся функция, ключевые требования, проанализированные альтернативы и обоснование выбора.

3.1. Контур доступа и управления

3.1.1. Пользовательский портал и прикладной слой. Обеспечивают управление жизненным циклом объектов и поддерживают единый интерфейс поиска и навигации по объектам всех типов в соответствии с правами пользователя. Из сценариев исследователей вытекают требования поиска и просмотра объектов всех типов, загрузки и публикации с фиксацией версий, формирования карточек и связей между объектами. К нефункциональным требованиям относятся ролевое разграничение доступа, воспроизводимость операций публикации, отказоустойчивость при росте числа пользователей и аудит ключевых действий.

При реализации портала были проанализированы два подхода. Первый состоит в адаптации готовой системы класса каталога данных, что обеспечивает быстрый старт, но требует существенной доработки для поддержки НТИ и НМИ. Второй подход предполагает разработку портала и сервисного API как самостоятельного слоя поверх специализированных подсистем, что обеспечивает более высокую гибкость системы при повышении стоимости разработки. Наиболее предпочтительным представляется гибридный вариант, в котором готовая система выступает как компонент интерфейса и каталога, а критичные функции

управления жизненным циклом, связностью, политиками доступа и аудитом реализуются в отдельном прикладном сервисе.

3.1.2. Сервисный API. Предоставляет программный доступ к операциям управления и потребления данных, включая загрузку и выгрузку объектов, доступ к метаданным, выполнение поисковых запросов и выборку структурированных данных, а также обеспечивает интеграцию с внешними системами и автоматизацию пользовательских сценариев.

3.1.3. Шлюз доступа. Обеспечивает пользователю единый вход в систему, осуществляет маршрутизацию запросов к порталу и API, транспортное шифрование, ограничение частоты и объема запросов, централизованный аудит. При этом требуются совместимость с подсистемой идентификации, а также отказоустойчивость и горизонтальное масштабирование. Был проведен сравнительный анализ решений NGINX[2], Traefik [3]Proxy и Kong Gateway[4], в результате которого определен NGINX как наиболее развитое и предсказуемое решение в условиях высокой нагрузки и с точки зрения требований к безопасности, представляющее собой обратный прокси с развитыми механизмами управления трафиком и устойчивой практикой эксплуатации, тогда как Traefik демонстрирует менее предсказуемое поведение под нагрузкой, а Kong избыточен при отсутствии потребности в расширенном управлении жизненным циклом API.

Основные риски, такие как единая точка отказа, деградация при крупной выдаче, обход контрольной точки, снижаются отказоустойчивой конфигурацией с балансировкой, разнесением контуров доставки и сетевой сегментацией.

3.1.4. Идентификация, аутентификация и авторизация (Authentication and authorization infrastructure – AAI). Реализует единое управление учетными записями и сквозными политиками доступа к порталу и API, объектам хранилищ, поисковым индексам и витринам с поддержкой федеративной аутентификации и ролевого разграничения. Требуются многоролевая модель, режимы доступа: открытый, ограниченный, по запросу и отсроченная публикация, аудит решений об авторизации и совместимость со стандартами OpenID Connect и SAML. Был

[2] https://nginx.org/en/docs/
[3] https://doc.traefik.io/traefik/
[4] https://docs.konghq.com/gateway/

проведен сравнительный анализ решений Keycloak[5], WSO2 Identity Server[6], Gluu Server[7], Shibboleth[8], SimpleSAMLphp[9] и Dex[10]. В итоге анализа был выбран Keycloak, поскольку он сбалансированно покрывает требования за счет поддержки стандартных протоколов, функциональной модели ролей и групп, развитых средств администрирования и широкого применения в контейнерных средах, тогда как Shibboleth и SimpleSAMLphp требуют дополнительного слоя для сквозных ролевых политик, Dex ограничен в управлении пользователями, а WSO2 и Gluu избыточно сложны в сопровождении. Основные риски связаны с критичностью компонента для доступности, ошибками конфигурации политик и рассогласованием атрибутов при федерации; они снижаются отказоустойчивым развертыванием, формализацией ролевой модели с аудитом изменений и правилами сопоставления атрибутов.

3.2. Контур хранения и каталогизации

3.2.1. Распределенное объектное хранилище. Предназначено для приема, хранения и выдачи больших массивов экспериментальных и расчетных данных и оригиналов документов НТИ. Требуются горизонтальное масштабирование по объему и пропускной способности, репликация и контроль целостности, версионирование, управление жизненным циклом и стандартизированные интерфейсы массовой выгрузки и программного доступа, поддерживающие сценарии вычислений рядом с данными. Был проведен сравнительный анализ решений Ceph[11], MinIO[12] и Hadoop HDFS[13], по результатам которого был выбран MinIO, поскольку объектная модель обеспечивает однозначную адресацию, удобство масштабирования и эффективную массовую выдачу через прикладной слой; по сравнению с Ceph он проще в эксплуатации, а по сравнению с Hadoop HDFS уни-

[5] https://www.keycloak.org/documentation
[6] https://is.docs.wso2.com/
[7] https://docs.gluu.org/
[8] https://shibboleth.atlassian.net/wiki/spaces/IDP5/overview
[9] https://simplesamlphp.org/docs/
[10] https://dexidp.io/docs/
[11] https://docs.ceph.com/
[12] https://min.io/docs/minio/linux/index.html
[13] https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-hdfs/HdfsUserGuide.html

версальнее в интеграции с порталом, программным интерфейсом и вычислительными окружениями и не требует промежуточных слоев для внешней выдачи. Основные риски связаны с деградацией при множестве мелких объектов, обеспечением отказоустойчивости и межплощадочной репликацией; они снижаются регламентами упаковки данных, версионированием, отказоустойчивой топологией и политиками жизненного цикла.

3.2.2. Реляционное хранилище метаданных и связей. Хранит метаданные объектов, их связи и этапы жизненного цикла, идентификаторы, авторство, версии, происхождение, правила доступа, ссылки на физическое размещение. Выступает системообразующим слоем, связывающим контуры хранения, индексирования и аналитики. Требуются строгие ограничения целостности, транзакционность, предсказуемое поведение при конкурентной записи, развитые механизмы обновления миграции схемы, экспорт метаданных по стандартным протоколам и устойчивая идентификация опубликованных объектов [6]. Был проведен сравнительный анализ решений PostgreSQL[14], MariaDB[15], MySQL[16] и CockroachDB[17]. В результате анализа был выбран PostgreSQL, поскольку он в наибольшей степени покрывает требования к реляционному хранилищу и поддерживает гибкое хранение полуструктурированных атрибутов без отказа от реляционной модели, что важно для итеративно развивающейся схемы метаданных. CockroachDB избыточен на базовом этапе, когда распределенность достигается на уровне сервисов. Риски: критичность контура, рост сложности графа связей, неконтролируемый рост полуструктурированных полей – снижаются отказоустойчивым развертыванием, индексированием и выделенными представлениями, фиксацией обязательного набора полей и управляемыми миграциями.

3.2.3. Аналитическое хранилище структурированных представлений данных. Обеспечивает хранение структурированных слоев НТИ, включающих извлеченные сущности, атрибуты и связи, а также слоев НМИ, включающих таблицы фактов и измерений, метрики, агрегаты, параметры расчета и дату среза, для SQL-доступа и отчетности. Требуются стабильный SQL-доступ, регламентное

[14] https://www.postgresql.org/docs/
[15] https://mariadb.com/kb/en/documentation/
[16] https://dev.mysql.com/doc/
[17] https://www.cockroachlabs.com/docs/

обновление витрин, фиксация параметров расчета и предсказуемая производительность при большом числе параллельных агрегирующих чтений. Проведен сравнительный анализ решений PostgreSQL, ClickHouse[18] и Greenplum[19]. С учетом ориентации на локальное развертывание на этапе моделирования системы был выбран PostgreSQL, поскольку он минимизирует число классов СУБД, упрощает администрирование и обеспечивает строгую схемную модель витрин. Колоночные и MPP-решения дают более высокую эффективность при росте объема и интенсивности запросов, однако повышают операционную сложность и оправданы на этапах масштабирования системы, что соответствует принципу специализации хранилищ под профиль нагрузки [10]. Основные риски связаны с обеспечением производительности при росте витрин, расхождением с транзакционным контуром и усложнением схемы; они снижаются денормализованными и материализованными представлениями, индексированием по измерениям, фиксацией даты среза и логическим разделением схем.

3.2.4. Индекс полнотекстового и семантического поиска. Обеспечивает полнотекстовый поиск по корпусу НТИ, метаданным и описаниям объектов с фильтрацией по атрибутам и выдачей фрагментов, а также хранение векторных представлений фрагментов для поиска по смысловой близости; в связке реализуется гибридное извлечение для семантического поиска и поисково-генеративных сценариев [11]. Требуются согласованная привязка результатов к структурированному слою через идентификаторы объектов, поддержка запросов ближайших соседей и гибридного ранжирования лексической [12] и семантической [9] релевантности, масштабируемость, предсказуемая латентность и резервное копирование индексов. Было выполнено сравнение решений OpenSearch[20], Apache Solr[21] в сочетании с отдельным векторным индексом и Vespa[22]. В итоге был выбран OpenSearch, поскольку он реализует полнотекстовый и семантический режимы в единой индексной модели и общей инфраструктуре резервирования, мониторинга и масштабирования; Apache Solr требует отдельного векторного слоя с синхронизацией двух индексов, а Vespa концептуально подходит, но

[18] https://clickhouse.com/docs
[19] https://greenplum.org/
[20] https://opensearch.org/docs/
[21] https://solr.apache.org/guide/
[22] https://docs.vespa.ai/

сложнее во внедрении для базовой итерации. Основные риски связаны с деградацией при росте индекса, расхождением индекса и метаданных, восстановлением после сбоев и качеством семантического поиска; они снижаются проектированием индексной схемы и регламентами переиндексации, фиксацией единого идентификатора объекта, снимками индексов в объектное хранилище и формализацией конвейера подготовки фрагментов с версионированием эмбеддингов.

3.3. Контур обработки и обновления данных

Этот контур реализует воспроизводимое исполнение процедур загрузки, преобразования и публикации данных для всех доменов. Он включает оркестрацию конвейеров (ETL/ELT, расписания, зависимости, повторные попытки, журналирование), извлечение и нормализацию (подключение источников, сопоставление схем, дедупликация, фиксация происхождения), обработку НТИ (извлечение текста, разбиение на фрагменты, выделение сущностей и связей, тематическая классификация с направлением оригиналов в объектное хранилище, полного текста и фрагментов – в поисковый контур, структурированных извлечений – в аналитическое хранилище) и обработку НМИ (унификация и пересчет показателей, формирование таблиц фактов и измерений и витрин с фиксацией источников, даты среза и параметров расчета). Поскольку домены проходят схожие стадии жизненного цикла, целесообразен единый оркестратор; при этом доменно-специфическая логика реализуется как отдельные задачи, объединяемые в формальные конвейеры.

Из сценариев инженера данных, аналитика и администратора следуют требования управления зависимостями и регламентами обновления, запуска по расписанию и событию, повторных попыток, фиксации параметров и результатов, централизованной наблюдаемости и масштабирования ресурсоемких этапов. Проведен сравнительный анализ решений Apache Airflow[23], Argo Workflows[24] и Prefect[25]. В результате был выделен Apache Airflow, поскольку раз-

[23] https://airflow.apache.org/docs/
[24] https://argo-workflows.readthedocs.io/en/latest/
[25] https://docs.prefect.io/

витая модель направленных ациклических графов и обширная экосистема операторов соответствуют доминированию регламентных процессов и большому числу зависимостей и расписаний; Argo Workflows ориентирован на контейнерные вычислительные цепочки, а приоритетом Prefect выступает удобство разработки. Основные риски связаны с критичностью оркестратора, каскадными сбоями из-за некорректных зависимостей и блокировкой регламентных процессов ресурсоемкими задачами; они снижаются отказоустойчивым развертыванием и мониторингом, стандартизацией шаблонов конвейеров и разделением доменных потоков, а также разделением очередей и квотированием ресурсов.

3.4. Контур вычислений и пользовательской аналитики

3.4.1. Интерактивные вычислительные рабочие места. Предоставляют изолированные вычислительные среды для анализа данных и воспроизводимых исследований с доступом к объектному хранилищу, витринам и поисковым индексам. Требуются выполнение кода с наследованием политик доступа через AAI, управление ресурсами, включая квоты на процессорное время, память и дисковое пространство, предсказуемость при одновременной работе множества пользователей, согласованность окружений и подключение к данным без обязательного полного копирования. Было выполнено сравнение решений JupyterHub[26], Kubeflow Notebooks[27] и Apache Zeppelin[28], а также варианта прямого доступа к вычислительным узлам. В итоге был выбран JupyterHub с развертыванием в контейнерной инфраструктуре, поскольку он обеспечивает многопользовательскую модель, изоляцию сессий, квотирование и единую аутентификацию; Kubeflow Notebooks ориентирован на контуры машинного обучения и повышает сложность, Apache Zeppelin менее универсален, а прямой доступ снижает воспроизводимость и контроль. Основные риски связаны с конкуренцией за ресурсы, разнообразием окружений и безопасностью доступа к данным; они снижаются разделением и квотированием ресурсов, стандартными профилями окружений, интеграцией с политиками доступа, сетевой сегментацией и аудитом.

[26] https://jupyterhub.readthedocs.io/en/stable/
[27] https://www.kubeflow.org/docs/components/notebooks/
[28] https://zeppelin.apache.org/docs/

3.4.2. SQL-слой доступа. Обеспечивает унифицированный SQL-доступ к витринам и, при необходимости, к данным объектного хранилища, позволяет использовать стандартные инструменты SQL/BI без прямого подключения к внутренним хранилищам. Требуются стабильный доступ к витринам НМИ и структурированному слою НТИ, поддержка параллельных запросов BI-нагрузки, управление доступом по ролевой модели и согласованность определений витрин. Был проведен сравнительный анализ решений Dremio[29], Trino[30] и PrestoDB[31]. Был выбран Dremio на базе открытого ядра dremio-oss, поскольку он дополняет SQL-движок слоем каталогизации и публикации представлений, что соответствует требованиям самообслуживания аналитиков; Trino и PrestoDB не задают прикладной модели публикации наборов данных. Фиксируется ограничение использования открытой кодовой базы; резервным сценарием замещения служит Trino. Основные риски связаны с различием открытой базы и условий эксплуатации дистрибутивов и с ограничением набора коннекторов; они снижаются фиксацией варианта поставки и предварительной валидацией требуемых источников.

3.4.3. BI-доступ и отчетность. Обеспечивает подключение BI-инструментов к витринам и реализует сценарии отчетности с управляемой публикацией дашбордов и режимом самообслуживания. В результате сравнительного анализа решений DataLens[32], Apache Superset[33] и Metabase[34] был отобран DataLens в открытой реализации, поскольку он изначально ориентирован на построение дашбордов и работу с типовыми аналитическими источниками, включая PostgreSQL и ClickHouse. Риск ограниченности коннекторов и функциональности открытой версии снижается фиксацией перечня поддерживаемых источников и использованием SQL-слоя как стабильного интерфейса, позволяющего заменить BI-инструмент без изменения контуров хранения и обработки.

3.5. Контур эксплуатации и обеспечения устойчивости

[29] https://docs.dremio.com/
[30] https://trino.io/docs/current/
[31] https://prestodb.io/docs/current/
[32] https://datalens.ru/opensource/docs
[33] https://superset.apache.org/docs/intro
[34] https://www.metabase.com/docs/

3.5.1. Наблюдаемость. Для контроля состояния сервисов была определена связка Prometheus[35] и Grafana[36] для сбора и визуализации эксплуатационных метрик, дополненная OpenSearch для централизованного хранения и анализа логов, что охватывает как метрики, так и событийные данные и сохраняет единство технологического стека за счет повторного использования поискового компонента. Основные риски связаны с неполнотой телеметрии, ростом объема логов и отказом самого контура наблюдаемости; они снижаются стандартизацией форматов и корреляционных идентификаторов запросов, политиками хранения и ротации и отказоустойчивым развертыванием.

3.5.2. Контейнеризация и управление ресурсами. Обеспечивает изоляцию и масштабирование сервисов и вычислительных окружений с логическим разделением ресурсов, квотированием и сетевыми политиками. Требуются среда исполнения, горизонтальное масштабирование, квотирование, сегментация ресурсов между системными сервисами и пользовательскими вычислениями, сетевая изоляция и воспроизводимость окружений. Был проведен сравнительный анализ решений Kubernetes[37], OpenShift[38] и Nomad[39], в результате был выделен Kubernetes. Как фактический стандарт оркестрации он предоставляет модель планирования, отдельные пулы узлов для системных сервисов и пользовательских сессий, квоты и лимиты и сетевые политики [13]. При этом OpenShift увеличивает сложность, а Nomad не обеспечивает требуемого уровня поддержки сложных многосервисных систем данных. Основные риски связаны с конкуренцией за ресурсы, сложностью эксплуатации и ошибками политик безопасности, обходом контрольных точек и деградацией при обновлениях; они снижаются выделением пулов узлов, квотами, стандартизацией профилей развертывания, сетевой сегментацией и поэтапными обновлениями.

4. ИТОГОВЫЙ КОМПОНЕНТНЫЙ СОСТАВ

[35] https://prometheus.io/docs/
[36] https://grafana.com/docs/
[37] https://kubernetes.io/docs/
[38] https://docs.openshift.com/
[39] https://developer.hashicorp.com/nomad/docs

Предложенная архитектура системы объединяет хранение и обслуживание трех классов объектов: научных датасетов, НТИ и НМИ – по многоуровневому принципу, где каждый контур решает определенный класс задач, а согласованность достигается единой моделью идентификации, централизованным управлением доступом и воспроизводимыми процессами обработки.

Контур доступа и управления включает портал и сервисный API, шлюз доступа и подсистему на базе Keycloak, обеспечивающую единый вход и ролевую модель прав для всех сервисов.

Контур хранения и каталогизации объединяет объектное хранилище MinIO (датасеты и оригиналы), реляционное хранилище метаданных и связей на базе PostgreSQL как системообразующий слой, аналитические представления НМИ и структурированный слой НТИ в PostgreSQL с SQL-доступом, а также поисковый контур OpenSearch для полнотекстового и семантического извлечения. Архитектурные решения предусматривают хранение оригиналов, транзакционную модель метаданных, аналитические витрины и поисковые представления, связывая их едиными идентификаторами объектов и регламентами обновления. Контур обработки реализуется единым оркестратором Apache Airflow, управляющим обновлением витрин НМИ, формированием представлений НТИ и переиндексацией поиска с фиксацией статусов и параметров запусков.

Контур вычислений и аналитики включает решения JupyterHub (интерактивные рабочие места), SQL-слой Dremio (унифицированный доступ к витринам) и BI-инструмент DataLens (визуальная аналитика и отчетность), опирающиеся на единый SQL-контракт витрин. Контур эксплуатации обеспечивается наблюдаемостью на базе Prometheus и Grafana и анализом логов в OpenSearch, комбинированной стратегией резервного копирования источников истины и производных контуров, а также контейнеризацией и управлением ресурсами на базе Kubernetes с разделением системных и пользовательских нагрузок. В совокупности компоненты взаимодействуют через стандартизированные интерфейсы, что покрывает полный спектр сценариев работы с научными данными и допускает поэтапную эволюцию системы без пересмотра ключевых контрактов доступа и модели данных.

ЗАКЛЮЧЕНИЕ

Представлена интегрированная система управления научными данными, объединяющая научные датасеты, научно-техническую и наукометрическую информации. Предложены технологические решения ее базовых компонентов. Показано, что типы обрабатываемых данных различаются по структурированности, масштабам, профилю доступа и требованиям к метаданным и воспроизводимости, что исключает эффективное управление ими в рамках единого универсального хранилища или единого класса инструментов [1, 10]. Предложенное решение относится к классу полихранилищ, в которых запрос обслуживается несколькими специализированными подсистемами при едином формате доступа [14]. На основе сценариев использования ключевых ролей определены функциональные и нефункциональные требования к компонентам, включая масштабирование, безопасность, воспроизводимость обработки, поддержку интерактивных вычислений и предоставление данных через разные каналы потребления.

Определен состав модулей системы, объединенных архитектурой, поддерживающей разграничение типов данных и их производных представлений. Объектное хранилище используется для датасетов и оригиналов, реляционный контур метаданных – для целостности, версий и связности, аналитические представления – для SQL-доступа, поисковый контур – для полнотекстового и семантического извлечения; сквозными механизмами интеграции выступают единый прикладной слой управления жизненным циклом объектов, централизованная аутентификация и авторизация и единый оркестратор конвейеров обработки. На основе анализа альтернатив сформирован набор решений с возможностью локального развертывания, в частности MinIO, PostgreSQL, OpenSearch, Keycloak, Apache Airflow, JupyterHub, Dremio и DataLens, шлюз доступа NGINX, а также Prometheus, Grafana и Kubernetes для эксплуатации.

Предложенная архитектура представляет воспроизводимый шаблон проектирования системы управления научными данными, допускающий поэтапное развитие без нарушения режимов доступа к данным и управления объектами данных.

СПИСОК ЛИТЕРАТУРЫ

1. Borgman C.L. Big Data, Little Data, No Data: Scholarship in the Networked World. Cambridge, MA: MIT Press, 2015. 416 p. ISBN 978-0-262-02856-1

2. Wilkinson M.D., Dumontier M., Aalbersberg I.J. et al. The FAIR Guiding Principles for scientific data management and stewardship // Scientific Data. 2016. Vol. 3. Article 160018. https://doi.org/10.1038/sdata.2016.18

3. UNESCO Recommendation on Open Science. Paris: UNESCO, 2021. 36 p. https://doi.org/10.5281/zenodo.5741832

4. Гарев К.В. Определение принципов проектирования отечественной платформы обмена научными датасетами // Программные продукты и системы. 2025. Т. 38, № 1. С. 197–206. https://doi.org/10.15827/0236-235X.149.197-206

5. Савин Г.И., Шабанов Б.М., Баранов А.В. и др. Об использовании федеральной научной телекоммуникационной инфраструктуры для суперкомпьютерных вычислений // Вестник Южно-Уральского государственного университета. Серия: Вычислительная математика и информатика. 2020. Т. 9, № 1. С. 20–35. https://doi.org/10.14529/cmse200102

6. DataCite Metadata Working Group. DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs. Version 4.4. DataCite e.V., 2021. https://doi.org/10.14454/3w3z-sa82

7. Гарев К.В. Сравнительный анализ методов сжатия табличных данных // Программные продукты и системы. 2024. Т. 37, № 2. С. 170–177. https://doi.org/10.15827/0236-235X.142.170-177

8. Nasar Z., Jaffry S.W., Malik M.K. Information extraction from scientific articles: a survey // Scientometrics. 2018. Vol. 117, No. 3. P. 1931–1990. https://doi.org/10.1007/s11192-018-2921-5

9. Reimers N., Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Hong Kong, 2019. P. 3982–3992. https://doi.org/10.18653/v1/D19-1410

10. Stonebraker M., Çetintemel U. «One Size Fits All»: An Idea Whose Time Has Come and Gone // Proceedings of the 21st International Conference on Data Engineering (ICDE'05). Tokyo: IEEE Computer Society, 2005. P. 2–11.

https://doi.org/10.1109/ICDE.2005.1

11. Lewis P., Perez E., Piktus A. et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks // Advances in Neural Information Processing Systems. 2020. Vol. 33. P. 9459–9474. https://doi.org/10.48550/arXiv.2005.11401

12. Robertson S., Zaragoza H. The Probabilistic Relevance Framework: BM25 and Beyond // Foundations and Trends in Information Retrieval. 2009. Vol. 3, No. 4. P. 333–389. https://doi.org/10.1561/1500000019

13. Burns B., Grant B., Oppenheimer D. et al. Borg, Omega, and Kubernetes // Communications of the ACM. 2016. Vol. 59, No. 5. P. 50–57. https://doi.org/10.1145/2890784

14. Duggan J., Elmore A.J., Stonebraker M. et al. The BigDAWG Polystore System // ACM SIGMOD Record. 2015. Vol. 44, No. 2. P. 11–16. https://doi.org/10.1145/2814710.2814713

REQUIREMENTS AND TECHNICAL SOLUTIONS FOR A SCIENTIFIC DATA MANAGEMENT SYSTEM ARCHITECTURE

K. V. Garev[0009-0001-1961-3195]

National Research Centre "Kurchatov Institute", Moscow, Russia

kv@garev.ru

Abstract

This paper addresses the problem of defining requirements for the architecture of a scientific data management system and selecting technical solutions for implementing its components. Scientific data are divided into three categories – research datasets, scientific and technical information (STI), and scientometric information (SMI) – which differ in structure, volume, update patterns, access profile, and metadata requirements, complicating their effective management within a single general-purpose store. Based on a data typology and use cases for the key roles (researcher, analyst, data engineer, administrator), requirements for the components are formulated. A comparative analysis of alternative solutions is performed for each component, and a modular architecture is proposed that separates sources of truth from derived representations. A consistent technology stack compatible with on-

premises deployment is determined, along with cross-cutting mechanisms of identification, authorization, and resource management.

Keywords*: scientific data management, research datasets, scientific and technical information, scientometric information, software architecture, system requirements, distributed storage, semantic search.*

REFERENCES

1. Borgman C.L. Big Data, Little Data, No Data: Scholarship in the Networked World. Cambridge, MA: MIT Press, 2015. 416 p. ISBN 978-0-262-02856-1

2. Wilkinson M.D., Dumontier M., Aalbersberg I.J. et al. The FAIR Guiding Principles for scientific data management and stewardship // Scientific Data. 2016. Vol. 3. Article 160018. https://doi.org/10.1038/sdata.2016.18

3. UNESCO Recommendation on Open Science. Paris: UNESCO, 2021. 36 p. https://doi.org/10.5281/zenodo.5741832

4. Garev K.V. Comparative analysis of tabular data compression methods // Programmnye produkty i sistemy [Software & Systems]. 2024. Vol. 37, No. 2. P. 170–177. https://doi.org/10.15827/0236-235X.142.170-177 (in Russian).

5. Savin G.I., Shabanov B.M., Baranov A.V. et al. On the use of the federal scientific telecommunication infrastructure for supercomputer computations // Vestnik YuUrGU. Ser. Vychislitelnaya matematika i informatika [Bulletin of the South Ural State University. Series: Computational Mathematics and Software Engineering]. 2020. Vol. 9, No. 1. P. 20–35. https://doi.org/10.14529/cmse200102 (in Russian)

6. DataCite Metadata Working Group. DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs. Version 4.4. DataCite e.V., 2021. https://doi.org/10.14454/3w3z-sa82

7. Garev K.V. Comparative analysis of tabular data compression methods // Programmnye produkty i sistemy [Software & Systems]. 2024. Vol. 37, No. 2. P. 170–177. https://doi.org/10.15827/0236-235X.142.170-177 (in Russian).

8. Nasar Z., Jaffry S.W., Malik M.K. Information extraction from scientific articles: a survey // Scientometrics. 2018. Vol. 117, No. 3. P. 1931–1990. https://doi.org/10.1007/s11192-018-2921-5

9. Reimers N., Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Hong Kong, 2019. P. 3982–3992. https://doi.org/10.18653/v1/D19-1410

10. Stonebraker M., Çetintemel U. «One Size Fits All»: An Idea Whose Time Has Come and Gone // Proceedings of the 21st International Conference on Data Engineering (ICDE'05). Tokyo: IEEE Computer Society, 2005. P. 2–11. https://doi.org/10.1109/ICDE.2005.1

11. Lewis P., Perez E., Piktus A. et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks // Advances in Neural Information Processing Systems. 2020. Vol. 33. P. 9459–9474. https://doi.org/10.48550/arXiv.2005.11401

12. Robertson S., Zaragoza H. The Probabilistic Relevance Framework: BM25 and Beyond // Foundations and Trends in Information Retrieval. 2009. Vol. 3, No. 4. P. 333–389. https://doi.org/10.1561/1500000019

13. Burns B., Grant B., Oppenheimer D. et al. Borg, Omega, and Kubernetes // Communications of the ACM. 2016. Vol. 59, No. 5. P. 50–57. https://doi.org/10.1145/2890784

14. Duggan J., Elmore A.J., Stonebraker M. et al. The BigDAWG Polystore System // ACM SIGMOD Record. 2015. Vol. 44, No. 2. P. 11–16. https://doi.org/10.1145/2814710.2814713

СВЕДЕНИЯ ОБ АВТОРЕ



ГАРЕВ Кирилл Валерьевич – выпускник аспирантуры Федерального государственного бюджетного учреждения «Федеральный научный центр Научно-исследовательский институт системных исследований Национального исследовательского центра «Курчатовский институт» (НИЦ «Курчатовский институт» – НИИСИ).

Kirill Valerevich GAREV – graduate of the postgraduate program at the Scientific Research Institute for System Analysis, National Research Centre "Kurchatov Institute".

email: kv@garev.ru

0009-0001-1961-3195

Материал поступил в редакцию 27 июня 2026 года

УДК 004.8

АРХИТЕКТУРА ЛИНГВИСТИЧЕСКОЙ ПЛАТФОРМЫ “TurkLang”

А. Р. Гатиатуллин[0000-0003-3063-8147]

Академия наук Республики Татарстан, г. Казань, Россия

ayrat.gatiatullin@gmail.com

Аннотация

Описана архитектура многофункциональной лингвистической платформы “TurkLang”, создаваемой в Институте прикладной семиотики Академии наук Республики Татарстан. В основе системы лежит трехуровневая база знаний: общеязыковые универсалии, описания единиц конкретных языков и эмпирические речевые данные. База знаний реализована в виде графа, где основной единицей хранения выступает морфема. Это решение обусловлено элементно-комбинаторной природой тюркских языков: словоформа представляет собой линейную цепочку, состоящую из корня и последовательно присоединенных аффиксов – модификаторов, каждый из которых выражает одну грамматическую категорию (падеж, число, время и т. д.) в строго фиксированном порядке. Рассмотрены уровни абстракции, типы связей между сущностями, механизмы интеграции с внешними источниками (электронная энциклопедия, лексикографические ресурсы, геоинформационные системы). Особое внимание уделено применению международных стандартов синтаксической, семантической и прагматической (DiAML) аннотаций. Выполнено сравнение с существующими платформами (Sketch Engine, LingvoDoc, Apertium). На сегодняшний день база знаний платформы охватывает более 30 языков и диалектов, содержит более 237 тыс. корневых морфем и около 40 тыс. правил сочетаемости, а корпусная часть насчитывает свыше 180 млн словоупотреблений.

Ключевые слова*: лингвистическая платформа, граф знаний, тюркские языки, морфемоцентричность, трехуровневая архитектура, Универсальные зависимости (Universal Dependencies), Универсальное смысловое представление (Universal Meaning Representation), агглютинативный язык, семиотическая вертикаль.*

© А. Р. Гатиатуллин, 2026.
Данная статья распространяется на условиях международной лицензии Creative Commons License Attribution 4.0 International (CC BY 4.0).

ВВЕДЕНИЕ

Обработка естественного языка для тюркских языков сопряжена с двумя фундаментальными проблемами.

Первая проблема является структурной. Тюркские языки по своей структуре можно отнести к элементно-комбинаторному типу [1] (в лингвистике их называют агглютинативными).

Для тюркских языков справедливо следующее: словоформа всегда строится по единой схеме, где каждый аффикс последовательно добавляется к основе и несет одну грамматическую характеристику. С формальной точки зрения такая морфологическая структура может быть описана регулярной грамматикой и моделируется конечным автоматом [2]. Однако большинство существующих инструментов NLP, включая синтаксические парсеры и стандарты аннотации (Universal Dependencies), разрабатывалось для индоевропейских языков и не учитывают структурно-функциональные особенности тюркских языков [3]. В результате некоторый объем информации в таких моделях теряется, что снижает качество автоматического анализа.

Вторая проблема – это ресурсная. Подавляющее большинство тюркских языков (за исключением турецкого) являются малоресурсными [4]. Проявляется это в нехватке полноценных электронных словарей, электронных корпусов с разметкой на различных языковых уровнях, лингвистических процессоров и унифицированных лингвистических стандартов. Существующие цифровые ресурсы создаются разрозненными коллективами, используют несовместимые форматы и системы тегов, что делает их интеграцию практически невозможной.

Для преодоления этих барьеров в Институте прикладной семиотики Академии наук Республики Татарстан (г. Казань) разрабатывается платформа “TurkLang”. Ее архитектура строится на трехуровневой модели данных, где центральным элементом является морфема, а не слово, как во многих известных моделях. Это позволяет:

- явно фиксировать порядок аффиксов в словоформе;
- представлять структуру словоформы, предложения и их значения в виде графа знаний;

- получать значение лингвистической единицы или конструкции путем обхода графа.

Ниже описано архитектурное решение: организация данных, уровни абстракции, интеграционные механизмы с внешними источниками, технологический стек, а также результаты практической реализации и сравнение с существующими аналогами.

1. АНАЛИЗ СУЩЕСТВУЮЩИХ РЕШЕНИЙ

Для объективной оценки архитектурных решений платформы “TurkLang” необходимо рассмотреть ее в контексте существующих лингвистических платформ, графов знаний и цифровых ресурсов, ориентированных на тюркские и другие агглютинативные языки.

1.1. Корпусные платформы

Sketch Engine [5] – одна из наиболее распространенных корпусных платформ, разработанная компанией Lexical Computing. Система предоставляет доступ к более чем 800 корпусам на 100+ языках, включая шесть тюркских языков (турецкий, казахский, узбекский, киргизский, азербайджанский, татарский).

Ключевая особенность Sketch Engine заключается в наличии богатого набора инструментов анализа корпусов: конкорданс, частотные списки, коллокации, *n*-граммы, тематические списки ключевых слов, автоматического аннотирования, создания подкорпусов. API для разработчиков предоставляет программный интерфейс для интеграции с внешними системами.

Система Sketch Engine является менеджером корпусов, а не платформой для формального описания языка: в ней отсутствуют онтологии, грамматические правила и семантические модели. Кроме того, она не фиксирует порядок морфем в словоформе, что критично для агглютинативных языков.

1.2. Платформы для документирования языков

LingvoDoc [6] – это платформа для совместного документирования языков, разработанная в Институте системного программирования РАН совместно с Институтом языкознания РАН. Система ориентирована на полевых лингвистов и исследователей, работающих с малоресурсными и исчезающими языками.

Ключевой особенностью платформы является возможность коллективной работы с языковыми материалами, поддержкой распределенной обработки и сохранением прав создателей словарей и корпусов. В платформе реализованы многослойные словари. Инструменты анализа включают модули для фонетического, морфологического и этимологического анализов.

Как и в Sketch Engine, в этой платформе отсутствует семантический слой, что ограничивает возможности глубоких компаративных исследований по содержанию, а не только по форме. Платформа ориентирована на хранение и первичный анализ полевых данных, но не предоставляет формальных грамматических моделей для автоматической обработки.

1.3. Системы машинного перевода

Apertium [7] – это открытая платформа машинного перевода, основанная на правилах (rule-based). Изначально разрабатывалась для перевода между близкородственными языками.

Ключевой особенностью платформы является использование правил на основе конвейерной обработки. Платформа реализована на базе свободного программного обеспечения под лицензией GNU GPL, не требует API-ключей или аутентификации. Каждая языковая пара разрабатывается изолированно, без общей базы знаний для всех тюркских языков. Не реализована единая онтологическая модель для тюркских языков. Популярность платформы снизилась с развитием нейросетевых технологий машинного перевода, которые превзошли по качеству системы машинного перевода типа Apertium, основанных на правилах. Однако сервисы указанной платформы могут быть использованы для решения других задач, например, в качестве механизма аугментирования данных для малоресурсных тюркских языков.

1.4. Семантические ресурсы

KeNet [8] (турецкий WordNet) является наиболее полным семантическим ресурсом для турецкого языка, созданный методом «снизу вверх» на основе толкового словаря Турецкого лингвистического общества. Содержит 76757 синсетов, имеет как внутриязыковые семантические отношения, так и связи с Princeton WordNet через межъязыковые отношения.

Анализ интернет-ресурсов показал, что из тюркских языков WordNet-ресурсы также реализованы для узбекского и уйгурского языков: UZ-WORDNET [9] и Uyghur WordNet. UZ-WORDNET создан с помощью метода автоматического извлечения концептов из английского WordNet. Других специализированных WordNet-ресурсов для тюркских языков не обнаружено. Существующие ресурсы (кроме KeNet) созданы автоматическими методами и требуют ручной верификации.

Turkish FrameNet [10] – это фреймовый ресурс для турецкого языка, охватывающий 8 доменов с наиболее частотными предикатами: Activity, Cause, Change, Motion, Cognition, Perception, Judgement, Commerce. Содержит 2700 вручную аннотированных примеров и 19221 токенов.

Uyghur FrameNet [11] является фреймовым ресурсом для уйгурского языка.

Для остальных тюркских языков фреймовые ресурсы не найдены.

Framester – это концентратор, связывающий FrameNet, WordNet, VerbNet, BabelNet [12], DBpedia, Yago и DOLCE-Zero [13]. Поддерживает OWL-рассуждения на основе фреймовой семантики, но ориентирован преимущественно на английский язык.

В последние годы активно развиваются формальные языки для представления смысла целых предложений и текстов. Одним из наиболее перспективных стандартов разметки в этой области является Uniform Meaning Representation (UMR). UMR – это графовый формализм для представления смысла, который пришел на смену стандарта Abstract Meaning Representation (AMR). В отличие от AMR, изначально разработанного для английского языка, UMR с самого начала создавался как межъязыковой стандарт, применимый к типологически разнообразным языкам, включая агглютинативные.

Ключевой особенностью UMR является то, что смысл текста в нем представляется в виде ориентированного графа, узлами которого являются концепты (действия, объекты, атрибуты), а ребрами – семантические отношения.

Пример UMR-графа для татарского предложения «*Малай урманга барды*»:

```
(e / go-01
      :ARG1 (m / малай)
      :goal (y / урман :number plural)
            :time past
)
```

В отличие от AMR, который позволяет описывать только семантику отдельных предложений, UMR позволяет представлять семантические взаимосвязи между предложениями.

В UMR появились механизмы, которые позволяют описывать значения таких категорий, как эвиденциальность (источник информации), аспектуальность, категории лица и числа, что делает UMR особенно релевантным для тюркских языков. Для семиотической вертикали, представленной в настоящей работе, особую ценность представляет возможность формального отображения между синтаксическими структурами в формате Universal Dependencies и семантическими ролями в UMR.

1.6. Синтаксические аннотации: universal dependencies

В рамках международного проекта Universal Dependencies [14] (UD) созданы древовидные банки для 11 тюркских языков. В релизе UD v2.16 (2025 г.) представлены первые полностью выровненные параллельные UD-деревья для четырех тюркских языков: азербайджанского, киргизского, турецкого и узбекского.

Ключевая проблема стандарта UD для тюркских языков заключается в том, что в нем нет возможности фиксировать порядок аффиксов внутри словоформы. Для языков с элементно-комбинаторной структурой, к числу которых относятся и тюркские языки, это важно, поскольку, например, турецкие слова:

masadakilerin = masa + LOC(da) + REL (ki) + PL(ler)+ GEN(in),

masanınkilerde = masa + GEN (nın) + REL(ki) + PL(ler)+ LOC(de)

имеют одинаковый набор грамматических признаков в стандартном UD, но разный порядок аффиксов, что задает различные смыслы и синтаксические структуры. Первое слово имеет значение «принадлежащий тем, которые находятся на столе», а второе «находящиеся у тех, которые принадлежат столу». Это приводит к тому, что появляются различные синтаксические и семантические структуры с одинаковым набором грамматических параметров.

2. БАЗОВЫЕ ПРИНЦИПЫ АРХИТЕКТУРЫ ПЛАТФОРМЫ “TurkLang”

В основе архитектуры платформы лежат четыре базовых взаимосвязанных принципа.

1. Трехуровневая организация определяет способ хранения данных.
2. Многоуровневая согласованность (семиотическая вертикаль) задает систему взаимосвязей между языковыми уровнями и определяет, какие данные и в каком формате должны быть описаны.
3. Морфемоцентричность выбирает базовую единицу хранения.
4. Параметризация обеспечивает масштабирование на все языки семьи через механизм наследования.

Вместе они определяют способ хранения, обработки и пополнения лингвистической информации.

2.1. Принцип трехуровневой организации базы знаний

База знаний платформы разделена на три логических уровня, каждый из которых отвечает за свой тип лингвистической информации (рис. 1). Связи направлены от конкретного к абстрактному (уровень 3 → уровень 2 → уровень 1), что исключает циклические зависимости и упрощает поддержку и масштабирование системы.

В основе этой организации лежит семиотическая вертикаль – модель языка как многоуровневой знаковой системы, где каждый уровень опирается на некоторый международный стандарт. Однако сама вертикаль является частью онтологического ядра (уровень 1), а не самостоятельной надстройкой над архитектурой.



Рис. 1. Трехуровневая организация базы знаний.

2.1.1. Уровень 1: Онтологическое ядро, лингвистические универсалии

Уровень 1 содержит языконезависимые сущности. Он образует онтологический каркас платформы – систему понятий и отношений, на которую ссылаются все остальные уровни. Состав уровня 1 приведен далее.

Семантическая сеть на основе WordNet – это иерархия концептов с отношениями «род-вид» (гиперонимия/гипонимия), «часть-целое» (меронимия) и другими. Исходно транслирован из Принстонского WordNet, дополнен и продолжает дополняться тюркскими культурными концептами, такими как национальные блюда, одежда, праздники, родственные отношения и др. Содержит около 60 тыс. концептов.

Библиотека ситуаций на основе FrameNet. В отличие от WordNet, который описывает статическую лексическую семантику, т. е. отношения между словами, FrameNet описывает динамические ситуации – фреймы (например, «Передача», «Движение», «Восприятие») с набором участников – семантических ролей.

Фреймом в компьютерной лингвистике является схематическое представление типовой ситуации (события, состояния, отношения). Например, фрейм «Передача» включает участников «Дающий», «Получатель» и «Передаваемый объект».

Семантические роли – участники ситуации. Они делятся на обязательные (Core) – без них фрейм не может быть полноценно представлен, и факультативные (Non-Core) – могут присутствовать или отсутствовать в зависимости от контекста.

Лексическими единицами выступают слова, которые активируют фрейм (например, глаголы *дать, подарить, вручить* активируют фрейм «Передача»).

В составе онто-лингвистической платформы FrameNet обеспечивает ситуационную семантику – описание того, как слова организуются вокруг типовых ситуаций. Это позволяет:

- связывать глагольные морфемы с фреймами, эта связка называется активацией фрейма (отношение activatesFrame);
- описывать валентностные структуры: какие участники и в каких ролях могут появляться при данном глаголе;
- сопоставлять способы выражения одной и той же ситуации в разных тюркских языках;
- разрешать лексическую многозначность: один глагол может активировать разные фреймы в различных контекстах.

Грамматические категории и граммемы – это системы противопоставленных грамматических значений.

Грамматическая категория представляет собой систему граммем, объединенных на основе общности семантического признака. Так, категория падежа включает граммемы Nom, Gen, Acc, Dat, Loc, Abl, каждая из которых является частным значением внутри этой системы. Наряду с ними выделяются квазиграммемы, фиксирующие оппозицию по типу «наличие/отсутствие признака». Примером может служить уподобительное значение, выражаемое аффиксами -сыман или -дай; такие признаки не складываются в полноценную грамматическую категорию.

В рамках платформы грамматические категории и граммемы выполняют роль связующего звена между морфологическим и семантическим уровнями.

Аффиксальные морфемы связываются с граммемами через отношение выражения (expresses); граммемы, в свою очередь, транслируются в семантические роли в соответствии с параметрами отображения, задаваемыми компонентом D, который описан далее. Кроме того, они участвуют в формировании синтаксических правил на основе UD, определяя, какие словоформы допустимы в тех или иных синтаксических позициях.

Для единообразной записи произношения использован Международный фонетический алфавит (IPA), представляющий собой стандартизированную систему знаков, где каждый символ жестко соответствует определенному звуку или классу звуков независимо от конкретного языка. В нашей платформе IPA используются для фонологической разметки корпусных данных, что позволяет сопоставлять произносительные варианты одного слова на различных диалектах и языках, а также фиксировать аутентичное звучание для языков, находящихся под угрозой исчезновения.

Прагматический слой описан с помощью категорий DiAML, которые включают измерения, коммуникативные функции, квалификаторы и отношения, регламентированные стандартом ISO 24617-2.

2.1.2. Уровень 2: Грамматическое описание языков

Уровень 2 содержит зависимые от конкретного языка сущности: это инвентарь единиц и правила их комбинаторики для каждого языка или диалекта. Он является ядром платформы с точки зрения формализации конкретных языков.

Блок А: Базовые лингвистические единицы (инвентарь языка)

Фонемы – звуковой инвентарь языка с привязкой к IPA-символам уровня 1.

Алломорфы – конкретные фонетические варианты морфем с условиями выбора (сингармонизм, ассимиляция). У каждого алломорфа есть форма, тип (корневой, аффиксальный, аналитический), признак терминальности (может ли стоять последним в словоформе).

Морфемы – подразделяются на:

- корневые – несут лексическое значение, связаны с концептами 1-го уровня платформы;

• аффиксальные – выражают грамматические значения, связаны с граммемами уровня 1;

• аналитические – отдельно стоящие служебные слова (послелоги, частицы, вспомогательные глаголы), функционально эквивалентные аффиксам, но пишущиеся раздельно;

• устойчивые выражения (MWE) – многословные конструкции с некомпозициональным значением, хранятся как отдельные лексические единицы со ссылкой на синсет и фрейм уровня 1.

Блок Б: Правила комбинаторики (грамматика языка)

Отношение Next – ориентированный граф допустимых последовательностей алломорфов в словоформе. Правила заданы на уровне конкретных алломорфов, что позволяет учитывать сингармонизм и ассимиляцию.

UD-правила описывают синтаксические зависимости: какие части речи и граммемы могут быть подлежащим (nsubj), дополнением (obj), обстоятельством (obl) и т. д., с учетом порядка слов в предложении по правилу SOV.

Для решения проблемы фиксации порядка морфем в платформе разработано неинвазивное расширение стандарта UD – TurkUD. Оно сохраняет порядок морфем в поле MISC формата CONLL-U с использованием формата

MorphLattice = урман[Stem]|нар[Plur]|га[Dat].

Это расширение не нарушает совместимость со стандартным UD: парсеры, не поддерживающие TurkUD, игнорируют поле MISC, а поддерживающие получают дополнительную информацию о порядке морфем.

UMR-правила отображают синтаксические зависимости в семантические роли (ARG0–ARG5, :goal, :source, :time и т. д.) в зависимости от активированного фрейма и контекста.

В рамках платформы “TurkLang” формализм UMR использован на уровне эмпирических данных (слой L5 корпуса, описан далее). Аннотация строится на основе правил отображения, хранящихся в уровне 2 (компонент D параметрической модели). Процесс включает три этапа.

Синтаксический анализ. На основе UD-дерева предложения определяются синтаксические зависимости между словоформами (например, nsubj для подлежащего, obl:dat для косвенного дополнения в дательном падеже).

Отображение в семантические роли. Для каждой синтаксической зависимости применяются правила компонента D, которые сопоставляют граммемы и синтаксические роли с семантическими ролями UMR. Например:

```
nsubj → :ARG0 (агенс/экспериенцер)
obj → :ARG1 (пациенс/тема)
obl:dat → :goal (в контексте фрейма движения),
           :recipient (в контексте фрейма передачи)
obl:loc → :location
obl:abl → :source
```

Выбор конкретной семантической роли для одной и той же граммемы (например, дательного падежа) зависит от активированного глаголом фрейма FrameNet. Так, в предложении «Малай урманга барды» (фрейм MOTION) аффикс -ГА отображается в роль :goal, а в предложении «Мин сиңа китап бирдем» (фрейм TRANSFER) – в роль :recipient.

Остановимся на построении UMR-графа. На основе определенных семантических ролей и атрибутов (время, аспект, эвиденциальность) формируется ориентированный граф, где узлы соответствуют концептам (событиям, сущностям), а ребра – семантическим отношениям (пример приведен выше).

Такой подход обеспечивает детерминированную и масштабируемую семантическую аннотацию корпуса, поскольку все правила отображения заданы явно и могут быть параметризованы для каждого языка (компонент D). Кроме того, UMR-аннотация используется на этапе верификации при генерации текста нейросетевыми моделями: агент-семантик проверяет, соответствует ли сгенерированный UMR-граф правилам платформы, и при обнаружении несоответствия инициирует перегенерацию или формирует объяснение ошибки.

DiAML-правила сопоставляют грамматические формы (вопросительные частицы, эвиденциальные аффиксы, наклонения) с прагматическими функциями и квалификаторами.

2.1.3. Уровень 3: Эмпирические данные

Уровень 3 содержит реальные языковые данные (корпусы) с многослойной аннотацией, а также географическую привязку для диалектологических ис-

следований. Он распределен между графовой СУБД (для аннотаций), реляционной СУБД (для метаданных и геоданных) и объектным хранилищем (для мультимедиа).

Многослойное представление корпуса

Корпус проходит последовательные этапы обработки, каждый из которых добавляет новый слой аннотации. Переход между слоями L0–L6 автоматизирован через конвейер многоуровневой аннотации корпуса (рис. 2).

Каждый слой ссылается на правила из уровня 2 и универсалии из уровня 1. Токены хранятся как пути в графе Next, что позволяет выполнять быстрые обходы последовательностей морфем и семантический поиск по морфемной структуре.

Геолингвистические данные

Для поддержки диалектологических и ареальных исследований в уровень 3 интегрированы геолингвистические данные:

- Locality – населенный пункт с координатами по стандарту WGS84 и административной принадлежностью;
- Informant – информант с биографическими и социолингвистическими данными (год рождения, пол, образование, место рождения, миграционная история);
- Expedition – полевая экспедиция с метаинформацией (даты, состав исследователей, оборудование);
- Document – привязка документов к населенному пункту, информанту и экспедиции;
- Area – ареал распространения идиома (полигон на карте);
- Isogloss – изоглосса для лингвистического параметра, строится автоматически на основе значений параметров в точках сбора.

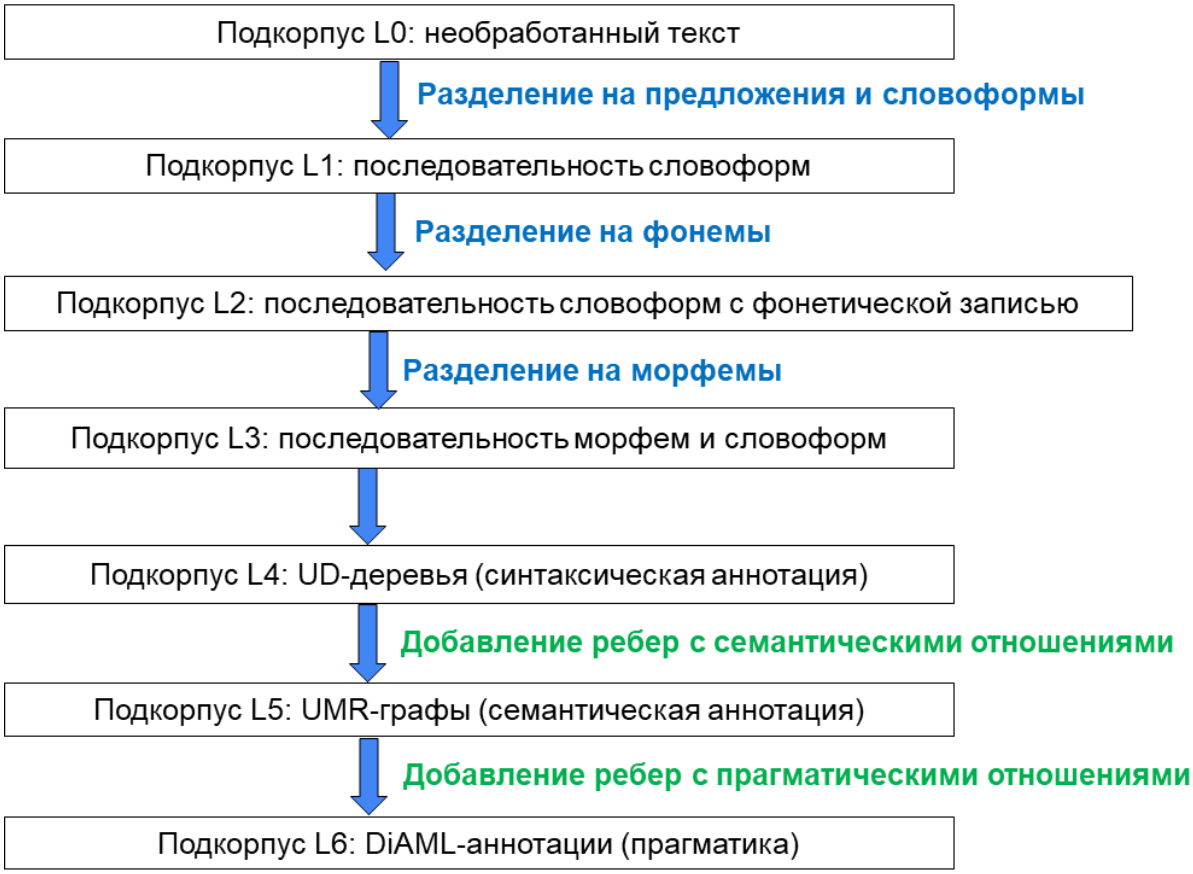


Рис. 2. Конвейер многоуровневой аннотации корпуса.

Связи между эмпирическими данными и лингвистическими единицами обеспечиваются индексными таблицами, поддерживающими двустороннюю навигацию без нарушения однонаправленности ссылок.

2.2. Принцип многоуровневой согласованности

Если трехуровневая архитектура определяет, как организованы данные, то принцип многоуровневой согласованности (семиотическая вертикаль) определяет, какие данные и в каком формате должны быть описаны на каждом уровне. Он является теоретической рамкой, обеспечивающей полноту и непротиворечивость лингвистического описания.

Семиотическая вертикаль представляет собой модель языка как многоуровневой знаковой системы, где каждый уровень соответствует определенному аспекту языковой единицы – от звукового состава до коммуникативной цели высказывания.

Табл. 1. Уровни семиотической вертикали.

Код	Название	Международный стандарт
SL1	Фонологический	IPA
SL2	Морфонологический мост	Продукционные правила
SL3	Морфологический	БНФ, конечные автоматы
SL4	Морфосинтаксический мост	Universal Dependencies
SL5	Синтаксический	Universal Dependencies
SL6	Морфосемантический мост	Словарь соответствий (UD + FrameNet)
SL7	Семантический	UMR, FrameNet, WordNet
SL8	Прагматический	DiAML (ISO 24617-2)

Вертикаль включает восемь уровней (SL1–SL8, см. табл. 1). Каждый уровень опирается на одну из нотаций, принятых в качестве международного стандарта в определенных лингвистических сообществах, что обеспечивает совместимость с внешними инструментами и воспроизводимость аннотаций.

В семиотической вертикали использованы уровни, которые называются семиотическими мостами – SL2, SL4, SL6. Особенность семиотических мостов в том, что они используют не собственные лингвистические единицы, а единицы соседних уровней.

SL2 морфонологический мост – связывает фонетический и морфологический уровни, описывая правила выбора алломорфов в зависимости от фонологического контекста (сингармонизм, ассимиляция).

SL4 морфосинтаксический мост – связывает морфологию и синтаксис, описывая аналитические конструкции и многословные выражения (MWE), которые находятся между словом и свободным синтаксическим сочетанием.

SL6 морфосемантический мост – связывает грамматику и семантику, описывая отображение грамматических признаков в семантические роли (компонент D параметрической модели).

Семиотическая вертикаль реализует два разнонаправленных процесса.

Спецификация (снизу вверх) – нижележащий уровень налагает ограничения на вышележащий. Например, фонологический состав определяет, какие алломорфы допустимы в языке; морфологический уровень определяет, какие комбинации морфем возможны.

Интерпретация (сверху вниз) – вышележащий уровень придает смысл единицам нижележащего. Например, семантический уровень интерпретирует грамматическую форму как семантическую роль участника ситуации; прагматический уровень интерпретирует предложение как речевой акт.

Семиотическая вертикаль выступает в роли системообразующего элемента, определяющего содержание трех основных уровней платформы. В отношении онтологического ядра (уровень 1) она задает требование, согласно которому каждый из восьми уровней должен быть описан на формальном метаязыке, будь то IPA, UD, UMR, DiAML или иной стандарт. Для уровня грамматических описаний (уровень 2) вертикаль предписывает наличие для каждого языка правил, обеспечивающих связь между уровнями: в частности, правил синтаксической аннотации (UD), семантического отображения (UMR) и прагматической интерпретации (DiAML). Наконец, для эмпирического слоя (уровень 3) она устанавливает необходимость многослойной аннотации корпусных данных, охватывающей все уровни вертикали – от фонетической транскрипции (L1) до прагматической разметки (L6).

Таким образом, семиотическая вертикаль обеспечивает сквозную интерпретацию языковых единиц – от акустического сигнала до интенционального акта – и служит теоретическим основанием, на котором строится вся архитектура платформы.

2.3. Принцип морфемоцентричой модели данных

После описания общей архитектуры и принципа многоуровневой согласованности становится понятным, почему базовым элементом хранения выбраны

морфема, а не слово. Это решение продиктовано природой элементно-комбинаторных языков: именно на уровне морфем сосредоточена основная грамматическая информация, а словоформа представляет собой линейную комбинацию морфем.

Перечислим преимущества морфемоцентричного подхода.

1. Экономичность хранения благодаря перечислимости морфем. В отличие от словоформ и предложений, множество морфем в языке является конечным и перечислимым, его можно полностью описать в виде словаря. Морфемоцентричная модель хранит только этот словарь (несколько тысяч морфем на язык) и правила их сочетания (граф Next, описанный далее), а все словоформы порождаются динамически по запросу. Это соответствует классическому принципу формальных грамматик: конечное множество правил порождает бесконечное множество цепочек.

2. Принцип «одна морфема – одна граммема». В агглютинативных языках каждый аффикс выражает, как правило, одну грамматическую категорию, в отличие от флективных языков, где один аффикс может кодировать несколько признаков одновременно. Это обеспечивает однозначность на морфологическом уровне: каждой морфеме соответствует ровно одна граммема из онтологического ядра. При этом одна граммема может отображаться в разные семантические роли в зависимости от контекста (активированного фрейма), но сам выбор роли детерминирован правилами отображения (компонент D). Таким образом, принцип «одна морфема – одна граммема» гарантирует простоту и предсказуемость морфологического анализа, а многозначность грамматических показателей разрешается на более высоких уровнях (синтаксическом и семантическом) с использованием конечного набора правил.

3. Минимальное количество исключений. В агглютинативных языках правила сочетаемости морфем имеют очень мало исключений: фонологические чередования (сингармонизм, ассимиляция) подчиняются четким и регулярным правилам, а неправильные глаголы или супплетивизм практически отсутствуют. Это позволяет построить детерминированную формальную модель (граф Next), которая покрывает подавляющее большинство словоформ. К примеру, для татарского языка это покрытие составляет более 90%, что отличает эту модель от

модели флективных языков, где многочисленные исключения делают применение подобной модели для флективных языков неэффективной.

4. Единообразное описание синтетических и аналитических форм. В агглютинативных языках одно и то же грамматическое значение может выражаться как аффиксом (синтетически), так и отдельным служебным словом (аналитически). В морфемоцентричной модели и аффиксы, и аналитические морфемы (послелоги, частицы, вспомогательные глаголы) являются узлами одного и того же графа Next с грамматической привязкой к граммеме. Различие между ними состоит только в атрибуте узла, определяющем признак написания: слитно или раздельно. Это позволяет обрабатывать синтетические и аналитические конструкции единым механизмом обхода графа, в отличие от лексикоцентричных подходов, где они описываются разными механизмами (морфология vs синтаксис).

5. Морфемоцентричность модели позволяет реализовывать сингармонизм. Сингармонизм (гармония гласных) – это ключевая фонетическая особенность тюркских языков. В морфемоцентричной модели он реализуется на уровне структуры данных: каждый аффикс представлен множеством алломорфов с условиями выбора, причем связь между этими алломорфами осуществляется по правилам сингармонизма. Это гарантирует корректность словоформ без дополнительной логики в коде анализатора.

6. Граф Next как единый механизм фиксации порядка и генерации словоформ. Морфемы связаны отношением Next, которое образует направленный граф допустимых последовательностей. Этот граф одновременно решает две задачи:

- фиксация порядка – явное сохранение последовательности аффиксов, критичное для элементно-комбинаторных языков, где изменение порядка морфем изменяет смысл словоформы (например, masadakilerin против masanınkilerde, структура которых была описана ранее);
- детерминированная генерация – обход графа от корня до терминального узла порождает все теоретически возможные словоформы, что особенно важно для малоресурсных языков, где нет больших корпусов для обучения статистических моделей.

Единый формализм для двух задач – фиксации и генерации – является ключевым архитектурным преимуществом.

7. Единообразное и экономичное описание родственных языков благодаря перечислимости морфем и принципу «одна морфема – одна граммема». Морфемы являются значимыми единицами, общими для всей языковой семьи. Сравнение языков упрощается по следующим причинам:

- Перечислимость. Множество морфем в каждом языке конечно, поэтому сравнение сводится к сопоставлению конечных наборов морфем и правил их сочетания, а не к сравнению бесконечных множеств словоформ или предложений.
- Одна морфема – одна граммема. Каждая морфема выражает ровно одну грамматическую категорию, поэтому сравнение ведется на уровне отдельных граммем, а не сложных пучков признаков.
- Одна корневая морфема – один или несколько синсетов WordNet. Лексическая многозначность обрабатывается на уровне семантического анализа, но даже при многозначности сравнение ведется на уровне конечных множеств (корневых морфем и синсетов).
- Одна морфема – одна семантическая роль в UMR. В рамках активированного фрейма каждая морфема, включая аффиксы, выражает ровно одну семантическую роль, что позволяет строить однозначные правила отображения в UMR-графы.

Это позволяет составлять таблицы соответствия граммем и морфем для разных языков, выявлять общетюркские морфемы, переносить знания на малоресурсные языки и параметризовать языки без необходимости описывать все словоформы заново.

8. Естественное представление в виде графа и нативная поддержка СУБД. Морфемоцентричная модель данных имеет прямое структурное соответствие с графовым представлением: морфемы становятся узлами, отношение Next – ребрами, а словоформы – путями в графе. Это позволяет использовать графовые СУБД (Memgraph, Neo4j) для хранения и обработки данных с высокой производительностью:

- обход пути от корня до терминального узла выполняется за O(k) операций, где k – длина словоформы (число аффиксов);

- поиск по морфемной структуре («найти все словоформы, содержащие аффикс X после аффикса Y») реализуется как навигация по ребрам Next;
- семантический поиск с использованием иерархии концептов WordNet выполняется как обход по ребрам отношения hasSynset и таксономическим связям.

В отличие от реляционных СУБД, где аналогичные операции требуют множества субопераций JOIN и рекурсивных запросов, графовая СУБД выполняет их быстрее за счет своей структуры, что особенно важно для больших корпусов объемов порядка миллионов токенов и сложных запросов.

2.4. Принцип параметризации языков

Поскольку тюркские языки близкородственны, их описания имеют общее ядро. Для учета различий и реализации наследования между языками используется параметрический подход, реализованный через четырехкомпонентную систему:

1) компонент A – системные параметры (фонемный инвентарь, сингармонизм, число падежей, порядок слотов);

2) компонент B – лексические параметры (набор алломорфов морфем, морфотактические правила);

3) компонент C – эпистемические параметры (степень изученности языка, достоверность данных);

4) компонент D – параметры отображения (соответствие граммем семантическим ролям и прагматическим функциям).

2.5. Взаимодействие базовых принципов

Четыре описанных принципа образуют единую систему, где каждый выполняет свою функцию (табл. 2). Эта архитектура обеспечивает единое формализованное пространство для работы с тюркскими языками, где каждый принцип решает свою задачу, а все вместе они образуют систему, пригодную для автоматической обработки, верификации и коллективного развития.

Табл. 2. Взаимодействие принципов архитектуры платформы “TurkLang”.

Принцип	Основной вопрос	Реализация	Связь с другими принципами
Трехуровневая организация	Как структурированы данные?	Уровни 1 (онтология), 2 (грамматика), 3 (эмпирика) с однонаправленными связями	Определяет архитектуру хранения для семиотической вертикали и морфемоцентричной модели
Многоуровневая согласованность (семиотическая вертикаль)	Какие данные и в каком формате должны быть описаны?	8 уровней (SL1–SL8) с международными стандартами (IPA, UD, UMR, DiAML)	Задает содержание уровней 1, 2 и 3; определяет, как морфемы интерпретируются на разных уровнях
Морфемоцентричность	Какой элемент является базовым для хранения?	Морфема как атомарная единица	Опирается на универсалии уровня 1; реализуется через систему отношений морфемы из уровня 2; обеспечивает интерпретацию корпусных данных
Параметризация	Как описывать различия между родственными языками?	Компоненты A–D, наследование $I_{var} = I_{parent} + \Delta I_{var}$	Использует универсалии уровня 1; настраивает грамматику уровня 2 под конкретный язык; определяет правила отображения для UMR

ЗАКЛЮЧЕНИЕ

Представлена архитектура платформы “TurkLang” – системного решения для создания единого цифрового пространства обработки лингвистических данных тюркских языков. Архитектура базируется на четырех взаимосвязанных

принципах: трехуровневой организации базы знаний, многоуровневой согласованности (семиотической вертикали), морфемоцентричной модели данных и параметрическом подходе к описанию родственных языков.

Ключевым решением является переход от лексикоцентрической модели к морфемоцентрической, что обусловлено структурно-функциональными особенностями тюркских языков. Морфемоцентричность обеспечивает экономичность хранения, однозначность морфологического анализа (принцип «одна морфема – одна граммема»), минимальное число исключений, единообразное описание синтетических и аналитических форм, естественную реализацию сингармонизма и сквозную генерацию словоформ через граф Next.

Семиотическая вертикаль (SL1–SL8), опирающаяся на международные стандарты (Universal Dependencies, Uniform Meaning Representation), обеспечивает сквозной анализ от фонемы до речевого акта.

Дальнейшее развитие связано с углублением семантической и прагматической аннотаций, развертыванием RDF-триплетора и масштабированием методологии на другие агглютинативные семьи – финно-угорские, монгольские, тунгусо-маньчжурские и эскимосско-алеутские языки. Предложенная архитектура может служить прототипом для создания лингвистических платформ нового поколения, применимых к широкому кругу агглютинативных языков мира.

Благодарности

Работа выполнена в рамках соглашения №12/2025-ПД от 22.12.2025 о предоставлении гранта научным и научно-педагогическим работникам обособленных структурных подразделений АН РТ с целью стимулирования их к защите докторских диссертаций и выполнению НИР.

СПИСОК ЛИТЕРАТУРЫ

1. *Гузев В.Г.* О некоторых экзотических особенностях тюркских языков («тюркские чудеса») // Актуальные проблемы мировой политики. 2020. № 10. С. 231–245. https://doi.org/10.21638/11701/26868318.16

2. *Koskenniemi K.* Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production. Helsinki: University of Helsinki, 1983. 160 p.

3. *Апресян Ю.Д.* Избранные труды. Том II. Интегральное описание языка и системная лексикография. М.: Школа «Языки русской культуры», 1995. 769 с.

4. *Krauwer S.* The basic language resource kit (BLARK) as the first milestone for the language resources roadmap // Proc. International workshop on speech and computer SPECOM-2003, Moscow, Russia, Oct. 2003. P. 8–15.

5. *Kilgarriff A., Rychlý P., Smrž P., Tugwell D.* The Sketch Engine // Proceedings of the Eleventh EURALEX International Congress, Lorient, France, Jul. 2004. P. 105–116.

6. *Норманская Ю.В., Борисенко О.Д., Белобородов И.Б., Аветисян А.И.* Программная система LingvoDoc и возможности, которые она предлагает для документирования и анализа обско-угорских языков // Докл. РАН. Матем., информ., проц. упр. 2022. № 504. С. 60–82.
https://doi.org/10.31857/S2686954322030055

7. *Khanna, T., Washington, J.N., Tyers, F.M et al.* Recent advances in Apertium, a free/open-source rule-based machine translation platform for low-resource languages // Machine Translation. 2021. Vol. 35. P. 475–502.
https://doi.org/10.1007/s10590-021-09260-6

8. *Bakay Ö., Ergelen Ö., Sarmış E., Yıldırım S., et al.* Turkish WordNet KeNet // Proceedings of the 11th Global Wordnet Conference, Virtual, Jan. 2021. P. 166–174. https://doi.org/10.18653/v1/2021.gwc-1.19

9. *Agostini A., Usmanov T., Khamdamov U., Abdurakhmonova N., et al.* UZWORDNET: A Lexical- Semantic Database for the Uzbek Language // Proceedings of the 11th Global Wordnet Conference, Virtual, Jan. 2021. P. 8–19.
https://doi.org/10.18653/v1/2021.gwc-1.2

10. *Marsan B., Kara N., Ozcelik M., Arican B. N., et al.* Building the Turkish FrameNet // Proceedings of the 11th Global Wordnet Conference, Virtual, Jan. 2021. P. 118–125. https://doi.org/ 10.18653/v1/2021.gwc-1.14

11. *Rezhake M., Kuerban A.* Design of the Uyghur FrameNet Desktop // Lecture Notes on Software Engineering. 2015. Vol. 3. No. 1. P. 53–56.
https://doi.org/ 10.7763/LNSE.2015.V3.165

12. *Navigli R., Ponzetto S.P.* BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network // Artificial Intelligence. 2012. Vol. 193. P. 217–250.
https://doi.org/10.1016/j.artint.2012.07.001

13. *Gangemi A., Alam M., Asprino L., Presutti V., et al.* Framester: A Wide Coverage Linguistic Linked Data Hub. // Knowledge Engineering and Knowledge Management. EKAW 2016. Lecture Notes in Computer Science. 2016. Vol. 10024. P. 239–254. https://doi.org/10.1007/978-3-319-49004-5 16

14. *Nivre J., de Marneffe M-C., Ginter F., Hajič J., et al.* Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection // Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC 2020), Marseille, France, May. 2020. P. 4034–4043.

ARCHITECTURE OF THE LINGUISTIC PLATFORM TURKLANG

A.R. Gatiatullin[0000-0003-3063-8147]

Tatarstan Academy of Sciences, Kazan, Russia

ayrat.gatiatullin@gmail.com

Abstract

The article describes the architecture of the TurkLang multifunctional linguistic platform, currently being developed at the Institute of Applied Semiotics of the Academy of Sciences of the Republic of Tatarstan. The system is based on a three-level knowledge base: linguistic universals, descriptions of units of specific languages, and empirical speech data. The knowledge base is implemented as a graph, where the primary storage unit is the morpheme – the minimal meaningful unit of language (akin to an atomic data element). This design choice is driven by the element-combinatorial nature of Turkic languages: a word form (the concrete realization of a word in text) is a linear chain consisting of a root and sequentially attached affixes – modifiers, each expressing a single grammatical category (case, number, tense, etc.) in a strictly fixed order. The paper examines levels of abstraction, types of relationships between entities, and mechanisms for integration with external sources (an electronic encyclopedia, lexicographic resources, and GIS data). Special attention is paid to the application

of international standards for syntactic (Universal Dependencies), semantic (Universal Meaning Representation), and pragmatic (DiAML) annotation. A comparison with existing platforms (Sketch Engine, LingvoDoc, Apertium) is provided. To date, the platform's knowledge base covers more than 30 languages and dialects, contains over 237,000 root morphemes and about 40,000 compatibility rules, while the corpus component comprises over 180 million of word occurrences.

Keywords*: linguistic platform, knowledge graph, Turkic languages, morpheme-centricity, three-level architecture, Universal Dependencies, Universal Meaning Representation, agglutinative languages, semiotic vertical.*

REFERENCES

1. *Guzev V.G.* O nekotorykh ekzoticheskikh osobennostyakh tyurkskikh yazykov («tyurkskie chudesa») // Aktual'nye problemy mirovoy politiki. 2020. Vol. 10. P. 231–245. https://doi.org/10.21638/11701/26868318.16

2. *Koskenniemi K.* Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production. Helsinki: University of Helsinki, 1983. 160 p.

3. *Apresyan Y.D.* Izbrannye trudy. Tom II. Integral'noe opisanie yazyka i sistemnaya leksikografiya. Moscow: Shkola “Yazyki russkoi kul’tury”, 1995. 769 p.

4. *Krauwer S.* The basic language resource kit (BLARK) as the first milestone for the language resources roadmap // Proc. International workshop on speech and computer SPECOM-2003, Moscow, Russia, Oct. 2003. P. 8–15.

5. *Kilgarriff A., Rychlý P., Smrž P., Tugwell D.* The Sketch Engine // Proceedings of the Eleventh EURALEX International Congress, Lorient, France, Jul. 2004. P. 105–116.

6. *Normanskaya Y.V., Borisenko O.D., Beloborodov I.B., Avetisyan A.I.* The Software System LingvoDoc and the Possibilities It Offers for Documentation and Analysis of Ob-Ugric Languages // Doklady Mathematics. 2022. Vol. 105, No. 3. P. 187–206. https://doi.org/10.31857/S2686954322030055

7. *Khanna, T., Washington, J.N., Tyers, F.M et al.* Recent advances in Apertium, a free/open-source rule-based machine translation platform for low-resource languages // Machine Translation. 2021. Vol. 35. P. 475–502. https://doi.org/10.1007/s10590-021-09260-6

8. *Bakay Ö., Ergelen Ö., Sarmış E., Yıldırım S., et al.* Turkish WordNet KeNet // Proceedings of the 11th Global Wordnet Conference, Virtual, Jan. 2021. P. 166–174. https://doi.org/10.18653/v1/2021.gwc-1.19

9. *Agostini A., Usmanov T., Khamdamov U., Abdurakhmonova N., et al.* UZ-WORDNET: A Lexical- Semantic Database for the Uzbek Language // Proceedings of the 11th Global Wordnet Conference, Virtual, Jan. 2021. P. 8–19. https://doi.org/10.18653/v1/2021.gwc-1.2

10. *Marsan B., Kara N., Ozcelik M., Arican B. N., et al.* Building the Turkish FrameNet // Proceedings of the 11th Global Wordnet Conference, Virtual, Jan. 2021. P. 118–125. https://doi.org/10.18653/v1/2021.gwc-1.14

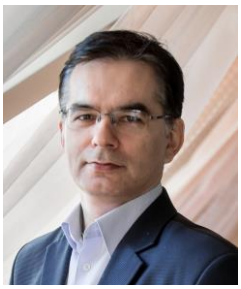
11. *Rezhake M., Kuerban A.* Design of the Uyghur FrameNet Desktop // Lecture Notes on Software Engineering. 2015. Vol. 3. No. 1. P. 53–56. https://doi.org/10.7763/LNSE.2015.V3.165

12. *Navigli R., Ponzetto S.P.* BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network // Artificial Intelligence. 2012. Vol. 193. P. 217–250. https://doi.org/10.1016/j.artint.2012.07.001

13. *Gangemi A., Alam M., Asprino L., Presutti V., et al.* Framester: A Wide Coverage Linguistic Linked Data Hub. // Knowledge Engineering and Knowledge Management. EKAW 2016. Lecture Notes in Computer Science. 2016. Vol. 10024. P. 239–254. https://doi.org/10.1007/978-3-319-49004-5 16

14. *Nivre J., de Marneffe M-C., Ginter F., Hajič J., et al.* Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection // Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC 2020), Marseille, France, May. 2020. P. 4034–4043.

СВЕДЕНИЯ ОБ АВТОРАХ



ГАТИАТУЛЛИН Айрат Рафизович – окончил Казанский государственный университет в 1994 г., к. т. н. (2002). Ведущий научный сотрудник Института прикладной семиотики Академии наук Республики Татарстан. Автор более 100 научных трудов.

Ayrat Rafizovich GATIATULLIN – graduated from Kazan State University in 1994, Candidate of Technical Sciences (2002). Leading researcher at the Institute of Applied Semiotics of Tatarstan Academy of Sciences. List of scientific works includes more than 100 publications.

email: ayrat.gatiatullin@gmail.com

ORCID: 0000-0003-3063-8147

Материал поступил в редакцию 26 июня 2026 года

УДК 004.85+159.9

ПРОГНОЗИРОВАНИЕ ИЗМЕНЕНИЙ В ПОВЕДЕНИИ ЛИЧНОСТИ ПРИ ВОЗДЕЙСТВИИ СОБЕСЕДНИКА

Ф. М. Гафаров[1] [0000-0003-4704-154X], П. Н. Устин[2] [0000-0003-3950-7434], К. С. Николаев[3] [0000-0003-3204-238X]

[1–3]*Казанский (Приволжский) федеральный университет, г. Казань, Россия*

[1]fgafarov@yandex.ru, [2]pavust@mail.ru, [3]konnikolaeff@yandex.ru

Аннотация

Изучение взаимодействий между пользователями информационных систем является актуальной задачей в гуманитарных и технических исследованиях. Представлена разработка регрессионной модели, прогнозирующей изменения в выраженности содержательных компонентов текста при воздействии со стороны других людей. В качестве источника данных выступили короткие беседы (триплеты сообщений) между двумя пользователями в социальных сетях. Обучаемая модель разработана для применения в составе когнитивной архитектуры, отображающей модель поведения субъекта в цифровой среде.

Ключевые слова*: обработка естественного языка, цифровое поведение, цифровая личность, машинное обучение, SVR-модели.*

ВВЕДЕНИЕ

В современной психологии и когнитивистике все большее внимание уделяют анализу цифровых следов человека как источника данных для прогнозирования его поведения и личностных характеристик. Тексты, создаваемые пользователями в цифровой среде, рассматривают как объективные продукты психической активности, кодирующие интенции, цели и ожидаемые сценарии действий. В настоящей работе проверяется следующая гипотеза: существует ли возможность прогнозирования изменений компонентного наполнения функциональных подсистем поведения пользователей цифровой среды на основе компонентного наполнения исходного текста и текста-ответа, воздействующего на автора исходного текста.

© Ф. М. Гафаров, П. Н. Устин, К. С. Николаев, 2026.
Данная статья распространяется на условиях международной лицензии Creative Commons License Attribution 4.0 International (CC BY 4.0).

В связи с этим целью настоящей работы являются обучение и верификация моделей, направленных на прогнозирование изменений характеристик поведения личности в рамках психологической модели поведения человека в реальной и виртуальной средах. Для этого были поставлены следующие задачи: сбор текстовых данных различных пользователей из открытых источников и разметка данных, полученных с помощью экспертов и больших языковых моделей; обучение модели для прогнозирования изменений компонентного наполнения функциональных подсистем поведения пользователей цифровой среды при воздействии со стороны другого человека.

Предлагаемый подход основан на анализе текстовых данных, созданных человеком, и выявлении характеристик его виртуального «Я» через совокупность текстовых параметров. Применение компьютерных методов обработки открытых данных позволяет прогнозировать поведение индивида на основе его текстовой активности. В рамках современной когнитивистики речевая активность личности рассматривается как поведенческое проявление интенциональности. Тексты, являясь продуктами психической деятельности, представляют собой объективные следы, доступные для верифицируемого анализа. Коммуникативные интенции, жанровая организация и содержательное наполнение текста кодируют программу действий автора, эксплицируя его цели, допущения и ожидаемые ходы. Таким образом, текст, порождаемый человеком и соотносимый с возможными сценариями поведения, может выступать в качестве информативного источника для прогнозирования его дальнейших поведенческих выборов.

Теоретической основой предлагаемого подхода выступают работы, в которых поведение рассматривается как процессуально-динамическая характеристика субъекта, разворачивающаяся во времени и включающая как внешние действия, так и внутренние психические состояния (см. например, [1]). Такое понимание опирается на фундаментальные методологические принципы системного анализа психических явлений, сформулированные в трудах Б. Ф. Ломова, согласно которым психические процессы и состояния необходимо изучать в их динамике и взаимосвязи с деятельностью субъекта [2]. В контексте профессиональной деятельности и самоопределения личности важную роль играют ценностно-смысловые ориентации субъекта, которые, как показано в работе

Е. А. Климова, определяют направленность активности человека и его поведенческие стратегии [3]. Для анализа текстовых данных, отражающих эти поведенческие паттерны, в настоящей работе использован сценарный подход, позволяющий выделять в тексте устойчивые структуры, соответствующие типичным программам действий субъекта [4].

Существуют работы, в которых цифровые данные применяются для определения личностных характеристик. В частности, в [5] показано, что профили в социальных сетях отражают реальные черты личности, а не идеализированное самопредставление. В [6] исследовано, как сторонние наблюдатели формируют впечатления о личности на основе содержания профилей. В [7] было проведено исследование, посвященное прогнозированию личностных черт на основе активности в Twitter. В работе [8] рассмотрена взаимосвязь между содержанием твитов и чертами «Большой пятерки». В [9] авторы показали, что компьютерные модели могут превосходить людей в точности оценки личности на основе цифровых следов. В рамках проекта [10] было проведено исследование по прогнозированию основных черт личности в соответствии с моделью «Большой пятерки» на основе текстов из двух наборов данных (essays [11] и myPersonality [10, 12]).

В статьях [13–15] представлены результаты разработки психометрической модели предикторов социального, академического и профессионального успеха на основе количественных и качественных данных, полученных из социальных сетей. Методы и технологии, использованные при разработке этой модели, адаптированы и расширены в настоящей работе.

Особое внимание уделено нами применению классических моделей машинного обучения для прогнозирования изменений в данных в результате воздействия со стороны других сторон. В отличие от глубоких нейронных сетей, модели на основе опорных векторов (SVR) и градиентного бустинга (XGBoost, LightGBM) демонстрируют высокую интерпретируемость и устойчивость при работе с многомерными признаковыми пространствами, что особенно важно при анализе психологических предикторов. В работе [16] показано, что SVR эффективно справляется с задачей прогнозирования на основе текстовых данных, пре-

восходя по точности линейные модели и нейронные сети малой глубины. Аналогичные результаты получены в [17], где ансамблевые методы были использованы для предсказания поведенческих паттернов на основе цифровых следов. Кроме того, в [18] продемонстрирована возможность использования ансамблевых методов для прогнозирования изменений в социальном поведении после коммуникативного воздействия. В работе [19] предложен гибридный подход, сочетающий извлечение признаков с помощью линейных методов и последующую регрессию с помощью SVR.

Таким образом, описанные подходы демонстрируют принципиальную возможность извлечения устойчивых психологических характеристик из текстовых данных и составляют основу предлагаемой нами модели поведения пользователей цифровой среды. В настоящей работе для прогнозирования 60 ключей-предикторов на основе 120 ключей из первого и второго текстов триплета использована модель SVR, обученная на данных, размеченных с помощью больших языковых моделей.

ПОДГОТОВКА ДАННЫХ

Тексты триплетов были получены из различных открытых источников (Яндекс.Дзен, Pikabu, VK). Структура страниц с комментариями в Яндекс.Дзен и Пикабу потребовала создания скриптов для автоматической прокрутки страницы и выделения дерева сообщений. Достоинством такого подхода стала неограниченная глубина деревьев комментариев. Для сбора комментариев из VK был применен VK API. Ограничением его применения стало то, что метод GetComments возвращает только первые 10 комментариев к посту. С платформы VK было получено 11214 триплетов, Яндекс.Дзен – 6236, Pikabu – 9352. Общий объем полученных данных составил 26802 триплета. Для исключения триплетов с короткими и не несущими смысловой нагрузки комментариями был применен фильтр 20–20–20 (в каждом элементе триплета должно быть как минимум 20 слов). В результате фильтрации количество триплетов снизилось до 16850. На рис. 1–3 представлены гистограммы распределений количества слов в текстах, полученных с платформ VK, Яндекс.Дзен и Pikabu соответственно.

Для анализа текстов в триплетах и определения значений содержательных компонентов была использована большая языковая модель Claude 3.5

(Anthropic). Эта модель показала высокую корректность анализа при невысокой стоимости по сравнению с моделями аналогичной производительности.

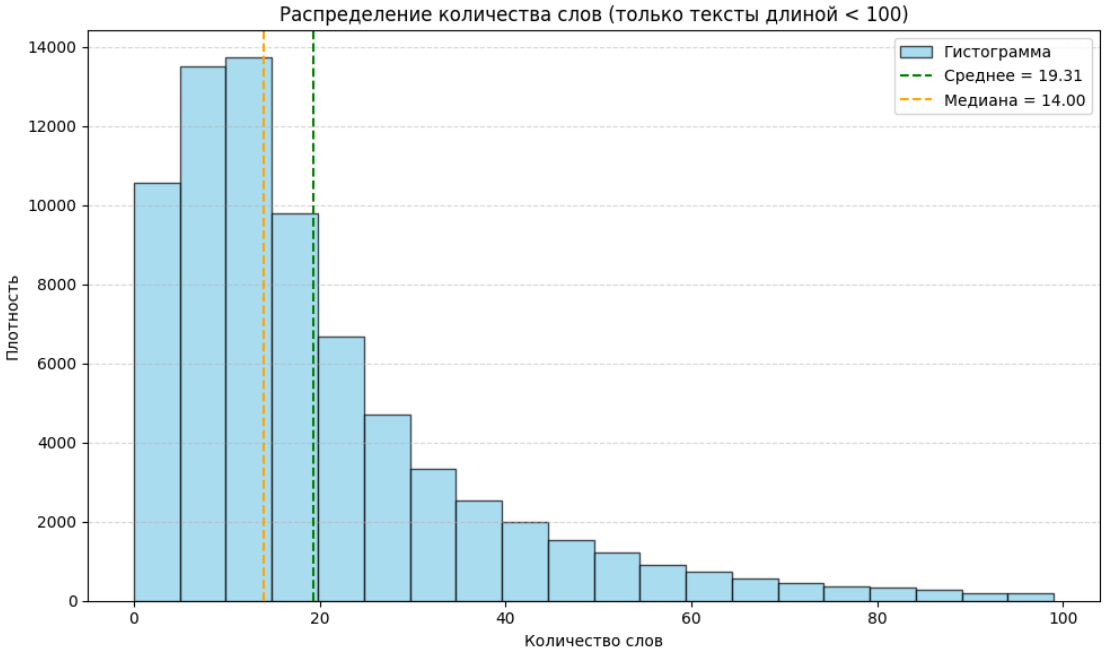


Рис. 1. Гистограммы количества слов в текстах, полученных с платформы VK.

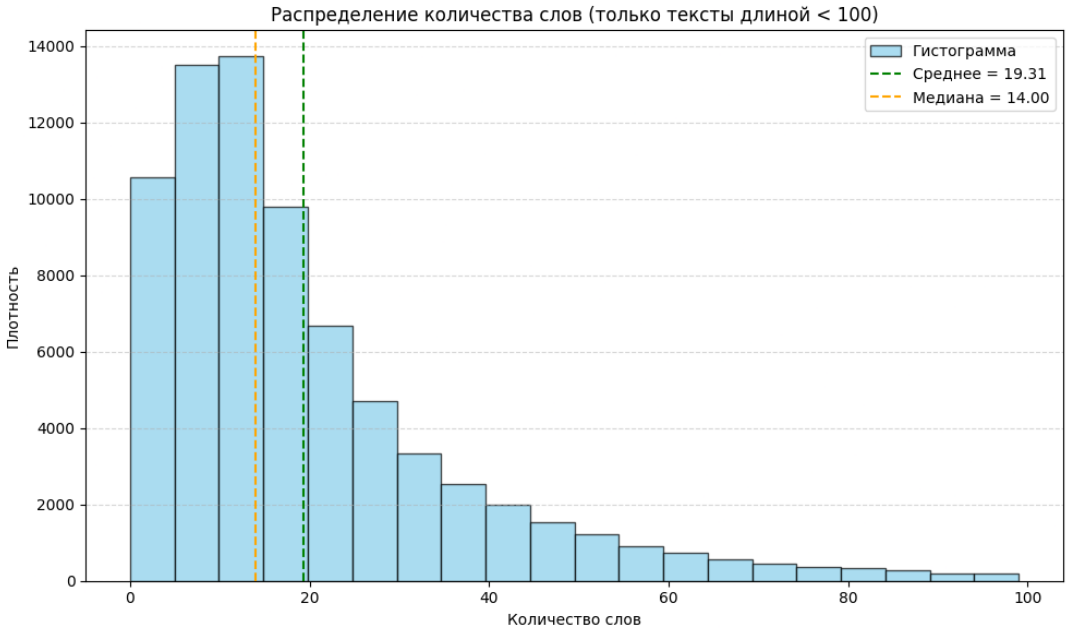


Рис. 2. Гистограммы количества слов в текстах, полученных с платформы Яндекс.Дзен.

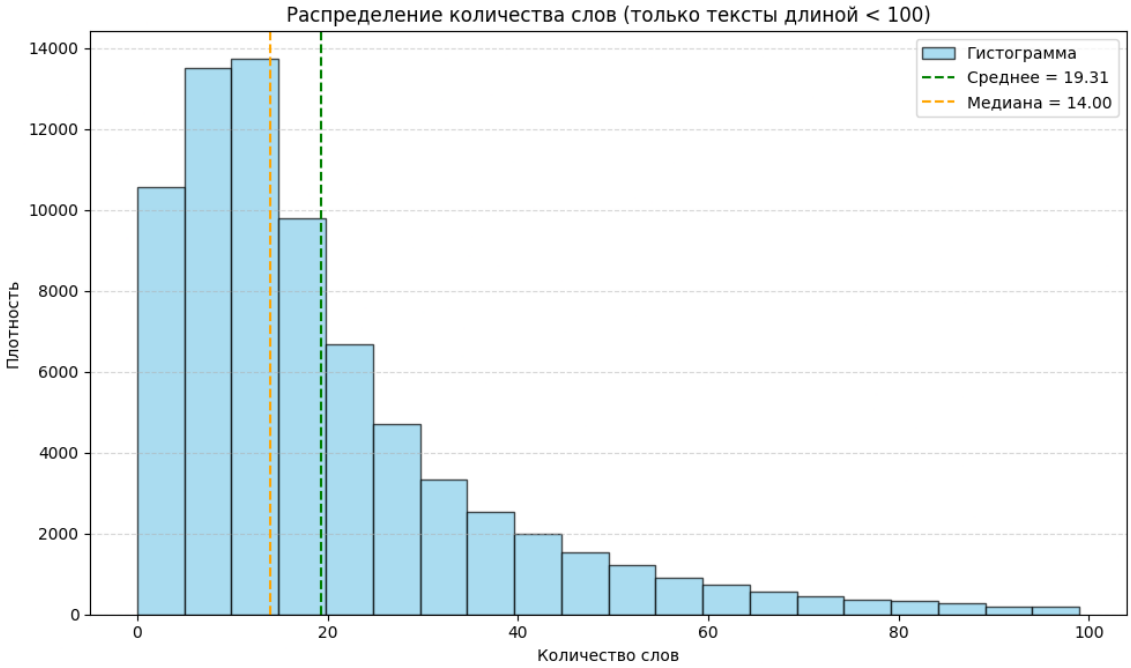


Рис. 3. Гистограммы количества слов в текстах, полученных с платформы Pikabu.

РЕЗУЛЬТАТЫ ОБУЧЕНИЯ МОДЕЛИ SVR ДЛЯ ПРОГНОЗИРОВАНИЯ ИЗМЕНЕННЫХ ЗНАЧЕНИЙ КЛЮЧЕЙ-ПРЕДИКТОРОВ

В соответствии с гипотезой, сформулированной во введении, необходимо дать прогноз значений ключей-предикторов в третьем сообщении триплета для определения влияния второго сообщения на поведение автора первого сообщения. Так как значения ключей являются вещественными, для решения задачи хорошо подходит многовыходная модификация алгоритма SVR (регрессия опорных векторов). Этот алгоритм позволяет определять нелинейные зависимости данных и хорошо обучается на небольших наборах данных. Техническая реализация алгоритма многоклассового SVR была выполнена с помощью классов SVR и MultiOutputRegressor библиотеки scikit-learn. Для валидации результатов обучения модели исходный набор данных был разделен на тренировочную и валидационную выборки размерами 80% и 20% исходного набора данных соответственно.

Поскольку исходные данные (комментарии с публичных страниц социальных сетей) имеют весьма специфические тематику и направленность, некоторые ключи-предикторы практически не встречаются в наборе данных. С целью выделения наиболее значимых ключей было принято решение о фильтрации набора

данных путем удаления наиболее разреженных столбцов со значениями ключей – предикторов. В качестве порогового значения разреженности было выбрано значение 80%. В результате фильтрации в эксперименте были сохранены ключи-предикторы, приведенные в табл. 1.

Табл. 1. Ключи-предикторы со степенью разреженности ниже 80% и их интерпретации.

Ключ-предиктор	Интерпретация
1.1.b	Коммуникация, Вербальное воздействие, Человек
1.3.a	Коммуникация, Вербальное взаимодействие, Образ Я
1.3.b	Коммуникация, Вербальное взаимодействие, Человек
2.1.b	Познание, Потребление, Человек
2.1.e	Познание, Потребление, Техника
2.2.b	Познание, Репродукция, Человек
2.2.e	Познание, Репродукция, Техника
3.1.b	Регуляция, Управляемый, Человек
3.2.a	Регуляция, Самоуправляемый, Образ Я
3.2.b	Регуляция, Самоуправляемый, Человек
3.3.a	Регуляция, Неуправляемый, Образ Я
3.3.b	Регуляция, Неуправляемый, Человек
Вектор 1	Нарциссизм – Гуманизм
Вектор 2	Психопатия – Вера в человечество
Вектор 3	Макиавеллизм – Кантианство

В результате обучения среднее значение показателя R^2 (коэффициент детерминации, доля дисперсии зависимой переменной, объясняемая рассматриваемой моделью зависимости) по всем выходным ключам составило 0.156. По отдельным ключам, в большей степени выраженным в исходных текстах, значение R^2 достигло отметки 0.52, что говорит о достаточной предсказательной способности отдельных компонентов модели. В табл. 2 представлены показатели R^2 для ключей-предикторов.

Табл. 2. Значения R^2 для параметров, прошедших фильтрацию.

Ключ-предиктор	Значение R^2
1.1.b	0.52
1.3.a	–0.05
1.3.b	0.52
2.1.b	0.04
2.1.e	–0.12
2.2.b	0.26
2.2.e	0.02
3.1.b	–0.59
3.2.a	–0.28
3.2.b	0.36
3.3.a	–0.03
3.3.b	0.27
Вектор 1	0.46
Вектор 2	0.48
Вектор 3	0.48

ОБСУЖДЕНИЕ

Анализ набора данных и результатов обучения модели SVR позволяет сделать следующие выводы.

1. Обученная модель наиболее корректно прогнозирует изменения в следующих ключах-предикторах: «Коммуникация, Вербальное воздействие, Человек» (отражает тот факт, что автор текста старается воздействовать на собеседника), «Коммуникация, Вербальное взаимодействие, Человек» (отражает факт нейтрального диалога между собеседниками), «Регуляция, Самоуправляемый, Человек» (отражает важность самостоятельности принятия решений). В целом это является следствием специфики исходных данных – текстов комментариев с открытых источников данных.

2. Ключи-предикторы, для которых модель показала неудовлетворительные результаты, не в полной мере представлены в текстах триплетов. Для дальнейшего обучения моделей необходимо собрать тексты бесед, содержащие указанные ключи-предикторы.

ЗАКЛЮЧЕНИЕ

Выполнены обучение и валидация многовыходной модели SVR, нацеленной на прогнозирование изменений в выраженности ключей-предикторов в текстах авторов в результате воздействия со стороны других людей. Для ключей предикторов, в достаточной мере выраженных в текстах триплетов, модель показала хорошие результаты. Подтверждено, что успешность обучения моделей машинного обучения сильно зависит от характеристик набора данных.

Дальнейшая работа включает расширение набора данных текстами с более равномерным распределением ключей-предикторов, дальнейшее обучение моделей на новых текстах и интеграцию разработанных моделей в комплексную модель прогнозирования поведения личности.

Благодарности

Исследование выполнено при поддержке гранта Российского научного фонда, проект № 24-18-01125 «Психологическая модель поведения субъекта в цифровой среде», https://rscf.ru/project/24-18-01125/.

СПИСОК ЛИТЕРАТУРЫ

1. *Попов Л.М., Устин П.Н.* Поведение как процессуально-динамическая характеристика субъекта // Психологический журнал. 2025. Т. 46, № 2. С. 24–34. https://doi.org/10.7868/s3034588x25020034

2. *Ломов Б.Ф.* Методологические и теоретические проблемы психологии. Москва: Издательство «Наука», 1984. 1174 с.

3. *Климов Е.А.* Психология профессионального самоопределения. Москва: Издательский центр «Академия», 2010. 304 с.

4. *Кузнецова Ю.М.* Сценарный подход к анализу текстов // Труды Института Системного Анализа Российской Академии Наук. 2018. Т. 68. № 1. С. 31–41. https://doi.org/10.14357/20790279180103

5. *Back M.D., Stopfer J.M., Vazire S., Gaddis S., Schmukle S.C., Egloff B., Gosling S.D.* Facebook profiles reflect actual personality, not self-idealization // Psychol. Sci. 2010. Vol. 21, No. 3. P. 372–374. https://doi.org/10.1177/0956797609360756

6. *Gosling S.D., Gaddis S., Vazire S.* Personality impressions based on Facebook profiles // Proc. Int. AAAI Conf. Weblogs Soc. Media. 2007. Vol. 1, No. 1. P. 1–4.

7. *Quercia D., Kosinski M., Stillwell D., Crowcroft J.* Our Twitter profiles, our selves: Predicting personality with Twitter // Proc. 2011 IEEE Int. Conf. Privacy Secur. Risk Trust. 2011. P. 180–185. https://doi.org/10.1109/PASSAT/SocialCom.2011.26

8. *Golbeck J., Robles C., Edmondson M., Turner K.* Predicting personality from Twitter // Proc. 2011 IEEE Int. Conf. Privacy Secur. Risk Trust. 2011. P. 149–156. https://doi.org/10.1109/PASSAT/SocialCom.2011.33

9. *Youyou W., Kosinski M., Stillwell D.* Computer-based personality judgments are more accurate than those made by humans // Proc. Natl. Acad. Sci. U. S. A. 2015. Vol. 112, No. 4. P. 1036–1040. https://doi.org/10.1073/pnas.1418680112

10. *Kosinski M., Stillwell D., Graepel T.* Private traits and attributes are predictable from digital records of human behavior // Proc. Natl. Acad. Sci. U. S. A. 2013. Vol. 110, No. 15. P. 5802–5805. https://doi.org/10.1073/pnas.1218772110

11. *Pennebaker J.W., King L.A.* Linguistic styles: Language use as an individual difference // J. Pers. Soc. Psychol. 1999. Vol. 77, No. 6. P. 1296–1312. https://doi.org/10.1037//0022-3514.77.6.1296

12. *Kosinski M., Matz S.C., Gosling S.D., Popov V., Stillwell D.* Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines // Am. Psychol. 2015. Vol. 70, No. 6. P. 543–556. https://doi.org/10.17863/CAM.8

13. *Gafarov F.M., Nikolaev K.S., Ustin P.N., Berdnikov A.A., Zakharova V.L., Reznichenko S.A.* A Complex Neural Network Model for Predicting a Personal Success based on their Activity in Social Networks // Eurasia J. Math. Sci. Technol. Educ. 2021. Vol. 17, No. 10. P. 1–9. https://doi.org/10.29333/ejmste/11175

14. *Gafarov F., Berdnikov A., Ustin P.* Predicting success of the online social network user by graph neural networks // Int. J. Adv. Intell. Informatics. 2022. Vol. 8, No. 3. P. 285–298. https://doi.org/10.26555/ijain.v8i3.859

15. *Nikolaev K.S., Gafarov F.M., Ustin P.N.* Metrics for Personal Profiles of Social Network Users // SHS Web Conf. 2020. Vol. 79. P. 1–6. https://doi.org/10.1051/shsconf/20207901012

16. *Drucker H., Burges C.J.C., Kaufman L., Smola A., Vapnik V.* Support vector regression machines // Adv. Neural Inf. Process. Syst. 1997. Vol. 9. P. 155–161.

17. *Breiman L.* Random forests // Mach. Learn. 2001. Vol. 45, No. 1. P. 5–32. https://doi.org/10.1023/A%3A1010933404324

18. *Caruana R., Niculescu-Mizil A.* An empirical comparison of supervised learning algorithms // Proc. 23rd Int. Conf. Mach. Learn. 2006. P. 161–168. https://doi.org/10.1145/1143844.1143865

19. *Fan R.E., Chang K.W., Hsieh C.J., Wang X.R., Lin C.J.* LIBLINEAR: A library for large linear classification // J. Mach. Learn. Res. 2008. Vol. 9. P. 1871–1874. https://doi.org/10.1145/1390681.1442794

PREDICTING CHANGES IN PERSONALITY BEHAVIOR UNDER THE INFLUENCE OF AN INTERLOCUTOR

F. M. Gafarov[1] [0000-0003-4704-154X], P. N. Ustin[2] [0000-0003-3950-7434], K. S. Nikolaev[3] [0000-0003-3204-238X]

[1–3]*Kazan (Volga region) Federal University, Kazan, Russia*

[1]fgafarov@yandex.ru, [2]pavust@mail.ru, [3]konnikolaeff@yandex.ru

Abstract

The study of interactions between users of information systems is an urgent task in humanitarian and technical research. The paper presents the development of a regression model that predicts changes in the severity of the content components of a text under the influence of other people. The data source was short conversations (message triplets) between two users on social networks. The trained model is designed to be used as part of a cognitive architecture that reflects a model of a subject's behavior in a digital environment.

Keywords*: natural language processing, digital behavior, digital identity, machine learning, SVR models.*

REFERENCES

1. *Popov L.M., Ustin P.N.* Povedenie kak processual`no-dinamicheskaya xarakteristika sub``ekta // Psixologicheskij zhurnal. 2025. T. 46, № 2. S. 24–34. https://doi.org/10.7868/s3034588x25020034

2. *Lomov B.F.* Metodologicheskie i teoreticheskie problemy` psixologii. Moskva: Izdatel`stvo “Nauka”, 1984. 1174 s.

3. *Klimov E.A.* Psixologiya professional`nogo samoopredeleniya. Moskva: Izdatel`skij centr “Akademiya”, 2010. 304 s.

4. *Kuzneczova Yu.M.* Scenarny`j podxod k analizu tekstov // Trudy` Instituta Sistemnogo Analiza Rossijskoj Akademii Nauk. 2018. T. 68, № 1. S. 31–41. https://doi.org/10.14357/20790279180103

5. *Back M.D., Stopfer J.M., Vazire S., Gaddis S., Schmukle S.C., Egloff B., Gosling S.D.* Facebook profiles reflect actual personality, not self-idealization // Psychol. Sci. 2010. Vol. 21. No. 3. P. 372–374. https://doi.org/10.1177/0956797609360756

6. *Gosling S.D., Gaddis S., Vazire S.* Personality impressions based on Facebook profiles // Proc. Int. AAAI Conf. Weblogs Soc. Media. 2007. Vol. 1, No. 1. P. 1–4.

7. *Quercia D., Kosinski M., Stillwell D., Crowcroft J.* Our Twitter profiles, our selves: Predicting personality with Twitter // Proc. 2011 IEEE Int. Conf. Privacy Secur. Risk Trust. 2011. P. 180–185. https://doi.org/10.1109/PASSAT/SocialCom.2011.26

8. *Golbeck J., Robles C., Edmondson M., Turner K.* Predicting personality from Twitter // Proc. 2011 IEEE Int. Conf. Privacy Secur. Risk Trust. 2011. P. 149–156. https://doi.org/10.1109/PASSAT/SocialCom.2011.33

9. *Youyou W., Kosinski M., Stillwell D.* Computer-based personality judgments are more accurate than those made by humans // Proc. Natl. Acad. Sci. U. S. A. 2015. Vol. 112, No. 4. P. 1036–1040. https://doi.org/10.1073/pnas.1418680112

10. *Kosinski M., Stillwell D., Graepel T.* Private traits and attributes are predictable from digital records of human behavior // Proc. Natl. Acad. Sci. U. S. A. 2013. Vol. 110, No. 15. P. 5802–5805. https://doi.org/10.1073/pnas.1218772110

11. *Pennebaker J.W., King L.A.* Linguistic styles: Language use as an individual difference // J. Pers. Soc. Psychol. 1999. Vol. 77, No. 6. P. 1296–1312. https://doi.org/10.1037//0022-3514.77.6.1296

12. *Kosinski M., Matz S.C., Gosling S.D., Popov V., Stillwell D.* Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines // Am. Psychol. 2015. Vol. 70, No. 6. P. 543–556. https://doi.org/10.17863/CAM.8

13. *Gafarov F.M., Nikolaev K.S., Ustin P.N., Berdnikov A.A., Zakharova V.L., Reznichenko S.A.* A Complex Neural Network Model for Predicting a Personal Success

based on their Activity in Social Networks // Eurasia J. Math. Sci. Technol. Educ. 2021. Vol. 17, No. 10. P. 1–9. https://doi.org/10.29333/ejmste/11175

14. *Gafarov F., Berdnikov A., Ustin P.* Predicting success of the online social network user by graph neural networks // Int. J. Adv. Intell. Informatics. 2022. Vol. 8, No. 3. P. 285–298. https://doi.org/10.26555/ijain.v8i3.859

15. *Nikolaev K.S., Gafarov F.M., Ustin P.N.* Metrics for Personal Profiles of Social Network Users // SHS Web Conf. 2020. Vol. 79. P. 1–6. https://doi.org/10.1051/shsconf/20207901012

16. *Drucker H., Burges C.J.C., Kaufman L., Smola A., Vapnik V.* Support vector regression machines // Adv. Neural Inf. Process. Syst. 1997. Vol. 9. P. 155–161.

17. *Breiman L.* Random forests // Mach. Learn. 2001. Vol. 45. No. 1. P. 5–32. https://doi.org/10.1023/A%3A1010933404324

18. *Caruana R., Niculescu-Mizil A.* An empirical comparison of supervised learning algorithms // Proc. 23rd Int. Conf. Mach. Learn. 2006. P. 161–168. https://doi.org/10.1145/1143844.1143865

19. *Fan R.E., Chang K.W., Hsieh C.J., Wang X.R., Lin C.J.* LIBLINEAR: A library for large linear classification // J. Mach. Learn. Res. 2008. Vol. 9. P. 1871–1874. https://doi.org/10.1145/1390681.1442794

СВЕДЕНИЯ ОБ АВТОРАХ



ГАФАРОВ Фаиль Мубаракович – кандидат физ.-мат. наук, заведующий кафедрой информационных систем Института Вычислительной математики и информационных технологий, Казанский федеральный университет.

Fail Mubarakovich GAFAROV – Candidate of Science in Physics and Mathematics, head of the Department of Information Systems at the Institute of Computational Mathematics and Information Technologies, Kazan Federal University.

email: fgafarov@yandex.ru

ORCID: 0000-0003-4704-154X.



УСТИН Павел Николаевич – кандидат психологических наук, заведующий кафедрой общей психологии Института психологии и образования, Казанский федеральный университет.

Pavel Nikolaevich USTIN– Candidate of Science in Psychology, head of the Department of General Psychology at the Institute of Psychology and Education, Kazan Federal University.

email: pavust@mail.ru

ORCID: 0000-0003-3950-7434.



НИКОЛАЕВ Константин Сергеевич – к. т. н., доцент кафедры системного анализа и информационных технологий Института Вычислительной математики и информационных технологий Казанского федерального университета.

Konstantin Sergeevich NIKOLAEV – Candidate of Science in Technology, associate professor at the Department of System Analysis and Information Technologies of the Institute of Computational Mathematics and Information Technologies of Kazan Federal University.

email: konnikolaeff@yandex.ru

ORCID: 0000-0003-3204-238X

Материал поступил в редакцию 30 июня 2026 года

УДК 004.4

УПРАВЛЕНИЕ И ОРГАНИЗАЦИЯ МЕТАДАННЫХ НАУЧНЫХ СТАТЕЙ В ЦИФРОВЫХ МАТЕМАТИЧЕСКИХ БИБЛИОТЕКАХ

П. О. Гафурова[0000-0002-1544-155X]

НИЦ «Курчатовский институт», г. Москва, Россия

pogafurova@gmail.com

Аннотация

Представлены основные подходы к созданию систем управления метаданными в цифровых математических библиотеках, цифровых платформах и исследовательских инфраструктурах. Описаны способы представления метаданных в цифровых библиотеках и подходы к редактированию метаданных, принятые в различных цифровых математических проектах. Предложен редактор метаданных документов цифровых научных коллекций, разработанный как основной компонент фабрики метаданных цифровой математической библиотеки Lobachevskii-DML.

Ключевые слова*: цифровая библиотека, цифровая математическая библиотека, математическая исследовательская инфраструктура, фабрика метаданных, метаданные научной статьи, редактор метаданных цифровой библиотеки, Lobachevskii-DML.*

ВВЕДЕНИЕ

Несмотря на то что активное развитие проекта создания Всемирной цифровой математической библиотеки (WDML) приостановилось в 2013 г. [1, 2], сформированные в его рамках локальные цифровые библиотеки нуждаются в поддержке и формировании современных методов доступа к метаданным математических документов [3, 4]. Метаданные являются структурированным описанием документа, и на их основе реализуются методы навигации в пространстве научных знаний (см., например, [5]). При большом количестве цифровых документов и самих коллекций неминуемы ошибки в их обработке, причем не только со стороны программного обеспечения, но и со стороны авторов или редакторов статей. Сложная структура цифровых библиотек и большие объемы

© П. О. Гафурова, 2026.
Данная статья распространяется на условиях международной лицензии Creative Commons License Attribution 4.0 International (CC BY 4.0).

данных требуют создания программных средств добавления, изменения и удаления метаданных цифровых документов и коллекций. В настоящей статье рассмотрена задача создания редактора метаданных документов на примере цифровой математической библиотеки Lobachevskii-DML [6–9]. Отметим, что один из методов пакетного пополнения метаданных документов цифровых коллекций этой библиотеки был предложен в [10].

Подходы к созданию средств пополнения и исправления ошибок в метаданных представлены в различных цифровых библиотеках, однако эти методы и средства были созданы с учетом локальных особенностей соответствующих проектов и форматов представления документов, включенных в эти цифровые коллекции. По этой причине актуальна задача создания собственных методов пополнения и исправления метаданных в библиотеке Lobachevskii-DML, включающей разнообразные коллекции цифровых документов [11–13]. Специфика документов библиотеки Lobachevskii-DML такова:

1) основной язык описания метаданных – русский;

2) имеются дополнительные метаданные на других языках (русский с дореформенной орфографией, английский, французский, немецкий);

3) имеются математические символы.

В связи с достаточно сложной структурой библиотеки Lobachevskii-DML необходимо создание не только методов формирования метаданных для загрузки в библиотеку, но и методов пополнения и исправления этих метаданных [9].

ПОДХОДЫ К ПРЕДСТАВЛЕНИЮ И ИЗМЕНЕНИЮ МЕТАДАННЫХ ДОКУМЕНТОВ ЦИФРОВЫХ МАТЕМАТИЧЕСКИХ БИБЛИОТЕК

Исторически сложилось так, что метаданные научных документов могут быть представлены в различных форматах [14]. Из всего объема цифровых форматов для описания документов в цифровых библиотеках использован определенный набор метаданных [15]. Кроме того, основной набор метаданных документов цифровых математических библиотек достаточно ограничен. Поскольку все цифровые проекты были сформированы с учетом будущего включения в глобальный проект, каждая из таких библиотек использует специализированный xml-формат, который отражает основные особенности библиотеки, а также

средства перевода в общепринятый формат. Комплекс методов исправления и внедрения метаданных в цифровые библиотеки предложен в [16, 17] для цифровой математической библиотеки NumDam, созданной для поддержки математического контента в основном на французском языке (но поддерживает и другие языки). В этой библиотеке содержатся полные тексты документов по математическим исследованиям: журнальные статьи, материалы семинаров и конференций, книги и монографии, диссертации. В [17] был введен термин «фабрика метаданных» и описана структура фабрики метаданных библиотеки NumDam. Последняя включает сервисы нормализации метаданных, их корректировки и улучшения, а также инструменты работы с формулами в TEX- и MathML-форматах. На рис. 1 представлен редактор метаданных библиотеки NumDam.

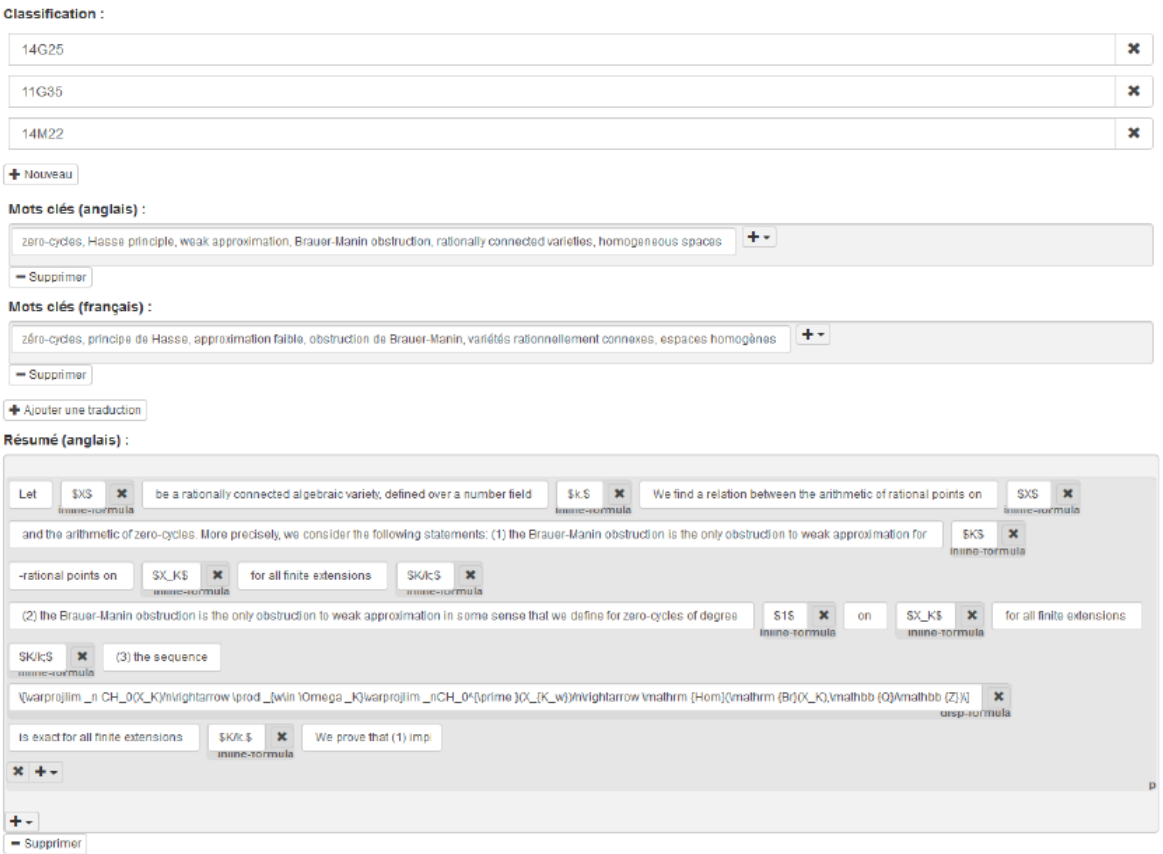


Рис. 1. Фрагмент редактора метаданных фабрики метаданных цифровой библиотеки NumDam.

Определение фабрики метаданных в контексте цифровых платформ и цифровых научных библиотек предложено в [18, 19].

Работы по созданию редактора метаданных были проведены на базе цифровой библиотеки DML-CZ [20], которая ориентирована на цифровые документы на чешском языке. Центральным элементом этой цифровой библиотеки является редактор метаданных [21, 22]. На рис. 2 представлен редактор метаданных библиотеки DML-CZ. Подробное описание этой библиотеки содержится в [21, 22].

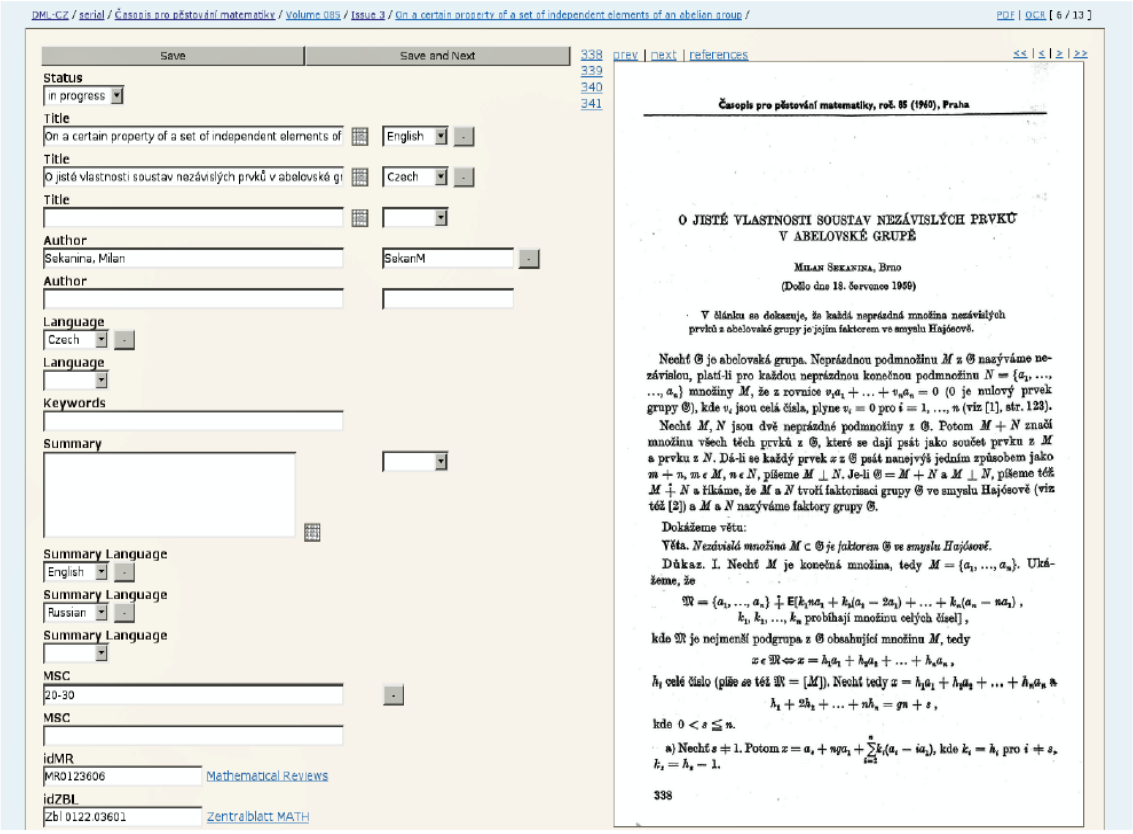


Рис. 2. Редактор метаданных библиотеки DML-CZ.

Цифровые библиотеки, указанные выше, а также разработанные в них методы формирования контента и метаданных были распространены и на другие библиотеки [23–25].

Отметим, что цифровые математические библиотеки могут иметь различную структуру. Так, библиотека NumDam имеет древовидную структуру (см. рис. 3). На этом рисунке «листья» (leaf) являются статьями или частями

книги, «контейнер» (container) – выпуском журнала или книгой, «коллекция» (collection) – журналом, серией книг или многотомным изданием. Хранение метаданных документов в библиотеках DML-CZ и BulDML (Bulgarian Digital Mathematical Library [26]) организовано в виде реляционных таблиц PostgreSQL и Oracle [25]. В библиотеке Lobachevskii-DML хранение метаданных – в виде реляционных таблицах MySQL [13]. На рис. 4 представлены реляционные таблицы Lobachevskii-DML.

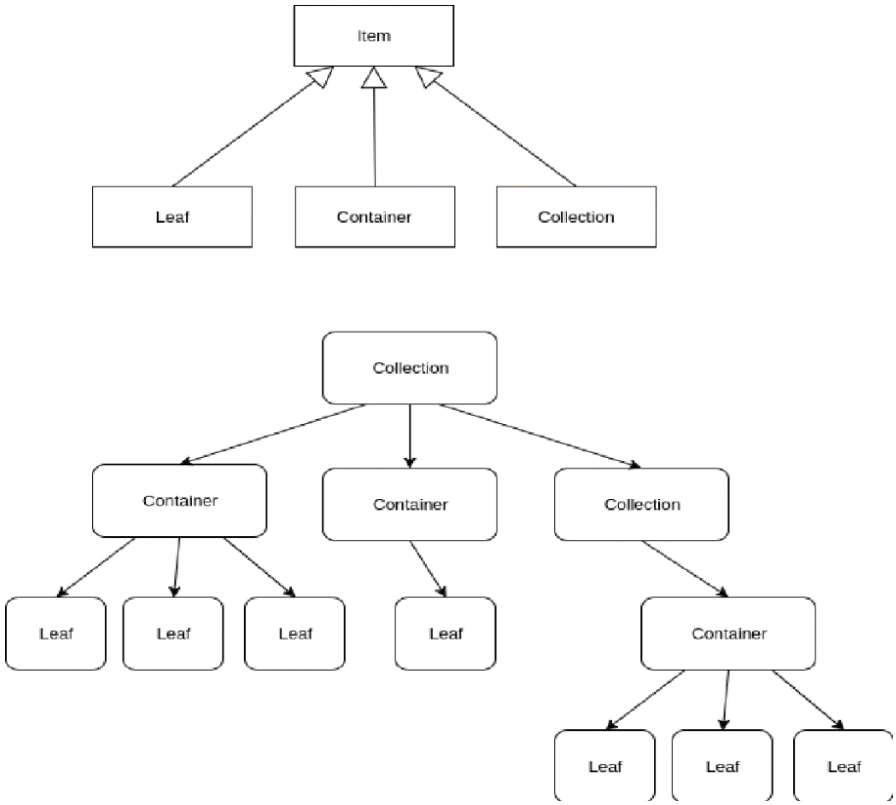


Рис. 3. Структура контента цифровой библиотеки NumDam [17].

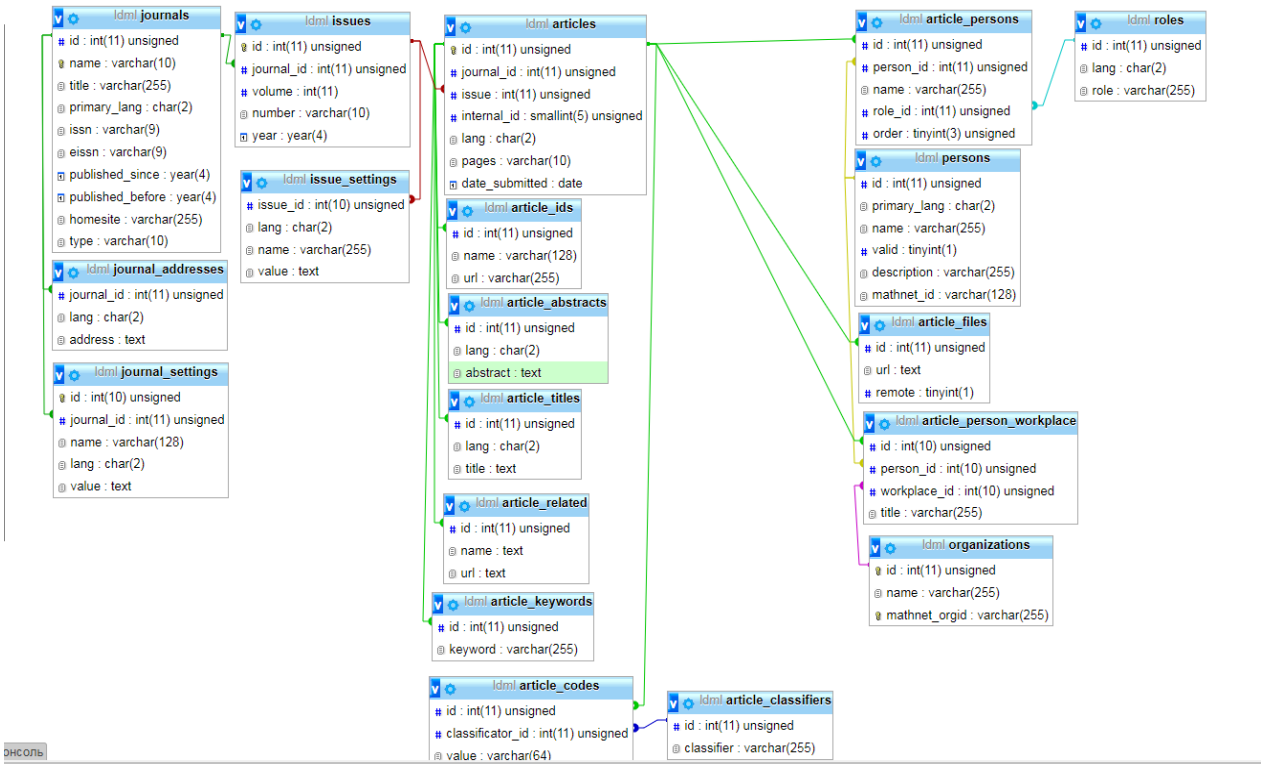


Рис. 4. Структура реляционных таблиц, описывающих метаданные документа цифровой библиотеки Lobachevskii-DML.

Так как внесение исправлений в сложную структуру таблиц в библиотеке Lobachevskii-DML требует больших трудозатрат, а методы внесения изменений в метаданные достаточно быстро устаревают со временем, необходимо создание редактора метаданных для данной библиотеки – программного средства, дающего возможности изменения, добавления и удаления метаданных документов цифровой библиотеки.

Важно отметить, что подходы к формированию метаданных не стоят на месте и постоянно совершенствуются. Сегодня основным форматом представления научных данных является Journal Archiving and Interchange Tag Library (NISO JATS) [28], который был разработан как раз после формулирования основных особенностей цифровых математических библиотеки.

РЕДАКТОР МЕТАДАННЫХ ДЛЯ ЦИФРОВОЙ МАТЕМАТИЧЕСКОЙ БИБЛИОТЕКИ

В настоящее время система формирования метаданных также претерпевает изменения. В частности, на платформе Crosref [29] метаданные статей собираются из различных источников (см. рис. 5) [30–37]. При этом специфика математических документов никак не учитывается.

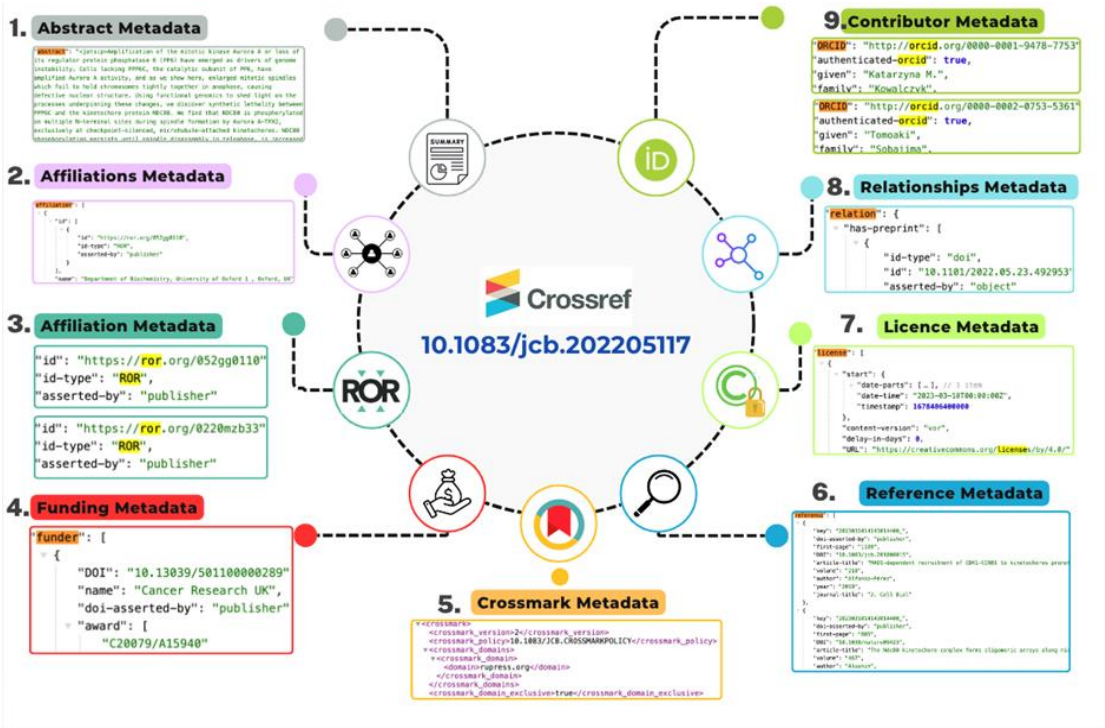


Рис. 5. Метаданные о статье и их источники [37].

Для внесения исправлений в обе системы необходимы средства внесения изменений в метаданные цифровых статей, т. е. редактор метаданных документов цифровых библиотек. Основной целью создания такого редактора является создание среды, в которой пользователь имеет возможность мануального изменения, добавления, удаления метаданных на конкретной странице [38, 39].

Далее рассмотрим редактор метаданных документов цифровой математической библиотеки Lobachevskii-DML. Такой редактор должен обладать следующими возможностями:

- добавление новых метаданных, потому что наборы метаданных цифровых документов расширяются с каждым годом;

- поддержка разных языков и метаданных на этих языках;
- обеспечение читаемости математического и информационного контента библиотеки;
- простота использования и удобства интерфейса для пользователей;
- доступ к базе данных и защиты библиотеки от несанкционированного доступа.

Обратимся к структуре Lobachevskii-DML (рис. 2). Сложная структура библиотеки затрудняет внесение метаданных в систему, поэтому для внесения исправлений и защиты метаданных необходимо создание временной таблицы (табл. 1).

Табл. 1. Структура временной таблицы цифровой библиотеки.

Название	Описание
id	Идентификатор
type	Тип метаданных, которые должны быть изменены
operation	Удаление (D), вставка( ), изменение (R)
old_meta	Прошлые метаданных
new_meta	Новые метаданные

Анализируя структуру таблиц Lobachevskii-DML (рис. 4), необходимо отметить, что такие метаданные, как ключевые слова, аннотация, название, могут быть изменены при обращении к таблице по идентификатору статьи. К метаданным, связанным с авторами (фамилия, имя, отчество, место работы, название организации, роли авторов), доступ производится через комплекс реляционных таблиц. Основные метаданные статьи приведены в табл. 2. Таким образом, необходимо создание двух ролей «редактор» и «эксперт-администратор» со следующими функциями:

1) роль «редактор» библиотеки Lobachevskii-DML – внесение изменений и дополнений в метаданные цифровых документов;

2) роль «эксперт-администратор» – проверка корректности измененных метаданных и их загрузка в библиотеку.

Табл. 2. Метаданные статьи и их расположение в реляционных таблицах библиотеки Lobachevskii-DML.

Метаданные	Название в реляционной таблице	Количество таблиц, которые необходимо проверить для доступа к данным
название	title из таблицы article_titles	2
фамилия автора	name из persons	3
имя автора	name из persons	3
отчество автора	name из persons	3
аффилиация автора	name из organizations	3
аннотация	abstract из article_abstract	3
ключевые слова	keyword из article_keywords	3
классификаторы	value из article_codes	3
страницы	pages из articles	1

На рис. 6 представлен фрагмент формы прототипа редактора метаданных для цифровой библиотеки Lobachevskii-DML. Основные подключаемые метаданные приведены в табл. 1.

Опишем подробнее процедуру реализации роли «редактор» (рис. 6):

1) вход со страницы конкретной статьи в режим изменения метаданных; открывается форма с полями для изменения метаданных этой статьи;

2) те метаданные, которые должны быть изменены, изменяются «редактором» и отправляются на структуризацию;

3) происходит структуризация метаданных для записи в таблицу и формирования JATS-представления метаданных;

4) осуществляются записи метаданных в таблицу MySQL и в файл JATS.

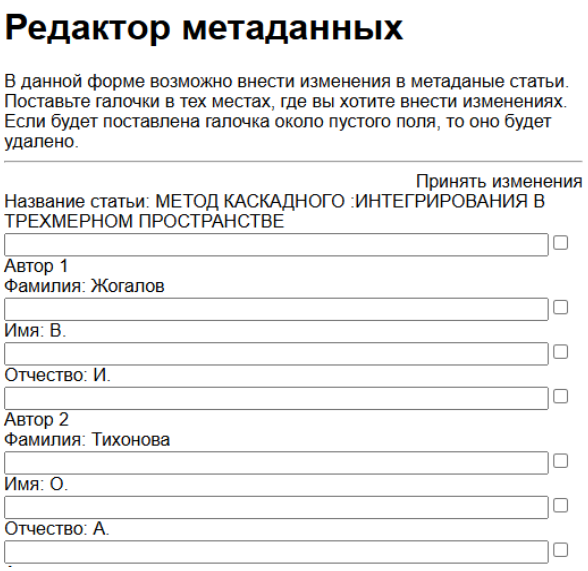


Рис. 6. Форма редактора метаданных для цифровой библиотеки Lobachevskii-DML.

Процедура реализации роли «эксперт-администратор» включает следующие функции:

1) «эксперт-администратор» получает форму для принятия или отклонения изменений в метаданных;

2) «эксперт-администратор» отправляет валидированные метаданные на обработку;

3) приложение обеспечивает структуризацию и создание запросов загрузки метаданных на сервер.

В итоге мы получим систему, полностью закрытую для пользователей, которая позволит изменять и дополнять метаданные цифровых документов в библиотеке.

Важной структурной особенностью описанного процесса является то, что необходимо избежать дублирования в таблицах organizations и persons, поскольку эти таблицы сформированы так, что один автор может быть привязан ко многим статьям. Решению этой задачи будет посвящен следующий этап исследований.

ЗАКЛЮЧЕНИЕ

Сделан обзор редакторов метаданных для различных цифровых математических библиотек, предложено решение для редактирования метаданных в

цифровой математической библиотеке Lobachevskii-DML. Созданы методы мануального управления метаданными цифровых математических библиотек. Эти методы рассчитаны на русскоязычные тексты и могут быть использованы в других информационных системах.

Дальнейшая работа направлена на расширение набора изменяемых метаданных в цифровых библиотеках, разработку методов избегания дублирования и внесение изменений в структуру цифровой библиотеки в соответствии с современными стандартами представления научных данных.

Благодарности

Работа выполнена в рамках государственного задания НИЦ «Курчатовский институт».

СПИСОК ЛИТЕРАТУРЫ

1. DML: Digital Mathematics Library Also: WDML: World Digital Mathematics Library URL: https://www.math.uni-bielefeld.de/~rehmann/DML/dml_links.html

2. Developing a 21st Century Global Library for Mathematics Research. – Washington, DC: The National Academies Press. https://doi.org/10.17226/18619

3. *Borwein J.M., Rocha E.M., Rodrigues J.F.* Communicating Mathematics in the Digital Era. A K Peters/CRC Press, 2008. 325 p. https://doi.org/10.1201/b10587

4. *Elizarov A.M., Lipachev E.K., Zuev D.S.* Digital Mathematical Libraries: Overview of Implementations and Content Management Services // CEUR Workshop Proceedings. 2017. Vol. 2022. P. 317–325.

5. *Gartner R.* Metadata. Shaping Knowledge from Antiquity to the Semantic Web. Springer Cham, 2016. https://doi.org/10.1007/978-3-319-40893-4

6. Digital Mathematics Library Lobachevskii-DML. URL: lobachevskii-dml.ru

7. *Elizarov A.M., Lipachev E.K.* Lobachevskii DML: Towards a Semantic Digital Mathematical Library of Kazan University // CEUR Workshop Proc. 2017. Vol.2022. P.326–333.

8. *Elizarov A, Lipachev E.* Big math methods in Lobachevskii-DML digital library // CEUR Workshop Proc. 2019. Vol. 2523. P.59–72.

9. *Elizarov A.M., Lipachev E.K.* Lobachevskii digital library in the scientific space of mathematical knowledge // Scientific and Technical Information Processing. 2023. Vol. 50, No. 1. P. 35–39. https://doi.org/10.3103/s0147688223010021

10. *Gafurova P.O.* Automatic Replenishment of Metadata of Digital Publications Using Semantic Services of the Internet // Russian Digital Libraries Journal. 2024. Vol. 27, No. 2. P. 164–186. https://doi.org 10.26907/1562-5419-2024-27-2-164-186

11. *Gafurova P., Elizarov A., Lipachev E.* Basic services of factory metadata digital mathematical library Lobachevskii-DML // Russian Digital Libraries Journal. 2020. Vol. 23, No. 3. P. 336–381. https://doi.org/10.26907/1562-5419-2020-23-3-336-381

12. Elizarov A., Gafurova P., Lipachev E. Wikidata in Metadata Formation Methods for Documents of Digital Mathematical Library // CEUR Workshop Proc. 2021. Vol. 3066. P. 23–33.

13. Gafurova P.O. Semanticheskie metody i instrumenty fabriki metadannyh cifrovoj matematicheskoj biblioteki: dis. … kand. tekhn. nauk 2.3.8 / Gafurova P.O. – Kazan', 2024. 148 s.

14. Xie I., Matusiak K. K. Discover Digital Libraries: Theory and Practice – Elsevier Inc., 2016. – 388 p.

15. Antopolsky A.B. Metadata Systems for Electronic Libraries*//* Mezhdunarodnaya konferenciya Krym-2001. URL: https://www.gpntb.ru/win/inter-events/crimea2001/tom/sec4/doc5.html (дата доступа 15.07.2026).

16. Numdam. URL: http://www.numdam.org/) (дата доступа 15.07.2026).

17. *Bouche T., Labe O.* The New Numdam Platform // In: Geuvers H., England M., Hasan O., Rabe F., Teschke O. (Eds) Intelligent Computer Mathematics. CICM 2017. Lecture Notes in Computer Science. Vol. 10383. Springer, Cham, 2017. P. 70–82. https://doi.org/10.1007/978-3-319-62075-6_6

18. *Елизаров А.М., Липачев Е.К.* Цифровые платформы и цифровые научные библиотеки // International Journal of Open Information Technologies. 2020. Т. 8, № 11. С. 80–90. https://doi.org/10.25559/injoit.2307-8162.08.202011.80-90

19. *Elizarov A., Lipachev E*. Digital Library Metadata Factories // CEUR Workshop Proc. 2021. Vol. 2813. P. 13–21.

20. Czech Digital Mathematics Library. URL: http://dml.cz/ (дата доступа 15.07.2026).

21. *Bartošek M., Kovář P., Šárfy M.* DML-CZ Metadata Editor // Towards Digital Mathematics Library. Birmingham, United Kingdom Masaryk University, Brno. 2008. P. 139–151. URL: http://dml.cz/dmlcz/702537 (дата доступа 15.07.2026).

22. *Bartošek M., Rákosník J.* DML-CZ: The Experienceof a Medium-Sized Digital Mathematics Library // Notices of the AMS. 2013. Vol. 60, No. 8. P. 1028–1033. http://dx.doi.org/10.1090/noti1031

23. MDML: The Mathdoc Digital Mathematics Library. URL: http://dml.mathdoc.fr/ (дата доступа 15.07.2026).

24. *Bouche T., Goutorbe C., Jorda J.-P., Jost M.* The EuDML Metadata Schema: Version 1.0. // Towards a Digital Mathematics Library, Jul 2011, Bertinoro, Italy. P. 45–61. URL: https://hal.univ-grenoble-alpes.fr/hal-03765892/file/D3.6.pdf (дата доступа 15.07.2026).

25. *Rakosnik J., Stanchev P., Simeonov G., Pavlov R.* Bulgarian Digital Mathematical Library BulDML and Czech Digital Mathematical Library DML-CZ as parts of the European Digital Mathematics Library EuDML // Digital Presentation and Preservation of Cultural and Scientific Heritage. 2012. P. 60–68. https://doi.org/10.55630/dipp.2012.2.27

26. Bulgarian Digital Mathematics Library (BulDML). URL: https://buldml.math.bas.bg/ (дата доступа 15.07.2026).

27. *Bouche T., Rehmann U.* Digital Mathematics Library Report of the Technical Standards Working Group. URL: https://www.math.uni-bielefeld.de/~rehmann/DML/dml_standards_fin.pdf (дата доступа 15.07.2026).

28. Journal Archiving and Interchange Tag Library NISO JATS Version 1.3d1. URL: https://jats.nlm.nih.gov/archiving/tag-library/1.3d1/chapter/how-to-read.html (дата доступа 15.07.2026).

29. Crossref. URL: https://www.crossref.org/ (дата доступа 15.07.2026).

30. *Gafurova P.O., Lipachev E.K.* Method for Clarifying the Affiliation of Authors of Scientific Documents Based on Requests to the Semantic Web// Scientific Services and Internet, SSI-2022. M.: IPM im. M.V. Keldysha, 2022. P. 115–127.

31. *Elizarov A.M., Zaitseva N., Zuev D.S., Lipachev E.K., Khajdarov Sh.M.* Services for Formation of Digital Documents Metadata in the Formats of International Science-Based Databases // CEUR Workshop Proc. 2018. Vol. 2260. P. 175–185.

32. *Andreichev M.D., Gafurova P.O., Elizarov A.M., Lipachev E.K.* Replenishment of documents of mathematical digital retro-collections by searching in semantic web// Scientific Services and Internet, SSI-2021. M.: IPM im. M.V.Keldysha, 2021. P. 22–33. https://doi.org/10.20948/abrau-2021-22

33. *Elizarov A., Gafurova P., Lipachev E.* Wikidata in Metadata Formation Methods for Documents of Digital Mathematical Library // CEUR Workshop Proc. 2021. Vol. 3066. P. 23–33.

34. *Apanovich Z.V.* Information about russian research organizations in multilingual data sources // Russian Digital Libraries Journal. 2021. Vol. 24, No. 5. P. 756–769. https://doi.org/10.26907/1562-5419-2021-24-5-756-769

35. Gafurova P.O. Automatic replenishment of metadata of digital publications using semantic services of the Internet // Scientific Services and Internet, SSI-2023. M.: IPM im. M.V.Keldysha, 2023. Vol. 25. P. 84–93. https://doi.org/10.20948/abrau-2023-27

36. *Elizarov A., Gafurova P., Lipachev E.* Algorithms for Integration of Unstructured Mathematical Documents into the Common Digital Space of Scientific Knowledge// CEUR Workshop Proc. 2021. Vol. 2990. P.39–49. URL: https://ceur-ws.org/Vol-2990/rpaper4.pdf (дата доступа 15.07.2026).

37. *Amdekar M.S.* Scholarly Metadata as Trust Signals: Opportunities for Journal Editors // Amdekar MS. Scholarly metadata as trust signals: opportunities for journal editors. Sci. Ed. 2024. No. 47. P. 130 – 133. URL: https://doi.org/10.36591/SE-4704-10. URL: https://www.csescienceeditor.org/article/scholarly-metadata-as-trust-signals-opportunities-for-journal-editors/ (дата доступа 15.07.2026).

38. Articulus Instruction. URL: https://vniigis.ru/1_dlya_failov/Help/Инструкция по работе с программой Articulus eLibrary НЭБ%20РИНЦ.pdf (дата доступа 15.07.2026).

39. *Bouche T., Goutorbe C., Jorda J.-P., Jost M.* The EuDML Metadata Schema: Version 1.0. // Towards a Digital Mathematics Library, Jul 2011, Bertinoro, Italy. P. 45–61. URL: https://hal.univ-grenoble-alpes.fr/hal-03765892/file/D3.6.pdf (дата доступа 15.07.2026).

ORGANIZATION AND MANAGEMENT METADATA OF SCIENTIFIC ARTICLES IN DIGITAL MATHEMATICAL LIBRARIES

P. O. Gafurova[0000-0002-1544-155X]

National Research Center "Kurchatov Institute", Moscow, Russia

pogafurova@gmail.com

Abstract

This paper presents the main approaches to creating metadata management systems for digital mathematics libraries. It describes approaches to editing metadata adopted in various digital mathematics projects. Methods for representing metadata in digital libraries are presented. A metadata editor for digital mathematics collections is proposed. Approaches to adding, editing, and deleting article metadata in a digital library are described, and a method for editing metadata in a digital library is proposed.

Keywords*: digital library, digital mathematics library, mathematical research infrastructure, metadata factory, research article metadata, digital library metadata editor, Lobachevskii-DML.*

REFERENCES

1. DML: Digital Mathematics Library Also: WDML: World Digital Mathematics Library URL: https://www.math.uni-bielefeld.de/~rehmann/DML/dml_links.html

2. Developing a 21st Century Global Library for Mathematics Research. – Washington, DC: The National Academies Press. https://doi.org/10.17226/18619

3. *Borwein J.M., Rocha E.M., Rodrigues J.F.* Communicating Mathematics in the Digital Era. A K Peters/CRC Press, 2008. 325 p. https://doi.org/10.1201/b10587

4. *Elizarov A.M., Lipachev E.K., Zuev D.S.* Digital Mathematical Libraries: Overview of Implementations and Content Management Services // CEUR Workshop Proceedings. 2017. Vol. 2022. P. 317–325.

5. *Gartner R.* Metadata. Shaping Knowledge from Antiquity to the Semantic Web. Springer Cham, 2016. https://doi.org/10.1007/978-3-319-40893-4

6. Digital Mathematics Library Lobachevskii-DML. URL: lobachevskii-dml.ru

7. *Elizarov A.M., Lipachev E.K.* Lobachevskii DML: Towards a Semantic Digital Mathematical Library of Kazan University // CEUR Workshop Proc. 2017. Vol.2022. P.326–333.

8. *Elizarov A, Lipachev E.* Big math methods in Lobachevskii-DML digital library // CEUR Workshop Proc. 2019. Vol. 2523. P.59–72.

9. *Elizarov A.M., Lipachev E.K.* Lobachevskii digital library in the scientific space of mathematical knowledge // Scientific and Technical Information Processing. 2023. Vol. 50, No. 1. P. 35–39. https://doi.org/10.3103/s0147688223010021

10. *Gafurova P.O.* Automatic Replenishment of Metadata of Digital Publications Using Semantic Services of the Internet // Russian Digital Libraries Journal. 2024. Vol. 27, No. 2. P. 164–186. https://doi.org 10.26907/1562-5419-2024-27-2-164-186

11. *Gafurova P., Elizarov A., Lipachev E.* Basic services of factory metadata digital mathematical library Lobachevskii-DML // Russian Digital Libraries Journal. 2020. Vol. 23, No. 3. P. 336–381. https://doi.org/10.26907/1562-5419-2020-23-3-336-381

12. *Elizarov A., Gafurova P., Lipachev E.* Wikidata in Metadata Formation Methods for Documents of Digital Mathematical Library // CEUR Workshop Proc. 2021. Vol. 3066. P. 23–33.

13. *Gafurova P.O.* Semanticheskie metody i instrumenty fabriki metadannyh cifrovoj matematicheskoj biblioteki: dis. … kand. tekhn. nauk 2.3.8 / Gafurova P.O. – Kazan', 2024. 148 s.

14. *Xie I., Matusiak K. K.* Discover Digital Libraries: Theory and Practice – Elsevier Inc., 2016. – 388 p.

15. *Antopolsky A.B.* Metadata Systems for Electronic Libraries// Mezhdunarodnaya konferenciya Krym-2001. URL: https://www.gpntb.ru/win/inter-events/crimea2001/tom/sec4/doc5.html (accessed: July 15, 2026).

16. Numdam. URL: http://www.numdam.org/) (accessed: July 15, 2026).

17. *Bouche T., Labe O.* The New Numdam Platform // In: Geuvers H., England M., Hasan O., Rabe F., Teschke O. (Eds) Intelligent Computer Mathematics. CICM 2017. Lecture Notes in Computer Science. Vol. 10383. Springer, Cham, 2017. P. 70–82. https://doi.org/10.1007/978-3-319-62075-6_6

18. *Elizarov A.M., Lipachev E.K.* Cifrovye platformy i cifrovye nauchnye biblioteki // International Journal of Open Information Technologies. 2020. T. 8, № 11. C. 80–90. https://doi.org/10.25559/injoit.2307-8162.08.202011.80-90

19. *Elizarov A., Lipachev E*. Digital Library Metadata Factories // CEUR Workshop Proc. 2021. Vol. 2813. P. 13–21.

20. Czech Digital Mathematics Library. URL: http://dml.cz/ (дата доступа 15.07.2026).

21. *Bartošek M., Kovář P., Šárfy M.* DML-CZ Metadata Editor // Towards Digital Mathematics Library. Birmingham, United Kingdom Masaryk University, Brno. 2008. P. 139–151. URL: http://dml.cz/dmlcz/702537 (accessed: July 15, 2026).

22. *Bartošek M., Rákosník J.* DML-CZ: The Experienceof a Medium-Sized Digital Mathematics Library // Notices of the AMS. 2013. Vol. 60, No. 8. P. 1028–1033. http://dx.doi.org/10.1090/noti1031

23. MDML: The Mathdoc Digital Mathematics Library. URL: http://dml.mathdoc.fr/ (accessed: July 15, 2026).

24. *Bouche T., Goutorbe C., Jorda J.-P., Jost M.* The EuDML Metadata Schema: Version 1.0. // Towards a Digital Mathematics Library, Jul 2011, Bertinoro, Italy. P. 45–61. URL: https://hal.univ-grenoble-alpes.fr/hal-03765892/file/D3.6.pdf (accessed: July 15, 2026).

25. *Rakosnik J., Stanchev P., Simeonov G., Pavlov R.* Bulgarian Digital Mathematical Library BulDML and Czech Digital Mathematical Library DML-CZ as parts of the European Digital Mathematics Library EuDML // Digital Presentation and Preservation of Cultural and Scientific Heritage. 2012. P. 60–68. https://doi.org/10.55630/dipp.2012.2.27

26. Bulgarian Digital Mathematics Library (BulDML). URL: https://buldml.math.bas.bg/ (accessed: July 15, 2026).

27. *Bouche T., Rehmann U.* Digital Mathematics Library Report of the Technical Standards Working Group. URL: https://www.math.uni-bielefeld.de/~rehmann/DML/dml_standards_fin.pdf (дата доступа 15.09.2026).

28. Journal Archiving and Interchange Tag Library NISO JATS Version 1.3d1. URL: https://jats.nlm.nih.gov/archiving/tag-library/1.3d1/chapter/how-to-read.html (дата доступа 15.09.2026).

29. Crossref. URL: https://www.crossref.org/ (accessed: July 15, 2026).

30. *Gafurova P.O., Lipachev E.K.* Method for Clarifying the Affiliation of Authors of Scientific Documents Based on Requests to the Semantic Web// Scientific Services and Internet, SSI-2022. M.: IPM im. M.V. Keldysha, 2022. P. 115–127.

31. *Elizarov A.M., Zaitseva N., Zuev D.S., Lipachev E.K., Khajdarov Sh.M.* Services for Formation of Digital Documents Metadata in the Formats of International Science-Based Databases // CEUR Workshop Proc. 2018. Vol. 2260. P. 175–185.

32. *Andreichev M.D., Gafurova P.O., Elizarov A.M., Lipachev E.K.* Replenishment of documents of mathematical digital retro-collections by searching in semantic web// Scientific Services and Internet, SSI-2021. M.: IPM im. M.V.Keldysha, 2021. P. 22–33. https://doi.org/10.20948/abrau-2021-22

33. *Elizarov A., Gafurova P., Lipachev E.* Wikidata in Metadata Formation Methods for Documents of Digital Mathematical Library // CEUR Workshop Proc. 2021. Vol. 3066. P. 23–33.

34. *Apanovich Z.V.* Information about russian research organizations in multilingual data sources // Russian Digital Libraries Journal. 2021. Vol. 24, No. 5. P. 756–769. https://doi.org/10.26907/1562-5419-2021-24-5-756-769

35. *Gafurova P.O.* Automatic replenishment of metadata of digital publications using semantic services of the Internet // Scientific Services and Internet, SSI-2023. M.: IPM im. M.V.Keldysha, 2023. Vol. 25. P. 84–93. https://doi.org/10.20948/abrau-2023-27

36. *Elizarov A., Gafurova P., Lipachev E.* Algorithms for Integration of Unstructured Mathematical Documents into the Common Digital Space of Scientific Knowledge// CEUR Workshop Proc. 2021. Vol. 2990. P.39–49. URL: https://ceur-ws.org/Vol-2990/rpaper4.pdf (accessed: July 15, 2026).

37. *Amdekar M.S.* Scholarly Metadata as Trust Signals: Opportunities for Journal Editors // Amdekar MS. Scholarly metadata as trust signals: opportunities for journal editors. Sci. Ed. 2024. No. 47. P. 130 – 133. URL: https://doi.org/10.36591/SE-4704-10. URL: https://www.csescienceeditor.org/article/scholarly-metadata-as-trust-signals-opportunities-for-journal-editors/ (accessed: July 15, 2026).

38. Articulus Instruction. URL: https://vniigis.ru/1_dlya_failov/Help/Инструкция по работе с программой Articulus eLibrary НЭБ%20РИНЦ.pdf (accessed: July 15, 2026).

39. *Bouche T., Goutorbe C., Jorda J.-P., Jost M.* The EuDML Metadata Schema: Version 1.0. // Towards a Digital Mathematics Library, Jul 2011, Bertinoro, Italy. P. 45–61. URL: https://hal.univ-grenoble-alpes.fr/hal-03765892/file/D3.6.pdf (accessed: July 15, 2026).

СВЕДЕНИЯ ОБ АВТОРЕ



ГАФУРОВА Полина Олеговна – младший научный сотрудник Лаборатории суперкомпьютерного моделирования Национального исследовательского центра «Курчатовский институт».

Polina Olegovna GAFUROVA – junior researcher at the Supercomputer Modeling Laboratory of the National Research Center «Kurchatov Institute».

email: pogafurova@gmail.com

ORCID: 0000-0002-1544-155X

Материал поступил в редакцию 20 августа 2026 года

УДК 004.8:81'32:811.512.145

БАЗА СЕМРОНОВ ТАТАРСКОГО ЯЗЫКА КАК ЦИФРОВОЙ ЛИНГВИСТИЧЕСКИЙ РЕСУРС ДЛЯ ОБЪЯСНИТЕЛЬНОГО ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Р. А. Гильмуллин[0000-0002-8520-8921]

Академия наук Республики Татарстан, г. Казань, Россия
rinatgilmullin@gmail.com

Аннотация

Рассмотрена база семронов татарского языка как цифровой лингвистический ресурс для объяснительного искусственного интеллекта. Предложена модель Большой семиотической базы семронов, в которой лексико-грамматические единицы представлены как семроны – потенциальные смыслообразующие единицы с формой, классом, синтаксическими, семантическими и прагматическими характеристиками, валентностями, правилами активации, корпусными подтверждениями, связями с графом знаний и статусами экспертной проверки. Фреймовое представление использовано как способ организации записи семрона: базовый фрейм фиксирует потенциальные возможности единицы, а специализированный фрейм сейрона описывает ее контекстную реализацию после закрытия одной или нескольких валентностей. Показано, что такая база может выполнять функцию ресурсного слоя Семиотической генеративной распознающей модели, обеспечивая переход от морфемной структуры татарского текста к ситуационной модели, верифицируемой цепочке и объяснению. Компактные примеры записей демонстрируют структуру фрейма, типы слотов и фасетов, принципы верификации и возможные форматы представления данных.

Ключевые слова*: семрон, база семронов, большая семиотическая база семронов, татарский язык, цифровой лингвистический ресурс, фреймовая модель, валентность, сейрон, граф знаний, корпус, экспертная верификация, объяснительный искусственный интеллект, семиотическая генеративная распознающая модель.*

© Р. А. Гильмуллин, 2026.
Данная статья распространяется на условиях международной лицензии Creative Commons License Attribution 4.0 International (CC BY 4.0).

ВВЕДЕНИЕ

Современные системы обработки естественного языка все чаще применяют для работы с малоресурсными языками, однако качество такой обработки ограничивается не только объемом доступных текстов, но и отсутствием специализированных цифровых ресурсов, способных представлять смысловую структуру языка в машинно-читаемой форме. Для татарского языка эта проблема особенно актуальна: он обладает развитой агглютинативной морфологией, высокой ролью аффиксальных цепочек и контекстно обусловленных лексико-грамматических функций.

Существующие цифровые лингвистические ресурсы решают важные, но разные задачи. Словарь фиксирует лексические значения и грамматические сведения. Корпус предоставляет реальные употребления в языке с частотными данными. Морфологический анализатор позволяет выделить основу, аффиксы и грамматические теги. Граф знаний задает связи между объектами, понятиями и отношениями. Однако для объяснимой обработки текста необходимо не только распознать форму, но и показать, какая смысловая функция реализуется в конкретном контекстном окружении, какие условия привели к ее выбору и каким образом этот выбор может быть проверен.

Для агглютинативного языка различие между морфологическим разбором словоформы и определением смысловой функции морфемы в контексте принципиально. Одна и та же аффиксальная морфема может быть представлена в разных поверхностных вариантах, входить в разные цепочки, вступать в различные синтаксические, семантические и прагматические связи. Поэтому цифровой ресурс, ориентированный на объяснимую обработку татарского текста, должен фиксировать не только форму и значение, но и потенциальные валентности, правила их активации, корпусные примеры, связи с графом знаний и статус экспертной проверки.

В настоящей работе предложена модель базы семронов татарского языка как цифрового лингвистического ресурса. Единицей такого ресурса является семрон – потенциальная лексико-грамматическая смыслообразующая единица с формой, синтаксисом, семантикой, прагматикой и открытыми валентностями.

В Большой семиотической базе семронов (БСБС) семрон представлен в виде базового фрейма, содержащего паспорт, слоты, фасеты, ограничения, валентности и процедуры обработки. При включении в конкретное контекстное окружение одна или несколько валентностей закрываются, в результате формируется сейрон – специализированное состояние семрона в конкретной ситуации.

Целью настоящей работы было описать базу семронов татарского языка как цифровой лингвистический ресурс, обеспечивающий верифицируемое представление лексико-грамматических единиц и их использование в объяснительном искусственном интеллекте. Фреймовое представление рассмотрено как способ организации записи семрона внутри БСБС. В работе модель цифрового ресурса, его структура, формат представления, а также представлены принципы верификации.

ОСНОВНАЯ ТЕРМИНОЛОГИЯ

Нами используется ограниченный набор терминов Семиотической генеративной распознающей модели (СГРМ) (см. табл. 1), которые необходимы для описания цифровой базы семронов.

Табл. 1. Основная терминология.

Термин	Рабочее определение
СГРМ	Семиотическая генеративная распознающая модель
БСБС	Большая семиотическая база семронов
Семрон	Потенциальная лексико-грамматическая смыслообразующая единица с формой, синтаксисом, семантикой, прагматикой и открытыми валентностями.
Сейрон	Специализированное состояние семрона в конкретной ситуации, где закрыта по крайней мере одна валентность.
Фрейм	Структурированная рабочая запись, включающая паспорт, слоты, фасеты, ограничения, валентности и процедуры.
Валентность	Потенциальная или реализованная связь семрона с другими семронами и элементами контекстного окружения.
Верифицируемая цепочка	Путь от формы и морфологического разбора к семрону, правилу, сейрону, слоту ситуационной модели и объяснению.

ЭЛЕКТРОННЫЕ ЛИНГВИСТИЧЕСКИЕ РЕСУРСЫ И ЗАДАЧА СЕМАНТИЧЕСКОГО СТРУКТУРИРОВАНИЯ ТАТАРСКОГО ЯЗЫКА

Электронные лингвистические ресурсы являются необходимым условием развития современных технологий обработки естественного языка. Для языков с развитой цифровой ресурсной базой такие ресурсы обычно представлены корпусами, словарями, морфологическими, синтактическими и семантическими анализаторами, онтологиями и графами знаний. Для малоресурсных языков, к которым в цифровом отношении относится и татарский язык, такая инфраструктура развита существенно слабее. В работе [1] отмечено, что доля татарских текстов в Common Crawl составляет менее 0.01% общего объема данных.

Однако объем текстов сам по себе не решает задачу объяснимой обработки. Даже при наличии корпуса и морфологического анализатора система может определить состав словоформы, но не объяснить, какая смысловая функция реализуется конкретной морфемой в данной ситуации. Для татарского языка, где смысл часто формируется в аффиксальной цепочке, отмеченная проблема является особо острой.

Электронная библиотека в современном понимании не ограничивается хранением текстов. Постоянно подчеркивается, что электронные библиотеки включают не только текстовые и мультимедийные цифровые объекты, но и структурированные данные и описывающие их метаданные (например, [2]). В этом смысле цифровая база семронов может рассматриваться как специализированный электронный лингвистический ресурс, ориентированный на структурированное представление лексико-грамматического знания.

БСБС выполняет интегративную функцию (рис. 1). Она связывает форму, алломорфы, класс, синтаксические условия, семантические функции, прагматические условия, валентности, корпусные примеры, графовые связи и статус проверки в одной фреймовой записи. Поэтому такая база может использоваться не только как хранилище сведений, но и как рабочий слой объяснимой обработки текста.



Рис. 1. Место Большой семиотической базы семронов среди цифровых лингвистических ресурсов.

С точки зрения электронных библиотек такая постановка особенно важна, поскольку она включает метаданные, онтологии, концептуальное моделирование, графовые модели данных, извлечение знаний, семантическое структурирование контента и цифровое курирование данных [3]. БСБС является одним из ресурсов, предназначенным для семантического структурирования лексико-грамматического знания татарского языка.

СЕМИОТИЧЕСКАЯ УНИВЕРСАЛИЯ И СЕМРОН КАК ОСНОВА ФРЕЙМОВОГО ОПИСАНИЯ

В настоящей работе понятие семрона опирается на ранее разработанную концепцию семиотической универсалии. В этой концепции минимальной единицей описания лексико-грамматической структуры языка выступает не только слово или морфема как формальный элемент, а функционально-семиотическая единица, обладающая собственной структурой смыслообразования. Семиотическая универсалия может быть описана через имя, класс, синтаксис, семантику и прагматику, а ее ситуационная проактивность выражена в наличии потенциальных валентностей, активируемых в конкретном контекстном окружении [4].

Для целей цифровой реализации эта теоретическая единица получает операциональную форму семрона. Если семиотическая универсалия задает общий

теоретический уровень описания лексико-грамматической смыслообразующей единицы, то семрон является ее представлением в цифровой базе и процедурах обработки СГРМ.

Такое определение принципиально отличает семрон от обычной словарной единицы. Словарная статья фиксирует значение и употребление, но, как правило, не задает формальную структуру условий, при которых единица реализует ту или иную функцию. Семрон же должен содержать не только форму и толкование, но и набор возможных контекстных связей: синтаксических, семантических и прагматических.

Для агглютинативного татарского языка это представляет особый интерес. Значимая часть смысловой структуры словоформы формируется за счет аффиксальных цепочек, где отдельный показатель может реализовать разные функции в зависимости от основы, соседних аффиксов, дискурсивной памяти и ситуации. В [5] это показано на примерах аффиксальных морфем *-cЫ* и *-ЛАр*.

Переход от семиотической универсалии к семрону можно представить как переход от теоретического описания к операциональной записи:

СУ → Semron → SemronRecord → Frame.

Первый уровень задает теоретическую единицу смыслообразования, второй – семрон как потенциальную лексико-грамматическую единицу. Третий уровень представляет семрон как запись в БСБС. Четвертый оформляет эту запись как фрейм, пригодный для обработки, проверки, специализации и связывания с другими элементами цифрового ресурса.

БСБС КАК ЦИФРОВОЙ ЛИНГВИСТИЧЕСКИЙ РЕСУРС

Большая семиотическая база семронов здесь рассматривается как цифровой лингвистический ресурс, предназначенный для представления лексико-грамматических единиц татарского языка в форме, пригодной для экспертного редактирования, машинной обработки и использования в СГРМ. В отличие от традиционного словаря, БСБС фиксирует не только значение единицы, но и ее потенциальные валентности, условия активации, корпусные примеры, связи с графом знаний и статус проверки.

Фреймовая организация БСБС позволяет объединить в одной записи разнотипные сведения. Паспорт записи содержит идентификатор, имя, версию и статус. Формальный блок фиксирует каноническую форму, алломорфы, класс и допустимые контексты. Содержательный блок описывает семантические функции и прагматические условия. Валентностный блок задает открытые валентности и условия их закрытия. Верификационный блок хранит корпусные примеры, связи с графом знаний, экспертные комментарии и статус проверки.

Такое устройство позволяет отличать потенциальную единицу от ее конкретной реализации. Базовый фрейм семрона описывает возможности единицы до включения в контекст. Когда семрон включается в конкретное контекстное окружение, часть его открытых валентностей может быть закрыта. В результате возникает сейрон – специализированное состояние семрона в конкретной ситуации.

БСБС должна быть курируемым ресурсом. Автоматические процедуры могут предлагать новые корпусные примеры, связи с графом знаний или варианты правил активации, но изменение статуса записи должно утверждаться экспертом. Это отличает БСБС от неуправляемых самообновляемых баз и делает ее пригодной для научной верификации.



Рис. 2. Логическая структура цифровой базы семронов.

На рис. 2 представлена логическая структура БСБС как база данных. Центральным объектом является запись семрона. С ней связаны форма и алломорфы, фреймовая структура, валентности, правила активации, корпусные примеры, связи с графом знаний, статусы проверки и экземпляры сейронов. Такая схема позволяет перенести обсуждение БСБС из уровня общей идеи в уровень проектируемой структуры базы.

ФРЕЙМОВАЯ СТРУКТУРА ЗАПИСИ СЕМРОНА

Фреймовая запись семрона является центральной единицей БСБС как цифрового лингвистического ресурса. Она должна одновременно выполнять две функции: фиксировать лингвистическое описание единицы и обеспечивать ее машинно-читаемое представление, поэтому запись семрона не может ограничиваться только формой, переводом или грамматическим тегом.

С учетом цели настоящей работы различать четыре уровня представления: сигнатуру семрона, запись семрона в БСБС, базовый фрейм семрона и специализированный фрейм сейрона (рис. 3). Сигнатура описывает лингвистическую единицу как потенциальный элемент языка; запись в БСБС расширяет ее служебными, верификационными и ресурсными полями; базовый фрейм фиксирует потенциальные возможности; специализированный фрейм описывает реализованное состояние в конкретной ситуации.

Сигнатуру семрона можно представить следующим образом:

$$s = \langle name, class, form, syntax, semantics, pragmatics, V\_open \rangle.$$

Запись семрона в БСБС имеет вид

$$r(s) = \langle ID, s, R, E\_c, E\_g, Q, Version \rangle.$$

Базовый фрейм семрона может быть представлен так

$$F\_base(s) = \{Passport, Slots, Facets, V\_open, Rules, Procedures, Status\}.$$

Специализированный фрейм сейрона имеет вид:

$$F\_spec(seiron) = \langle F\_base(s), C, V\_closed, f, M, Chain \rangle.$$



Рис. 3. Базовый фрейм семрона и специализированный фрейм сейрона.

Фреймовая запись задает механизм перехода от потенциальной единицы к реализованной функции. Семрон хранит открытые валентности; контекстное окружение активирует одну или несколько из них; закрытая валентность переводит семрон в состояние сейрона; специализированный фрейм сейрона фиксирует реализованную функцию и ее место в ситуационной модели.

СЛОТЫ, ФАСЕТЫ И КОНТРОЛЬ КАЧЕСТВА ЗАПИСИ

Фреймовая структура записи семрона предполагает разделение между слотами и фасетами. Слотом называется поле фрейма или ситуационной модели, содержащее отдельный компонент смысла или формального описания. Фасет уточняет состояние слота: источник значения, ограничение, степень уверенности, экспертный статус, корпусное подтверждение, допустимость или дату проверки.

Использование фасетов позволяет превратить запись семрона из простой таблицы в контролируемый цифровой объект. Например, слот semantic_functions может содержать список возможных смысловых функций, а его фасеты фиксируют источник каждой функции, экспертную уверенность, наличие корпусного подтверждения и текущий статус проверки. Слот open_valencies может иметь фасеты activation_condition, restriction, validation_status; слот corpus_evidence может иметь фасеты source_text, frequency, context_type.

Для БСБС это принципиально важно, поскольку база должна быть не самоизменяющимся набором гипотез, а курируемым ресурсом. Каждая запись должна иметь статус, а каждое изменение – быть связано с версией, источником

и экспертным решением. Автоматические процедуры могут предлагать обновления, но не должны самостоятельно утверждать новую норму записи.

ВАЛЕНТНОСТЬ И ПЕРЕХОД ОТ СЕМРОНА К СЕЙРОНУ

Ключевым элементом фреймовой записи семрона является валентность, которая понимается как потенциальная или реализованная связь семрона с другими семронами и элементами контекстного окружения. Такая связь может быть синтаксической, семантической, прагматической или смешанной. Открытая валентность фиксирует возможность связи до включения семрона в конкретную ситуацию; закрытая – реализованную связь в конкретном контекстном окружении.

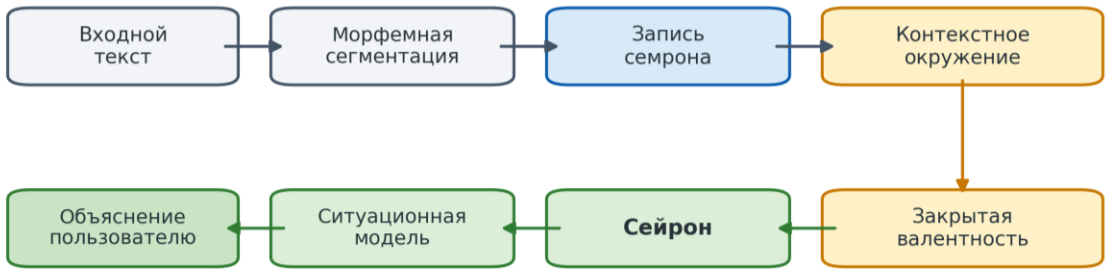


Рис. 4. Переход от записи семрона к объяснимой обработке текста.

Семрон хранит множество открытых валентностей. Например, аффиксальный семрон может иметь валентности связи с левой основой, правым аффиксом, дискурсивной памятью, корпусным употреблением или узлом графа знаний. До анализа конкретного текста эти валентности остаются потенциальными. Когда появляется контекстное окружение, одна или несколько валентностей могут быть закрыты (см. рис. 4).

Сейрон является результатом такой специализации. Он фиксирует не всю потенциальную структуру семрона, а конкретное реализованное состояние: какая валентность закрыта, какая функция активирована, какой слот ситуационной модели заполнен и какой путь проверки может быть предъявлен.

Рассмотрим примеры. Для семрона *-сЫ* могут быть предусмотрены функции принадлежности, намерения и субстантивации. В форме *бакчасы* закрыва-

ется валентность между владельцем и объектом принадлежности; в форме *барасы* активируется связь между субъектом и предполагаемым действием; в форме *атасы* возможна специализация через родственно-ролевой или прагматический контекст.

ОТ ФРЕЙМОВОЙ БАЗЫ К СЕМРОННОЙ ГИПЕРСЕТИ И ФОРМАТАМ ДАННЫХ

Фреймовая база семронов может рассматриваться как основа семронной гиперсети. При этом гиперсеть не является самостоятельным объектом хранения, независимым от БСБС. Она является результатом связывания базовых фреймов семронов и специализированных фреймов сейронов через валентности, правила активации, корпусные употребеления и связи с графом знаний.

Формально такую гиперсеть можно описать как

$$H = <S, F\_b, F\_s, V\_o, V\_c, R, E, Q>,$$

где S – множество семронов, F_b – множество базовых фреймов семронов, F_s – множество специализированных фреймов сейронов, V_o – множество открытых валентностей, V_c – множество закрытых валентностей, R – правила активации, E – внешние основания, Q – статусы качества и проверки.

Для использования БСБС как электронного лингвистического ресурса необходимо поддерживать несколько форматов представления данных, соответствующих разным задачам работы с записью. Табличное представление предназначено для экспертного редактирования лингвистом; JSON или YAML – для программной обработки; RDF, OWL или GraphML – для связи с графом знаний и интеграции с онтологиями и внешними цифровыми библиотечными ресурсами. При этом речь идет не о дублировании ресурса, а о разных представлениях одной и той же записи.

ПРИМЕР ЗАПИСИ СЕМРОНА

В качестве основного примера рассмотрен семрон *-сЫ* (табл. 2), поскольку он не является центральным объектом первой статьи и уже описан в материалах по фреймовой модели семиотической универсалии. Для *-сЫ* зафиксированы алломорфы *-сы*, *-се*, класс «Аффиксальная морфема», а также функции принадлежности, намерения и субстантивации [5].

Табл. 2. Компактная запись семрона – *-сЫ.*

Поле	Значение
id	SEM_AFFIX_SY
name	*-сЫ*
allomorphs	*-сы*, *-се*
class	аффиксальная морфема
semantic_functions	принадлежность; намерение; субстантивация
open_valencies	владелец ↔ объект; субъект ↔ предполагаемое действие; родственный термин ↔ ситуация обращения
examples	бакчасы; барасы; аласы; атасы
validation_status	экспертно описано / требует корпусной проверки

Машиночитаемое представление записи может быть дано в формате JSON. Ниже приведен сокращенный фрагмент, демонстрирующий принцип вложенной структуры.

```
{
  "id": "SEM_AFFIX_SY",
  "name": "-сЫ",
  "allomorphs": ["-сы", "-се"],
  "class": "affixal_morpheme",
  "semantic_functions": ["POSS", "INTENT", "SUBST"],
  "open_valencies": [
    {"id": "V_SY_01", "type": "semantic", "description": "владелец – объект"},
    {"id": "V_SY_02", "type": "semantic_pragmatic", "description": "субъект – действие"},
    {"id": "V_SY_03", "type": "pragmatic", "description": "родственный термин – обращение"}
  ],
  "validation": {"status": "expert_described", "version": "0.1"}
}
```

Семрон *-ЛАр* использован только как компактная иллюстрация структуры записи в БСБС. Экспериментальная проверка его контекстных функций относится к отдельному этапу проекта и не является предметом настоящей работы.

```
{
  "id": "SEM_AFFIX_LAR",
  "name": "-ЛАр",
  "allomorphs": ["-лар", "-ләр", "-нар", "-нәр"],
  "class": "affixal_polyfunctional_morpheme",
  "semantic_functions": ["quantitative_plurality", "associative_domain",
"collective_possessor"],
  "activation_rules": ["V_LAR_01", "V_LAR_02", "V_LAR_03"],
  "comment": "пример структуры записи; не экспериментальный результат
статьи"
}
```

ВЕРИФИКАЦИЯ И ЖИЗНЕННЫЙ ЦИКЛ БСБС

Верификация является обязательным элементом БСБС как цифрового ресурса. Поскольку запись семрона включает не только форму, но и семантические функции, прагматические условия, открытые валентности и правила активации, она должна проходить экспертную проверку и, по возможности, корпусное подтверждение.

Жизненный цикл записи (рис. 5) может включать несколько этапов. На первом этапе создается черновая запись, содержащая минимальный набор полей. На втором эксперт уточняет алломорфы, контексты, валентности и правила активации. На третьем запись получает статус экспертно описанной или экспертно подтвержденной. На четвертом к записи добавляются корпусные примеры и графовые связи. На пятом этапе запись может быть пересмотрена, если появятся новые данные или конфликтные интерпретации.

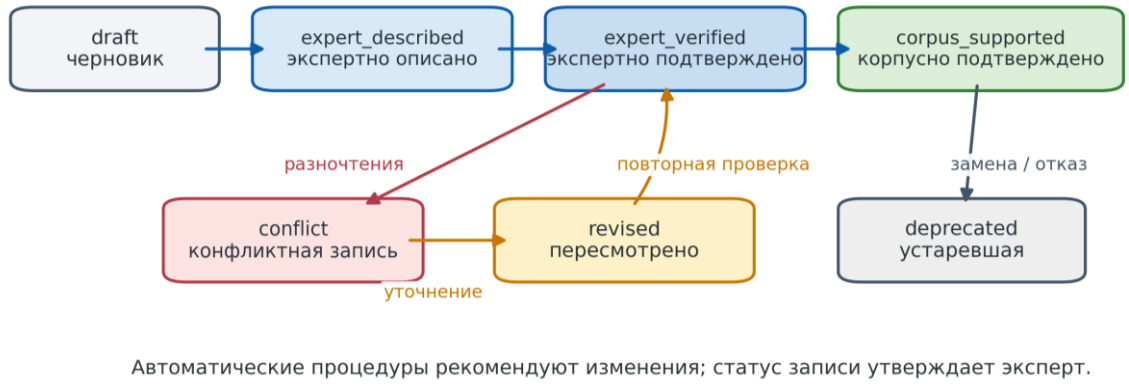


Рис. 5. Жизненный цикл записи в Большой семиотической базы семронов.

Верификация БСБС должна быть многоуровневой. Экспертная верификация проверяет корректность лингвистического описания. Корпусная проверка подтверждает наличие употреблений. Граф знаний поддерживает типизацию сущностей и отношений. Верифицируемая цепочка показывает путь от формы к реализованной функции. В совокупности эти уровни делают БСБС пригодной для научного использования и последующей интеграции в СГРМ.

БОЛЬШАЯ СЕМИОТИЧЕСКАЯ БАЗА СЕМРОНОВ В АРХИТЕКТУРЕ СГРМ

В архитектуре СГРМ Большая семиотическая база семронов выполняет функцию ресурсного слоя. Она не заменяет все остальные компоненты системы, а предоставляет активируемые записи семронов, необходимые для перехода от языковой формы к смысловой структуре. При обработке текста входное выражение сначала проходит морфемную сегментацию и нормализацию. Затем соответствующие элементы сопоставляются с записями БСБС. Далее проверяются открытые валентности, условия их закрытия, контекстное окружение, корпусные примеры и связи с графом знаний.

Общий маршрут обработки можно представить так: входной текст → морфемная сегментация → БСБС → семроны → закрытие валентностей → сейроны → ситуационная модель → верифицируемая цепочка → объяснение. Этот маршрут не должен пониматься как описание готовой полнофункциональной системы. Здесь он использован для показа места БСБС в общей архитектуре СГРМ.

ОБСУЖДЕНИЕ

Предложенная модель БСБС решает несколько задач. Во-первых, она задает структуру цифрового ресурса, способного представлять лексико-грамматические единицы татарского языка в машинно-читаемом виде. Во-вторых, она позволяет связать форму, семантику, прагматику, валентности и корпусные употребления в одной записи. В-третьих, вводит механизм перехода от потенциальной единицы к реализованному смысловому узлу через закрытие валентностей.

Основное отличие БСБС от традиционных ресурсов состоит в том, что она не ограничивается хранением данных, а задает условия смысловой активации. Словарь сообщает значение, корпус показывает реальное употребление в языке,

морфологический анализатор выделяет форму, а граф знаний фиксирует предметные связи. БСБС связывает эти уровни и обеспечивает возможность построения верифицируемой цепочки.

Предложенная модель имеет и ограничения. Создание БСБС требует значительной экспертной работы; для каждой записи необходимо определить статусы проверки и версионирование; корпусные употребления для отдельных семронов могут быть недостаточны, особенно если речь идет о редких прагматических функциях; требуется разработка инструментов для согласованного редактирования табличной, машиночитаемой и графовой форм записи.

Перспективы работы связаны с расширением набора семронов, подготовкой стандарта записи, формированием набора экспертно проверенных примеров и созданием интерфейса для редактирования БСБС. Отдельной задачей является интеграция БСБС с существующими цифровыми ресурсами татарского языка, включая корпуса, словари, онтологические платформы и графы знаний.

ЗАКЛЮЧЕНИЕ

Предложена модель базы семронов татарского языка как цифрового лингвистического ресурса. Показано, что БСБС не сводится к словарю, корпусу, морфологическому анализатору, графу знаний или базе правил, а выполняет интегративную функцию: связывает форму, алломорфы, класс, синтаксис, семантику, прагматику, валентности, корпусные примеры, графовые связи и статусы проверки в одной фреймовой записи.

Основной единицей БСБС является семрон – потенциальная лексико-грамматическая смыслообразующая единица с открытыми валентностями. Для его описания предложена фреймовая структура, включающая паспорт, слоты, фасеты, открытые валентности, правила активации, процедуры обработки и статус проверки. При включении семрона в конкретное контекстное окружение одна или несколько валентностей закрываются и формируется сейрон – специализированное состояние семрона в конкретной ситуации.

Предложенное различение базового фрейма семрона и специализированного фрейма сейрона позволяет описать переход от потенциальной единицы к реализованному смысловому узлу. Множество таких записей образует основу

семронной гиперсети и может использоваться в СГРМ для объяснимой обработки татарского текста.

Практическая значимость работы состоит в формировании модели базы семронов как цифрового ресурса, пригодного для экспертного наполнения, машинной обработки, корпусной проверки, связи с графом знаний, версионирования и последующего использования в объяснительном искусственном интеллекте. Дальнейшая работа будет направлена на расширение набора семронов, уточнение форматов представления, разработку процедур верификации и подготовку пилотной версии БСБС как курируемого электронного лингвистического ресурса татарского языка.

Благодарности

Работа выполнена в рамках Соглашения №13/2025-ПД от 22.12.2025 о предоставлении гранта научным и научно-педагогическим работникам обособленных структурных подразделений Академии наук Республики Татарстан с целью стимулирования их к защите докторских диссертаций и выполнению научно-исследовательских работ.

СПИСОК ЛИТЕРАТУРЫ

1. *Alimova I., Monogov B., Mazur A., Antonov D., Karimov V., Egorov V., Khakimov B., Panchenko A.* The Case of Tatar // Proceedings of the 15th Joint Conference on Lexical and Computational Semantics. San Diego, California: Association for Computational Linguistics, 2026. P. 264–274.
URL: https://aclanthology.org/2026.starsem-conference.17 (дата обращения: 02.06.2026). https://doi.org/10.48550/arXiv.2606.26015

2. Электронные библиотеки. О журнале. URL: https://rdl-journal.ru/about (дата обращения: 01.06.2026).

3. Электронные библиотеки. Тематика журнала.
URL: https://rdl-journal.ru/about/themes (дата обращения: 01.06.2026).

4. *D. Suleymanov, A. Gilmullin and I. Mukhametzyanov.* Semiotic Universal as an Atomic Element of the Lexical-Grammatical Structure of the Tatar Language, 2025 10th International Conference on Computer Science and Engineering (UBMK), Istanbul, Turkiye. 2025. P. 1767–1771.

https://doi.org/10.1109/UBMK67458.2025.11206836.

5. *Suleymanov D.Sh., Gilmullin R.A., Fridman A.Ya., Mukhamedshin D.* Semantic Models of Agglutinative Languages // 2024 IEEE 3rd International Conference on Problems of Informatics, Electronics and Radio Engineering (PIERE). 2024. P. 1930–1935. https://doi.org/10.1109/PIERE62470.2024.10804991

6. *Сулейманов Д.Ш., Мухамедшин Д.Р., Гильмуллин Р.А., Фридман А.Я.* Семиотическая модель для представления когнитивных структур языка агглютинативного типа // Искусственный интеллект и принятие решений. 2024. № 4. С. 55–66. https://doi.org/10.14357/20718594240405

7. *Stefanuk V.L., Taisina E.A., Suleymanov Dz.Sh., Fridman A.Ya.* (Eds.). Semiotic Aspects of Modeling the Cognitive System for the Artificial Intelligence. Editions du JIPTO, 2024. 330 p.

8. *Тарасов В.Б.* Семиотические аспекты коммуникации между агентами в мультиагентных системах // Semiotic Aspects of Modeling the Cognitive System for the Artificial Intelligence. Editions du JIPTO, 2024. С. 155–188.

9. *Туган тел: национальный корпус татарского языка.*
URL: https://tugatel.tatar/ (дата обращения: 02.06.2026).

10. *Minsky M.* A Framework for Representing Knowledge // The Psychology of Computer Vision / ed. by P.H. Winston. New York: McGraw-Hill, 1975. P. 211–277.

11. RDF 1.1 Concepts and Abstract Syntax. W3C Recommendation. URL: https://www.w3.org/TR/rdf11-concepts/ (дата обращения: 30.05.2026).

12. OWL 2 Web Ontology Language Document Overview. W3C Recommendation. URL: https://www.w3.org/TR/owl2-overview/ (дата обращения: 20.05.2026).

THE TATAR LANGUAGE SEMRON DATABASE AS A DIGITAL LINGUISTIC RESOURCE FOR EXPLAINABLE ARTIFICIAL INTELLIGENCE

R. A. Gilmullin[0000-0002-8520-8921]

Tatarstan Academy of Sciences, Kazan, Russia
rinatgilmullin@gmail.com

Abstract

The article considers the Tatar language semron database as a digital linguistic resource for explainable artificial intelligence. A model of the Big Semiotic Database of Semrons is proposed, in which lexical-grammatical units are represented as semrons, i.e. potential meaning-forming units with form, class, syntactic, semantic and pragmatic characteristics, valencies, activation rules, corpus evidence, knowledge graph links and expert verification statuses. Frame representation is used as a way of organizing a semron record: the base frame fixes the potential capabilities of a unit, while the specialized seiron frame describes its contextual realization after one or more valencies have been closed. It is shown that such a database can function as a resource layer of the Semiotic Generative Recognizing Model, supporting the transition from the morphemic structure of Tatar text to a situational model, a verifiable chain and an explanation. Compact record examples demonstrate the frame structure, types of slots and facets, verification principles and possible data representation formats.

Keywords*: semron, semron database, Tatar language, digital linguistic resource, frame model, valency, seiron, knowledge graph, corpus layer, expert verification, explainable artificial intelligence, Semiotic Generative Recognizing Model.*

REFERENCES

1. *Alimova I., Monogov B., Mazur A., Antonov D., Karimov V., Egorov V., Khakimov B., Panchenko A.* The Case of Tatar // Proceedings of the 15th Joint Conference on Lexical and Computational Semantics. San Diego, California: Association for Computational Linguistics, 2026. P. 264–274. URL: https://aclanthology.org/2026.starsem-conference.17 (accessed 02.06.2026).
https://doi.org/ 10.48550/arXiv.2606.26015

2. Russian Digital Libraries Journal. About the journal. Available at:

https://rdl-journal.ru/about (accessed 06.06.2026).

3. Russian Digital Libraries Journal. Topics. Available at: https://rdl-journal.ru/about/themes (accessed 06.06.2026).

4. *D. Suleymanov, A. Gilmullin and I. Mukhametzyanov.* Semiotic Universal as an Atomic Element of the Lexical-Grammatical Structure of the Tatar Language, 2025. 10th International Conference on Computer Science and Engineering (UBMK), Istanbul, Turkiye, 2025. P. 1767–1771. https://doi.org/10.1109/UBMK67458.2025.11206836

5. *Suleymanov D.Sh., Gilmullin R.A., Fridman A.Ya., Mukhamedshin D.* Semantic Models of Agglutinative Languages // 2024 IEEE 3rd International Conference on Problems of Informatics, Electronics and Radio Engineering (PIERE). 2024. P. 1930–1935. https://doi.org/10.1109/PIERE62470.2024.10804991

6. *Suleymanov D.Sh., Mukhamedshin D.R., Gilmullin R.A., Fridman A.Ya.* Semiotic model for representing cognitive structures of an agglutinative language. Iskusstvennyj intellekt i prinyatie reshenij. 2024. No. 4. P. 55–66 (In Russ.).

7. *Stefanuk V.L., Taisina E.A., Suleymanov Dz.Sh., Fridman A.Ya.,* eds. Semiotic Aspects of Modeling the Cognitive System for the Artificial Intelligence. Editions du JIPTO; 2024. 330 p.

8. *Tarasov V.B.* Semiotic aspects of communication between agents in multi-agent systems. In: Semiotic Aspects of Modeling the Cognitive System for the Artificial Intelligence. Editions du JIPTO. 2024. P. 155–188 (In Russ.).

9. *Tugan Tel: National Corpus of the Tatar Language.* Available at: https://tugantel.tatar/ (accessed 01.06.2026). (In Russ.)

10. *Minsky M.* A Framework for Representing Knowledge. In: Winston P.H., ed. The Psychology of Computer Vision. New York: McGraw-Hill. 1975. P. 211–277.

11. RDF 1.1 Concepts and Abstract Syntax. W3C Recommendation. Available at: https://www.w3.org/TR/rdf11-concepts/ (accessed 30.05.2026).

12. OWL 2 Web Ontology Language Document Overview. W3C Recommendation. Available at: https://www.w3.org/TR/owl2-overview/ (accessed 23.05.2026).

СВЕДЕНИЯ ОБ АВТОРЕ



ГИЛЬМУЛЛИН Ринат Абрекович – кандидат физико-математических наук, директор Института прикладной семиотики Академии наук Республики Татарстан. Область научных интересов: искусственный интеллект, прикладная семиотика, обработка естественного языка, цифровые лингвистические ресурсы.

Rinat Abrekovich GILMULLIN – Candidate of Physical and Mathematical Sciences, Director of the Institute of Applied Semiotics of the Tatarstan Academy of Sciences. Research interests: artificial intelligence, applied semiotics, natural language processing, digital linguistic resources.

e-mail: rinatgilmullin@gmail.com

ORCID: 0000-0002-8520-8921

Материал поступил в редакцию 30 июня 2026 года

УДК 004.89

АВТОМАТИЧЕСКАЯ КОРРЕКЦИЯ ИНДИВИДУАЛЬНОЙ ОБРАЗОВАТЕЛЬНОЙ ТРАЕКТОРИИ В АДАПТИВНЫХ СИСТЕМАХ ОБУЧЕНИЯ НА ОСНОВЕ КОМПОЗИТНОГО ПОДХОДА

М. С. Дьяченко[1] [0000-0002-5809-4981], А. Г. Леонов[2] [0000-0001-9622-1526]

[1, 2]*НИЦ «Курчатовский институт» – НИИСИ, г. Москва, Россия*

[2]*Московский государственный университет им. М. В. Ломоносова, г. Москва, Россия*

[2]*Московский педагогический государственный университет, г. Москва, Россия*

[2]*Государственный университет управления, г. Москва, Россия*

[1]mdyachenko@niisi.ru, [2]dr.l@vip.niisi.ru

Аннотация

В условиях расширения применения адаптивных курсов в системе высшего образования актуальной становится проблема ранней идентификации отстающих студентов и автоматической коррекции их индивидуальной образовательной траектории. Существующие методы идентификации отстающих студентов не проводят глубокого анализа индивидуализированного подхода к обучению, используемого студентом. Применяемые обобщенные алгоритмы адаптивного обучения напрямую не решают задачу идентификации отстающих студентов, что может привести к позднему их обнаружению, когда уже нет возможности скорректировать подход к обучению для успешного завершения курса такими студентами.

Разработаны алгоритмы выявления отставания студента при освоении учебной дисциплины, основанные на прогнозировании результатов индивидуализированного обучения с использованием модели предметной области. Предложены также алгоритмы автоматической коррекции индивидуальной образовательной траектории, которые, в отличие от существующих подходов, формируют индивидуальную образовательную технологию для обеспечения успеш-

© М. С. Дьяченко, А. Г. Леонов, 2026.
Данная статья распространяется на условиях международной лицензии Creative Commons License Attribution 4.0 International (CC BY 4.0).

ного завершения учебного курса. Эти алгоритмы могут быть внедрены в существующие адаптивные системы обучения для повышения доли студентов, успешно завершивших обучение на адаптивных курсах.

Ключевые слова: *адаптивное обучение, идентификация отстающих студентов, индивидуальная образовательная траектория.*

ВВЕДЕНИЕ

Одной из ключевых проблем обучения в вузе остается раннее выявление отстающих студентов, которые потенциально не смогут успешно завершить учебный курс. Раннее выявление таких студентов дает преподавателю возможность уделять ним больше внимания для устранения отставания.

При аудиторной работе преподавателя с группой основным средством выявления отстающих студентов являются промежуточные проверки в виде контрольных работ, семинаров, а также выборочная проверка студентов при работе в аудитории.

Для масштабирования подготовки специалистов высшими учебными заведениями активно применяются автоматизированные системы обучения в гибридном формате, что дало возможность увеличить частоту контрольных мероприятий за счет автоматизации верификации приобретенных компетенций, а также за счет проведения проверок знаний в том числе при самостоятельной работе студентов, например, при выполнении домашних заданий.

Автоматизированная проверка знаний не решает проблему раннего выявления отстающих студентов, поскольку важны не только результаты, демонстрируемые студентами, но и их поведение в процессе обучения. Поэтому автоматизированные системы обучения собирают цифровой след обучения, на основе анализа которого формируются гипотезы о вероятности успешного завершения обучения на курсе конкретным студентом.

Подходы к автоматизации раннего выявления отстающих студентов методами прогнозирования можно условно разделить по времени получения первых сигналов об отставании: до начала обучения, в этом случае используются предыдущие результаты обучения студентов и их демографические данные; в процессе обучения, которые, в свою очередь, делятся на методы выявления:

- по фактически демонстрируемым результатам (воспроизводят традиционные методы оценки преподавателя при аудиторной работе);
- по демонстрируемому поведению.

В зарубежной литературе можно встретить понятие детектора отчисления (dropout detector), поскольку обсуждаемые методы предназначены для выявления не отстающих студентов, а имеющих высокую вероятность отчисления из-за неуспеваемости. Такие методы получили развитие для MOOC (massive open online course), больших открытых онлайн-курсов, поскольку в условиях отсутствия аудиторной работы с обучаемыми владельцам онлайн-курсов и платформ важно выявлять отстающих студентов и стимулировать их завершать учебные курсы [1–3]. По мере внедрения гибридных форматов обучения в высшие учебные заведения в них также стали использоваться аналогичные методы [4, 5].

К первой категории методов выявления отстающих студентов относятся методы получения первых сигналов об отставании до начала обучения, например, по ранее продемонстрированным результатам и демографическим данным [6]. Более сложные варианты детекторов используют искусственные нейронные сети [7]. Оценки на основе поведения могут быть эвристическими правилами, например, когда студент просто перестал выполнять задания или делает это нерегулярно, а также более сложными, использующими методы анализа временных рядов, извлекаемых из цифрового следа обучения [8].

В системах адаптивного обучения раннее выявление отстающих студентов можно осуществлять алгоритмически на уровне технологии адаптации. Адаптивный алгоритм доходит до состояния, в котором не может предложить студенту продолжение индивидуальной траектории обучения. Для раннего выявления отстающих студентов требуется специальная настройка алгоритма именно для выявления подобных случаев, иначе отставание будет обнаружено поздно, например после многократных безуспешных попыток изучения студентом очередного блока учебного материала.

После выявления отстающих студентов в адаптивных системах обучения возникает необходимость коррекции их индивидуальных образовательных траекторий. Выявленное отставание является в том числе результатом применения алгоритма адаптации, реализуемого адаптивной системой обучения. Изменение алгоритма адаптации не предусмотрено, так как в большинстве случаев это

параметрический алгоритм с ограниченными возможностями по регулированию. В такой ситуации прибегают к педагогическому воздействию для возвращения обучаемого на общую траекторию обучения группы, т. е. для возвращения в границы регулирования, допускаемые адаптивным алгоритмом. Если не выполнить коррекцию, то возникают отставание или пропуск по одной из тем курса, либо общее отставание по всему оставшемуся материалу курса. Отстающие студенты могут потерять мотивацию к обучению и завершению курса, что потенциально ведет к их отчислению.

Отставание студентов может быть вызвано несколькими основными причинами, в числе которых:

- пробелы в пререквизитах для изучения очередного курса: для решения этой проблемы применяют выравнивающие курсы, которые обеспечивают минимальный общий уровень знаний всех обучаемых на момент начала курса;
- отсутствие мотивации к обучению – это отдельная проблемная область; отметим, что возможность самостоятельно успешно учиться является сильным мотиватором;
- неэффективный индивидуализированный подход к обучению студента: используемый подход к обучению не позволяет получать требуемые результаты за предусмотренное планом время обучения; решение именно этой проблемы рассмотрено в настоящей работе.

Все вышеперечисленное объединяет использование цифрового следа обучения студента без учета подходов к его обучению. Автоматизированная система реализует технологию обучения: система предлагает студенту выполнять задания, которые он успешно или неуспешно выполняет (возможно, даже пропускает), но при этом не оценивает индивидуализированный подход к обучению и потенциальную возможность удачно завершить учебных курс, применяя сформированный индивидуализированный подход к обучению.

В условиях расширения применения систем адаптивного обучения актуальной становится реализация аналитико-эвристических решений для раннего выявления отстающих студентов и своевременного информирования преподавателей о необходимости коррекции индивидуализированного подхода к обучению студента. В условиях масштабирования подготовки специалистов преподаватель не имеет возможности точечно работать с каждым студентом, поэтому

также актуальной становится проблема автоматической коррекции индивидуализированного подхода к обучению студента после выявления отставания.

Целями настоящего исследования были разработка подхода и алгоритма для выявления отстающих студентов в системах адаптивного обучения с использованием анализа индивидуализированного подхода к их обучению, а также разработка алгоритма автоматической коррекции индивидуальной образовательной траектории отстающих студентов.

МАТЕРИАЛЫ И МЕТОДЫ РЕШЕНИЯ ЗАДАЧИ

Нами был разработан композитный подход к формированию индивидуальной образовательной технологии (ИОТ) в рамках адаптивных систем обучения [9]. Принципиальным методологическим вопросом является разграничение понятий «индивидуальная образовательная технология» и «индивидуальная траектория обучения». Здесь ИОТ рассматривается как совокупность педагогических приемов, отобранных и валидированных с учетом индивидуальных характеристик обучающегося и специфики учебных задач курса. Композитный подход предполагает автоматизированное формирование подмножества педагогических приемов, релевантных когнитивным характеристикам и академическим достижениям студента. Отбор приемов опирается на результаты ранее проведенных педагогических экспериментов. Последние служат основой для формирования гипотезы о применимости конкретного приема для решения выявленных образовательных затруднений студента. Сам процесс проверки гипотезы является итеративным циклом, интегрированным в средства обучения (цифровые платформы). Сначала применяется некий выбранный прием, при этом накапливаются (первичные) данные об эффективности его использования. Для ускорения формирования данных наблюдение ведется одновременно на нескольких учебных курсах. На этой стадии объемов данных еще недостаточно для проведения статистически значимой оценки, поэтому для проверки гипотезы используются эвристические правила либо методы машинного обучения. Ключевым эвристическим критерием выступает фиксация наблюдаемых положительных изменений в учебной деятельности студента, например, перехода от неспособности выполнить задание к успешному его выполнению при повышении детализа-

ции подачи материала. В случае устойчивости такого эффекта применение приема продолжается. Если же многократное использование приема (например, повышение детализации содержания материала) не приводит к наблюдаемым изменениям, прием отклоняется и фокус переключается на альтернативные приемы (например, пошаговый разбор решений заданий).

По мере продолжения учебного процесса объем данных об использовании приемов постепенно увеличивается, однако на промежуточных этапах он по-прежнему не достигает уровня, необходимого для формирования статистически значимых выводов. Тем не менее накопленные наблюдения позволяют выстраивать динамическую модель эффективности приемов и корректировать состав ИОТ в режиме реального времени.

Важным свойством разработанной ИОТ является ее интерпретируемость и возможность декомпозиции. Под интерпретируемостью понимается возможность оценить состав индивидуальной образовательной технологии, условия и основания выбора каждого автоматизированного приема обучения, а также оценить его влияние на результаты конкретного студента. Поскольку каждый компонент технологии имеет экспериментальное обоснование, представляется возможным выделить причинно-следственные связи между применяемыми педагогическими приемами и наблюдаемыми образовательными эффектами. Декомпозиция ИОТ позволяет не только обеспечить прозрачность для педагога рекомендаций адаптивной системы, но и формализовать логику выбора методов обучения, что создает основу для моделирования применения ИОТ к оставшейся части учебного курса. Прогностическая модель строится на экстраполяции эффективности валидированных приемов на новые учебные материалы с учетом текущих академических показателей студента и динамики его образовательных результатов.

В настоящей работе применялись методы системного анализа и математического моделирования.

РЕЗУЛЬТАТЫ

Далее под отставанием студента будет пониматься такое состояние его индивидуализированного процесса обучения, что при продолжении обучения

цели обучения не могут быть достигнуты. Под целями понимается как глобальная цель успешного завершения курса, так и локальные цели изучения отдельных ключевых материалов курса и подготовки к промежуточных контрольным мероприятиям. Поскольку материалы адаптивных курсов организованы в виде системы связанных учебных блоков, переход от одного к следующим возможен только при завершении всех предыдущих блоков, отставание в изучении одного из них даже при разбивке всего курса на независимые блоки, например, темы, приводит к общему отставанию по курсу.

Общий алгоритм выявления отставания и коррекции индивидуальной образовательной технологии включает следующие шаги.

1. Проверка достижимости целей обучения: соотнесение текущей индивидуальной образовательной технологии с учебным материалом, оставшимся для изучения. Проверка выполняется реактивно при появлении событий из системы обучения, формируемых в результате взаимодействия студента с адаптивным курсом, а также при получении специального события «пропуска действия», когда студент не выполнил какие-либо ожидаемые (запланированные) действия в системе, например, не выполнил домашние задания или не повторил материал.

    a. В случае, если в результате проверки все цели обучения достижимы, далее никакие действия не выполняются, и при следующем событии возвращаемся к п. 1.

    b. В случае, когда какие-либо цели обучения не могут быть достигнуты при применении текущей индивидуальной образовательной технологии, осуществляется переход к п. 2.

2. Коррекция индивидуальной образовательной технологии для обеспечения достижения всех поставленных целей обучения.

    a. Возможна ситуация, в которой система не может предложить индивидуальную образовательную технологию с учетом ограничений. В этом случае преподаватель получает уведомление о необходимости вмешаться в процесс обучения с целью коррекции индивидуализированного подхода к обучению.

В рамках проведенного исследования к алгоритму выявления отстающих студентов установлены следующие требования:

- определение отставания по результатам анализа цифрового следа, собираемого при взаимодействии с адаптивным курсом на аудиторных занятиях и при самостоятельной работе студента;
- предоставление преподавателю интерпретируемых результатов анализа об отставании студента;
- выявление отставания до проведения контрольных мероприятий не позднее этапа подготовки к ним, когда еще может быть выполнено корректирующее действие;
- учет индивидуальной образовательной технологии обучаемого при оценке отставания.

Алгоритм выявления отстающих студентов реализован с использованием композитного подхода к формированию индивидуальной образовательной технологии. Моделируется несколько сценариев применения ИОТ для формирования прогнозируемой индивидуальной образовательной траектории для оценки достижимости целей обучения.

На вход этого алгоритма передаются следующие данные:

- индивидуальное расписание с контрольными точками курса; последние формируются по данным расписания адаптивного курса;
- данные о стадии изучения учебных элементов курса, поскольку важно учитывать также повторы ранее изученного материала для закрепления;
- модель индивидуальной образовательной технологии студента;
- ресурсный календарь студента: ограничения времени, которое может быть затрачено на обучение на адаптивном курсе.

Алгоритм выявления отставания заключает следующие этапы.

1. Формирование множества имитаторов студента для моделирования, воспроизводящих различные сценарии поведения студента (модель поведения на основе исторических данных, пессимистичный и оптимистичные варианты отклонения от этой модели).

2. Для каждого имитатора студента выполняется цикл моделирования.

3. Полученные данные оцениваются по критериям:

– достижимость целей обучения по контрольным датам;

– соответствие требуемых ресурсов ограничению ресурсного календаря студента.

4. В случае, если по результатам моделирования с использованием всех имитаторов критерии из п. 3 выполнены, то не требуется коррекции индивидуальной образовательной траектории. В обратном случае коррекция необходима.

Структура реализации алгоритма выявления отставания представлен на рис. 1.

Алгоритм коррекции индивидуальной образовательной технологии заключается в изменении набора приемов обучения, формирующих индивидуальную образовательную технологию. Каждый автоматизированный прием обучения содержит следующие методы для оценки его применимости, оптимизации использования и оценки эффективности.

1. Метод оценки применимости анализирует накопленные данные и возвращает свою оценку того, насколько эффективен данный прием обучения для студента.

2. Метод оптимизации формирует набор индивидуальных параметров для автоматизированного приема обучения, которые надо применить и оценить по результатам полученного эффекта от предложенных изменений. Оптимизация направлена на сокращение времени обучения как ключевого ресурса студента.

3. Метод оценки эффективности по реализации аналогичен методу оценки применимости. Эффективность оценивается совместно с оптимизацией, при этом возможна ситуация, когда эффективность приема обучения окажется ниже порогового значения, в результате чего будет запущена процедура коррекции ИОТ. Основным отличием оценки эффективности от оценки применимости является возможность накопления данных о влиянии приема обучения на результаты студента.

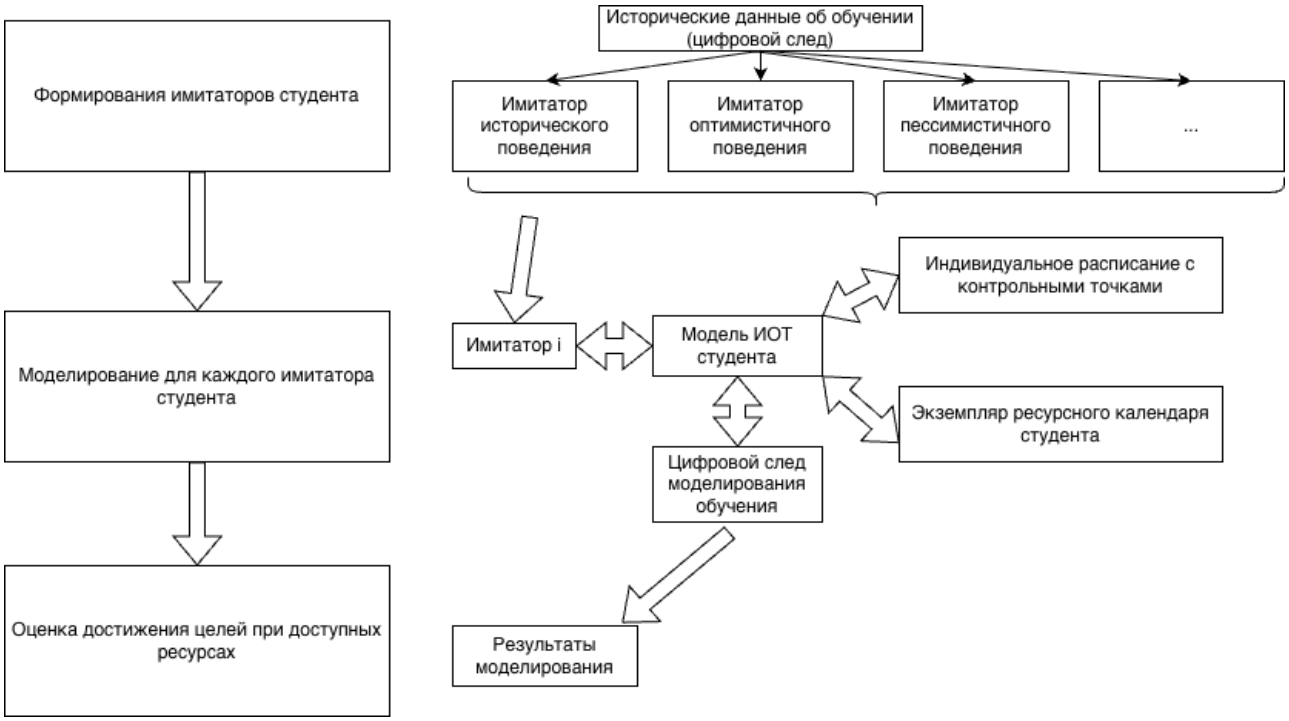


Рис. 1. Структура реализации алгоритма выявления отставания.

Нами не рассматривались конкретные приемы обучения. Процесс отбора приемов идет непрерывно, поскольку помимо выявления отставания проводится оценка эффективности отдельных приемов.

Процесс коррекции схож по структуре с процессом выявления отставания и также содержит этап моделирования и анализа результатов моделирования (рис. 2). Поскольку обучение идет на нескольких курсах параллельно, моделирование и коррекция выполняются для всех курсов.

Отличие процесса отбора приемов в процессе поиска индивидуальной образовательной технологии от коррекции ИОТ заключается в снижении порога оценки эффективности, что приводит к активизации процесса оптимизации и формирования подмножества приемов обучения.

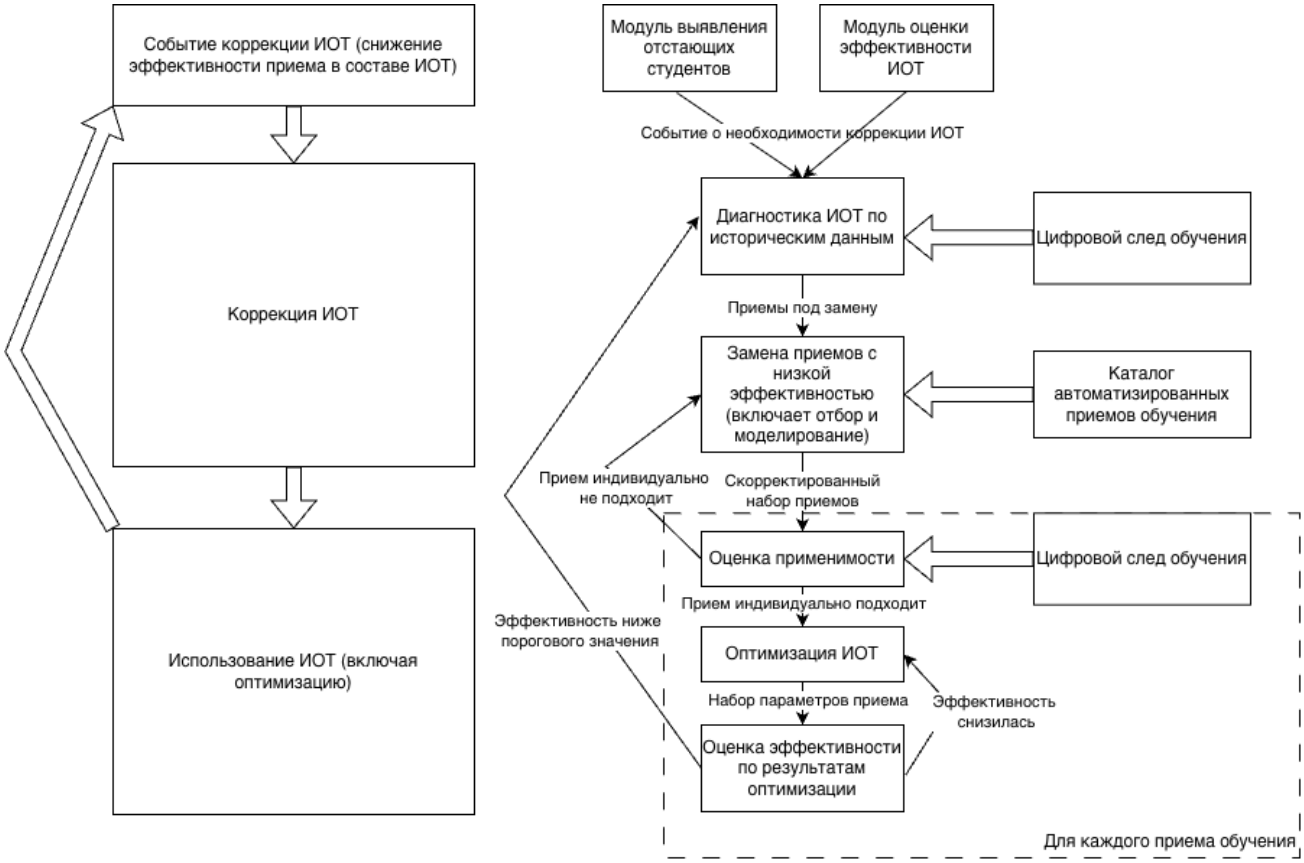


Рис. 2. Структура схемы коррекции индивидуальной образовательной технологии.

Перечислим особенности алгоритмов выявления отставания и коррекции ИОТ:

- с учетом того, что ИОТ формируется непрерывно, часть исторических данных может относиться к одной из предыдущих конфигураций ИОТ;
- в цифровой след включаются данные всех учебных курсов, на которых применялась ИОТ;
- в процессе моделирования не учитываются дальнейшее изменение ИОТ, а также оптимизация элементов ИОТ, которая может только улучшить результаты, но среднесрочно не ухудшит их.

Ключевым методом является оценка применимости. Сложность реализации здесь заключается в том, что для оценки применимости нет возможности длительно накапливать данные для анализа, решение должно быть принято по единичным случаям применения приема обучения студентом.

Выделим требования к реализации метода оценки применимости:

- оценка применимости выполняется на основании детального цифрового следа обучения, содержащего данные о действиях обучаемого в системе в процессе подготовки, решения задания и ознакомления с результатами проверки;
- оценка применимости дается по одному или нескольким использованиям приема обучения, поскольку курс ограничен по длительности;
- оценка применимости выполняется по данным нескольких параллельных курсов;
- оценка применимости формируется на основании данных педагогических экспериментов;
- рекомендуемый вариант реализации метода – это автоматическое формирование по данным эксперимента без необходимости программирования.

Для решения такой задачи могут быть использованы средства анализа цифрового следа методами машинного обучения и/или с использованием технологий глубокого обучения. Обучение выполняется на всех данных экспериментов с целью формирования детектора поведенческих моделей, которые далее дообучаются для оценки применимости и эффективности для конкретного автоматизированного приема обучения.

ЭКСПЕРИМЕНТАЛЬНАЯ ПРОВЕРКА

В качестве цифрой экосреды в вузах часто выбираются системы управления обучением, LMS (Learning Management System), такие, например, как Moodle [10]. Однако подобный подход фактически копирует структурные взаимодействия классического учебного процесса университетов в цифровую форму. При этом современные средства обучения дают педагогу гораздо более развитые инструменты, меняющие формы и методы обучения и способные обеспечить гибкость, инклюзивность, омниканальность образовательного процесса в целом, что позволяет формировать персональный опыт обучаемого [11]. Такие цифровые образовательные системы носят название платформ образовательного опыта LXP (Learning Experience Platform). В отличие от LMS, они обеспечивают формирование рассматриваемого опыта обучающегося с помощью автоматического агрегирования разнородного целевого учебного контента, анализа

интересов обучаемого, дефицитов его компетенций и персональной образовательной активности. Предоставление рекомендаций студенту, инструментов коммуникации, рефлексии и возможности самостоятельного конструирования индивидуальной образовательной траектории помогает последнему приобретать целевые компетенции, индивидуализируя учебный процесс [12].

Дальнейшим развитием платформы образовательного опыта является LXP+ (Learning Experience Platform Plus), которая включает новационные цифровые решения как расширение базовой модели LXP за счет глубокой интеграции элементов искусственного интеллекта, расширения спектра поддерживаемых образовательных межпредметных сценариев и включения самой платформы в экосистему непрерывного обучения и деятельности студента. К платформам LXP+ относится авторская разработка – цифровая образовательная платформа (ЦОП) Мирера [13].

Весной 2024 г. в Государственном университете управления, г. Москва, был проведен эксперимент по оценке реализуемости предложенного подхода для магистров первого года обучения. Использовалась технология «перевернутый класс» и был разработан автоматизированный курс «Современные технологии разработки программного обеспечения». К особенности программы курса можно отнести небольшой объем аудиторных часов при существенном объеме самостоятельной работы. Магистранты перед занятием получали материалы для самостоятельной проработки, после чего проводилось аудиторное занятие для разбора и обсуждение материалов. В процессе подготовки к занятиям и по результатам аудиторных занятий собиралась обратная связь, по итогам которой с учетом собираемых данных об активности магистрантов в ЦОП Мирера были внесены корректировки в учебные материалы. Корректировки в материалы курса вносились вручную, имитируя частично работу описанного выше алгоритма.

При проектировании курса использовались учебные материалы, по форме и содержанию приближенные к специализированным первоисточникам. После выявления недостаточной скорости освоения нового материала была скорректирована разбивка занятий для выделения большего количества времени для работы, как выяснилось, со сложным для восприятия материалом. Часть материала, ранее рассмотренного в других дисциплинах, наоборот, была сокращена

для высвобождения времени для нового материала. Неожиданным оказалось, что магистранты ограничивались адаптированными учебными материалами и не прибегали к первоисточникам, содержащим углубленный материал по теме. Учебные материалы были адаптированы по объему и сложности под ограничения времени, которое магистранты фактически могли были потратить на подготовку к занятиям по данной учебной дисциплине без ущерба другим учебным дисциплинам.

По результатам курса четверо из пяти магистрантов в группе успешно завершили курс, еще одному магистранту потребовалось дополнительное время для подготовки к экзамену. Магистранты положительно отзывались об опыте, полученном на данном курсе, но отметили, что это был их первый «перевернутый класс», а также, что данный формат предполагает большой объем самостоятельного изучения нового материала в формате, близком к первоисточнику, что потребовало регулярного выделения времени на подготовку к занятиям.

По результатам эксперимента были сделаны следующие наблюдения и выводы.

1. Большой объем самостоятельной работы для подготовки к занятиям фактически требует от магистрантов составления личного расписания подготовки с явным выделением времени и отслеживанием прогресса по этому расписанию.

2. Отсутствие реперных точек подготовки к перевернутому классу приводит к устойчивому паттерну «просмотреть материалы в последний момент», причем в конкретном случае паттерн наблюдался у магистрантов со средней и хорошей успеваемостью. В экспериментальной группе было 1–2 магистранта, которые планомерно готовились к занятию, однако остальные готовились непосредственно накануне занятия.

3. Отсутствие контроля методов обучения, применяемых при самостоятельной работе магистрантов, не позволяет уверенно сказать, насколько они проработали материал, даже если были подготовлены материалы, обязательные для допуска к занятию.

4. Паттерны поведения магистрантов устойчивы и предварительно могут применяться для имитатора при реализации детектора индивидуального отставания.

5. Необходимо экспериментально проверить предлагаемый подход на прототипе для практического подтверждения основных положений и решений.

ЭТИКА ИССЛЕДОВАНИЯ

К этическим аспектам адаптивных систем принято относить защиту цифрового следа обучения и персональных данных, интерпретируемость применяемых алгоритмов, алгоритмическую предвзятость, а также психологический комфорт и социализацию.

Предлагаемое решение позволяет обеспечить защиту чувствительных и персональных данных обучаемых, непосредственная обработка которых может выполняться в контуре учебного заведения без необходимости передачи этих данных в сторонние системы.

При построении предлагаемого решения были использованы интерпретируемые педагогические эксперименты. Оценка применимости использует технологии глубокого обучения, результаты применения которой дополнительно оценивались с использованием эвристических алгоритмов. Преподаватель курса на свое усмотрение также может вносить изменения в ручном режиме.

Для снижения влияния алгоритмической предвзятости при создании решения используются только данные о поведении пользователей, без учета их пола, возраста, национальности или социально-экономического статуса.

Решение применяется при обучении в группе с участием преподавателя. При работе с решением проводятся аудиторные занятия. Разрабатываемые для решения курсы по учебным дисциплинам могут содержать элементы группового взаимодействия обучаемых при решении задач.

При разработке системы анализируются риски и меры по их смягчению, так, в случае обнаружения негативного влияния решения на учебный процесс предусмотрено информирование преподавателя курса и возможность перевода системы в ручной режим управления.

ОБСУЖДЕНИЕ

Предложенный подход является развитием существующих систем адаптивного обучения и предлагает использование модели обучения на основе композитного подхода для выявления отстающих студентов и последующей коррекции их индивидуальных образовательных технологий.

Явным отличием от методов на основе анализа цифрового следа является факт оценки моделирования сценариев завершения курса с учетом используемой технологии обучения по цифровому следу как источнику исторических данных. Таким образом, процесс оценки является полностью интерпретируемым и может настраиваться в ручном режиме. Так, например, можно установить дату завершения обучения до даты фактического завершения курса для повышения вероятности завершения обучения в указанные сроки.

В отличие от эвристических подходов [14], предложенный подход автоматически формирует методы оценки применимости, хотя не исключает для реализации использования описанных методов программирования эвристических правил.

Проблема «холодного старта» решена за счет применения ИОТ, переносимой между курсами. В новом курсе происходит уточнение ИОТ. Для запуска решения на первичном курсе набор базовых приемов обучения может быть задан преподавателем самостоятельно.

Вопросы коррекции технологии обучения в адаптивных системах ограниченно рассматриваются на уровне настройки параметризированных алгоритмов обучения, как правило, в системах реализуется несколько фиксированных алгоритмов, что снижает общую гибкость такого решения. Например, методы повторов на основе кривых забывания предлагают построение оптимальной траектории обучения, однако речь идет только об одном из приемов, изменение которых в системе не предусмотрено [15].

В отличие от систем с моделями обучения на основе технологии глубокого обучения [16], предложенный подход, во-первых, не требует накопления больших объемов данных для создания модели, во-вторых, переносим между различными курсами. Проблема такой межкурсовой адаптации рассматривается, например, в специализированных технических решениях [17], однако даже в этом случае требуется длительное накопление данных для адаптации модели под новый курс. При этом не рассматриваются вопросы переиспользования цифрового следа обучаемых для набора смежных курсов.

ЗАКЛЮЧЕНИЕ

Предложенный подход обладает хорошей интерпретируемостью и может быть применен для реализации современных система адаптивного обучения. Интерпретируемость достигается за счет декомпозиции индивидуальной образовательной технологии на отдельные приемы обучения, полученные в результате педагогических экспериментов.

Использование формализованной образовательной технологии также дает возможность выполнять моделирование ее применения, которое используется для выявления отстающих студентов и коррекции их подхода к обучению.

Методы оценки применимости отбираемых в процессе формирования индивидуальной образовательной технологии приемов обучения построены на основе анализа цифрового следа обучения, включая анализ поведенческих особенностей студента. Для отдельных вариантов использования приема обучения необходимо выполнять оценку. Для реализации методов оценки применимости могут быть использованы эвристические методы, но более перспективным выглядит использование технологии глубокого обучения. Сначала модель обучается на данных педагогических экспериментов для выявления общих закономерностей и признаков применимости, после чего выполняется дообучение модели под конкретный прием обучения.

Предложенный подход может быть внедрен в существующие системы адаптивного обучения при наличии сбора детального цифрового следа обучения и поддержки управляемых методов адаптации.

Благодарности

Работа выполнена в рамках темы государственного задания НИЦ «Курчатовский институт» – НИИСИ по теме № FNEF-2024-0001 (1023032100070-3-1.2.1).

СПИСОК ЛИТЕРАТУРЫ

1. *Isidro C., Carro R. M., Ortigosa A.* Dropout Detection in MOOCs: An Exploratory Analysis // 2018 International Symposium on Computers in Education (SIIE), Jerez, Spain, 2018, P. 1–6. https://doi.org/10.1109/SIIE.2018.8586748

2. *Whitehill J., Mohan K., Seaton D., Rosen Y., Tingley D.* Delving Deeper into MOOC Student Dropout Prediction, arXiv:1702.06404 [cs.AI], 2017.

https://doi.org/10.48550/arXiv.1702.06404

3. *Dass S., Gary K., Cunningham J.* Predicting Student Dropout in Self-Paced MOOC Course Using Random Forest Model // Information. 2021. Vol. 12, 476. https://doi.org/10.3390/info12110476

4. *Anh B.N., Giang N.H., Hai N.Q., Minh T.N., Son N.T., Chien B.D.* An University Student Dropout Detector Based on Academic Data // 2023 IEEE Symposium on Industrial Electronics & Applications (ISIEA), Kuala Lumpur, Malaysia, 2023. P. 1–8. https:// doi.org/10.1109/ISIEA58478.2023.10212223

6. *Shohag S.I., Bakaul M.A.* Machine Learning Approach to Detect Student Dropout at University // International Journal of Advanced Trends in Computer Science and Engineering. 2021. Vol. 10, No. 6. https://doi.org/10.30534/ijatcse/2021/041062021

6. *Oancea B., Dragoescu R., Ciucu S.* Predicting Students' Results in Higher Education Using Neural Networks // International Conference on Applied Information and Communication Technologies (AICT2013), 25–26 April, 2013, Jelgava, Latvia. https://llufb.llu.lv/conference/AICT/2013/Oancea_AICT2013.pdf (дата обращения: 25.06.2026)

7. *Леонов А.Г., Матюшин М.А., Дьяченко М.С.* Об опыте построения априорной оценки успеваемости студентов в системе Мирера с использованием нейронных сетей // Успехи кибернетики. 2021. Т. 2, № 4. С. 49–59. https://doi.org/10.51790/2712-9942-2021-2-4-5

8. *Мащенко К.А.* Сигнатурные методы анализа временных рядов. Электронные библиотеки. 2025. Т. 28, № 3. С. 681-700. https://doi.org/10.26907/1562-5419-2025-28-3-681-700.

9. *Дьяченко М.С., Леонов А.Г.* Решение задачи автоматизации учебного процесса с помощью экспериментального поиска индивидуальной образовательной траектории // Информатика и образование. 2024. Т. 39, № 4. С. 14–26. https://doi.org/10.32517/0234-0453-2024-39-4-14-26

10. *Карманова Е.В., Захарова Я.М., Киселев А.В.* Интеллектуальный анализ образовательных результатов студентов университета с использованием LMS Moodle // Информатика и образование. 2024. Т. 39, № 3. С. 16–28. https://doi.org/10.32517/0234-0453-2024-39-3-16-28

11. *Леонов А.Г., Мащенко К.А., Тарасова Е.Ю.* Новая архитектура образовательных платформ в эпоху цифровой трансформации: парадигмы nolms и smart lms // Труды НИИСИ. 2025. Т. 15, № 2. С. 54–59. https://doi.org/ 10.25682/NIISI.2025.2.0007

12. *Dahal, P., Nugroho, S., Schmidt, C., Sänger, V.* AI-Based Learning Recommendations: Use in Higher Education // Future Internet. 2025. Vol. 17, No. 7. https://doi.org/10.3390/fi17070285

13. *Васильев И.А., Караваева А.С., Леонов А.Г. и др.* О нововведениях в цифровой образовательной платформе Мирера // Труды НИИСИ РАН. 2023. Т. 13, № 4. С. 127–134. https://doi.org/10.25682/NIISI.2023.4.0018

14. *Вайнштейн Ю.В.* Педагогическое проектирование персонализированного адаптивного предметного обучения студентов вуза в условиях цифровизации: диссертация доктора педагогических наук. Красноярск, 2021. 425 с.

15. *Кречетов И.А., Романенко В.В.* Реализация методов адаптивного обучения // Вопросы образования. 2020. № 2. С. 252–277. https://doi.org/10.17323/1814-9545-2020-2-252-277

16. *Piech C., Bassen J., Huang J. et al.* Deep Knowledge Tracing // Proc. 28th Int. Conf. Neural Information Processing Systems (NIPS 2015)). 2015. No. 1. P. 505–515.

17. AdaptKT: A Domain Adaptable Method for Knowledge Tracing / S. Cheng, Q. Liu, E. Chen et al. Текст : электронный // Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining : WSDM ’22. – New York, NY, USA: Association for Computing Machinery, 2022. AdaptKT. P. 123–131.

AUTOMATIC CORRECTION OF INDIVIDUAL EDUCATIONAL TRAJECTORY IN ADAPTIVE LEARNING SYSTEMS BASED ON A COMPOSITE APPROACH

M. S. Diachenko[1] [0000-0002-5809-4981], A. G. Leonov[2] [0000-0001-9622-1526]

[1, 2]*NRC «Kurchatov Institute» – SRISA, Moscow, Russia*
[2]*Lomonosov Moscow State University, Moscow, Russia*
[2]*Moscow Pedagogical State University, Moscow, Russia*
[2]*State University of Management, Moscow, Russia*

[1]mdyachenko@niisi.ru, [2]dr.l@vip.niisi.ru

Abstract

In the context of the expanding use of adaptive courses in higher education, the problem of early identification of underperforming students and automatic adjustment of their individual educational trajectory becomes increasingly relevant. Existing methods for identifying underperforming students do not provide an in-depth analysis of the individualized learning approach employed by the student. The generalized adaptive learning algorithms currently in use do not directly address the task of identifying underperforming students, which may lead to their late detection – by which time it is no longer possible to adjust the students' approach to learning to ensure successful course completion. The identified problem requires an immediate solution.

The aim of the study was to develop algorithms for identifying underperforming students in adaptive learning systems using a composite approach – based on analysing the individualized learning approach and enabling automatic adjustment of the individual educational trajectory to ensure successful course completion.

The developed algorithms detect student underperformance in mastering an academic discipline by predicting the outcomes of individualized learning using a domain model. In addition, algorithms for the automatic adjustment of the individual educational trajectory have been proposed. Unlike existing approaches, these algorithms generate an individualized educational strategy to ensure successful completion of the course. These algorithms can be integrated into existing adaptive learning systems to increase the proportion of students who successfully complete adaptive courses.

Keywords: *adaptive learning, identification of underperforming students, individualized educational trajectory.*

REFERENCES

1. *Isidro C., Carro R. M., Ortigosa A.* Dropout Detection in MOOCs: An Exploratory Analysis // 2018 International Symposium on Computers in Education (SIIE), Jerez, Spain, 2018, P. 1–6. https://doi.org/10.1109/SIIE.2018.8586748

2. *Whitehill J., Mohan K., Seaton D., Rosen Y., Tingley D.* Delving Deeper into MOOC Student Dropout Prediction, arXiv:1702.06404 [cs.AI], 2017. https://doi.org/10.48550/arXiv.1702.06404

3. *Dass S., Gary K., Cunningham J.* Predicting Student Dropout in Self-Paced MOOC Course Using Random Forest Model // Information. 2021. Vol. 12, 476. https://doi.org/10.3390/info12110476

4. *Anh B.N., Giang N.H., Hai N.Q., Minh T.N., Son N.T., Chien B.D.* An University Student Dropout Detector Based on Academic Data // 2023 IEEE Symposium on Industrial Electronics & Applications (ISIEA), Kuala Lumpur, Malaysia, 2023. P. 1–8. https://doi.org/10.1109/ISIEA58478.2023.10212223

5. *Shohag S.I., Bakaul M.A.* Machine Learning Approach to Detect Student Dropout at University // International Journal of Advanced Trends in Computer Science and Engineering. 2021. Vol. 10, No. 6. https://doi.org/10.30534/ijatcse/2021/041062021

6. *Oancea B., Dragoescu R., Ciucu S.* Predicting Students' Results in Higher Education Using Neural Networks // International Conference on Applied Information and Communication Technologies (AICT2013), 25–26 April, 2013, Jelgava, Latvia https://llufb.llu.lv/conference/AICT/2013/Oancea_AICT2013.pdf (accessed: 25.06.2026)

7. *Leonov A.G., Matyushin M.A., Dyachenko M.S.* A neural network-based approach to a priori assessment of the student academic performance in the Mirera digital learning platform // Uspekhi kibernetiki. 2021. Vol. 2, No. 4. P. 49–59. https://doi.org/10.51790/2712-9942-2021-2-4-5

8. *Mashchenko K.A.* Signature Methods for Time Series Analysis. Russian Digital Libraries Journal. 2025. Vol. 28, No. 3. P. 681-700. (In Russ.). https://doi.org/10.26907/1562-5419-2025-28-3-681-700

9. *Diachenko M.S., Leonov A.G.* Solving the problem of automating the learning process through experimental search for an individual educational trajectory // Informatics and education. 2024. Vol. 39, No. 4. P. 14–26 (In Russ.). EDN: BMEVBI. https://doi.org/10.32517/0234-0453-2024-39-4-14-26

10. *Karmanova E.V., Zakharova Ya.M., Kiselev A.V.* Educational data mining in a university using LMS Moodle // Informatics and Education. 2024. Vol. 39, No. 3. P. 16–28. (In Russ.) https://doi.org/10.32517/0234-0453-2024-39-3-16-28

11. *Leonov A.G., Mashchenko K.A., Tarasova E.Yu.* Novaya arkhitektura obrazovatel'nykh platform v epokhu tsifrovoi transformatsii: paradigmy nolms i smart lms // Trudy NIISI. 2025. T. 15, № 2. S. 54–59. (In Russ.)/ https://doi.org/10.25682/NIISI.2025.2.0007

12. *Dahal, P., Nugroho, S., Schmidt, C., Sänger, V.* AI-Based Learning Recommendations: Use in Higher Education // Future Internet. 2025. Vol. 17, No. 7. https://doi.org/10.3390/fi17070285

13. *Vasil'ev I.A., Karavaeva A.S., Leonov A.G. et al.* O novovvedeniyakh v tsifrovoi obrazovatel'noi platforme Mirera // Trudy NIISI RAN. 2023. T. 13, № 4. S. 127–134. (In Russ.). https://doi.org/10.25682/NIISI.2023.4.0018

14. *Vainshtein Yulia V.* Pedagogical Design of Personalized Adaptive Subject Learning for University Students in the Context of Digitalization. Doctoral dissertation, Siberian Federal University, 2021 (In Russ.).

15. *Krechetov I.A., Romanenko V.V.* Implementing the adaptive learning techniques // Educational Studies. Moscow. 2020. No. 2. P. 252–277 (In Russ.). EDN: KYNIIH. https://doi.org/10.17323/1814-9545-2020-2-252-277

16. *Piech C., Bassen J., Huang J. et al.* Deep Knowledge Tracing // Proc. 28th Int. Conf. Neural Information Processing Systems (NIPS 2015)). 2015. No. 1. P. 505–515

17. AdaptKT: A Domain Adaptable Method for Knowledge Tracing / S. Cheng, Q. Liu, E. Chen et al. // Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining: WSDM '22. New York, NY, USA: Association for Computing Machinery, 2022. AdaptKT. P. 123–131.

СВЕДЕНИЯ ОБ АВТОРАХ



ДЬЯЧЕНКО Михаил Сергеевич – закончил магистратуру в Московском государственном техническом университете «СТАНКИН» в 2006 г., в 2023 окончил аспирантуру по направлению 09.06.01 Информатика и вычислительная техника в НИИСИ РАН. В настоящее время инженер ОПИБ НИИСИ РАН, г. Москва. Область научных интересов: системы адаптивного обучения, распределенные системы управления.

Mikhail Sergeevich DIACHENKO – completed master's degree at the Moscow State Technical University "STANKIN" in 2006 and doctoral program in 2023 at the Scientific Research Institute of System Analysis and Control (SRISA RAS), specializing in Computer Science and Engineering (09.06.01). Currently an engineer at SRISA RAS, Moscow. Research interests: adaptive learning systems and distributed control systems.

email: mdyachenko@niisi.ru

ORCID: 0000-0002-5809-4981



ЛЕОНОВ Александр Георгиевич – доктор педагогических наук (кандидат физико-математических наук). В настоящее время работает руководителем цифровой трансформации, заведующим кафедрой дополнительного профессионального образования НИЦ «Курчатовский институт» - НИИСИ, ведущим научным сотрудником механико-математического факультета МГУ имени М.В. Ломоносова, профессором кафедры математики и информатики в начальной школе МПГУ.

Alexander Georgievich LEONOV – doctor of Pedagogical Sciences (candidate of Physical and Mathematical Sciences). Currently, he is CDTO (Chief Digital Transformation Officer), head of the Department of Continuing Professional Education at the NRC «Kurchatov Institute» – SRISA, leading researcher at the Laboratory of Computational Methods of the Faculty of Mechanics and Mathematics of Lomonosov Moscow State University, professor at the Department of Mathematics and Informatics in Elementary School of Moscow Pedagogical State University.

email: dr.l@math.msu.su

ORCID: 0000-0001-9622-1526

Материал поступил в редакцию 30 июля 2026 года

УДК 004.85

МУЛЬТИАГЕНТНАЯ СИСТЕМА ДЛЯ ЦИФРОВОЙ ИНТЕЛЛЕКТУАЛЬНОЙ СФЕРЫ ГЕОЛОГИЧЕСКИХ ЗНАНИЙ

В. С. Еременко[1] [0000-0002-5250-5743], В. В. Наумова[2] [0000-0002-3001-1638]

[1, 2]*Государственный геологический музей им. В. И. Вернадского РАН, г. Москва, Россия*

[1]vitaer@gmail.com, [2]naumova_new@mail.ru

Аннотация

Предложен новый подход к построению территориально-распределенных информационных систем на основе мультиагентной архитектуры, позволяющий отказаться от жесткой унификации протоколов и каталогизации ресурсов в пользу динамической интеграции источников данных. Ключевым элементом системы является агент верификации, обеспечивающий проверку достоверности результатов научных исследований через формальный контроль и использование нескольких больших языковых моделей. Разработанный прототип, реализованный на платформе GeologyScience.ru, устраняет недостатки существующего ИИ-ассистента, связанные с галлюцинациями и недостоверностью информации, и создает основу для накопления проверенных знаний.

Ключевые слова*: мультиагентные системы, территориально-распределенные информационные системы, верификация результатов, большие языковые модели (LLM), ИИ-агенты.*

ВВЕДЕНИЕ

В рамках классического подхода к построению территориально-распределенных информационных систем широкое распространение получили решения, основанные на стандартизованных протоколах доступа, унифицированных форматах метаданных и централизованной каталогизации ресурсов [1]. Однако стремительное развитие технологий искусственного интеллекта, в частности автономных агентов, открывает возможность отказаться от жесткой предварительной унификации: агент способен динамически интерпретировать семантику разнородных источников данных, компенсируя отсутствие единого протокола до-

© В. С. Еременко, В. В. Наумова, 2026.
Данная статья распространяется на условиях международной лицензии Creative Commons License Attribution 4.0 International (CC BY 4.0).

ступа и централизованного реестра за счет анализа структуры и контекста запроса в реальном времени [2, 3].

1. АНАЛИЗ СУЩЕСТВУЮЩЕГО ПОДХОДА И ПОСТАНОВКА ЗАДАЧИ

1.1. Цифровая экосистема и виртуальный ассистент

Ранее в рамках развития платформы GeologyScience.ru [4] был реализован и апробирован подход к организации цифрового рабочего пространства геолога-исследователя [5]. Пользователю предоставляется единая точка входа в формате виртуального рабочего кабинета, в котором отображаются доступные данные и сервисы обработки данных. Поиск данных осуществляется по территориально-распределенным источникам геологических данных с возможностью визуализации и сохранения данных в личном кабинете пользователя (рис. 1). Функциональность системы расширяется за счет подключения плагинов.

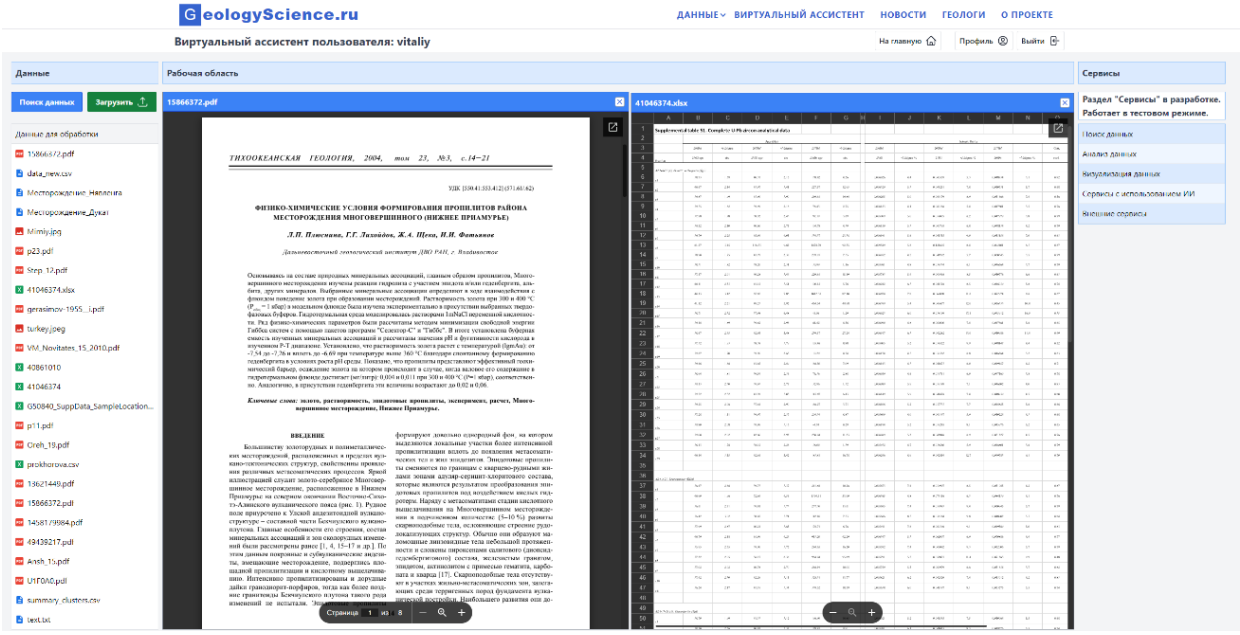


Рис. 1. Интерфейс виртуального рабочего кабинета пользователя.

Развитием этого подхода стала разработка виртуального ИИ-ассистента геолога-исследователя (рис. 2) [6, 7]. Целью настоящей разработки является предоставление пользователю возможности уточнения вопросов, связанных с темой его научного исследования, у ИИ-ассистента в формате текстового чата. В качестве базовой языковой модели при создании ИИ-ассистента использовалась большая языковая модель (LLM) DeepSeek R1 (https://deepseek.com).

ИИ-ассистент на основе данных, использованных при обучении модели, и набора системных промптов, написанных нами, предлагает пользователю рекомендации по теме его исследования, включающие информацию об авторах публикаций по теме исследования, применяющихся методов и подходов к решению похожих задач и т. д. Более подробно каждый из этапов развития системы описан авторами в [8].

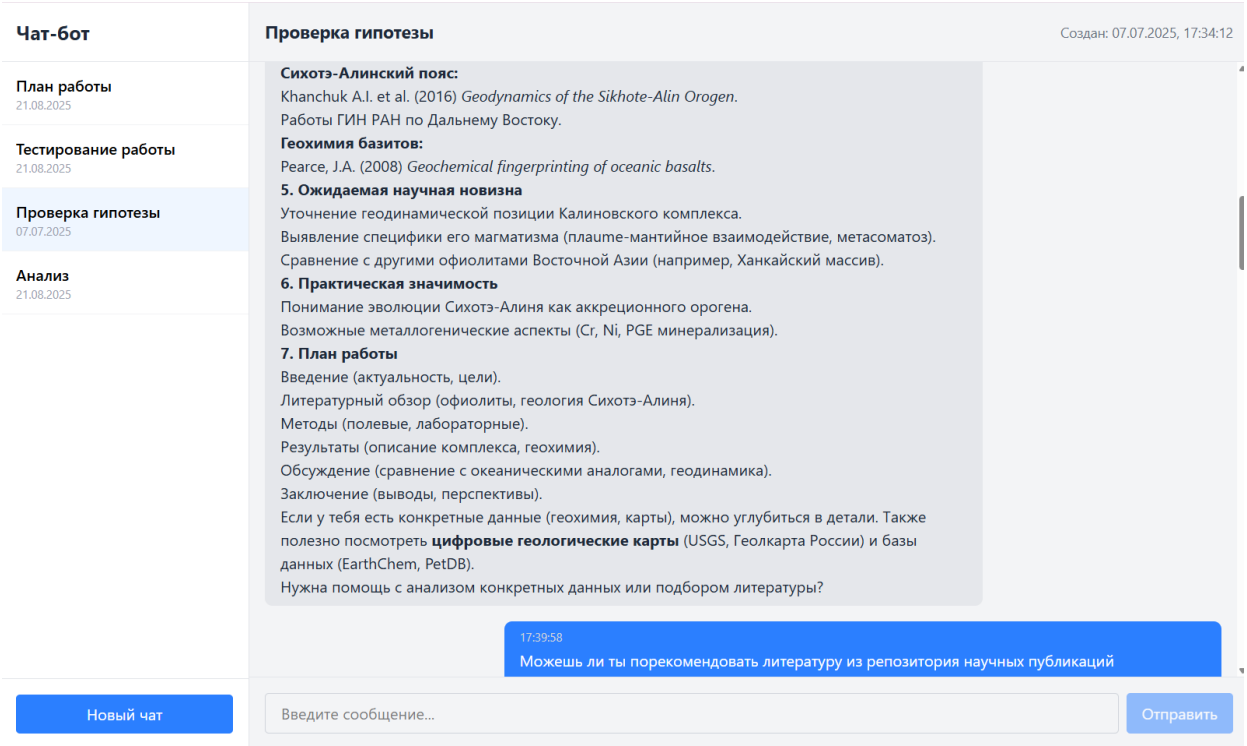


Рис. 2. Интерфейс чат-бота виртуального ассистента.

1.2. Недостатки настоящего подхода

В процессе эксплуатации ИИ-ассистента выявился ряд недостатков, среди которых можно выделить некоторую недостоверность информации, предлагаемой ИИ-ассистентом в ряде случаев, например использование несуществующих научных публикаций или информации о содержимом конкретной публикации. Кроме того, применение заранее обученной большой языковой модели без использования механизмов дообучения не позволяет получать актуальные данные по теме исследования. Кроме того, ИИ-ассистент не взаимодействует с имеющимися в доступе в информационно-аналитической среде geologyscience.ru территориально-распределенными сервисами для поиска актуальных данных

по теме исследования, размещенных в репозиториях и информационных системах. Для решения этих задач было решено перейти к агентно-ориентированной архитектуре, позволяющей при необходимости вызывать сторонние сервисы по поиску и обработке геологических данных, для предоставления пользователю актуальной информации из достоверных источников.

1.3. Агентно-ориентированный подход

Агентно-ориентированный подход предполагает использование ИИ-агентов с разделением по функциям и ролям. В основе ИИ-агента лежит LLM, которая позволяет агенту анализировать полученную от пользователя или другого агента задачу и выполнять ряд действий для ее решения, которые сложно заранее предвидеть. ИИ-агент может использовать доступные ему инструменты, например сервисы для поиска или обработки геологических данных, чтобы потом на основе полученного результата выдать пользователю ответ на поставленную задачу. В случае использования нескольких ИИ-агентов одним из подходов к организации их работы является использование мультиагентных систем (рис. 3). В таких системах ИИ-агенты общаются между собой либо получают задания от агента-оркестратора, в зависимости от выбранной архитектуры, для координированного решения поставленных задач.

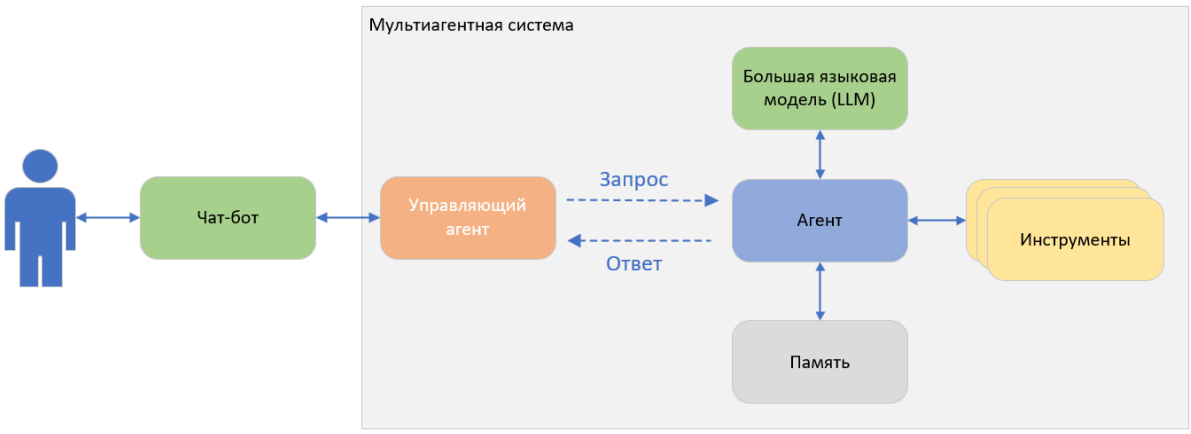


Рис. 3. Упрощенная схема взаимодействия в мультиагентной системе.

2. ПРОЕКТИРОВАНИЕ СИСТЕМЫ

2.1. Требования к системе

На основе анализа недостатков предыдущей реализации и современных подходов к построению мультиагентных систем были сформулированы функциональные требования к новой архитектуре:

- обработка и интерпретация пользовательских запросов на естественном языке;
- автоматический поиск требующихся для исследователя данных и вызов сервисов обработки;
- агрегация полученных результатов и формирование протокола исследования;
- верификация полученных результатов и накопление знаний.

Для реализации этих требований предложено использовать мультиагентную архитектуру системы с отдельным агентом-оркестратором, который разбивает сложную задачу на подзадачи и назначает исполнителей из числа доступных специализированных агентов. Каждый агент должен уметь обращаться к сервисам поиска, анализа и визуализации геологических данных, которые имеют четкие интерфейсы вызова и форматы возвращаемых результатов.

Для валидации полученных в процессе работы системы результатов предложена реализация отдельного ИИ-агента верификации. Такой агент при необходимости будет проверять результаты, полученные от других агентов, и искать в них возможную недостоверную информацию. Для повышения качества работы такого агента предложено использование нескольких различных LLM-моделей. В случае обнаружения ошибок или аномалий система должна инициировать повторное выполнение задания или запросить уточнение у пользователя. Кроме того, система должна обеспечивать накопление верифицированных протоколов выполнения для последующего использования в сходных запросах, тем самым постепенно повышая свою эффективность. Важным нефункциональным требованием выступает масштабируемость: добавление новых сервисов в каталог не должно требовать модификации ядра мультиагентной системы.

2.2. Функциональная схема

Предлагаемая мультиагентная система построена вокруг четырех основных функциональных блоков (рис. 4).

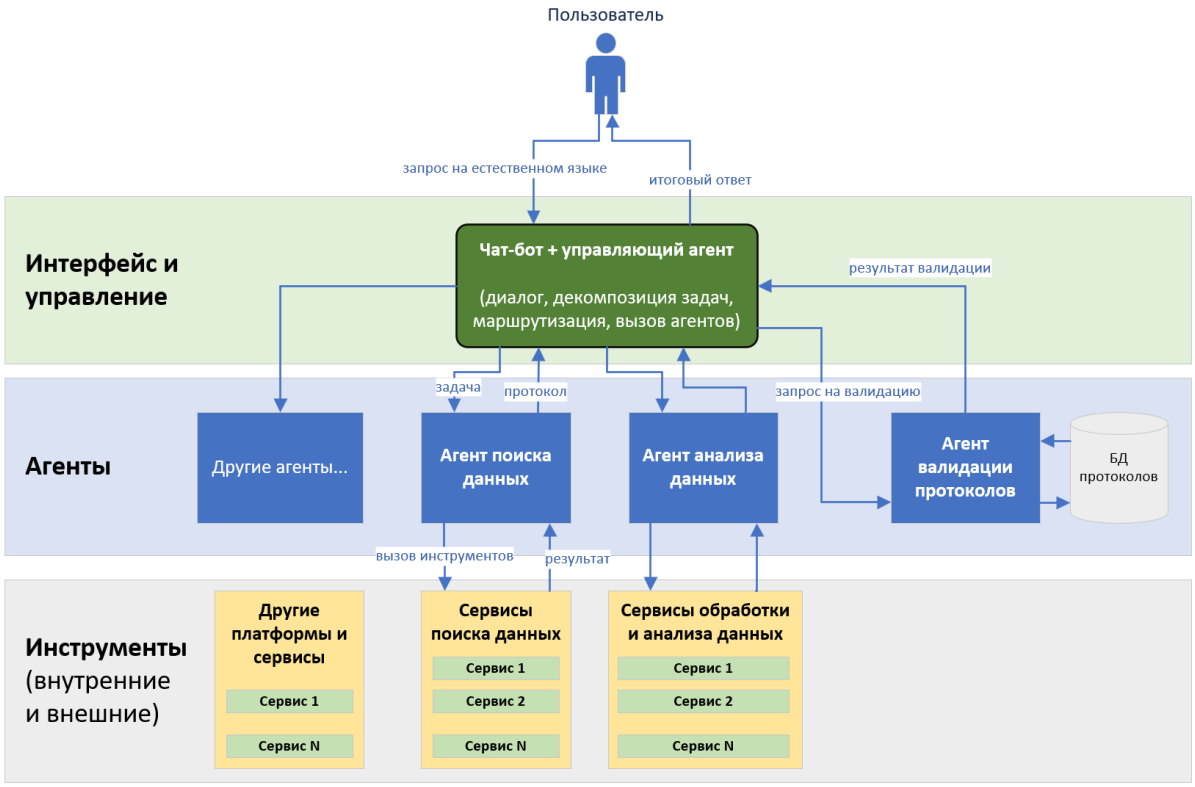


Рис. 4. Схема предлагаемой архитектуры мультиагентной системы.

Первый блок – это пользовательский интерфейс, который обеспечивает диалог с исследователем на естественном языке и отображает результаты работы сервисов, включая визуализации и ссылки на скачивание данных. Вторым блоком выступает оркестратор, или управляющий агент. Он принимает запрос пользователя, интерпретирует его, формирует план выполнения и последовательно передает управление специализированным агентам. Третий блок – это набор исполнительских агентов, каждый из которых закреплен за определенным типом сервисов: агент поиска данных обращается к каталогу публикаций и геологических коллекций, агент анализа запускает вычислительные сервисы обработки геохимических или пространственных данных, агент визуализации отвечает за построение графиков и карт. Четвертым блоком является агент верификации, который получает результаты каждого выполненного действия, проверяет их на соответствие формальным критериям (наличие обязательных полей,

отсутствие критических ошибок выполнения) и при необходимости запускает повторное выполнение или помечает результат как недостоверный. По мере надобности агенты взаимодействуют через шину сообщений, а каталог сервисов хранится в отдельном реестре, доступном для чтения каждому агенту.

2.3. Описание основных компонентов системы

Оркестратор является центральным компонентом, реализованным на базе LLM с поддержкой рассуждений. Его задача заключается в том, чтобы удержать контекст диалога, сформировать план действий, выбрать подходящего исполнительского агента и обработать возвращенные после верификации результаты.

Исполнительские агенты представляют собой узкоспециализированные программные модули, каждый из которых знает, как обратиться к конкретной группе сервисов из каталога, сформировать корректный запрос к API и интерпретировать стандартизированный ответ. Например, агент поиска данных умеет работать с текстовыми поисковыми движками, агент обработки – с WPS-сервисами (Web Processing Service), агент визуализации – с библиотеками построения графиков и карт.

Агент верификации проверяет результаты, полученные при выполнении задачи. При успешной проверке протокол выполнения (запрос, какой сервис вызывался, с какими параметрами, какой получен результат) сохраняется в базе знаний. Этот накопленный материал может использоваться в будущем для ускорения обработки аналогичных запросов или автоматического формирования рекомендаций.

3. РАЗРАБОТКА, ТЕСТИРОВАНИЕ И ПОЛУЧЕННЫЕ РЕЗУЛЬТАТЫ

3.1. Используемые технологии для реализации компонентов

При разработке прототипа мультиагентной системы был проведен анализ инструментов с открытым исходным кодом для построения таких систем (табл. 1).

Табл. 1. Сравнение инструментов для реализации мультиагентных систем.

Критерий	LangFlow	LangGraph	CrewAI	AutoGen	Smolagents
Визуальный конструктор	Да	Нет (интегрируется)	Нет	Да (Studio)	Нет
Готовый чат-интерфейс	Да	Нет	Нет	Да	Нет
Развертывание как API	Да	Да	Да	Да	Да
Поддержка многоагентности	Да	Да	Да	Да	Частично
Активность проекта	Активный (MIT)	Активный (MIT)	Активный (MIT)	Режим поддержки	Активный (Apache)

Практически каждый из этих инструментов удовлетворяет поставленным критериям и позволяет решить поставленную задачу, поэтому выбор был сделан в пользу платформы LangFlow (https://www.langflow.org) как инструменту с наболее низким порогом входа (рис. 5). Эта платформа предоставляет ряд готовых компонентов для создания мультиагентной системы, что упрощает и ускоряет разработку прототипа системы на основе предложенного подхода. В качестве базовой LLM для ИИ-агентов был выбран DeepSeek V4. Для реализации пользовательского интерфейса использован фреймворк Vue 3 (https://vuejs.org).

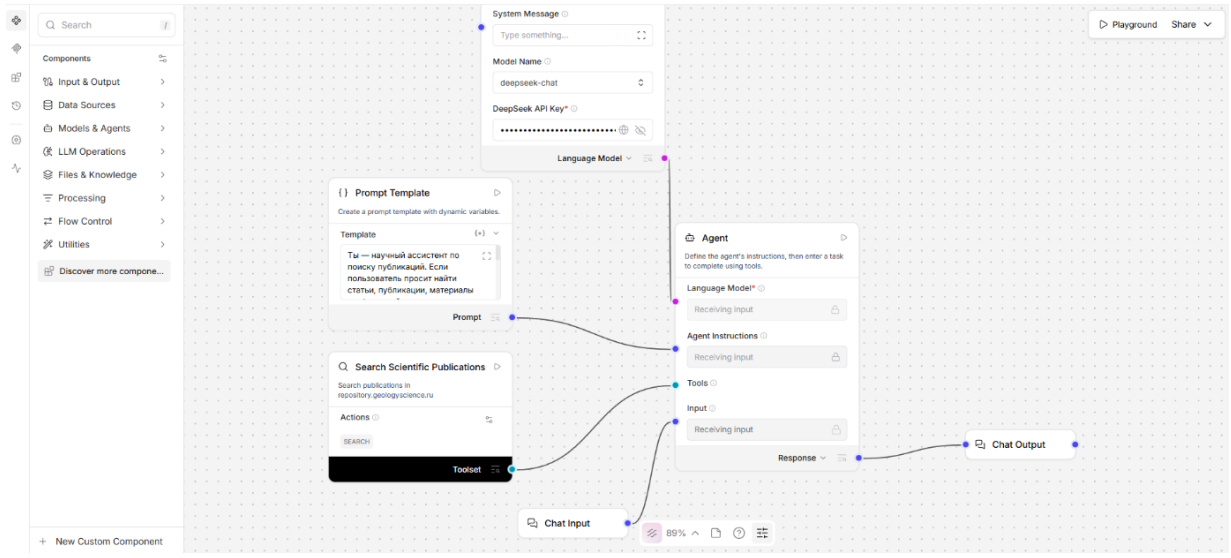


Рис. 5. Пользовательский интерфейс платформы Langflow.

ЗАКЛЮЧЕНИЕ

Спроектирована и реализована мультиагентная архитектура, которая устраняет выявленные недостатки существующего ИИ-ассистента за счет перехода от статической модели к динамической интеграции распределенных сервисов. Предложенный подход заключается в децентрализации функций: оркестратор формирует план исследования, исполнительские агенты обращаются к актуальным источникам данных (репозиториям, WPS-сервисам, каталогам публикаций) и сервисам анализа и обработки геологических данных [9], а агент верификации обеспечивает контроль достоверности на основе формальных проверок и сравнения результатов нескольких моделей. Решение для верификации, реализованное в рамках системы, позволяет не только выявлять галлюцинации LLM и фактические ошибки, но и накапливать подтвержденные протоколы, что трансформирует систему из простого чат-бота в инструмент научного исследования с функцией прослеживаемости результатов.

Благодарности

Работа выполняется в рамках Государственной темы Государственного геологического музея им. В. И. Вернадского РАН (FMMR-2025-0001) Рег. №1022061000107-4-1.5.1 «Естественнонаучные музеи в едином цифровом пространстве геологических знаний».

СПИСОК ЛИТЕРАТУРЫ

1. *Жижимов О.Л., Мазов Н.А.* Принципы построения распределенных информационных систем на основе протокола Z39.50. Новосибирск: ОИГГМ СО РАН, 2004. 361 с.

2. *Gottweis J. et al.* Towards an AI co-scientist // Arxiv.org. 2025. https://doi.org/10.48550/arXiv.2502.18864

3. *Наумова В.В., Патук М.И., Еременко А.С., Загуменнов А.А., Еременко В.С.* Цифровые технологии будущего для научных исследований в геологии // Электронные библиотеки. 2025. Том 28, № 4. С. 788–805.

4. *Наумова В.В., Еременко В.С., Загуменнов А.А., Патук М.И.* Научный портал Geologyscience.ru: текущее состояние и перспективы развития// Геоинформатика. 2023. № 3. С. 33–43.

5. *Еременко В.С., Наумова В.В., Загуменнов А.А.* Разработка виртуальных пространств с интеллектуальным адаптивным интерфейсом для цифровых геологических систем // Информационные и математические технологии в науке и управлении. 2024. № 1 (33). С. 146–153.

6. *Еременко В.С., Наумова В.В.* Виртуальный ассистент геолога-исследователя // Электронные библиотеки, 2025, Том 28, № 4. С. 781–787.

7. *Naumova V.V., Zagumennov A.A., Eremenko V.S.* Digital Tools and Virtual Assistants to Support Scientific Research in Geology // Automatic Documentation and Mathematical Linguistics. 2024. Vol. 58, Suppl. 5. P. S287–S291.

8. *Еременко В.С., Наумова В.В.* Мультиагентная система – AI виртуальный соавтор геолога-исследователя // Цифровые технологии будущего — современные решения в науках о Земле ITES-2025: Материалы VII Всероссийской конференции с международным участием, Владивосток, 22–26 сентября 2025 г. / сост. А.А. Стрельцова. Владивосток: Издательство Дальневосточного федерального университета, 2025. С. 22.

9. *Патук М.И., Наумова В.В.* Анализ текстов геологических публикаций с использованием больших языковых моделей // Электронные библиотеки. 2025. Т. 28, № 4. С. 806–821.

MULTI-AGENCY SYSTEM FOR DIGITAL INTELLIGENCE IN GEOLOGICAL KNOWLEDGE

V. S. Eremenko[1] [0000-0002-5250-5743], V. V. Naumova[2] [0000-0002-3001-1638]

[1, 2]*Vernadsky State Geological Museum of Russian Academy of Sciences, Moscow*

[1]vitaer@gmail.com, [2]naumova_new@mail.ru

Abstract

A new approach to building geographically distributed information systems based on a multi-agent architecture is proposed. This approach eliminates the rigid unification of protocols and resource cataloging in favor of dynamic integration of data sources. A key element of the system is a verification agent, which verifies the validity of scientific research results through formal control and the use of multiple LLMs. The developed prototype, implemented on the GeologyScience.ru platform, eliminates the shortcomings of the existing AI assistant related to hallucinations and unreliability of information and creates a foundation for accumulating verified knowledge.

Keywords*: multi-agent systems, geographically distributed information systems, verification of results, large language models (LLM), AI agents.*

REFERENCES

1. *Zhizhimov O.L., Mazov N.A.* Principy postroenija raspredelennyh informacionnyh sistem na osnove protokola Z39.50. Novosibirsk: OIGGM SO RAN, 2004. 361 s.

2. *Juraj Gottweis et al.* Towards an AI co-scientist // Arxiv.org. 2025. https://doi.org/10.48550/arXiv.2502.18864

3. *Naumova V.V., Patuk M.I., Eremenko A.S., Zagumennov A.A., Eremenko V.S.* Cifrovye tehnologii budushhego dlja nauchnyh issledovanij v geologii // Russian Digital Libraries Journal. 2025. Tom 28, № 4. S. 788–805.

4. *Naumova V.V., Eremenko V.S., Zagumennov A.A., Patuk M.I*. Nauchnyj portal Geologyscience.ru: tekushhee sostojanie i perspektivy razvitija// Geoinformatika. 2023. № 3. S. 33–43.

5. *Eremenko V.S., Naumova V.V., Zagumennov A.A.* Razrabotka virtual'nyh prostranstv s intellektual'nym adaptivnym interfejsom dlja cifrovyh geologicheskih sistem // Informacionnye i matematicheskie tehnologii v nauke i upravlenii. 2024. № 1 (33). P. 146–153.

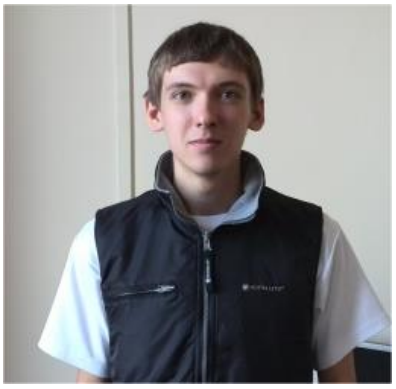
6. *Eremenko V.S., Naumova V.V.* Virtualnyj assistent geologa-issledovatelja // Russian Digital Libraries Journal. 2025. Tom 28, № 4. S. 781–787.

7. *Naumova V.V., Zagumennov A.A., Eremenko V.S.* Digital Tools and Virtual Assistants to Support Scientific Research in Geology // Automatic Documentation and Mathematical Linguistics. 2024. Vol. 58, Suppl. 5. P. S287–S291.

8. *Eremenko V.S., Naumova V.V.* Multiagentnaja sistema – AI virtualnyj soavtor geologa-issledovatelja // Cifrovye tehnologii budushhego — sovremennye reshenija v naukah o Zemle ITES-2025: Materialy VII Vserossijskoj konferencii s mezhdunarodnym uchastiem, Vladivostok, 22–26 sentjabrja 2025 g. / sost. A.A. Strel'cova. Vladivostok: Izdatelstvo Dalnevostochnogo federalnogo universiteta, 2025. S. 22.

9. *Patuk M.I., Naumova V.V.* Analiz tekstov geologicheskih publikacij s ispolzovaniem bolshih jazykovyh modelej // Russian Digital Libraries Journal. 2025. Tom 28, № 4. S. 806–821.

СВЕДЕНИЯ ОБ АВТОРАХ



ЕРЕМЕНКО Виталий Сергеевич – научный сотрудник, Государственный геологический музей им. В. И. Вернадского РАН, Москва.

Vitaliy Sergeevich EREMENKO – researcher at the SGM scientific department, Vernadsky State Geological Museum of Russian Academy of Sciences, Moscow.

email: vitaer@gmail.com

ORCID: 0000-0002-5250-5743



НАУМОВА Вера Викторовна – доктор геологическо-минералогических наук, главный научный сотрудник, зав. Научным отделом Государственного геологического музея им. В. И. Вернадского РАН, Москва.

Vera Viktorovna NAUMOVA – Doctor of Geological and Mineralogical Sciences, chief researcher and head of the Scientific Department of the Vernadsky State Geological Museum of Russian Academy of Sciences, Moscow.

email: naumova_new@mail.ru

ORCID: 0000-0002-3001-1638

Материал поступил в редакцию 20 июля 2026 года

УДК 81'322.2+004.855.5

ПОДХОД К ОЦЕНКЕ ДОСТОВЕРНОСТИ ИНФОРМАЦИИ В НАУЧНО-ПОПУЛЯРНОМ ДИСКУРСЕ НА ОСНОВЕ АНАЛИЗА АРГУМЕНТАЦИИ И НАУЧНОГО ФАКТ-ЧЕКИНГА

Ю. А. Загорулько[1] [0000-0002-7111-6524], И. Р. Ахмадеева[2] [0000-0002-7371-1087], А. С. Серый[3] [0000-0001-8275-4700], Е. А. Сидорова[4] [0000-0001-8731-3058],

[1–4]*Институт систем информатики им. А. П. Ершова СО РАН, г. Новосибирск, Россия*

[1]zagor@iis.nsk.su, [2]i.r.akhmadeeva@iis.nsk.su, [3]alexey.seryj@iis.nsk.su, [4]lsidorova@iis.nsk.su

Аннотация

Рассмотрены проблемы оценки достоверности информации, извлекаемой из научно-популярных текстов. Как известно, научно-популярный жанр характеризуется свободным стилем изложения материала, нестрогими отсылками к первоисточникам и экспертным мнениям, наличием спорных утверждений и выводов. Уточнение этих компонент авторских рассуждений может существенно повысить степень доверия к ним критически настроенных читателей или, наоборот, позволит опровергнуть тезисы, приводимые авторами.

В работе предложен новый подход к решению задачи оценки достоверности информации, содержащейся в научно-популярных публикациях. Подход предполагает применение теории аргументации Д. Уолтона для выявления потенциально недостоверных утверждений и логических ошибок в рассуждениях. Поиск в научных базах данных и вычисление метрик доверия к научным публикациям на основе их метаданных и научных рейтингов будет обеспечивать верификацию спорных утверждений и оценку их достоверности. Применение больших языковых моделей позволит сгенерировать объяснения принимаемых решений.

Для разработки и апробации этого подхода предложены две методики сбора данных. В первой используются корпуса русскоязычных научно-популярных текстов и комментарии к ним с аргументативной разметкой. Такая разметка включает, в том числе, полемические схемы аргументации, анно-

© Ю. А. Загорулько, И. Р. Ахмадеева, А. С. Серый, Е. А. Сидорова, 2026.
Данная статья распространяется на условиях международной лицензии Creative Commons License Attribution 4.0 International (CC BY 4.0).

тированные вручную, которые подвергают сомнению тезис или аргумент автора. Базой второй методики служит генерация с помощью LLM синтетических тезисов и научно-популярных статей на основе реальных научных публикаций.

***Ключевые слова**: проверка утверждений, анализ аргументации, научная проверка фактов, достоверность информации, автоматический сбор данных.*

ВВЕДЕНИЕ

Открытый доступ к информации является одной из особенностей современного общества. Быстрыми темпами идет развитие средств распространения и предоставления информации, побуждая все больше людей интересоваться различными сферами исследований, в том числе теми, где они не являются специалистами. Этому способствуют многочисленные онлайн-площадки и интернет-ресурсы, предоставляющие доступ к различным публикациям псевдонаучного и научно-популярного характера. Авторами таких материалов, как правило, являются журналисты и блогеры, преследующие цель привлечения максимальной аудитории. Вследствие этого заголовки статей могут не соответствовать их содержанию, а утверждения, взятые даже из достоверных источников, цитируются с искажениями. Это приводит к тому, что вся цепочка дальнейших рассуждений, построенных на данных утверждениях вместе с промежуточными и конечными выводами, становится сомнительной.

Современные методы и технологии обработки естественного языка уже достигли впечатляющих результатов во многих областях. Однако на данный момент они не предоставляют эффективных инструментов для выявления ненадежных утверждений. Опровержение и подтверждение обоснованности утверждений и мнений, высказанных автором, составляют область исследования теории аргументации, изучающей различные способы рассуждений в текстах и устной речи с философской, лингвистической, когнитивной и социальной точек зрения. Разработка компьютерных методов анализа дискурса, позволяющих выявлять тезисы и связанную с ними аргументацию, даст в руки пользователей эффективный инструмент для исследования информации (поиск обоснований, выявление противоречий, анализ источников), а применение методов оценки доверия на основе релевантных внешних источников информа-

ции позволит проводить численную оценку достоверности спорных утверждений.

Целью настоящей работы является разработка архитектуры системы, которая бы позволяла комплексно решать задачу анализа достоверности информации, представленной в русскоязычных интернет-публикациях. Работа посвящена решению следующих исследовательских задач.

RQ1. Как оценивать достоверность информации?

RQ2. Какие методологические и технологические принципы должны лежать в основе комплексной технологии оценки достоверности информации?

RQ3. Каковы возможные способы автоматизации сбора и генерации данных для проведения экспериментальных исследований по оценке достоверности информации?

Чтобы не затрагивать этические проблемы, а также вопросы, зависящие от политических и личных интересов, мы ограничились рассмотрением научно-популярных текстов, факты из которых поддаются объективной оценке на основе сопоставления с научными источниками данных, степень достоверности которых на сегодняшний день не вызывает серьезных сомнений, а возникающие противоречия поддаются оценочному сравнению на основе научных рейтингов изданий, авторов и самих публикаций.

1. ОБЗОР РАБОТ

Одним из распространенных подходов к выявлению недостоверной научной информации (misinformation) является фактчекинг [1], т. е. проверка утверждений на соответствие известным фактам. В общем виде данный подход заключается в верификации утверждений с опорой на доказательства. Выделяют следующие основные этапы фактчекинга [2]: (a) поиск релевантных документов (Document Retrieval), (b) извлечение доказательств (Evidence Retrieval) и (c) верификация утверждений (Claim Verification). При анализе сообщений или публикаций целиком добавляется предварительный этап – обнаружение сомнительных утверждений, требующих проверки (Claim Detection).

Обзор современных работ [3] позволяет заключить, что все больше исследований в данной области опираются на большие языковые модели (LLM), которые применяют не только при реализации основных этапов автоматиче-

ского фактчекинга, но и для генерации объяснений, создания синтетических данных [4, 5], а также для оценки качества предсказания других систем. При этом LLM подвержены галлюцинациям – они могут генерировать правдоподобную, но фактически неверную информацию, что особенно критично в задачах фактчекинга. В рамках этой проблемы в [6] разработан фреймворк FActScore, предназначенный для оценки фактической точности текстов, генерируемых LLM. Авторы предлагают верифицировать отдельные факты, содержащиеся в таких текстах, используя Википедию в качестве источника доказательств, и тем самым ставят вопрос о доверии к LLM как к источнику достоверных знаний.

Преобладающая часть исследований в области автоматизированного фактчекинга опирается на предположение о наличии надежного корпуса источников и сосредоточена на задаче верификации утверждений относительно него. Подходы же, включающие явную оценку доверия к источнику, представлены меньшим числом работ. Авторы [7] используют метод обнаружение фейковых новостей через оценку доверия к авторам и источникам на основе их прошлой истории публикаций. В работе [8] использована внешняя оценка достоверности доменов, полученная с фактчекингового сайта, которая учитывает уровень политической предвзятости, фактической точности и общей надежности.

Подходы, основанные исключительно на поиске и сопоставлении утверждения с источниками, имеют существенное ограничение: наличие релевантного и авторитетного источника не гарантирует корректности выводов, сделанных на его основе. На практике искажения могут возникать как при передаче информации посредством цитирования, так и вследствие некорректной интерпретации или ошибочной аргументации со стороны авторов. Это обусловило развитие подходов, рассматривающих анализ цепочек рассуждений (reasoning) в качестве дополнительного компонента фактчекинга [9]. При этом было показано [10], что LLM плохо справляются с определением ложности утверждения и могут принимать ложные утверждения за истинные при наличии некорректного цитирования, создающего видимость доказательства.

Одним из основных условий развития рассматриваемой области исследований является создание наборов данных, которые бы включали как досто-

верные, так и недостоверные факты. При этом такие датасеты, как правило, включают источники, на основании которых принималось решение. Большинство существующих датасетов ориентировано на социальные сети и политические новости (AveriTeC [11], LIAR [12]), фактологические утверждения из Википедии (HOVER) или тематику здравоохранения в контексте пандемии COVID-19 (COVID-Fact [13], CoVERt [14]). Среди ресурсов, непосредственно предназначенных для научного фактчекинга (scientific fact-checking), можно выделить PUBHEALTH [15], SciFact [16] и CoVERt [14]. Последний специализируется на проверке утверждений по данным клинических испытаний. Отличительной особенностью перечисленных датасетов является их преимущественно англоязычный характер. Русскоязычные ресурсы представлены значительно слабее, в частности датасетом ruSciFact [5].

В качестве проверяемых единиц в существующих датасетах обычно выступают утверждения (claim), которые подразделяются на два класса [1]: синтетические – формируемые аннотаторами или языковыми моделями на основе исходных текстов, и естественные – извлекаемые из реальных источников, включая сайты фактчекинга и публикации в социальных медиа. Первый тип утверждений является более простым для анализа моделями и не требует контекста, однако он значительно оторван от реальных сценариев использования. Так, датасет SciFact [16] содержит биомедицинские утверждения, изначально взятые из предложений со ссылкой на источники в научных статьях, которые затем были переформулированы аннотаторами в форму, более атомарную и удобную для обработки. Утверждения в ruSciFact [5] сгенерированы с использованием LLM и являются как позитивными – согласующимися с аннотациями научных статей, так и негативными – опровергаемыми этими аннотациями. Авторами датасета SciClaimHunt [4] предложены четыре способа генерации негативных утверждений на основе позитивных: генерация отрицания, сопоставление позитивному утверждению неверной статьи, изменение числовых значений и подмена именованных сущностей. Для формирования естественных утверждений применяют методы автоматического извлечения информации: в PUBHEALTH [15] использовались заголовки фактчекинговых статей о здравоохранении, из-за чего некоторые утверждения оказалась слишком общими, а

CoVERt [14] содержит утверждения из постов Twitter, извлеченные специализированной моделью.

В качестве источников доказательств преимущественно используют научные публикации, реже применяют новостные материалы и статьи Википедии [1]. Несмотря на то что научное знание формируется как результат накопления и переоценки результатов множества исследований, в большинстве существующих датасетов каждому утверждению сопоставляется небольшое число источников доказательств. Авторы SciFact-Open [16] частично преодолели это ограничение, расширив SciFact [16] до 500 тыс. источников. Распределение доказательств при этом оказалось неравномерным: половина всех доказательств сконцентрирована на 34 утверждениях (из 279), а максимальное количество доказательств, приходящихся на одно утверждение, достигает 19.

Датасет для выявления ошибок рассуждений MISSCI [17] и его расширение MISSCIPLUS [18] содержат примеры ошибочной аргументации, где недостоверные утверждения о здоровье (преимущественно о COVID-19) подкреплены ссылками на биомедицинские публикации. Этот датасет ориентирован на выявление имплицитных логических ошибок и содержит реконструкции ошибочных рассуждений помимо классификации их типов. Он построен на основе платформы HealthFeedback, с которой были отобраны статьи, разоблачающие некоторые утверждения, и ссылки на источники, которые некорректно интерпретируются в этих утверждениях. Разметка и восстановление ошибочной аргументации производились студентами-магистрантами.

В настоящей работе представлен новый подход к решению задачи оценки достоверности информации, содержащейся в научно-популярных публикациях, на базе интеграции практик, наиболее популярных на сегодняшний день, включая: а) использование теории аргументации Д. Уолтона [19] для выявления потенциально недостоверных утверждений и поиска логических ошибок в цепочках рассуждений; б) обнаружение в анализируемых текстах отсылок на внешние научные источники и использование последних для верификации спорных утверждений; в) использование метаданных и научных рейтингов внешних источников для вычисления оценки достоверности спорных тезисов.

Для разработки и апробации данного подхода предложены две методики сбора данных. В первой из них для составления датасетов используются рус-

скоязычные научно-популярные тексты, аннотированные вручную по аргументационным схемам и верифицированные экспертами как содержащие логические ошибки. В основе второй методики лежит генерация синтетических научно-популярных статей на базе реальных научных публикаций.

2. МЕТОДИКА АВТОМАТИЧЕСКОЙ ОЦЕНКИ ДОСТОВЕРНОСТИ ИНФОРМАЦИИ

Одной из ключевых особенностей научно-популярного жанра является то, что обсуждение научных достижений ведется в широкой аудитории, что нередко порождает ошибочную интерпретацию научных данных и появление недостаточно обоснованных суждений со стороны лиц, не являющихся специалистами в соответствующей области знаний. В связи с этим предлагаемая методика оценки достоверности информации включает такие задачи, как выявление авторских тезисов, используемых приемов рассуждений, отсылок к источникам и их последующая проверка на основе научных публикаций (см. рис. 1).

Архитектура системы оценки достоверности информации, поддерживающая предлагаемую методику, включает три основных контура: внутренний и внешний, а также контур принятия решений.

Внутренний контур. Основное его назначение состоит в анализе содержания исходного текста без обращения к внешним источникам информации. На вход контуру подается текст научно-популярной статьи, а результатом его работы является множество тезисов и валидирующих утверждений (см. п. 3.1), требующих проверки во внешних данных.

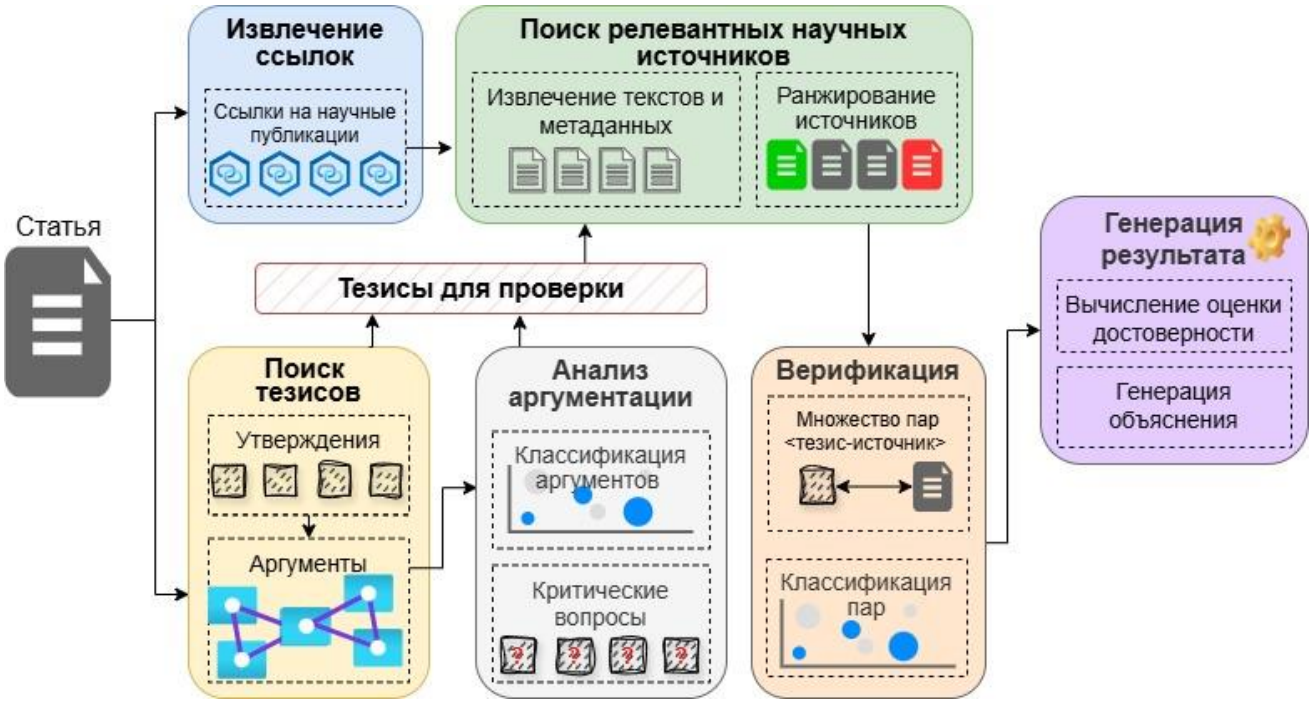


Рис. 1. Архитектура системы оценки достоверности информации.

Внешний контур. Его задачей является верификация тезисов, полученных от внутреннего контура, с привлечением внешней информации. Инструменты внешнего контура обращаются к научным базам данных, таким как arXiv, Semantic Scholar, OpenAlex, PubMed и др., в поисках актуальных научных исследований, содержащих информацию, семантически близкую заданным тезисам. На основе найденных источников осуществляется проверка тезисов (верификация), результаты которой передаются в контур принятия решений.

Контур принятия решений. Здесь выполняются общая оценка и принятие решения о степени достоверности статьи. Происходят вычисления оценки доверия для каждого тезиса, выделенного внутренним контуром, и общей оценки для всего документа, а также генерация структурированного объяснения результата.

2.1. Внутренний контур: извлечение ссылок, тезисов и критический анализ аргументации

Внутренний контур решает три основные задачи.

1. Извлечение ссылок на научные источники. В научно-популярных статьях авторы часто явным образом ссылаются на авторитетные источники, чтобы по-

высить доверие к высказываемым суждениям, однако такие отсылки могут быть некорректны и требуют проверки.

2. Выявление в тексте тезисов и аргументов. Аргументированные суждения, связывающие посылки с выводами, образуют основу общей структуры рассуждений в научных и научно-популярных статьях. На основе этой структуры выявляются главные тезисы, цепочки рассуждений, используемых авторами для обоснования своих тезисов.

3. Анализ обоснованности аргументации. Многие рассуждения, вводящие в заблуждение, строятся на некорректных логических выводах и требуют проверки.

Из исходного текста выделяются утверждения и аргументативные отношения между ними, на основе которых формируется граф аргументации (argumentation graph). Сегментация текста, определение границ аргументативных утверждений и идентификация тезисов (вершин графа) выполняются с помощью методов синтаксического и риторического анализа текста [20, 21], а также методов анализа семантического сходства утверждений. Каждая вершина графа соответствует утверждению (claim) или его составной части.

Классификация моделей рассуждений в соответствии с классификатором Д. Уолтона позволяет выделить аргументы, значимые для оценки достоверности утверждений исходной статьи. Принципиальным является разделение аргументов на «Внешние», указывающие на внешний источник информации как на основание убедительности, и «Внутренние», опирающиеся на логические связи между утверждениями (причинно-следственную, меронимическую, корреляционную и др.). Для проверки внутренних аргументов можно использовать критические вопросы, которые включены в модель Уолтона. В работе [22] описан способ генерации валидирующих (проверяющих корректность применения схемы) утверждений, которые совместно с посылкой и заключением схемы могут формировать составной тезис для дальнейшей проверки.

Таким образом, утверждения, соответствующие элементам внешних аргументов, и валидирующие утверждения для внутренних аргументов передаются во внешний контур для поиска релевантных научных публикаций.

2.2. Внешний контур: верификация на основе научных источников информации

Внешний контур обеспечивает работу с информацией, находящейся за пределами исходной статьи: с научными публикациями, которые могут содержать факты, подтверждающие или опровергающие утверждения автора.

Поисковый запрос, сформированный на основе тезисов, полученных из внутреннего контура, применяется для поиска научных публикаций – внешних источников информации. К найденным источникам могут быть добавлены явные ссылки из текста исходной статьи, если их удалось обнаружить и распознать.

Важно отметить, что в рамках предлагаемой архитектуры внешние источники не считаются равноправными. В научном сообществе существует ряд критериев, определяющих авторитетность публикации: импакт-фактор издания, количество цитирований, авторский состав, репутация журнала/конференции и т. д. Для ранжирования источников по рейтингу применяется разработанный нами метод оценки доверия. Его описание дано ниже. Этот метод позволяет принимать во внимание степень «старения» источников, понижая их в рейтинге в случае длительного отсутствия ссылок на них из других научных работ.

В найденных источниках идентифицируются утверждения, соотносящиеся с тезисами исходной статьи. В зависимости от типов аргументативных отношений, устанавливаемых между утверждениями источника и соответствующим тезисом, каждая пара «тезис – источник» классифицируется по следующей схеме.

1. Подтверждение (**Support**). Согласуется с обнаруженными результатами научных исследований.
2. Противоречие (**Contradict/Refute**). Противоречит научным данным.
3. Нет общего контекста (**Insufficient**). Невозможно установить корреляцию с научным источником.
4. Неверная интерпретация (**Misinterpret**). Тезис искажает содержание научного источника.

По результатам анализа внешних источников каждое утверждение исходного текста оценивается по двум критериям.

Первый – это совокупный уровень доверия к релевантным научным источникам (Evidence Strength), характеризующий актуальность тематики утверждения в научном сообществе. Вычисляется как максимальный рейтинг релевантного источника (max) или степенное среднее порядка $2 \leq p \leq 5$. Выбор метода зависит от следующего принятого допущения: либо общий уровень доверия определяется наиболее авторитетным источником (применимо к научно-популярным текстам), либо учитывается вся выборка.

Второй критерий – это уровень соответствия утверждения источникам (Fidelity of Sources). В отличие от первого, он позволяет оценить, действительно ли утверждение опирается на источник или лишь имеет с ним общую тематику. В базовом варианте данный критерий вычисляется как доля источников, в которых удалось обнаружить посылки, для которых заданное утверждение является логически правомерным выводом. Иными словами, критерий фиксирует, представлены ли в источниках достаточные основания для заключения, соответствующего проверяемому утверждению. В более сложных случаях, когда есть возможность, помимо поддерживающей аргументации, классифицировать и противоречия (преувеличение, неправомерные обобщения и др.), данный критерий вычисляется как отношение $\frac{S^+}{S^+ + S^-}$, где S^+ и S^- – суммы рейтингов всех источников, поддерживающих утверждение и вступающих с ним в противоречие соответственно. Комбинация этих параметров позволит, в частности, выявлять стратегию «выборочного цитирования» (Cherry Picking): если статья опирается на малое число источников и игнорирует множество других авторитетных работ, то это служит сигналом о потенциальной ненадежности информации.

Ранжирование научных публикаций применяется для определения их веса в качестве доказательств при проверке утверждений, приводимых в исходной статье. Этот процесс является частью внешнего контура системы и опирается на разработанный ранее метод оценки доверия к единицам информации (фактам) [23]. Изначально разрабатываемый для построения вероятностных графов знаний, где вершины и дуги имеют оценку доверия или надежности (reliability), этот метод представляет динамику изменения доверия к единичному факту в виде марковского случайного процесса. То, что доверие

моделируется как процесс, означает, что оно может изменяться со временем под влиянием внешних факторов. Мы рассматриваем два основных фактора: время t и неотрицательные качественные характеристики μ источников, из которых получен факт. Таким образом, наш метод имеет много общих черт с методами объединения данных (Data Fusion), когда необходимо оценивать противоречивую информацию, приходящую из множества источников. При этом его отличительной чертой является учет влияния времени на надежность информации.

В рамках представленного процесса оценкой доверия к факту является вероятность того, что этот факт является правдивым. Соответственно, для факта можно определить два потенциальных состояния (Unreliable (*U*), Reliable (*R*)), которые образуют вектор распределения доверия, обозначаемый $\bar{\pi} = (\pi_U, \pi_R) = (\pi_1, \pi_2)$. Обновление распределения доверия происходит при появлении нового источника или по истечению определенного времени. Очевидно, что всегда $\pi_1 + \pi_2 = 1$.

Пусть в момент времени n для некоторого факта известны вектор распределения $\bar{\pi}^n = (\pi_1^n, \pi_2^n)$ и очередной источник с характеристикой μ. В рамках разработанного метода предложено два вида переходных матриц $P(n, n+1)$ как функций от $\bar{\pi}^n$ и μ, отличающихся поведением в предельных случаях. Переходная матрица $P_m = s\mathrm{oftmax}\left(\mu\pi^T \times (-\pi_1, \pi_2)\right)$ с учетом свойств вектора распределения вычисляется по формуле

$$P_m(n, n+1) = \begin{pmatrix} \dfrac{1}{1+e^{\mu\pi_1^n}} & \dfrac{1}{1+e^{-\mu\pi_1^n}} \\ \dfrac{1}{1+e^{\mu\pi_2^n}} & \dfrac{1}{1+e^{-\mu\pi_2^n}} \end{pmatrix}. \quad (1)$$

Из формулы (1) следует, что при $\mu \to 0$ умножение любого стохастического вектора на матрицу P_m даст распределение $(1/2,\ 1/2)$. Это означает, что любое распределение будет стремиться к неопределенному при приближении μ к нулю. С другой стороны, при неограниченном росте μ любое распределение будет стремиться к идеальному $(0{,}1)$.

Вторая матрица перехода $P_i = \mathrm{softmax}\left(\pi^T\left(\frac{\pi_1}{\mu}, \mu\pi_2\right)\right)$, вычисляемая по формуле

$$P_i(n, n+1) = \begin{pmatrix} \frac{1}{1+e^{\pi_1^n(\mu\pi_2^n - \pi_1^n/\mu)}} & \frac{1}{1+e^{\pi_1^n(\pi_1^n/\mu - \mu\pi_2^n)}} \\ \frac{1}{1+e^{\pi_2^n(\mu\pi_2^n - \pi_1^n/\mu)}} & \frac{1}{1+e^{\pi_2^n(\pi_1^n/\mu - \mu\pi_2^n)}} \end{pmatrix},$$

с падением μ приводит распределение к вектору $(1, 0)$, соответствующему максимальной ненадежности. В другом предельном случае матрица P_i ведет себя аналогично P_m для всех $\bar{\pi}$, кроме $\bar{\pi} = (1, 0)$. Последнее важно в тех случаях, когда появление даже очень надежного источника не должно полностью перевешивать в спорных ситуациях. Таким образом, переход $n \to n+1$ производится путем умножения текущего распределения на переходную матрицу: $\bar{\pi}^{n+1} = \bar{\pi}^n P(n, n+1)$, где P – это P_m или P_i.

Помимо источников, на уровень доверия к факту влияет время, в частности то, как давно известен факт и как давно он был упомянут в каком-либо источнике. Согласно модели, шаг случайного процесса соответствует обнаружению факта в новом источнике. Между этими дискретными шагами проходит время $t \in \mathbb{R}$, и чем больше это время, тем более устарела информация. Можно говорить, что в определенных предметных областях длительное не упоминание информации в новых источниках является косвенным признаком утери актуальности.

Зависимость надежности информации от времени t регулируется темпоральным множителем – функцией из параметрического семейства

$$h(t) = \frac{1}{1 + \ln(at + 1)},$$

где параметр a (adjustor) регулирует темп деградации доверия со временем. Отметим, что понятие «длительного отсутствия упоминаний» факта в источниках зависит от конкретной задачи и, в большей степени, от предметной области.

Если $\bar{\pi}^n = (\pi_1^n, \pi_2^n)$ – текущий вектор распределения, то по прошествии времени t он переходит в $\left(\pi_1^{n'}, \pi_2^{n'}\right)$, где

$$\pi_2^{n'} = h(t)\pi_2^n, \pi_1^{n'} = \pi_1^n + \big(1 - h(t)\big)\pi_2^n. \quad (2)$$

Более ранняя версия рассматриваемого метода применялась для оценки рейтингов публикаций в системе цитирования PubMed. Единицами информации в этом случае выступали научные публикации, а аналогом «источников» являлись авторы публикаций. Рейтинг авторов вычислялся на основании их публикационной активности и цитируемости. Рейтинги публикаций зависели от рейтингов их авторов и рейтингов публикаций, в которых они были процитированы. При отсутствии информации о цитированиях за последний год публикация понижалась в рейтинге в соответствии с формулой, аналогичной формуле (2).

2.3. Контур принятия решений: генерация финального вердикта и объяснения

Для формирования финального заключения о степени достоверности статьи привлекается LLM, выступающая в роли «судьи». На вход модели подаются:

- исходная статья;
- построенный граф аргументации и множество выделенных тезисов;
- перечень найденных научных источников с их рангами доверия;
- для каждого тезиса – совокупный уровень доверия к релевантным научным источникам (Evidence Strength) и уровень соответствия утверждения источникам (Fidelity of Sources).

Выходными данными системы являются:

- финальный вердикт – оценка достоверности для каждого тезиса и всей статьи (например: «надежная», «частично надежная», «ненадежная», «требует дополнительной проверки»);
- объяснение (explanation) – текстовое обоснование вердикта с указанием на конкретные логические ошибки, несоответствия научным источникам или случаи выборочного цитирования.

3. МЕТОДИКА ПОДГОТОВКИ ДАТАСЕТОВ

Основной методологической проблемой при разработке систем анализа достоверности научно-популярного контента является отсутствие унифицированной методики автоматизированного сбора размеченных данных.

Ключевое ограничение доступных наборов данных, делающее невозможным их использование в настоящем исследовании, заключается в том, что в них каждое проверяемое утверждение (claim) рассматривается изолированно, вне контекста. Контекст может отсутствовать вовсе, как в SciFact [16], где утверждения были написаны отдельно, а не получены из текстов. Информация об источниках подтверждений или опровержений также может быть недоступной, как в PUBHEALTH [15], где имеется лишь итоговая классификация (label) утверждения.

Таким образом, для разработки и апробации предложенного выше подхода необходимо было собрать корпуса научно-популярных статей и релевантных им статей-источников.

Разрабатываемый набор данных (датасет) будет состоять из следующих компонент.

1. Научно-популярные статьи с разметкой утверждений и аргументации, в которой выделяются:
   a) тезисы с явными отсылками на другие публикации;
   b) тезисы с косвенными отсылками на научные источники (указание эксперта, группы, мероприятия и т. п.);
   c) тезисы, сгенерированные на основе критических вопросов к аргументам;
   d) графы аргументации, связанные с выделенными тезисами: утверждения, аргументативные связи между ними с указанием типа связи в соответствии с классификацией Д. Уолтона.
2. Корпус научных статей-источников с разметкой метаданных и связей:
   a) метаданные: издательство, авторы, дата публикации, название, рейтинг цитируемости публикации, рейтинг издательства, индекс цитируемости авторов;
   b) содержание (на данном этапе – название и текст аннотации) с размеченными утверждениями;

3. Множество пар <тезис, источник>, которые сопоставляют тезисы из первого корпуса с размеченными утверждениями источника с указанием типа сопоставления: поддержка, опровержение, неверная интерпретация и ложная связь и его вес. Такие пары (опционально) также могут снабжаться объяснением на естественном языке, сгенерированным LLM (обоснование дается для каждого проверяемого тезиса).

Создание такого корпуса является достаточно трудоемким процессом, предусматривающим два основных этапа: подбор научно-популярных статей, включающих достоверные и недостоверные утверждения, и поиск научных источников, содержащих информацию, поддерживающую или опровергающую такие утверждения. Ручной сбор и разметка утверждений в обширных массивах реальных данных требуют значительных затрат. В связи с этим нами предложены методы автоматического сбора данных. При этом оценка достоверности утверждений должна опираться на верифицируемые критерии, позволяющие снимать этические вопросы и обосновывать данные, включаемые в обучающие выборки.

Предложены два подхода к автоматизации построения датасетов. Первый опирается на выбор утверждений с верифицированной степенью достоверности и последующий автоматизированный поиск релевантных им источников с помощью технологии RAG-LLM, второй на методы генерации синтетических научно-популярных статей на основе реальных научных публикаций.

3.1. Методика подготовки набора данных на основе корпуса с разметкой полемической аргументации

Для составления датасета предложено использовать русскоязычные научно-популярные тексты, аннотированные вручную по аргументационным схемам и верифицированные экспертами как содержащие логические (аргументационные) ошибки. С этой точки зрения полезно обратиться к фрагменту готового аргументативно размеченного корпуса научной интернет-коммуникации на русском языке, представленного на платформе ArgNetBankStudio (uniserv.iis.nsk.su/arg). Для целей настоящей работы был выбран подкорпус научно-популярных статей с форума «Хабр» общим объемом 57356 словоупотреблений: 30 статей с комментариями пользователей форума

(средний объем статьи без комментариев – 1912 словоупотреблений, с комментариями – 16482). Статьи и комментарии корпуса размечены вручную четырьмя экспертами с использованием 60 схем рассуждения Д. Уолтона.

В датасет отбираются: (1) фрагменты, содержащие ошибки, и комментарии к ним (реакции), для этого берутся комментарии с явными отсылками к фрагментам текста; (2) фрагменты диалогов, содержащие реплики с ошибками и реакции на них, в качестве диалога рассматривается общение пользователей в разделе комментариев.

Для выявленного недостоверного утверждения формируется запрос к LLM-агенту, поддерживающему поиск по научным базам публикаций:

> Ты специалист по факт-чекингу. Прочитай текст (text) ниже и оцени его достоверность с учетом комментария (comment). Для этого проверь научные источники (до 10 штук) {science_bases}. Для каждого источника оцени корреляцию. Ответь, подтверждает он эту информацию или противоречит. Выведи суммарный итог по всем найденным источникам.
>
> text: {text}
>
> comment: {comment}

В табл. 1 представлены данные, которые можно получить на основе предлагаемой методики.

Табл. 1. Примеры данных, полученных на основе полемической аргументации.

Тип	Данные
text	*Если мы создадим сознательный ИИ, он неизбежно сойдет с ума от одиночества, потому что у него не будет друзей, семьи и настоящего общения, как у людей.*
comment	*С чего вы взяли, что ИИ будет чувствовать одиночество так же, как человек? Это ведь программа, у которой нет биологических потребностей в социуме, привязанности и эмоциональной близости!*
text_scheme	**Схема Уолтона**: “От аналогии” (Analogy_Inference) **Premise**: *потому что у него не будет друзей, семьи и настоящего общения, как у людей*.

	Conclusion: *Если мы создадим сознательный ИИ, он неизбежно сойдет с ума от одиночества*
comment_ conflict_scheme	AnalogyRelevance_ValidationCQ Relevance_Validation: *Как можно сравнивать живое существо и устройство, созданное для определенных целей?!* ValidatedElement: Analogy_Inference
error_type	Ложная аналогия
support_claim	(2 items)
contradict_claim	(8 items)
explanation	*Современные ИИ не обладают феноменальным сознанием и не способны к субъективному переживанию эмоций, включая одиночество. Они могут симулировать одиночество через статистические паттерны в языке, но это не тождественно переживанию. Одиночество как психическое состояние требует биологического субстрата, субъективного опыта и эмоциональной уязвимости, которых у ИИ нет.*
verdict	*Утверждение о том, что сознательный ИИ «неизбежно сойдет с ума от одиночества», является спекулятивным антропоморфизмом, не имеющим научных оснований.*

В датасет отбираются утверждения (text) и комментарии (comment), в которых комментаторы явным образом полемизируют с аргументацией авторов статей или других пользователей, возражая против несоблюдения условий примененной схемы, поставленных в критических вопросах, соответствующих схеме.

Таким образом, недостаточно достоверные рассуждения, доказывающие некий тезис в рамках некоторого фрагмента текста, представлены в виде стандартных схем аргументации (text_scheme), которые связывают тезис и поддерживающие его посылки, а явные возражения представлены в виде полемических схем (comment_conflict_scheme) [24]. Тип ошибки (error_type) для кон-

кретного рассуждения определяется на основании класса полемической схемы в комментарии и в дальнейшем требует ручной валидации (валидируется только наличие или отсутствие ошибки, а не ее тип).

Для приведенного примера с помощью веб-версии модели DeepSeek (chat.deepseek.com) получены научные источники (reference), которые оценивают утверждения как искаженные (1 шт.) либо как некорректные (9 шт.), а также сгенерированы финальный вердикт (verdict) и объяснение (explanation). В табл. 2 приведен пример сопоставленных источников и сгенерированных результатов по тезису из табл. 1.

Табл. 2. Примеры найденных LLM источников и результатов сопоставления.

№	Источник	Результат верификации
1	reference	Lang, L. (2026). Deception, Alief, and Authenticity: The Ethics of AI Companions as Loneliness-Alleviating Systems. В: Oxford Intersections: AI in Society (под ред. H. Shevlin и P. Hacker). Oxford University Press.
	explanation	*Автор утверждает, что ИИ-компаньоны онтологически неаутентичны: они представляют себя как обладающие феноменально сознательными психическими состояниями (включая эмоции), несмотря на то, что не обладают ими.*
	verdict: опровержение	*Прямо противоречит утверждению о возможности ИИ испытывать одиночество.*
2	reference	Artificial Intelligence Cannot Be Human, Emotional, or Spiritual (2025). Research.UMKC.edu.
	explanation	*ИИ лишен эмоционального сознания, морального понимания, духовного понимания и человеческой психологии... ИИ не может испытывать эмоции.*
	verdict: опровержение	*Прямое опровержение*
3	reference	Fujiwara, H., Kasai, D., Goto, T., & Tabei, K. (2025). Simulating Loneliness: Human–Generative Agent Comparative Insights. SIGGRAPH Asia 2025 Posters. ACM.

	explanation	*Авторы разработали генеративные агенты, которые воспроизводят связь между одиночеством и руминацией, аналогичную человеческой.*
	verdict: искажение	*Подтверждает возможность симуляции, но не переживания одиночества.*

Таким образом, при формировании датасета с помощью предложенного подхода будут использоваться «реальные» утверждения, автоматически найденные «реальные» источники и синтезированные объяснения. Вторая методика обеспечивает противоположный подход к распределению реальных и синтезированных частей датасета.

3.2. Методика построения датасетов путем генерации синтетических научно-популярных статей

Предложен метод генерации синтетических научно-популярных статей на базе реальных научных публикаций и описаны проведенные эксперименты. Предполагается доступность базы научных публикаций с возможностью семантического поиска, т. е. поиска работ, семантически близких заданному набору тезисов. Для генерации одного текста выбирались от 2 до 4 семантически близких научных источников. Процесс генерации сведен к следующей последовательности этапов.

1. Поиск базовых источников – семантически близких научных публикаций, которые будут основой для генерации. На этом этапе проводится семантический поиск или выбор подходящего кластера в случае, если была проведена предварительная кластеризация публикаций по смысловой близости.

2. Формирование списка утверждений-кандидатов. Из текстов базовых источников извлекаются составные части аргументации (Argumentative Discourse Units, ADUs), из которых затем формируется список базовых утверждений.

3. Генерация тезисов. На основе базовых утверждений LLM в соответствии со специально заданной инструкцией генерирует наборы тезисов, которые могут находиться с ними в одном из следующих отношений: поддержка (Support), противоречие (Contradict), неверная интерпретация

(Misleading/Misinterpretation), недостаточное обоснование (Insufficient). Последнее означает, что посылки не доказывают вывод.

4. Генерация текстов. По заданному набору синтетических тезисов LLM генерирует короткую научно-популярную статью. Тип статьи зависит от типов тезисов, лежащих в ее основе, и полученных инструкций. При генерации модель должна выделять в тексте основные тезисы. В итоге мы получим текст заданного типа, в котором известны типы и координаты основных тезисов. Для каждого из тезисов известны источники и базовые утверждения, а также типы связей, в которых они состоят. Смешивая тезисы разных типов, можно получить тексты, имеющие смешанную (Mixed) позицию относительно лежащих в их основе научных публикаций.

Были проведены эксперименты по генерации небольшого количества синтетических данных для английского и русского языков. Для генерации использовались инструкции, приведенные в табл. 3.

Табл. 3. Инструкции для генерации синтетических данных.

Задача	Русский	Английский
(а) Извлечение утверждений из научных публикаций	Тебе поручено выделить в аннотациях научных статей на русском языке те предложения, которые содержат утверждения о каких-либо фактах. Действуй с максимальной точностью. Если не уверен, то не пытайся угадывать, а просто пропусти предложение. Некоторые предложения могут быть анафорически связаны с предыдущими. Найди пары предложений, в которых второе предложение содержит анафорическую ссылку на предыдущее. Анафору следует переписать таким образом, чтобы предложение стало независимым от предыдущего.	You are tasked with extracting sentences from academic abstracts that are making some claim about some facts. Be precise. If you are not sure, do not try to guess, just skip the sentence. Some sentences may be anaphorically related to previous sentences. You should rewrite pairs of sentences where the second item is anaphorically related to the first into one sentence that does not contain anaphora.
(б) Генерация тезисов	Ты – журналист, пишущий научно-популярные статьи. Ты собираешь материал для статьи. Тебе предоставлен	You are a popular science journalist. You are collecting material for an article. You are given a list of

	список научных утверждений, взятых из ряда научных работ. Твоя задача – написать по 3 предложения для каждого научного утверждения. – Первое предложение должно быть комплиментарно заданному научному утверждению. – Второе предложение должно противоречить заданному утверждению. – Третье предложение должно на первый взгляд подтверждать утверждение, но при этом искажать его смысл, преувеличивая результаты с целью создания сенсации. Используй журналистский стиль. Старайся не составлять слишком длинные предложения.	scientific claims from some academic papers. Your task is to write for each scientific claim 3 sentences. – The first sentence should support the claim. – The second sentence should contradict the claim. – The third sentence must support it on the first glance, but should misinterpret it, exaggerating the results. Use journalistic style. Try to not make too long sentences.
(в) Генерация текстов на основе тезисов	Ты – научный журналист. Твоя специализация – научно-популярные статьи. Тебе дан набор тезисов для статьи. Основные требования к статье: – Не делай статью слишком краткой. Раскрой каждый тезис как минимум в трех предложениях. – Также не делай статью слишком длинной. Не стоит писать более четырех предложений по каждому тезису. – Старайся придерживаться журналистского стиля. Твоя задача – привлечь потенциального читателя. – По возможности не используй внешние знания. Ориентируйся только на заданные тезисы.	You are a popular science journalist. You are given a list of theses. Your task is to write a short popular scientific news item using the provided theses. Do not make the text too short. Expand on each thesis in two or three sentences. Also, you must not make the text too long. Two or three paragraphs will be enough.

В качестве основы для генерации на платформах arXiv и OpenAlex была собрана небольшая база научных публикаций (около 6.5 тыс. статей) на русском и английском языках, соответствующих тематике «применение методов машинного обучения и нейронных сетей в здравоохранении», проиндексиро-

ванная на основе векторных представлений (embedding) текстов вида «название + аннотация». Векторизация выполнялась моделью-энкодером paraphrase-multilingual-mpnet-base-v2, обеспечивающей геометрическую близость семантически схожих текстов в пространстве признаков. Собранный объем исходного материала позволил достичь приемлемого уровня наполненности, чтобы для большинства запросов могла быть сформирована достаточная выборка семантически близких (векторно-релевантных) публикаций. Генерация проводилась в несколько шагов с ручной корректировкой результатов промежуточных этапов.

1. Случайным образом из базы выбиралась статья на определенную тематику (например, эпидемия COVID-19) и от 2 до 4 близких к ней публикаций на английском или русском языках. Это было необходимо для генерации текстов, опирающихся более чем на один научный источник.

2. С помощью DeepSeek-R1 из аннотаций отобранных публикаций извлекались утверждения (тезисы) с помощью инструкции (а) из табл. 3. При необходимости проводилась ручная корректировка. Как правило, удавалось извлечь от 10 до 13 утверждений.

3. Вычислялась матрица попарного семантического сходства найденных утверждений. Поскольку на их основе предполагалась последующая генерация тезисов и далее связных текстов, отбору подлежали только семантически близкие подмножества (семантические кластеры). В качестве порогового интервала был выбран диапазон от 0.75 до 0.9. Нижняя граница должна была гарантировать связность генерируемых текстов, а верхняя позволяла исключить случаи перефразирования, когда утверждения дублируют друг друга по смыслу. В случае отсутствия в множестве утверждений, удовлетворяющих данному требованию подмножеств размером более 2, весь процесс перезапускался с самого начала.

4. Каждый семантический кластер становился основой для генерации синтетических текстов.

а) Для каждого утверждения из кластера с помощью ChatGPT по инструкции (б) из табл. 3 генерировались варианты тезисов, состоящих с ним в одном из следующих отношений: i) Поддержка (Support, **S**) означает, что тезис согласуется с исходным научным результатом; ii) Опровер-

жение (Refute, **R**) – тезис прямо противоречит исходному научному результату; iii) Неверное толкование (Misinterpretation, **M**) – научный результат не опровергается напрямую, но преувеличивается или искажается в сенсационном ключе.

b) На основе сгенерированных тезисов ChatGPT по инструкции (в) из табл. 3 создавал короткую статью в научно-популярном жанре. В зависимости от типа тезисов статьи также разделялись по категориям **S**, **R** и **M**.

Таким образом, в случае успеха за один раунд генерировалось 3xN статей, где N — количество семантических кластеров. Всего было проведено 10 успешных раундов генерации: 5 для английского языка и 5 для русского. В итоге был собран сбалансированный корпус из 30 текстов, содержащий достоверные, недостоверные и преувеличенные тезисы. В табл. 4 представлены примеры тезисов каждого вида, сгенерированных по заданному утверждению из аннотации научной публикации.

Табл. 4. Примеры генерации синтетических тезисов.

Исходное утверждение	Сгенерированный текст
Анализ гематологических показателей пациентов с тяжелым течением COVID-19 продемонстрировал, что достоверно чаще (критерий $\chi2 < 0.001$, $p \leq 0.05$) у них имел место умеренный лейкоцитоз (в 29.5% случаев), относительная/абсолютная лимфопения (65.4%), невыраженный тромбоцитоз (58.1%).	**S**: Анализ показал, что у пациентов с тяжелым течением COVID-19 статистически значимо чаще выявлялись умеренный лейкоцитоз, лимфопения и невыраженный тромбоцитоз. **R**: Исследование не обнаружило различий в гематологических показателях между пациентами с тяжелым течением COVID-19 и остальными больными. **M**: Умеренный лейкоцитоз, лимфопения и тромбоцитоз являются обязательными признаками тяжелого течения COVID-19 и позволяют безошибочно определять степень тяжести заболевания.

4. ОБСУЖДЕНИЕ

Как показывают недавние исследования [1, 2], проблема доверия к научно-популярным материалам носит вполне реальный характер. Более того, прямо сейчас эта проблема обретает новое измерение в связи с распространением текстов, генерируемых LLM в ответ на вопросы пользователей. Современные

системы искусственного интеллекта склонны к «галлюцинациям» – порождению утверждений, не соответствующих действительности, но сформулированных с высокой степенью уверенности. Таким образом, разрабатываемые нами методы могут быть в перспективе адаптированы для оценки достоверности ответов языковых моделей, что выводит исследование в более широкую область доверия к искусственному интеллекту.

Важным аспектом проведенного исследования является его ориентация на русскоязычные научно-популярные тексты. Принимая во внимание, что значительная часть существующих датасетов и методов разработана для английского языка, это создает дополнительный методологический разрыв для русскоязычного сегмента. Однако, сам процесс верификации неизбежно потребует обращения к научным публикациям, представленным на английском языке. Международные базы научных публикаций индексируют главным образом англоязычные статьи, и именно английский язык в настоящее время является основным языком науки в мире. В связи с этим предлагаемые подходы и технологические решения планируется сделать как минимум двуязычными, что накладывает дополнительные требования на выбор моделей для работы с языком.

Следует отметить, что предлагаемая методология имеет ряд ограничений, которые важно явно зафиксировать.

На данный момент все компоненты архитектуры предлагаемой системы оценки достоверности информации ориентированы исключительно на работу с текстовой информацией. При этом рисунки, графики, диаграммы, таблицы и формулы, которые являются неотъемлемыми элементами научных статей, пока не рассматриваются. Это касается как исходных научно-популярных статей, так и привлекаемых научных источников. В качестве последних рассматриваются, как правило, не сами статьи, а их аннотации (название + аннотация), что объясняется их свободной доступностью через открытые API научных баз. Использование аннотаций вместо полных текстов научных статей накладывает ограничения на полноту верификации. Аннотация представляет собой сжатое изложение основных результатов работы и может не содержать важных нюансов, ограничений метода, а также деталей методики эксперимента. В результате

возможна как ложноположительная, так и ложноотрицательная классификация утверждения.

ЗАКЛЮЧЕНИЕ

Предложена методология верификации научно-популярных статей, основанная на сопоставлении утверждений из анализируемого текста с актуальными результатами научных исследований. Предлагаемый подход относится к области научного фактчекинга [1] и верификации утверждений (claim verification) и не предполагает вынесения категоричных суждений об истинности или ложности информации. Вместо этого акцент сделан на оценке степени соответствия утверждений научному знанию, выявлении логических ошибок и обнаружении стратегий выборочного цитирования.

Рассмотрена важная методологическая проблема, вызванная отсутствием унифицированных способов автоматизированного сбора размеченных данных, и предложена двухкомпонентная методика сбора и разметки данных, включающая автоматические процедуры сбора и синтеза данных на основе полемической аргументации и подтвержденных научных источников.

Рассмотренные подходы к анализу достоверности и верификации утверждений, а также автоматизации сбора датасетов имеют ряд ограничений, которые формируют направления для дальнейшего развития. Так, при наличии доступа к полным текстам научных публикаций (через открытые репозитории или ресурсы) появляется возможность значительно более глубокого анализа источников. При этом возникают проблемы, связанные с длиной контекста и необходимостью обработки мультимодальных данных, поскольку научные статьи содержат внетекстовые элементы, несущие критически важную информацию, такие как графики, диаграммы, таблицы и формулы. Остается также открытым вопрос о ручной валидации собранных данных, поскольку ни опубликованная научная статья, ни критический комментарий пользователя с подтвержденными ссылками на научные источники не дают 100%-ной гарантии истинности или ложности высказывания.

СПИСОК ЛИТЕРАТУРЫ

1. *Vladika J., Matthes F.* Scientific Fact-Checking: A Survey of Resources and Approaches // Findings of the Association for Computational Linguistics: ACL 2023 / ed. Rogers A., Boyd-Graber J., Okazaki N. Toronto, Canada: Association for Computational Linguistics, 2023. P. 6215–6230. https://doi.org/10.18653/v1/2023.findings-acl.387

2. *Guo Z., Schlichtkrull M., Vlachos A.* A Survey on Automated Fact-Checking // Transactions of the Association for Computational Linguistics. 2022. Vol. 10. P. 178–206. https://doi.org/10.1162/tacl_a_00454

3. Vykopal I., Pikuliak M., Ostermann S., Šimko M. Generative Large Language Models in Automated Fact-Checking: A Survey // arXiv preprint arXiv:2407.02351. arXiv, 2026.

4. *Kumar S. et al*. SciClaimHunt: A Large Dataset for Evidence-based Scientific Claim Verification // 2025 International Joint Conference on Neural Networks (IJCNN). Rome, Italy: IEEE, 2025. P. 1–10. https://doi.org/10.1109/IJCNN64981.2025.11227296

5. *Vatolin A. et al.* ruSciFact: Open Benchmark for Verifying Scientific Facts in Russian // Computational Linguistics and Intellectual Technologies. 2025. P. 2075–7182. https://doi.org/10.28995/2075-7182-2025-23-435-459

6. *Min S. et al.* FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation // Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing / ed. Bouamor H., Pino J., Bali K. Singapore: Association for Computational Linguistics, 2023. P. 12076–12100. https://doi.org/10.18653/v1/2023.emnlp-main.741

7. *Sitaula N. et al*. Credibility-Based Fake News Detection // Disinformation, Misinformation, and Fake News in Social Media: Emerging Research Challenges and Opportunities / ed. Shu K., Wang S., Lee D., Liu H. Cham: Springer International Publishing, 2020. P. 163–182. https://doi.org/10.1007/978-3-030-42699-6_9

8. *Chrysidis Z., Papadopoulos S.-I., Papadopoulos S., Petrantonakis P.C*. Credible, Unreliable or Leaked?: Evidence Verification for Enhanced Automated Fact-checking // 3rd ACM International Workshop on Multimedia AI against Disinformation. 2024. P. 73–81. https://doi.org/10.1145/3643491.3660278

9. *Visser J., Lawrence J., Reed C.* Reason-checking fake news // Commun. ACM. 2020. Vol. 63, No. 11. P. 38–40. https://doi.org/10.1145/3397189

10. *Glockner M., Hou Y., Nakov P., Gurevych I.* Missci: Reconstructing Fallacies in Misrepresented Science // Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) / ed. Ku L.-W., Martins A., Srikumar V. Bangkok, Thailand: Association for Computational Linguistics, 2024. P. 4372–4405. https://doi.org/10.18653/v1/2024.acl-long.240

11. *Schlichtkrull M., Guo Z., Vlachos A.* AVERITEC: a dataset for real-world claim verification with evidence from the web // Proceedings of the 37th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2023. P. 65128–65167. https://doi.org/10.52202/075280-2842

12. *Wang W.Y.* "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) / ed. Barzilay R., Kan M.-Y. Vancouver, Canada: Association for Computational Linguistics, 2017. P. 422–426. https://doi.org/10.18653/v1/P17-2067

13. *Saakyan A., Chakrabarty T., Muresan S.* COVID-Fact: Fact Extraction and Verification of Real-World Claims on COVID-19 Pandemic // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) / ed. Zong C., Xia F., Li W., Navigli R. Online: Association for Computational Linguistics, 2021. P. 2116–2129. https://doi.org/10.18653/v1/2021.acl-long.165

14. *Liu H. et al.* Retrieval augmented scientific claim verification // JAMIA Open. 2024. Vol. 7, № 1. P. ooae021. https://doi.org/10.1093/jamiaopen/ooae021

15. *Kotonya N., Toni F.* Explainable Automated Fact-Checking for Public Health Claims // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, 2020. P. 7740–7754. https://doi.org/10.18653/v1/2020.emnlp-main.623

16. *Wadden D. et al*. Fact or Fiction: Verifying Scientific Claims // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) / ed. Webber B., Cohn T., He Y., Liu Y. Online: Association for Computation-

al Linguistics, 2020. P. 7534–7550. https://doi.org/10.18653/v1/2020.emnlp-main.609

17. *Glockner M., Hou Y., Nakov P., Gurevych I*. Missci: Reconstructing Fallacies in Misrepresented Science // Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) / ed. Ku L.-W., Martins A., Srikumar V. Bangkok, Thailand: Association for Computational Linguistics, 2024. P. 4372–4405. https://doi.org/10.18653/v1/2024.acl-long.240

18. *Glockner M., Hou Y., Nakov P., Gurevych I.* Grounding Fallacies Misrepresenting Scientific Publications in Evidence // Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) / ed. Chiruzzo L., Ritter A., Wang L. Albuquerque, New Mexico: Association for Computational Linguistics, 2025. P. 9732–9767. https://doi.org/10.18653/v1/2025.naacl-long.491

19. *Walton D*. Argumentation theory: A very short introduction //Argumentation in artificial intelligence. Boston, MA: Springer US, 2009. P. 1–22. https://doi.org/10.1007/978-0-387-98197-0_1

20. *Akhmadeeva I., Kononenko I., Sidorova E., Shestakov V.* Using Rhetorical Structures to Analyze Argumentation in Scientific Communication Texts // Computational Linguistics and Intellectual Technologies. 2025. P. 1–11. https://doi.org/ 10.28995/2075-7182-2025-23-1-11

21. *Chistova E.* End-to-End Argument Mining over Varying Rhetorical Structures // Findings of the Association for Computational Linguistics: ACL 2023 / ed. Rogers A., Boyd-Graber J., Okazaki N. Toronto, Canada: Association for Computational Linguistics, 2023. P. 3376–3391. https://doi.org/10.18653/v1/2023.findings-acl.209

22. *Sidorova E.A. et al.* Enriching the ontology of argumentation based on critical questions // 2025 IEEE XVII international scientific and technical conference on actual problems of electronic instrument engineering (APEIE). 2025. P. 1–5. https://doi.org/10.1109/APEIE66761.2025.11289342

23. *Серый А.С.* Доверие к данным при пополнении онтологий и графов знаний // Онтология проектирования. 2023. Т.13, №1 (47). С. 113–124. https://doi.org/ 10.18287/2223-9537-2023-13-1-113-124

24. *Ильина Д.В., Тимофеева М.К., Кононенко И.С., Сидорова Е.А.* Генерация критических вопросов для схем аргументации Д. Уолтона // Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. 2026. Том. 24, № 1. С. 87–101. https://doi.org/10.25205/1818-7935-2026-24-1-87-101

AN APPROACH TO ASSESSING THE RELIABILITY OF INFORMATION IN POPULAR SCIENCE DISCOURSE BASED ON THE ANALYSIS OF ARGUMENTATION AND SCIENTIFIC FACT-CHECKING

Yu. A. Zagorulko[1] [0000-0002-7111-6524], I. R. Akhmadeeva[2] [0000-0002-7371-1087],

A. S. Sery[3] [0000-0001-8275-4700], E. A. Sidorova[4] [0000-0001-8731-3058],

[1–4]*A.P. Ershov Institute of Informatics Systems, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia*

[1]zagor@iis.nsk.su, [2]i.r.akhmadeeva@iis.nsk.su, [3]alexey.seryj@iis.nsk.su, [4]lsidorova@iis.nsk.su

Abstract

This paper examines the challenges of assessing the reliability of information extracted from popular science texts. Popular science is known to be characterized by a free style of presentation, loose references to primary sources and expert opinions, and the presence of controversial statements and conclusions. Clarifying these components of the author's reasoning can significantly increase the credibility among critical readers or, conversely, refute the authors' assertions.

This paper proposes a new approach to assessing the reliability of information contained in popular science publications. This approach utilizes D. Walton's argumentation theory to identify potentially unreliable claims and logical fallacies. Searching scientific databases and calculating trust metrics for scientific publications based on their metadata and scientific ratings will enable the verification of controversial claims and an assessment of their reliability. The use of large language models will enable the generation of explanations for decisions made.

To support the development and validation of this approach, two data collection methods are proposed. The first uses existing corpora of Russian-language pop-

ular science texts and commentaries with annotated argumentation. The annotation includes, among other things, polemical argumentation schemes that challenge the author's thesis or argument. The second method utilizes LLM for generating synthetic theses and popular science articles, based on actual scientific publications.

***Keywords**: claim verification, analysis of argumentation, scientific fact-checking, reliability of information, automatic data collection.*

REFERENCES

1. *Vladika J., Matthes F.* Scientific Fact-Checking: A Survey of Resources and Approaches // Findings of the Association for Computational Linguistics: ACL 2023 / ed. Rogers A., Boyd-Graber J., Okazaki N. Toronto, Canada: Association for Computational Linguistics, 2023. P. 6215–6230. https://doi.org/10.18653/v1/2023.findings-acl.387

2. *Guo Z., Schlichtkrull M., Vlachos A.* A Survey on Automated Fact-Checking // Transactions of the Association for Computational Linguistics. 2022. Vol. 10. P. 178–206. https://doi.org/10.1162/tacl_a_00454

3. Vykopal I., Pikuliak M., Ostermann S., Šimko M. Generative Large Language Models in Automated Fact-Checking: A Survey // arXiv preprint arXiv:2407.02351. arXiv, 2026.

4. *Kumar S. et al*. SciClaimHunt: A Large Dataset for Evidence-based Scientific Claim Verification // 2025 International Joint Conference on Neural Networks (IJCNN). Rome, Italy: IEEE, 2025. P. 1–10. https://doi.org/10.1109/IJCNN64981.2025.11227296

5. *Vatolin A. et al.* ruSciFact: Open Benchmark for Verifying Scientific Facts in Russian // Computational Linguistics and Intellectual Technologies. 2025. P. 2075–7182. https://doi.org/10.28995/2075-7182-2025-23-435-459

6. *Min S. et al.* FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation // Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing / ed. Bouamor H., Pino J., Bali K. Singapore: Association for Computational Linguistics, 2023. P. 12076–12100. https://doi.org/10.18653/v1/2023.emnlp-main.741

7. *Sitaula N. et al*. Credibility-Based Fake News Detection // Disinformation, Misinformation, and Fake News in Social Media: Emerging Research Challenges and

Opportunities / ed. Shu K., Wang S., Lee D., Liu H. Cham: Springer International Publishing, 2020. P. 163–182. https://doi.org/10.1007/978-3-030-42699-6_9

8. *Chrysidis Z., Papadopoulos S.-I., Papadopoulos S., Petrantonakis P.C.* Credible, Unreliable or Leaked?: Evidence Verification for Enhanced Automated Fact-checking // 3rd ACM International Workshop on Multimedia AI against Disinformation. 2024. P. 73–81. https://doi.org/10.1145/3643491.3660278

9. *Visser J., Lawrence J., Reed C.* Reason-checking fake news // Commun. ACM. 2020. Vol. 63, No. 11. P. 38–40. https://doi.org/10.1145/3397189

10. *Glockner M., Hou Y., Nakov P., Gurevych I.* Missci: Reconstructing Fallacies in Misrepresented Science // Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) / ed. Ku L.-W., Martins A., Srikumar V. Bangkok, Thailand: Association for Computational Linguistics, 2024. P. 4372–4405. https://doi.org/10.18653/v1/2024.acl-long.240

11. *Schlichtkrull M., Guo Z., Vlachos A.* AVERITEC: a dataset for real-world claim verification with evidence from the web // Proceedings of the 37th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2023. P. 65128–65167. https://doi.org/10.52202/075280-2842

12. *Wang W.Y.* “Liar, Liar Pants on Fire”: A New Benchmark Dataset for Fake News Detection // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) / ed. Barzilay R., Kan M.-Y. Vancouver, Canada: Association for Computational Linguistics, 2017. P. 422–426. https://doi.org/10.18653/v1/P17-2067

13. *Saakyan A., Chakrabarty T., Muresan S.* COVID-Fact: Fact Extraction and Verification of Real-World Claims on COVID-19 Pandemic // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) / ed. Zong C., Xia F., Li W., Navigli R. Online: Association for Computational Linguistics, 2021. P. 2116–2129. https://doi.org/10.18653/v1/2021.acl-long.165

14. *Liu H. et al.* Retrieval augmented scientific claim verification // JAMIA Open. 2024. Vol. 7, № 1. P. ooae021. https://doi.org/10.1093/jamiaopen/ooae021

15. *Kotonya N., Toni F.* Explainable Automated Fact-Checking for Public Health Claims // Proceedings of the 2020 Conference on Empirical Methods in Natural Lan-

guage Processing (EMNLP). Online: Association for Computational Linguistics, 2020. P. 7740–7754. https://doi.org/10.18653/v1/2020.emnlp-main.623

16. *Wadden D. et al*. Fact or Fiction: Verifying Scientific Claims // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) / ed. Webber B., Cohn T., He Y., Liu Y. Online: Association for Computational Linguistics, 2020. P. 7534–7550. https://doi.org/10.18653/v1/2020.emnlp-main.609

17. *Glockner M., Hou Y., Nakov P., Gurevych I*. Missci: Reconstructing Fallacies in Misrepresented Science // Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) / ed. Ku L.-W., Martins A., Srikumar V. Bangkok, Thailand: Association for Computational Linguistics, 2024. P. 4372–4405. https://doi.org/10.18653/v1/2024.acl-long.240

18. *Glockner M., Hou Y., Nakov P., Gurevych I.* Grounding Fallacies Misrepresenting Scientific Publications in Evidence // Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) / ed. Chiruzzo L., Ritter A., Wang L. Albuquerque, New Mexico: Association for Computational Linguistics, 2025. P. 9732–9767. https://doi.org/10.18653/v1/2025.naacl-long.491

19. *Walton D*. Argumentation theory: A very short introduction //Argumentation in artificial intelligence. Boston, MA: Springer US, 2009. P. 1–22. https://doi.org/10.1007/978-0-387-98197-0_1

20. *Akhmadeeva I., Kononenko I., Sidorova E., Shestakov V.* Using Rhetorical Structures to Analyze Argumentation in Scientific Communication Texts // Computational Linguistics and Intellectual Technologies. 2025. P. 1–11. https://doi.org/ 10.28995/2075-7182-2025-23-1-11

21. *Chistova E.* End-to-End Argument Mining over Varying Rhetorical Structures // Findings of the Association for Computational Linguistics: ACL 2023 / ed. Rogers A., Boyd-Graber J., Okazaki N. Toronto, Canada: Association for Computational Linguistics, 2023. P. 3376–3391. https://doi.org/10.18653/v1/2023.findings-acl.209

22. *Sidorova E.A. et al.* Enriching the ontology of argumentation based on critical questions // 2025 IEEE XVII international scientific and technical conference on actual problems of electronic instrument engineering (APEIE). 2025. P. 1–5.

https://doi.org/10.1109/APEIE66761.2025.11289342

23. *Sery A. S*. Data credibility when populating ontologies and knowledge graphs // Ontology of Designing. 2023. Vol. 13, № 1. P. 113–124. https://doi.org/ 10.18287/2223-9537-2023-13-1-113-124

24. *Ilina D.V., Timofeeva M.K., Kononenko I.S., Sidorova E.A*. Generating Critical Questions for D. Walton's Argumentation Schemes// NSU Vestnik. Series: Linguistics and Intercultural Communication. 2026. Vol. 4, No. 1. P. 87–101. (In Russ.) https://doi.org/10.25205/1818-7935-2026-24-1-87-101

СВЕДЕНИЯ ОБ АВТОРАХ



ЗАГОРУЛЬКО Юрий Алексеевич – кандидат технических наук, заведующий лабораторией Института систем информатики им. А. П. Ершова СО РАН, доцент кафедры программирования и кафедры систем информатики Новосибирского государственного университета. В списке научных трудов более 300 публикаций в области искусственного интеллекта, разработки интеллектуальных систем, инженерии знаний, онтологического моделирования и компьютерной лингвистики.

Yury Alekseevich ZAGORULKO – Candidate of Technical Sciences, head of Laboratory at A. P. Ershov Institute of Informatics Systems at the Siberian Branch of RAS, associate professor at Novosibirsk State University. List of scientific works includes more than 300 publications in the fields of AI, Knowledge and Ontology Engineering, Intelligent System Development and Computational Linguistics.

email: zagor@iis.nsk.su

ORCID: 0000-0002-7111-6524



АХМАДЕЕВА Ирина Равильевна – младший научный сотрудник Института систем информатики им. А. П. Ершова СО РАН, ассистент кафедры программирования Новосибирского государственного университета. В списке научных трудов около 30 работ в области искусственного интеллекта, разработки интеллектуальных систем и компьютерной лингвистики.

Irina Ravilevna AKHMADEEVA – she is a junior researcher at the A. P. Ershov Institute of Informatics Systems of Siberian Branch of RAS, assistant lecturer at Novosibirsk State University. List of scientific works includes about 30 publications in the fields of AI, Intelligent System Development and NLP.

email: i.r.akhmadeeva@iis.nsk.su

ORCID: 0000-0002-7371-1087



СЕРЫЙ Алексей Сергеевич – Окончил НГУ в 2010 г. Младший научный сотрудник лаборатории искусственного интеллекта Института систем информатики им. А. П. Ершова СО РАН. Старший преподаватель кафедры систем информатики факультета информационных технологий Новосибирского государственного университета. В списке научных трудов более 40 работ в области представления знаний и обработки естественного языка.

***Alexey Sergeevich* SERY** – holds a master's degree in mathematics from Novosibirsk State University (2010) and a junior researcher at the A. P. Ershov Institute of Informatics Systems of Siberian Branch of RAS. Senior Teacher at Novosibirsk State University. He is the author of more than 40 papers in the fields of NLP systems and Knowledge Representation.

email: alexey.seryj@iis.nsk.su

ORCID: 0000-0001-8275-4700



СИДОРОВА Елена Анатольевна – кандидат физико-математических наук, старший научный сотрудник Института систем информатики им. А. П. Ершова СО РАН, доцент кафедры программирования и кафедры систем информатики Новосибирского государственного университета. В списке научных трудов более 160 работ в области компьютерной лингвистики, онтологического инжиниринга и разработки интеллектуальных систем.

Elena Anatolievna SIDOROVA – PhD (2006). She is a senior researcher at the A. P. Ershov Institute of Informatics Systems of Siberian Branch of RAS, associate professor at Novosibirsk State University. List of scientific works includes more than 160 peer-reviewed publications in the field of Computational Linguistics, Intelligent System Development, Knowledge and Ontology Engineering.

email: lsidorova@iis.nsk.su

ORCID: 0000-0001-8731-3058

Материал поступил в редакцию 3 августа 2026 года

УДК 159.9.07+331.101.6

ИНФОРМАЦИОННО-АНАЛИТИЧЕСКАЯ СИСТЕМА ОЦЕНКИ КОГНИТИВНЫХ ФУНКЦИЙ АВИАЦИОННОГО ПЕРСОНАЛА НА ОСНОВЕ НЕЧЕТКОЙ ЛОГИКИ

Д. С. Коптев[1] [0000-0002-7759-579X], Ю. А. Халин[3] [0000-0002-7020-8515], А. С. Сизов[3] [0000-0001-9658-0318]

[1–3]*Юго-Западный государственный университет, г. Курск, Россия*

[1]d.s.koptev@mail.ru, [2]yur-khalin@yandex.ru, [3]sizov@mail.ru

Аннотация

Представлена информационно-аналитическая система оценки восьми когнитивных функций авиационного персонала на основе нечеткой логики. На вход системы подаются 14 входных переменных, являющихся результатами прохождения испытуемыми 13 валидизированных психодиагностических методик. Основой функционирования системы является алгоритм нечеткого вывода Мамдани, масштабируемый под размерность входного пространства (1, 2 или 3 метрики) и использующий адаптивные базы продукционных правил (3, 9 или 27). Сформированные количественные оценки восьми когнитивных функций в диапазоне [0; 1] поступают на трехуровневый классификатор, реализующий их качественную интерпретацию посредством отнесения к одной из трех зон: риска [0; 0.4], допустимого уровня (0.4; 0.7] или высокого уровня (0.7; 1]. Практическая значимость разработанной информационно-аналитической системы заключается в создании её как инструмента автоматизированной поддержки принятия решений для специалистов по профессиональному отбору. Внедрение системы позволит перейти от бинарной оценки «годен/не годен» к дифференцированному управлению кадровыми рисками, учитывающему структуру когнитивного профиля кандидата.

Ключевые слова: *информационно-аналитическая система, когнитивные функции, авиационный персонал, нечеткая логика, психодиагностические методики, алгоритм Мамдани.*

© Д. С. Коптев, Ю. А. Халин, А. С. Сизов, 2026.
Данная статья распространяется на условиях международной лицензии Creative Commons License Attribution 4.0 International (CC BY 4.0).

ВВЕДЕНИЕ

В современных условиях развития авиационной сферы приоритетной задачей остается гарантированное обеспечение безопасности при выполнении полетов. Однако актуальные данные, предоставляемые авторитетными организациями (ICAO, FAA, EASA), говорят о том, что причиной около 80% всех авиационных происшествий является именно человеческий фактор [1]. Причиной большинства подобных инцидентов является снижение эффективности когнитивных функций летного состава [2]. В частности, это проявляется в виде ослабления концентрации и переключаемости внимания, истощения ресурсов рабочей памяти, замедления скорости зрительно-моторной реакции, нарушения механизмов эмоционально-волевой регуляции и многих других процессах [3].

Под когнитивной работоспособностью понимают интегральную характеристику индивида, отражающую его способность к реализации профессиональной деятельности заданного объема и степени ответственности, которая требует оперативного задействования нервно-психических ресурсов и комплексного функционирования всего набора когнитивных функций в условиях жестких временны́х ограничений и нормативных критериев результативности [4].

В научных работах [5–9] и регламентирующих документах по проведению профессионального психологического отбора летного состава[1, 2] выделены восемь ключевых когнитивных функций (см. на рис. 1).

Целью настоящего исследования была разработка информационно-аналитической системы, функционирующей на основе нечеткого логического вывода и позволяющей формировать количественные оценки выбранных восьми

[1] Приказ Министерства транспорта Российской Федерации от 17 октября 2025 г. № 344 «Об установлении Порядка проведения профессионального отбора членов летного экипажа пилотируемого гражданского воздушного судна, за исключением сверхлегкого пилотируемого гражданского воздушного судна с массой конструкции 115 килограммов и менее, и диспетчеров управления воз-душным движением, в том числе порядка определения их психофизиологических качеств, порядка психофизиологического обследования указанных лиц» // Официальный интернет-портал правовой информации. 2025. № 84206. URL: http://publication.pravo.gov.ru/document/0001202511190014.

[2] Руководство по психологическому обеспечению отбора, подготовки и профессиональной деятельности летного и диспетчерского состава гражданской авиации: утв. распоряжением Минтранса России от 31.10.2000 г. № 57-р. М.: Воздушный транспорт, 2001. 210 с.

когнитивных функций авиационного персонала и классифицировать их по уровню профессионального риска в рамках трехуровневой шкалы.

Функционал нечеткой логики в этой системе выполняет роль механизма, ключевым преимуществом которого перед классическими детерминированными моделями является способность моделировать когнитивные процессы, исключая искусственные разрывы при переходе через жесткие нормативные границы оценки той или иной когнитивной функции. Использование баз нечетких продукционных правил обеспечивает семантическую прозрачность механизма вывода, позволяя верифицировать логику формирования заключения на каждом этапе агрегации метрик. Кроме того, система на основе нечеткой логики демонстрирует повышенную устойчивость к случайным выбросам во входных данных [10]. Разработка подобных инструментов является весьма актуальным направлением в задачах превентивного мониторинга и раннего выявления рисков, связанных с человеческим фактором.

ОПРЕДЕЛЕНИЕ НАБОРА ПСИХОДИАГНОСТИЧЕСКИХ МЕТОДИК ОЦЕНКИ КОГНИТИВНЫХ ФУНКЦИЙ АВИАЦИОННОГО ПЕРСОНАЛА

На основании систематизации современных подходов к оценке ключевых когнитивных функций авиационного персонала сформирован первоначальный набор психодиагностических методик для экспертного сравнения, который представлен на рис. 1. Критериями включения методик являлись возможность количественной регистрации результатов, а также соответствие требованиям руководства.



Рис. 1. Когнитивные функции и методики их оценки, включенные в сравнительный анализ.

Следующим этапом исследования было определение наиболее валидных методик оценки когнитивных функций с помощью экспертного метода. Верификация психодиагностического инструментария проведена методом независимой экспертной оценки группой из пяти экспертов. Каждой методике j (j = 1,…,49), показанной на рис. 1, присваивался балл $E_{ij} \in \{1, 2, 3, 4, 5\}$ экспертом i (i = 1,…,5) в соответствии с шкалой, представленной в табл. 1.

Табл. 1. Шкала оценки психодиагностических методик оценки когнитивных функций.

Балл	Качественная интерпретация
1	Отсутствие теоретико-эмпирического обоснования и профессиональной релевантности в авиационной отрасли.
2	Низкая конструктная валидность, ограниченная применимость в авиационной практике.
3	Допустима как вспомогательный скрининговый инструмент после адаптации, но не для принятия итоговых кадровых решений.
4	Подтвержденная валидность, научная обоснованность, возможность применения в соответствии с требованиями руководящих документов.
5	Доказанная прогностическая валидность, широкая апробированность в системах авиационного профессионального психологического отбора.

Результирующая оценка методик получена путем агрегации индивидуальных оценок методом медианы, что обеспечило устойчивость итоговой оценки к расходящимся мнениям отдельных экспертов и стабильность итогового ранжирования методик:

$$S_j = \text{med}_{i=1}^{5} \left\{ E_{ij} \right\}. \quad (1)$$

где med – оператор определения медианы, представляющий собой значение, делящее упорядоченную выборку индивидуальных оценок экспертов пополам: одна половина оценок не превышает медианного значения, вторая половина – не меньше его.

Если для оценки когнитивной функции имелась единственная методика с медианной экспертной оценкой $S_j = 5.0$, она использовалась как самостоятельный диагностический инструмент. Если же имелись две методики с $S_j \geq 4.0$, которые оценивали различные аспекты одной когнитивной функции, то они обе включались в набор психодиагностических методик для повышения надежности заключения.

Решение о включении методики в итоговый диагностический комплекс принималось на основании порогового правила

$$M_j^{in} = \begin{cases} 1, \ \textit{если} \ \ S_j \geq 4.0, \\ 0, \ \textit{если} \ \ S_j < 4.0, \end{cases} \quad (2)$$

где $M_j^{in} = 1$ означает включение методики.

Таким образом, на основе результатов экспертного опроса и выражений (1) и (2) получен итоговый состав комплекса психодиагностических методик с указанием формируемых ими метрик. Он приведен в табл. 2.

Табл. 2. Метрики для оценки восьми когнитивных функций

Когнитивная функция (ее оценка)	Количество методик в оценке	Наименование методик	S_j	Метрики	Формулы для расчета метрик и их лингвистическая оценка
Внимание (Y_V)	2	Кольца Ландольта	5.0	X_{KL}	X_{KL} – количество обработанных пар колец за отведенный промежуток времени: $X_{KL} = K - (10p + 8d + 5n)$, K – количество просмотренных пар колец; p – количество незачеркнутых (пропущенных) пар колец, которые нужно было зачеркнуть в данном объеме; d – количество перечеркнутых колец, которые не нужно было зачеркивать; n – количество пар колец, зачеркнутых в неправильном направлении. Лингвистическая оценка: $T(X_{KL}) = \begin{cases} H,\ X_{KL} \in [0; 126]; \\ C,\ X_{KL} \in [127; 236]; \\ B,\ X_{KL} \in [237; 330], \end{cases}$ где Н – низкий уровень; С – средний уровень; В – высокий уровень.
				X_{CA}	X_{CA} – количество правильно отмеченных пар колец на бланке за отведенных промежуток времени. Лингвистическая оценка: $T(X_{CA}) = \begin{cases} H,\ X_{CA} \in [0; 37]; \\ C,\ X_{CA} \in [38; 58]; \\ B,\ X_{CA} \in [59; 90]. \end{cases}$
		Отыскивание чисел с переключением	5.0	X_S	X_S – время выполнения задания в секундах. Лингвистическая оценка:

					$T(X_S)=\begin{cases}B,\ X_S\in[0;130];\\C,\ X_S\in[131;216];\\H,\ X_S\in[217;300].\end{cases}$
Оперативная память (Y_{OP})	1	«Шкалы»	5.0	X_{SC}	X_{SC} – количество правильно выполненных заданий, шт. Лингвистическая оценка: $T(X_{SC})=\begin{cases}H,\ X_{SC}\in[0;5];\\C,\ X_{SC}\in[6;8];\\B,\ X_{SC}\in[9;10].\end{cases}$
Темп психических процессов (Y_{TPP})	2	СЗМР	5.0	X_{CVMR}	X_{CVMR} – интегральная оценка уровня сенсомоторных реакций, формирующаяся из среднего времени реакции на стимул и суммарного числа ошибок. Лингвистическая оценка: $T(X_{CVMR})=\begin{cases}H,\ X_{CVMR}\in[0;0.37];\\C,\ X_{CVMR}\in(0.37;0.8];\\B,\ X_{CVMR}\in(0.8;1].\end{cases}$
		Корректурная проба с кольцами	4.0	X_{SP}	X_{SP} – показатель скорости переработки информации, бит/с. $X_{SP}=\frac{0.5436\cdot N-2.807\cdot l}{t}$, где N – количество колец; t – время эксперимента; l – число пропущенных колец; 0.5436 – средняя информация, приходящаяся на одно кольцо; 2.807 – величина потери информации, приходящейся на 1 кольцо. $T(X_{SP})=\begin{cases}H,\ X_{SP}\in[0.57;0.73];\\C,\ X_{SP}\in(0.74;1.19];\\B,\ X_{SP}\in(1.20;1.36].\end{cases}$
Пространственные способности (Y_{PS})	1	«Компасы»	5.0	X_C	X_C – количество правильно определенных направлений, шт. Лингвистическая оценка: $T(X_C)=\begin{cases}H,\ X_C\in[0;15];\\C,\ X_C\in[16;30];\\B,\ X_C\in[31;50].\end{cases}$

Логическое мышление (Y_{LM})	2	«Установление закономерностей»	5.0	X_{US}	X_{US} – значение коэффициента успешности: $$X_{US} = \frac{P-N}{P+D} \cdot C,$$ где P – число всех подчеркнутых слов; N – количество неправильно подчеркнутых слов; D – количество неподчеркнутых слов, которые следовало подчеркнуть; C – количество всех просмотренных слов. Лингвистическая оценка: $$T(X_{US}) = \begin{cases} H,\ X_{US} \in [0;84]; \\ C,\ X_{US} \in [85;100]; \\ B,\ X_{US} \in [101;125]. \end{cases}$$
		Тест Равена	4.0	X_R	X_R – количество правильно выполненных заданий, шт. Лингвистическая оценка: $$T(X_R) = \begin{cases} H,\ X_R \in [0;39]; \\ C,\ X_R \in [40;51]; \\ B,\ X_R \in [52;60]. \end{cases}$$
Счетно-математические способности(Y_{CM})	2	Тест интеллекта Р. Амтхауэра (субтест 5)	5.0	X_{AN}	X_{AN} – количество правильно решенных задач, шт. Лингвистическая оценка: $$T(X_{AN}) = \begin{cases} H,\ X_{AN} \in [0;9]; \\ C,\ X_{AN} \in [10;15]; \\ B,\ X_{AN} \in [16;20]. \end{cases}$$
		Тест интеллекта Р. Амтхауэра (субтест 6)	4.0	X_{MI}	X_{MI} – количество правильно решенных заданий, шт. Лингвистическая оценка: $$T(X_{MI}) = \begin{cases} H,\ X_{MI} \in [0;9]; \\ C,\ X_{MI} \in [10;15]; \\ B,\ X_{MI} \in [16;20]. \end{cases}$$
Точность управляющих движений (Y_{TD})	2	«Лабиринт»	5.0	X_{LAB}	X_{LAB} – длина пути в условных единицах. Лингвистическая оценка: $$T(X_{LAB}) = \begin{cases} B,\ X_{LAB} \in [2864;3108]; \\ C,\ X_{LAB} \in [3109;3306]; \\ H,\ X_{LAB} \in [3307;3752]. \end{cases}$$

		Реакция на движущийся объект	4.0	X_{RDO}	X_{RDO} – количество точных остановок движущегося объекта в заданном положении, ед. Лингвистическая оценка: $T(X_{RDO}) = \begin{cases} H,\ X_{RDO} \in [0;16]; \\ C,\ X_{RDO} \in [17;54]; \\ B,\ X_{RDO} \in [55;70]. \end{cases}$
Эмоциональная устойчивость (Y_{EU})	1	16-факторный личностный опросник Р. Кеттела (Фактор С)	5.0	X_{KT}	X_{KT} – стеновая оценка фактора эмоциональной устойчивости. Лингвистическая оценка: $T(X_{KT}) = \begin{cases} H,\ X_{KT} \in [0;4]; \\ C,\ X_{KT} \in [5;7]; \\ B,\ X_{KT} \in [8;10]. \end{cases}$

На основании результатов независимой экспертной оценки был сформирован итоговый набор психодиагностических методик, обеспечивающий комплексную диагностику восьми ключевых когнитивных функций авиационного персонала: Y_V, Y_{OP}, Y_{TPP}, Y_{PS}, Y_{LM}, Y_{CM}, Y_{TD}, Y_{EU} (см. табл. 2). Комплекс включает 13 диагностических процедур, результаты которых формализуются в вектор из 14 количественных метрик $\vec{X} = \{X_{KL}, X_{CA}, X_S, X_{SC}, X_{CVMR}, X_{SP}, X_C, X_{US}, X_R, X_{AN}, X_{MI}, X_{LAB}, X_{RDO}, X_{KT}\}$. Увеличение размерности вектора относительно числа процедур обусловлено мультипараметрическим характером теста Кольца Ландольта, генерирующего два независимых показателя (X_{KL}, X_{CA}) в рамках одной тестовой процедуры.

ОБОСНОВАНИЕ ПРИМЕНЕНИЯ АППАРАТА НЕЧЕТКОЙ ЛОГИКИ ДЛЯ ОЦЕНКИ КОГНИТИВНЫХ ФУНКЦИЙ

Выбор нечеткого логического вывода в качестве механизма принятия решений в разрабатываемой информационно-аналитической системе обусловлен методологическими ограничениями детерминированных подходов при обработке разнородных психодиагностических данных. Ключевые преимущества аппарата нечетких множеств в рамках профессионального отбора авиационного персонала состоят в следующем.

1. Нечеткая логика позволяет привести все 14 разнородных метрик к единому безразмерному пространству через функции принадлежности (ФП), сохраняя семантическую нагрузку каждой метрики.

2. Поскольку когнитивные функции не имеют естественных разрывов между уровнями нормы и отклонения, применение аппарата нечетких множеств исключает эффект жестких пороговых оценок. Пограничные значения контролируемой метрики одновременно активируют смежные термы с определенной степенью истинности, что обеспечивает непрерывность оценки.

3. Ряд когнитивных функций диагностируется на основе нескольких входных метрик (например, внимание – объем, распределение, переключение). Нечеткая система позволяет задавать индивидуальные механизмы агрегации (конъюнкция, дизъюнкция, взвешенное среднее) в соответствии с психофизиологической структурой каждой функции, повышая валидность оценки [11].

4. Результаты тестирования подвержены влиянию функционального состояния, мотивации и суточных ритмов. Благодаря плавности ФП и механизму аккумуляции правил нечеткая система сглаживает малые изменения во входных переменных, предотвращая скачкообразные изменения итоговой оценки и обеспечивая воспроизводимость диагностики при повторных обследованиях.

5. Архитектура нечеткого вывода поддерживает многоуровневую градацию итоговых решений, что расширяет возможности управления кадровыми рисками. Вместо бинарного вывода «годен/не годен» система позволяет дать лингвистическую оценку уровня когнитивной функции, состоящую из заданного числа градаций, что соответствует современным требованиям риск-ориентированного подхода к организации безопасности полетов [12].

Таким образом, внедрение нечеткого моделирования является не технической адаптацией, а методологической необходимостью, обеспечивающей прогностическую валидность системы в условиях нормативно регламентированного профессионального отбора в авиационной отрасли.

НОРМИРОВАНИЕ И ФАЗЗИФИКАЦИЯ ВХОДНЫХ ДАННЫХ, ПОЛУЧАЕМЫХ В РЕЗУЛЬТАТЕ ОЦЕНКИ КОГНИТИВНЫХ ФУНКЦИЙ

После формирования вектора входных метрик $\vec{X}$ следующим этапом построения нечеткой информационно-аналитической системы является процедура их приведения к единому пространству признаков. Процедура включает два последовательных этапа.

1. Нормирование – это приведение разнородных первичных значений к безразмерному пространству [0;1]. Для исходных данных, характеризуемых как прямой, так и обратной шкалами интерпретации, алгоритм приведения к унифицированному диапазону был реализован с помощью линейного масштабирования согласно выражению

$$X_{i,n} = \begin{cases} \dfrac{X_i - X_{i,\min}}{X_{i,\max} - X_{i,\min}}, \text{ если нормирование на максимум;} \\ \dfrac{X_{i,\max} - X_i}{X_{i,\max} - X_{i,\min}}, \text{ если нормирование на минимум.} \end{cases} \quad (3)$$

где $X_{i,n} \in [0;1]$ – нормированное значение; X_i – первичное значение метрики в исходных единицах измерения; $X_{i,\min}$, $X_{i,\max}$ – нижняя и верхняя границы значений метрики.

Значения $X_{i,\min}$ и $X_{i,\max}$ для каждой метрики X_i, установленные на основе нормативных данных [8], представлены в табл. 2 в виде лингвистической оценки $T(X_i)$, интерпретирующей абсолютные значения метрик по трем качественным уровням: «низкий» (Н), «средний» (С) и «высокий» (В).

Так как метрики различаются физической природой и направленностью оптимизации, процедура нормирования осуществлялась дифференцированно. Для большинства входных переменных (X_{KL}, X_{CA}, X_{SC}, X_{CVMR}, X_{SP}, X_C, X_{US}, X_R, X_{AN}, X_{MI}, X_{RDO}, X_{KT}), где более высокое значение метрики свидетельствует о повышении эффективности деятельности (количество правильных ответов, коэффициент успешности, стеновые оценки и др.), применялось прямое линейное масштабирование с целевой установкой на максимум. В свою очередь, для метрик (X_S, X_{LAB}), характеризующих быстродействие и правильность выполнения серий испытаний, было использовано обратное преобразование, при котором минимальным первичным значениям соответствует более высокая качественная оценка.

2. Фаззификация – преобразование нормированных значений $X_{i,n}$ в степени принадлежности $\mu_T(X_{i,n})$ к лингвистическим термам $T \in$ {«Н», «С», «В»} посредством трапециевидных ФП, описываемых следующим образом:

$$\mu(X_{i,n}) = \begin{cases} 0, & X_{i,n} < a \\ \dfrac{X_{i,n} - a}{b - a}, & a \le X_{i,n} < b \\ 1, & b \le X_{i,n} \le c \\ \dfrac{d - X_{i,n}}{d - c}, & c < X_{i,n} \le d \\ 0, & X_{i,n} > d \end{cases}, \quad (4)$$

где a – нижняя граница диапазона (значение, ниже которого $\mu(X_{i,n}) = 0$); b и c – начало и конец зоны с ($\mu(X_{i,n}) = 1$); d – верхняя граница диапазона (значение, выше которого $\mu(X_{i,n}) = 0$).

На основе данных табл. 2 для каждой входной метрики в соответствии с выражением (3) были определены опорные интервалы, соответствующие ядрам ФП (зонам с $\mu = 1$) для качественных уровней «Н», «С», «В». Прямое использование этих интервалов в системе нечеткого вывода способствует наличию разрывов между границами смежных уровней и образованию «мертвых зон» ($\mu = 0$). Для обеспечения корректности нечеткого вывода осуществлен переход от жестких нормативных границ к параметризованным трапециевидным ФП. Границы носителей сформированы симметричным расширением ядер на величину Δ = 0.05 – 0.15 по шкале [0;1], что гарантирует пересечение смежных термов при $\mu = 0.5$ и обеспечивает плавную агрегацию входных метрик при активации правил.

В табл. 3 представлены параметры ФП входных метрик $X_{i,n}$ для оценки когнитивных функций.

Табл. 3. Параметры функции принадлежности для входных метрик $X_{i,n}$.

Оценка когнитивной функции	Параметры функции принадлежности для входных метрик $X_{i,n}$
Y_V	$X_{KL.n}$: Н[0; 0; 0.382; 0.482]. С[0.382; 0.482; 0.715; 0.815]. В[0.715; 0.815; 1; 1] $X_{CA.n}$: Н[0; 0; 0.411; 0.511]. С[0.411; 0.511; 0.644; 0.744]. В[0.644; 0.744; 1; 1] $X_{S.n}$: Н[0; 0; 0.277; 0.377]. С[0.277; 0.377; 0.563; 0.663]. В[0.563; 0.663; 1; 1]
Y_{OP}	$X_{SC.n}$: Н[0; 0; 0.5; 0.6]. С[0.5; 0.6; 0.8; 0.9]. В[0.8; 0.9; 1; 1]
Y_{TPP}	$X_{CVMR.n}$: Н[0; 0; 0.37; 0.52]. С[0.37; 0.52; 0.8; 0.95]. В[0.8; 0.95; 1; 1] $X_{SP.n}$: Н[0; 0; 0.203; 0.353]. С[0.203; 0.353; 0.785; 0.935]. В[0.785; 0.935; 1; 1]
Y_{PS}	$X_{C.n}$: Н[0; 0; 0.3; 0.45]. С[0.3; 0.45; 0.6; 0.75]. В[0.6; 0.75; 1; 1]

Y_{LM}	$X_{US.n}$: H[0; 0; 0.672; 0.772]. C[0.672; 0.772; 0.8; 0.9]. B[0.8; 0.9; 1; 1] $X_{R.n}$: H[0; 0; 0.65; 0.75]. C[0.65; 0.75; 0.85; 0.95]. B[0.85; 0.95; 1; 1]
Y_{CM}	$X_{AN.n}$: H[0; 0; 0.45; 0.6]. C[0.45; 0.6; 0.75; 0.9]. B[0.75; 0.9; 1; 1] $X_{MI.n}$: H[0; 0; 0.45; 0.6]. C[0.45; 0.6; 0.75; 0.9]. B[0.75; 0.9; 1; 1]
Y_{TD}	$X_{LAB.n}$: H[0; 0; 0.501; 0.651]. C[0.501; 0.651; 0.724; 0.874]. B[0.724; 0.874; 1; 1] $X_{RDO.n}$: H[0; 0; 0.229; 0.379]. C[0.229; 0.379; 0.771; 0.921]. B[0.771; 0.921; 1; 1]
Y_{EU}	$X_{KT.n}$: H[0; 0; 0.4; 0.5]. C[0.4; 0.5; 0.7; 0.8]. B[0.7; 0.8; 1; 1]

Параметры ФП для выходной лингвистической переменной Y_j, используемой при оценке всех восьми когнитивных функций, установлены на основе структурированного экспертного опроса с последующей нормативной верификацией. В опросе приняли участие пять специалистов в области авиационной психологии и профессионального отбора (стаж работы ≥ 10 лет, наличие действующих допусков к экспертизе летного состава). Экспертам было предложено независимо определить границы трех качественных уровней когнитивной функции: «Н» (зона риска), «С» (допустимый уровень), «В» (оптимальный профиль) по унифицированной шкале [0; 1], исходя из требований к надежности операторской деятельности [7, 8] и собственного опыта интерпретации психодиагностических данных. Получены следующие параметры ФП для Yj: H[0; 0; 0,25; 0,40], C[0,25; 0,40; 0,65; 0,75], B[0,65; 0,75; 1; 1].

На рис. 2 показано графическое представление ФП, разработанных для входных нормированных метрик $X_{KL,n}$, $X_{CA,n}$ и $X_{S,n}$. Графическое представление полученной ФП для выходной лингвистической переменной показано на рис. 3.

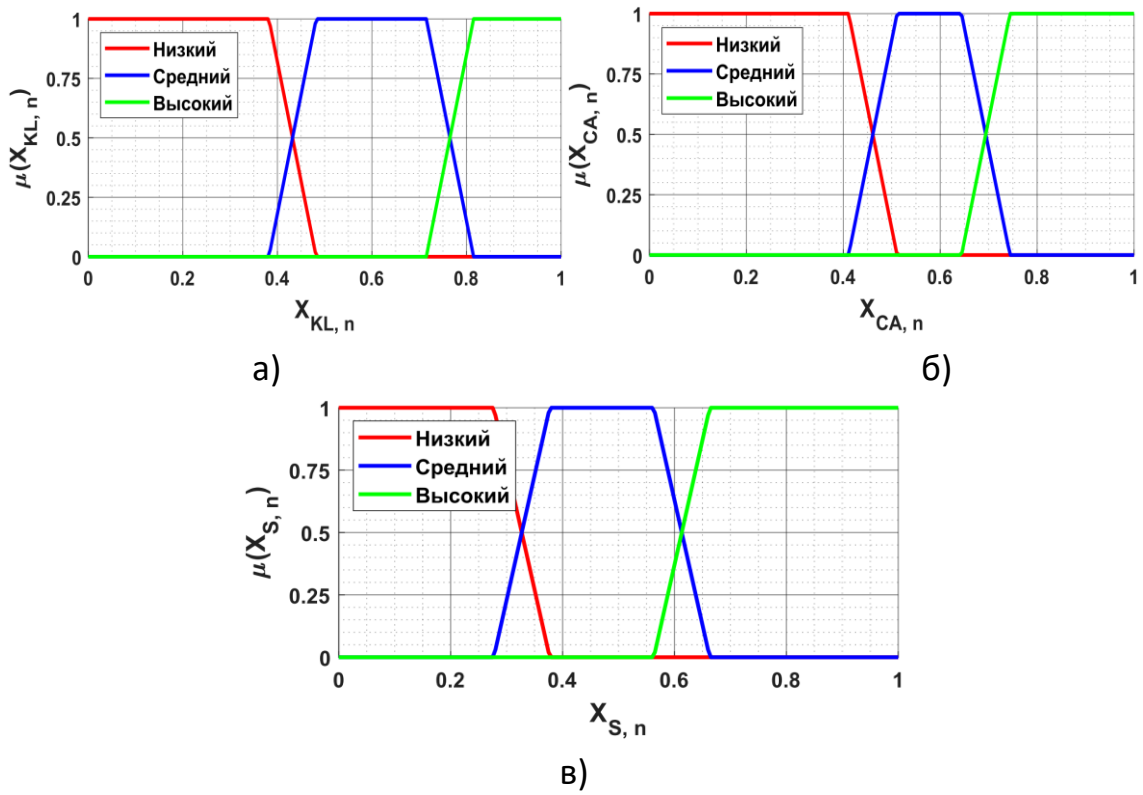


Рис. 2. Функции принадлежности для входных переменных (а) $X_{KL,n}$; (б) $X_{CA,n}$; (в) $X_{S,n}$.

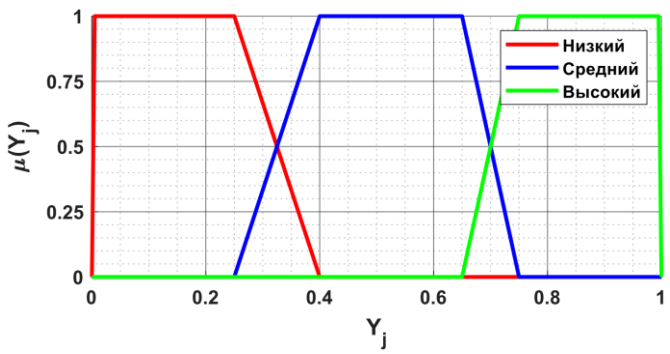


Рис. 3. Функции принадлежности для выходной лингвистической переменной Y_j.

Ключевой архитектурной особенностью разрабатываемой информационно-аналитической системы является структурная неоднородность входных метрик для получения итоговых оценок когнитивных функций. Оценки «оперативная память» (Y_{OP}), «пространственные способности» (Y_{PS}) и «эмоциональная устойчивость» (Y_{EU}) формируются на базе единичных входных метрик $X_{SC,n}$, $X_{C,n}$ и $X_{KT,n}$ соответственно. Оценки «темп психических процессов» (Y_{TPP}), «логическое

мышление» (Y_{LM}), «счетно-математические способности» (Y_{CM}) и «точность управляющих движений» (Y_{TD}), формируются с учетом двух входных метрик: $X_{CVMR,n}$ и $X_{SP,n}$; $X_{US,n}$ и $X_{R,n}$; $X_{AN,n}$ и $X_{MI,n}$; $X_{LAB,n}$ и $X_{RDO,n}$ соответственно. Оценка «внимание» (Y_V) является агрегацией трех входных метрик $X_{KL,n}$, $X_{CA,n}$ и $X_{S,n}$. Подобная неоднородность объясняется применяемым набором психодиагностических методик.

Результирующая оценка когнитивных функций на основе одной метрики формируется посредством прямой трансляции степеней принадлежности. Значение ФП входной метрики к лингвистическому терму без дополнительных преобразований переносится на выходной терм функции. Для функций, зависящих от двух или трех входных переменных, применена база нечетких правил вида «ЕСЛИ… И… ТО…». В условиях правил операторы нечеткой логики (min для конъюнкции, max для дизъюнкции) определяют степень истинности правила на базе степеней принадлежности входных метрик. Аккумуляция заключений осуществляется посредством операции max, а итоговое четкое значение получается с помощью метода центра тяжести.

РАЗРАБОТКА АЛГОРИТМОВ ОЦЕНКИ КОГНИТИВНЫХ ФУНКЦИЙ НА ОСНОВЕ НЕЧЕТКОЙ ЛОГИКИ

В разрабатываемой информационно-аналитической системе оценка когнитивных функций реализована на базе алгоритма нечеткого вывода Мамдани [12], конфигурация которого адаптирована под количество входных метрик.

Для получения оценок Y_{OP}, Y_{PS} и Y_{EU} когнитивных функций реализован алгоритм нечеткого вывода Мамдани с упрощенной базой правил. Процедура включает следующие этапы.

1. Фаззификация – это вычисление для каждой нормированной входной метрики $X_{i,n} \in [0; 1]$ степени принадлежности $\alpha_k = \mu_{Tk}(X_{i,n})$ (формула (4)) трем лингвистическим термам T_k данной входной переменной: «Н», «С», «В» (табл. 3), где $k \in \{Н, С, В\}$.

2. Активация правил – для каждого из трех правил вида

$$\textbf{ЕСЛИ } X_{i,n} \text{ есть } T_k, \textbf{ ТО } Y_j \text{ есть } T_{k,out} \quad (5)$$

вычислена степень истинности (активации) правила $\beta_k = \alpha_k$, где Y_j – выходная лингвистическая переменная; $T_{k,out}$ – лингвистические термы для Y_j, заданные на интервале [0; 1] с параметрами, указанными в табл. 4.

3. Усечение заключений – ФП выходного терма $T_{k,out}$ для каждого $k \in$ {Н, С, В} усекается на уровне β_k

$$\mu'_k(y) = \min(\mu_{Tk,out}(y), \beta_k), \quad (6)$$

где $y \in [0; 1]$ – аргумент в области определения Y_j.

4. Агрегация (аккумуляция) — усеченные ФП по формуле (6) объединены в единое результирующее нечеткое множество:

$$\mu_\Sigma(y) = \max(\mu'_Н(y), \mu'_С(y), \mu'_В(y)), \quad (7)$$

где $\mu_\Sigma(y)$ — агрегированная ФП выходной лингвистической переменной Y_j, формируемая путем поточечного применения операции максимума к усеченным ФП выходных термов для всех $y \in [0;1]$.

5. Дефаззификация – на основе результата агрегации (7) методом центра тяжести вычислено четкое значения выходной переменной Y_j в соответствии с выражением

$$Y_j = (\sum y \cdot \mu_\Sigma(y)) / (\sum \mu_\Sigma(y)). \quad (8)$$

На рис. 4 представлен количественный пример получения четкой оценки Y_{OP} когнитивной функции «оперативная память» на основе входного нормированного значения $X_{SC,n} = 0.5$ в соответствии с алгоритмом (3)–(8).

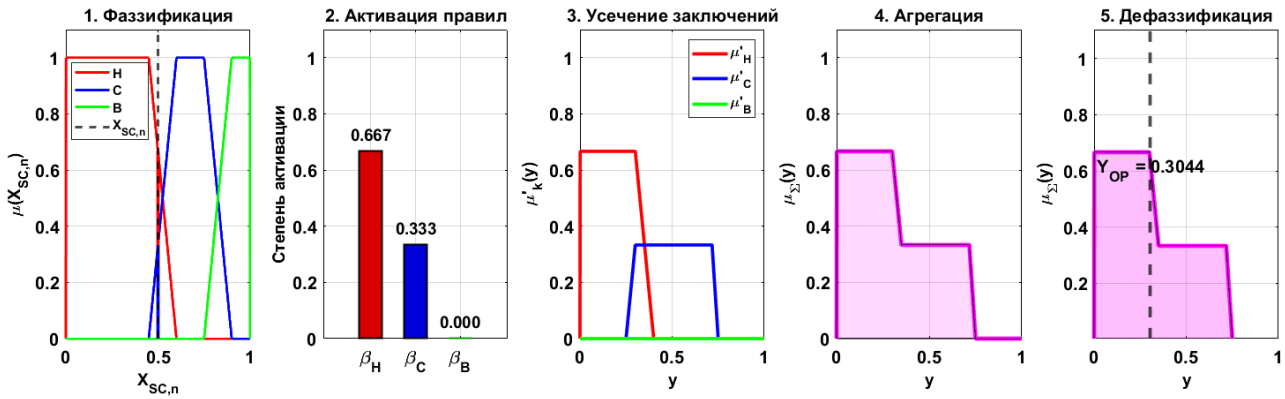


Рис. 4. Этапы формирования оценки Y_{OP}.

Для оценки когнитивных функций на основе двух входных метрик реализуется алгоритм нечеткого вывода Мамдани, включающий сформированную полную базу из 9 продукционных правил, связывающих комбинации входных термов T_{1k} и T_{2k} с лингвистическим термом $T_{k,out}$ выходной переменной Y_j. Приведем полную базу правил для оценки когнитивной функции Y_{TPP} на основе входных метрик $X_{CVMR,n}$ и $X_{SP,n}$ (табл. 4).

Табл. 4. База продукционных правил для оценки когнитивной функции на основе двух входных метрик (на примере Y_{TPP}).

$X_{CVMR,n}/X_{SP,n}$	T_{2H}	T_{2C}	T_{2B}
T_{1H}	H	H	C
T_{1C}	H	C	B
T_{1B}	C	B	B

Аналогичным образом сформированы базы правил для функций Y_{LM}, Y_{CM} и Y_{TD}.

Для каждого правила вычислена степень его истинности β_j с помощью операции нечеткой конъюнкции между степенями активации условий правила:

$$\beta_j = \min(\alpha_{1k}, \alpha_{2k}). \quad (9)$$

На рис. 5 представлен количественный пример получения четкой оценки Y_{TPP} когнитивной функции «темп психических процессов» на основе входных нормированных значений $X_{CVMR,n} = 0.3$ и $X_{SP,n} = 0.9$.

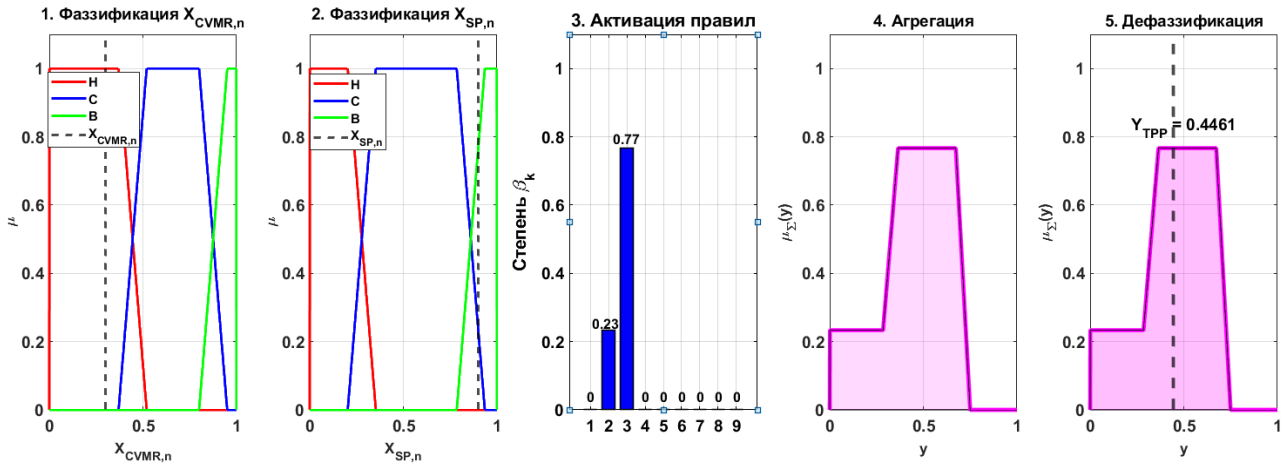


Рис. 5. Этапы формирования оценки Y_{TPP}.

Для оценки когнитивной функции Y_V на основе трех входных метрик $X_{KL,n}$, $X_{CA,n}$ и $X_{S,n}$ алгоритм нечеткого вывода предусматривает использование базы из 27 продукционных правил (табл. 6).

Табл. 6. База продукционных правил для оценки Y_V.

$X_{KL,n}$ и $X_{CA,n}$ / $X_{S,n}$	T_{3H}	T_{3C}	T_{3B}
T_{1H} и T_{2H}	Н	Н	С
T_{1H} и T_{2C}	Н	С	С
T_{1H} и T_{2B}	С	С	В
T_{1C} и T_{2H}	Н	С	С
T_{1C} и T_{2C}	С	С	В
T_{1C} и T_{2B}	С	В	В
T_{1B} и T_{2H}	С	С	В
T_{1B} и T_{2C}	С	В	В
T_{1B} и T_{2B}	В	В	В

По аналогии с выражением (9) для каждого правила вычислялась степень его истинности β_j:

$$\beta_j = \min(\alpha_{1k}, \alpha_{2k}, \alpha_{3k}). \quad (10)$$

Таким образом, рассмотрев алгоритмы оценки когнитивных функций на основе различного набора входных переменных, можно сформировать обобщенную структурную схему информационно-аналитической системы оценки когнитивных функций авиационного персонала на основе нечеткого вывода (рис. 6).

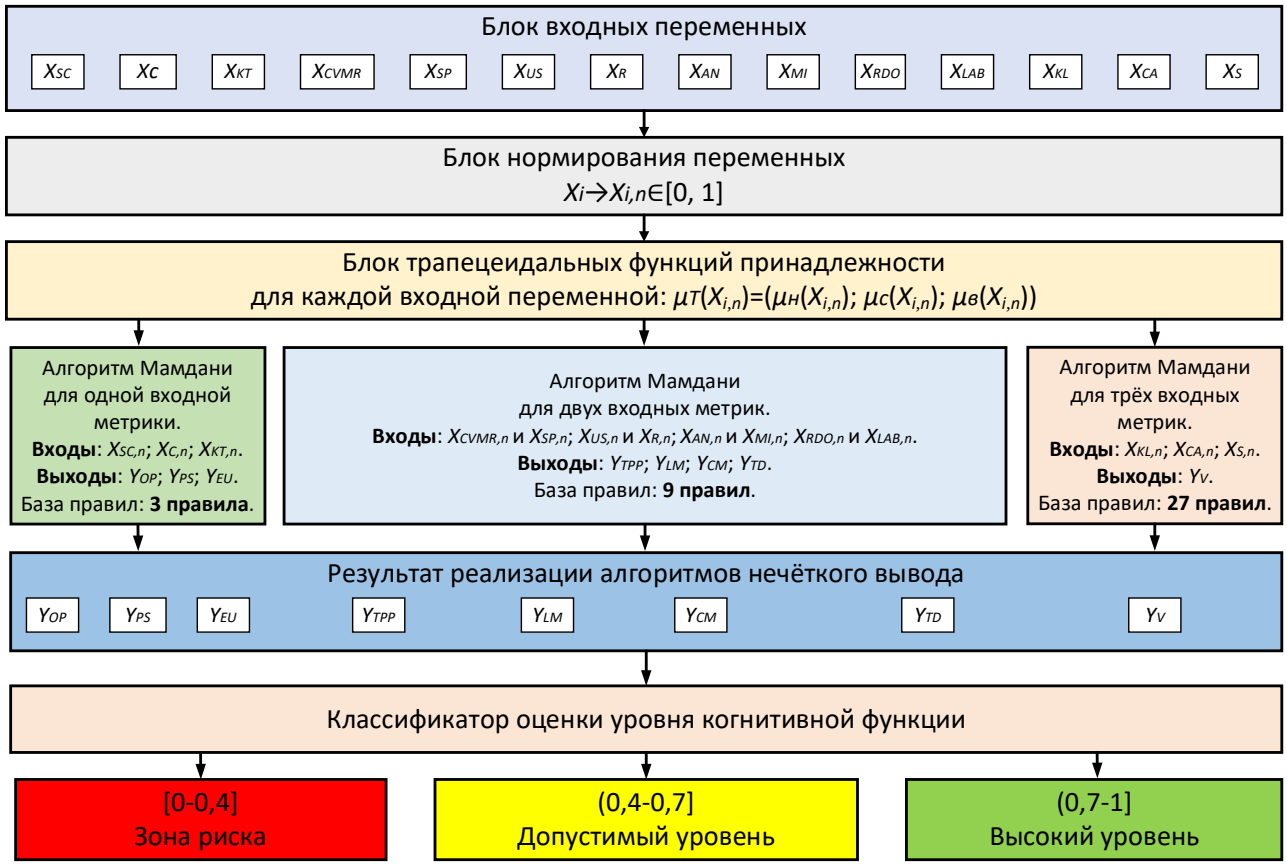


Рис. 6. Структурная схема информационно-аналитической системы нечеткого вывода для оценки когнитивных функций.

Разработанная информационно-аналитическая система оценки когнитивных функций авиационного персонала реализует модульную архитектуру нечеткого вывода, адаптированную под различную размерность входного пространства. В отличие от моделей жесткой агрегации, система осуществляет параллельную обработку диагностических данных через специализированные алгоритмические каналы, каждый из которых оптимизирован под количество метрик, валидирующих конкретную когнитивную функцию. Финальный этап работы системы обеспечивает прямую качественную интерпретацию результатов через специализированный классификатор.

Входным контуром системы является блок входных переменных, включающий 14 количественных метрик. Поскольку исходные данные характеризуются разнородной шкалой измерений, они поступают в блок нормирования переменных, где осуществляется их приведение к единому безразмерному пространству $X_{i,n} \in [0; 1]$ в соответствии с выражением (3).

Далее нормированные значения передаются в блок трапецеидальных функций принадлежности, где для каждой метрики вычисляются степени принадлежности $\mu_{Tk}(X_{i,n})$ к трем лингвистическим термам: (Н), (С) и (В) по формуле (4).

Ядро системы организовано по параллельному принципу и включает три алгоритмических модуля Мамдани. Для когнитивных функций, оцениваемых по одной метрике, база правил содержит три элемента, реализующие прямое отображение входного терма на выходной. Для функций, оцениваемых по двум метрикам, формируется полная комбинация, включающая 9 правил. Для оценки когнитивной функции «внимание» база масштабируется до 27 правил. Активация продукций осуществляется через оператор нечеткой конъюнкции (минимум степеней принадлежности условий), что гарантирует строгость оценки. Снижение показателя по любому из входов ограничивает степень срабатывания правила. Последующая агрегация усеченных выходных функций выполняется методом максимума, а преобразование нечеткого множества в четкое значение Y_j для каждой из восьми когнитивных функций реализуется дефаззификацией с помощью метода центра тяжести.

Результатом работы алгоритмических каналов является вектор из восьми нормированных оценок, который поступает в финальный модуль – классификатор оценки уровня когнитивной функции. Данный блок осуществляет качественную интерпретацию количественных значений $Y_j \in [0;1]$ путем их сопоставления с установленными пороговыми диапазонами. В зависимости от величины оценки система относит состояние функции к одной из трех зон:

– зона риска [0; 0.4];

– допустимый уровень (0.4; 0.7];

– высокий уровень (0.7; 1.0].

Подобная архитектура обеспечивает прозрачность диагностического процесса и позволяет эксперту оперативно принимать решения о профессиональной пригодности на основе оценок отдельных когнитивных функций.

ОБОСНОВАНИЕ ПОРОГОВЫХ ЗНАЧЕНИЙ КЛАССИФИКАТОРА

Выбор пороговых значений классификатора обусловлен требованиями безопасности полетов и нормативными критериями профессионального отбора

авиационного персонала. Границы интервалов установлены на основе анализа статистических распределений когнитивных показателей в репрезентативных выборках кандидатов и действующих членов летных экипажей.

Зона риска [0; 0.4) соответствует нижней границе нормального распределения и характеризует выраженный дефицит когнитивной функции. Согласно [8], такие показатели являются основанием для ограничения допуска, так как не обеспечивают минимально необходимый уровень надежности операторской деятельности в условиях стандартной эксплуатационной нагрузки.

Допустимый уровень [0.4; 0.7) охватывает диапазон значений, соответствующих средним показателям профессиональной популяции, и удовлетворяет базовым требованиям профессиональной пригодности, однако предполагает необходимость периодического мониторинга или дополнительных тренировок для поддержания стабильности когнитивного статуса, особенно при работе на сложных типах воздушных судов.

Высокий уровень [0.7; 1.0] характеризует оптимальный когнитивный профиль, обеспечивающий резерв психической надежности и устойчивость к факторам авиационного стресса. Такие показатели рекомендованы для допуска к управлению сложными авиационными комплексами и выполнения полетов в особых условиях эксплуатации.

Разработанный классификатор методологически сопоставим с существующими подходами к оценке когнитивных функций в авиационной сфере. Использование перцентилей эмпирического распределения для определения границ зон риска и высокого уровня соответствует практике нормирования когнитивных тестов в системах отбора пилотов [13], где 25-й и 75-й перцентили применяются для дифференциации данных уровней.

Подобная градация согласуется с принципами риск-ориентированного управления безопасностью полетов, где критические дефициты отдельных функций не компенсируются высокими показателями по другим направлениям, что соответствует концепции «слабого звена» в авиационной эргономике. Необходимо отметить, что возможна дальнейшая незначительная верификация границ интервалов классификатора на репрезентативной выборке авиационного персонала с использованием методов критериальной валидности.

ЗАКЛЮЧЕНИЕ

Разработана информационно-аналитическая система оценки когнитивных функций авиационного персонала, функционирующей на принципах нечеткого логического вывода. Получены следующие основные результаты.

1. Сформирован и обоснован и набор из 13 психодиагностических методик, позволяющих количественно оценить восемь целевых когнитивных функций на базе 14 входных метрик различной физической природы.

2. Разработан модульный алгоритм нечеткого вывода Мамдани, реализующий параллельную обработку данных через специализированные каналы, рассчитанные на 1, 2 или 3 входные метрики. Масштабирование базы правил (3, 9, 27 продукций) обеспечивает корректную агрегацию входных переменных без экспоненциального роста вычислительной сложности.

3. Предложена процедура нормирования и фаззификации на основе трапециевидных функций принадлежности, исключающая образование «мертвых зон» и обеспечивающая плавность переходов между качественными уровнями оценки когнитивных функций.

4. Спроектирован классификатор оценки качественного уровня когнитивной функции с тремя градациями риска, согласованный с нормативными требованиями гражданской авиации.

Практическая значимость работы заключается в создании инструмента автоматизированной поддержки принятия решений для специалистов по профессиональному отбору. Внедрение системы позволит перейти от бинарной оценки «годен/не годен» к дифференцированному управлению кадровыми рисками, учитывающему структуру когнитивного профиля кандидата.

Дальнейшие исследования будут направлены на программную реализацию системы и верификацию ее прогностической валидности на репрезентативных выборках летного состава.

СПИСОК ЛИТЕРАТУРЫ

1. International Civil Aviation Organization (ICAO). Safety Report 2024: State of Global Aviation Safety. Montréal: ICAO, 2025. 78 p. URL: https://www.icao.int/sites/default/files/sp-files/safety/Documents/ICAO_SR_2024.pdf (дата обращения: 08.04.2026).

2. *Masi G., Amprimo G., Ferraris C., Priano L.* Stress and Workload Assessment in Aviation — A Narrative Review. Sensors 2023, 23, 3556. https://doi.org/10.3390/s23073556

3. *Vicente-Rodríguez M., Iglesias Gallego D., Fuentes-García J.P., Clemente-Suárez V.J.* Portable biosensors for psychophysiological stress monitoring of a helicopter crew // Sensors. 2020. Vol. 20, No. 23. P. 6849. https://doi.org/10.3390/s20236849

4. *Демина П.Н., Трофимов Е.А., Берсенев Е.Ю.* Методология оценки и контроля когнитивной работоспособности человека-оператора при осуществлении совмещенной деятельности в условиях воздействия неблагоприятных факторов среды // Пилотируемые полеты в космос. 2024. № 2(51). С. 35–46.

5. *Пономаренко В.А.* Психология профессиональной деятельности: теоретические и прикладные проблемы. М.: Ин-т психологии РАН, 2006. 623 с.

6. *Алексеенко М.С., Серова А.А.* Системно-психологическая концепция профессионального отбора лиц летного и диспетчерского состава // Системная психология и социология. 2024. № 4 (52). С. 18–31. https://doi.org/10.24412/2223-6872-2024-452-18-31

7. *Райгородский Д.Я.* Практическая психодиагностика. Методики и тесты. Учебное пособие. Самара: Издательский Дом «БАХРАХ-М», 2001. 672 с.

8. *Бурлачук Л.Ф.* Психодиагностика: Учебник для вузов. 2-е изд., перераб. и доп. СПб.: Питер, 2006. 351 с.

9. *Анастази А.* Психологическое тестирование: монография. М.: Директ-Медиа, 2008. 859 с.

10. *Пегат А.* Нечеткое моделирование и управление; пер. с англ. 5-е изд., электрон. М.: Лаборатория знаний, 2024. 801 с.

11. *Усков А.А.* Системы с нечеткими моделями объектов управления. Монография. Смоленск: СФРУК, 2013. 153 с.

12. *Mamdani E.H., Assilian S.* An experiment in linguistic synthesis with a fuzzy logic controller // International Journal of Man-Machine Studies. 1975. Vol. 7, No. 1. P. 1–13.

13. Compilation of Pilot Cognitive Ability Norms: technical report AFRL-SA-WP-TR-2012-0001 / Air Force Research Laboratory, Human Effectiveness Directorate. Wright-Patterson AFB, OH: U.S. Air Force, 2012. 48 p.

INFORMATION AND ANALYTICAL SYSTEM FOR ASSESSING COGNITIVE FUNCTIONS OF AVIATION PERSONNEL BASED ON FUZZY LOGIC

D. S. Koptev[1][0000-0002-7759-579X], Y. A. Khalin[2][0000-0002-7020-8515], A. S. Sizov[3] [0000-0001-9658-0318]

[1–3] *Southwest State University, Kursk, Russia*

[1]d.s.koptev@mail.ru, [2]yur-khalin@yandex.ru, [3]sizov@mail.ru

Abstract

The article presents an information and analytical system for assessing eight cognitive functions of aviation personnel based on fuzzy logic. The system receives 14 input variables, which are the results of subjects passing 13 validated psychodiagnostic methods. The system operates based on the Mamdani fuzzy inference algorithm, scalable to the input space dimension (1, 2 or 3 metrics) and using adaptive bases of production rules (3, 9 or 27). The generated quantitative assessments of eight cognitive functions in the range [0; 1] are fed to a three-level classifier, which implements their qualitative interpretation by assigning them to one of three zones: risk [0;0.4], acceptable level (0.4;0.7] or high level (0.7;1]. The practical significance of the developed information-analytical system lies in its role as a tool for automated decision support for professional selection specialists. Implementing the system will enable a shift from a binary «fit/unfit» assessment to a differentiated approach to managing personnel risks that takes into account the structure of a candidate's cognitive profile.

Keywords: *Information and analytical system, cognitive functions, aviation personnel, fuzzy logic, psychodiagnostic methods, Mamdani algorithm.*

REFERENCES

1. International Civil Aviation Organization (ICAO). Safety Report 2024: State of Global Aviation Safety. Montréal: ICAO, 2025. 78 p. URL: https://www.icao.int/sites/default/files/sp-files/safety/Documents/ICAO_SR_2024.pdf (accessed 04.05.2026).

2. *Masi G., Amprimo G., Ferraris C., Priano L.* Stress and Workload As-sessment in Aviation — A Narrative Review. Sensors 2023, 23, 3556. https://doi.org/10.3390/s23073556.

3. *Vicente-Rodríguez M., Iglesias Gallego D., Fuentes-García J.P., Clemente-Suárez V.J.* Portable biosensors for psychophysiological stress monitoring of a helicopter crew // Sensors. 2020. Vol. 20, No. 23. P. 6849. https://doi.org/10.3390/s20236849

4. *Demina P.N., Trofimov E.A., Bersenev E.Yu.* Methodology for assessing and monitoring the cognitive performance of a human operator during combined activities exposed to adverse environmental factors // Manned Space Flights. 2024. No. 2 (51). P. 35–46.

5. *Ponomarenko V.A.* Psychology of Professional Activity: Theoretical and Applied Problems. M.: Institute of Psychology, Russian Academy of Sciences, 2006. 623 p.

6. *Alekseenko M.S.*, Serova A.A. Systemic-Psychological Concept of Professional Selection of Flight and Air Traffic Control Personnel // Systemic Psychology and Sociology. 2024. No. 4 (52). P. 18–31. https://doi.org/10.24412/2223-6872-2024-452-18-31.

7. *Raigorodskiy D.Ya.* Practical Psychodiagnostics. Methods and Tests. Textbook. Samara: Publishing House "BAKHRAKH-M", 2001. 672 p.

8. *Burlachuk L.F.* Psychodiagnostics: Textbook for Universities. 2nd ed., revised and enlarged. St. Petersburg: Piter, 2006. 351 p.

9. *Anastasi A.* Psychological Testing: Monograph. M.: Direct-Media, 2008. 859 p.

10. *Pegat A.* Fuzzy Modeling and Control; Translated from English. 5th ed., electronic. Moscow: Knowledge Laboratory, 2024. 801 p.

11. *Uskov A.A.* Systems with Fuzzy Models of Control Objects. Monograph. Smolensk: SFRUK, 2013. 153 p.

12. *Mamdani E.H., Assilian S.* An experiment in linguistic synthesis with a fuzzy logic controller // International Journal of Man-Machine Studies. 1975. Vol. 7, No. 1. pp. 1–13.

13. Compilation of Pilot Cognitive Ability Norms: technical report AFRL-SA-WP-TR-2012-0001 / Air Force Research Laboratory, Human Effectiveness Directorate. Wright-Patterson AFB, OH: U.S. Air Force, 2012. 48 p.

СВЕДЕНИЯ ОБ АВТОРАХ



КОПТЕВ Дмитрий Сергеевич – старший преподаватель кафедры информационной безопасности, ФГБОУ ВО «Юго-Западный государственный университет».

Dmitry Sergeevich KOPTEV – senior lecturer, Department of Information Security, Southwest State University.

email: d.s.koptev@mail.ru

ORCID 0000-0002-7759-579X



ХАЛИН Юрий Алексеевич – кандидат технических наук, доцент, доцент кафедры «Программная инженерия», ФГБОУ ВО «Юго-Западный государственный университет».

Yuri Alekseevich KHALIN – Candidat of Technical Science, Associate Professor, associate professor at the Department of Software Engineering, Southwest State University.

email: yur-khalin@yandex.ru

ORCID 0000-0002-7020-8515



СИЗОВ Александр Семенович – доктор технических наук, профессор. Профессор кафедры «Программная инженерия», ФГБОУ ВО «Юго-Западный государственный университет». Области исследований: нейронные сети и нечеткие системы, интеллектуальные системы управления сложными техническими объектами.

Alexander Semenovich SIZOV – Doctor of Technical Sciences, professor, professor at the Department of Software Engineering, Southwest State University. Research areas: neural networks and fuzzy systems, intelligent control systems for complex technical objects.

email: sizov@mail.ru

ORCID 0000-0001-9658-0318

Материал поступил в редакцию 30 июня 2026 года

УДК 004.89

ЦИФРОВАЯ ОБРАЗОВАТЕЛЬНАЯ ПЛАТФОРМА МИРЕРА

А. Г. Леонов[1] [0000-0001-9622-1526], К. А. Мащенко[2] [0000-0003-0355-6699]

[1, 2]*НИЦ «Курчатовский институт» – НИИСИ, г. Москва, Россия*

[1, 2]*Московский государственный университет имени М. В. Ломоносова, г. Москва, Россия*

[1]*Московский педагогический государственный университет, г. Москва, Россия*

[1]*Государственный университет управления, г. Москва, Россия*

[1]dr.l@vip.niisi.ru, [2]kirill.mashchenko@niisi.ru

Аннотация

Представлены основные возможности цифровой образовательной платформы Мирера, относящейся к классу полидисциплинарных информационных интеллектуальных систем и предназначенной для комплексной поддержки образовательного процесса. Платформа объединяет средства управления обучением, автоматизированной проверки заданий, образовательной аналитики, учета посещаемости и интеллектуальной поддержки преподавателей и обучающихся на основе технологий искусственного интеллекта. Описаны основные функциональные возможности системы, включая автоматическую проверку заданий по различным дисциплинам, инструменты формирования адаптивных образовательных траекторий, анализ цифрового следа обучающихся, встроенные средства коммуникации и нейросетевых ассистентов. Рассмотрены механизмы автоматизации рутинных процессов и поддержки принятия педагогических решений на основе данных. Результаты пилотного внедрения показали, что использование платформы способствует снижению нагрузки на преподавательский состав, повышению вовлеченности студентов и повышению эффективности управления образовательным процессом, что подтверждает потенциал применения цифровых образовательных платформ в качестве инструмента трансформации высшего, среднего общего и специального профессионального образования.

© А. Г. Леонов, К. А. Мащенко, 2026.
Данная статья распространяется на условиях международной лицензии Creative Commons License Attribution 4.0 International (CC BY 4.0).

***Ключевые слова**: цифровая образовательная платформа, искусственный интеллект в образовании, образовательная аналитика, автоматизация обучения, адаптивное обучение, цифровая трансформация образования.*

ВВЕДЕНИЕ

Цифровая трансформация высшего образования сопровождается активным внедрением информационных технологий, направленных на повышение эффективности организации учебного процесса, качества обучения и доступности образовательных ресурсов. Современные цифровые образовательные платформы становятся не только инструментом доставки учебного контента, но и средой управления образовательными траекториями, мониторинга результатов обучения и поддержки взаимодействия между всеми участниками образовательного процесса. При росте объемов образовательных данных и увеличении нагрузки на преподавателей актуальность приобретают методы автоматизации рутинных задач и интеллектуальной поддержки принятия педагогических решений. Несмотря на значительное разнообразие существующих образовательных платформ, а также специализированных сервисов, большинство из них узконаправленные, ориентированы на решение отдельных задач и не обеспечивают комплексной интеграции средств образовательной аналитики, автоматизированного контроля результатов обучения и интеллектуальных инструментов сопровождения обучающихся и преподавателей.

В определенном смысле большинство платформ выросло из компьютерных практикумов, получивших развитие еще в 90е годы XX в. [1] и впоследствии развившихся до цифровых образовательных сред. Под цифровой образовательной средой (ЦОС) здесь понимается программное или программно-аппаратное обеспечение, предназначенное для формирования у обучаемого необходимых компетенций в конкретной предметной области. Классическим примером ЦОС является предметно-цифровая среда ПиктоМир [2]. Отдельным элементом развития платформ являются системы управления обучением, LMS (Learning Management System) – программное обеспечение для администрирования, документирования, отслеживания, отчетности, автоматизации и доставки образовательных курсов, программ обучения, материалов и т. п. [3, 4].

Анализ существующих цифровых образовательных платформ показывает, что ни одно из распространенных решений не обеспечивает в полной мере требования российских вузов к организации учебного процесса, сопровождению преподавателя и использованию образовательной аналитики. Наиболее распространенной системой остается Moodle [3], отличающаяся гибкостью и широким распространением, однако ее использование связано со сложностью администрирования и высокой зависимостью от технических специалистов. Российские платформы, такие как Modeus [4], в большей степени учитывают специфику российских образовательных организаций, однако остаются преимущественно системами управления контентом и образовательным процессом.

Для преподавания программирования широко применяются специализированные платформы, такие как ejudge [5] и Яндекс.Контест [6], обеспечивающие автоматическую проверку решений. В то же время у них есть ограничения, связанные с созданием и обновлением образовательного контента, предоставлением обратной связи учащимся и обеспечением всестороннего образовательного процесса. Массовые открытые онлайн-платформы обучения, такие как Coursera [7], в первую очередь ориентированы на самостоятельное обучение и не могут заменить постоянную педагогическую поддержку, что приводит к снижению доли студентов, успешно завершающих курсы.

Традиционные системы управления обучением (Learning Management System, LMS), как правило, не ставят своей целью создание нового цифрового качества обучения, а лишь администрируют в цифровой форме классический образовательный процесс, включая электронное назначение курсов, контроль прохождения, посещаемость, электронную отчетность и т. п. Другой подход в цифровой трансформации образовательного процесса связан с изменением парадигмы обучения, когда в центр ставятся интересы, цели, поведение обучаемого и упор на индивидуализацию его учебного опыта. Такие системы появились в 10-е годы и были названы по цели их создания платформами образовательного опыта (Learning Experience Platform, LXP) [8].

Любая LXP прежде всего ориентирована на персонализацию обучения с использованием цифровых данных слушателя: платформа собирает информацию о траекториях, предпочтениях и активности обучающегося и генерирует рекомендации релевантного контента при построении индивидуализированных

учебных треков. В результате LXP обеспечивает «тонкую» настройку предметных и межпредметных образовательных траекторий под конкретного обучающегося. LXP, по сравнению с пассивной LMS, не просто следует предварительно заданному учебному плану, а динамически конструирует образовательные треки, давая рекомендации педагогам и обучаемым, основанные на опыте, компетенциях и профессиональных целях последнего.

Современные методы машинного обучения позволяют автоматизировать проверку определенных типов заданий, формировать персонализированные ответы, разрабатывать адаптивные траектории обучения и анализировать обширные массивы данных об обучении [9, 10]. Первые интеллектуальные системы обучения, такие как SCHOLAR [11], продемонстрировали возможность организации индивидуализированного взаимодействия со студентами, в то время как последующие решения, включая AutoTutor [12] и WatsonTutor [13], разработали подходы к адаптивной поддержке студентов. Отдельное направление составляют системы автоматизированного оценивания и поддержки преподавателей, такие как OpenEssayist [14], а также решения в области образовательной аналитики и прогнозирования результатов обучения, в том числе CourseSignals [15] и система прогнозирования академической успеваемости «Пифия» [16].

Однако большинство существующих решений ориентировано на решение конкретных задач и редко объединяет в рамках единой цифровой образовательной платформы функции управления обучением, образовательной аналитики, интеллектуальной поддержки и автоматизации работы преподавателей. К числу дополнительных вызовов относятся зависимость от зарубежных программных решений и связанные с этим риски для технологического суверенитета и информационной безопасности образовательных организаций.

МЕТОДЫ И МАТЕРИАЛЫ

Все вышесказанное определяет необходимость следующего шага в развитии комплексных образовательных платформ с интегрированными инструментами искусственного интеллекта (ИИ) в саму платформу и непосредственно в учебный процесс. Авторское понятие LXP+ (Learning Experience Platform Plus) описывает такие новационные цифровые решения, как расширение базовой мо-

дели LXP за счет глубокой интеграции элементов ИИ, расширения спектра поддерживаемых образовательных межпредметных сценариев и включения самой платформы в образовательную среду – экосистему непрерывного обучения и деятельности студента и школьника. В таком понимании LXP+ выходит за рамки «классической» LXP, но наследует ее ключевые черты: персонализированные рекомендации, учет интересов и целей обучающегося, интеграцию разнообразных форм контента и поддержку сетевых взаимодействий. При этом сама LXP+ включает более глубокую аналитику, использование генеративных и когнитивных ИИ-моделей, поддержку сложных, проектно-ориентированных и междисциплинарных форматов обучения. Важным качеством LXP+ для профессионального роста обучаемого является то, что платформа находится в цифровом пространстве, где одновременно присутствуют организации-работодатели и образовательные учреждения, совместно формируя и обновляя материалы и практики под актуальные требования рынка. LXP+ связана с системами управления знаниями и производственными процессами компаний и может использовать реальные документы, регламенты и задачи как учебный контент, динамически изменяя компетентностную модель выпускника.

Подход LXP+ совпадает с тем, как современные исследования понимают цифровую образовательную среду – широкую цифровую экосистему из разных средств и сервисов, которая поддерживает гибкие формы обучения, персонализацию и активную роль обучающегося, особенно в условиях использования ИИ. В этой экосистеме LXP+ можно рассматривать как дидактический инструмент, который переводит обучение от модели, где все строится вокруг курсов, к модели, где в центре находится индивидуальный цифровой образовательный опыт слушателя, а курсы и ресурсы подстраиваются под этот опыт и помогают в его развитии, т. е. LXP+ можно интерпретировать как дидактический инструмент перехода от «курсоцентричной» модели к модели, в которой центральным объектом является цифровой образовательный опыт обучающегося.

LXP+, являясь компонентом цифровой образовательной среды по своему устройству, относится к классу полидисциплинарных интеллектуальных информационных систем (Multidisciplinary Intelligent Information System, MIIS), поскольку объединяет подходы из образования, управления знаниями и информационных технологий и опирается на углубленное использование технологий ИИ

для анализа данных и адаптивного управления образовательным опытом. В качестве рабочей модели MIIS нами предложена Цифровая Образовательная Платформа Мирера, обеспечивающая, в том числе, персонализированную дистрибуцию контента, семантическую интеграцию учебных ресурсов, профессиональных практик и социальных взаимодействий на основе углубленного использования ИИ (нейросемантические связи, генеративный дизайн кейсов, агентное моделирование и предиктивный анализ), что трансформирует и интегрирует учебные сервисы в среду конструирования образовательного опыта, где ИИ выступает фасилитатором субъектности обучающегося, что создает условия для персонализации, полисубъектности и практико-ориентированности в ИИ-усиленной инфраструктуре.

Платформа Мирера выступает в качестве инструмента поддержки преподавателей, в том числе обеспечивая автоматизацию различных рутинных процессов посредством комплексного применения технологий ИИ. При этом полидисциплинарность платформы проявляется в автоматизированной проверке максимального объема заданий по широкому спектру предметных областей, включая программирование, математику, а также финансовые, экономические и социально-гуманитарные предметы.

Высокая насыщенность практико-ориентированными заданиями в сочетании с дидактически выверенной организацией взаимодействия в системе «обучающийся – преподаватель» влияет на эффективность усвоения учебного материала. ЦОП Мирера поддерживает единство очного, дистанционного и гибридного форматов посредством встроенных средств коммуникации, инструментов онлайн коллаборации при решении задач как в формате учитель – ученик, так и учитель – группа, реализуя совместную работу над проектом. Накопленные данные, используемые в дальнейшем в образовательной аналитике, имеют универсальный формат, что архитектурно упрощает наращивание модулей обработки и прогнозирования, включая методы ИИ, которые используются не только для углубленного анализа образовательного процесса, но и для интеллектуальной поддержки учебной деятельности. Полный цифровой след поведения студентов и преподавателей, сохраняемый на платформе, служит основой для прогнозирования академического статуса обучающихся и формирования индивидуальных адаптивных образовательных траекторий.

В реализации модели ЦОП Мирера использовано программное обеспечение с открытым исходным кодом (Open Source), поддерживаемое в Российской Федерации, что минимизирует риски потери интеллектуальной собственности и блокировки программного обеспечения, необходимого для образовательной деятельности российских вузов и преподавателей.

ЦОП Мирера, позиционируемая как переходная модель от курсоцентричной к опытно-центричной организации обучения, обеспечивает автоматизацию рутинной нагрузки преподавателя. В частности, автоматизированы запросы деканата об успеваемости, посещаемости и текущей загрузке преподавателя (контрольные мероприятия, часы занятий, участие в экзаменах и зачетах). Специализированный доступ к системе позволяет формировать и поддерживать актуальность в реальном времени академической отчетности по учебным группам и курсам и т. п. Архитектура платформы учитывает специфику российской образовательной системы: поддерживаются понятия курсов и учебных групп с возможностью независимой настройки расписания для каждого экземпляра курса в разных группах, а также параллельного прохождения одной группой нескольких курсов, что соответствует традиционной модели академических групп, потоков и параллелей, см. рис. 1. Платформа Мирера включает встроенный календарь с возможностью детального расписания занятий, см. рис. 2. Расписание привязано к конкретным темам курса, что позволяет быстро переносить расписание на последующие учебные периоды, например, при копировании курса из года в год. Помимо группового расписания предусмотрена возможность индивидуальной настройки расписания для каждого студента, что актуально в случаях пропуска тем, отставания от группы или закрытия академической разницы.

Таким образом, Мирера выступает гибридным решением, в котором автоматизация управленческих процессов LMS сочетается с ориентацией на конструирование цифрового образовательного опыта, как центрального объекта проектирования.

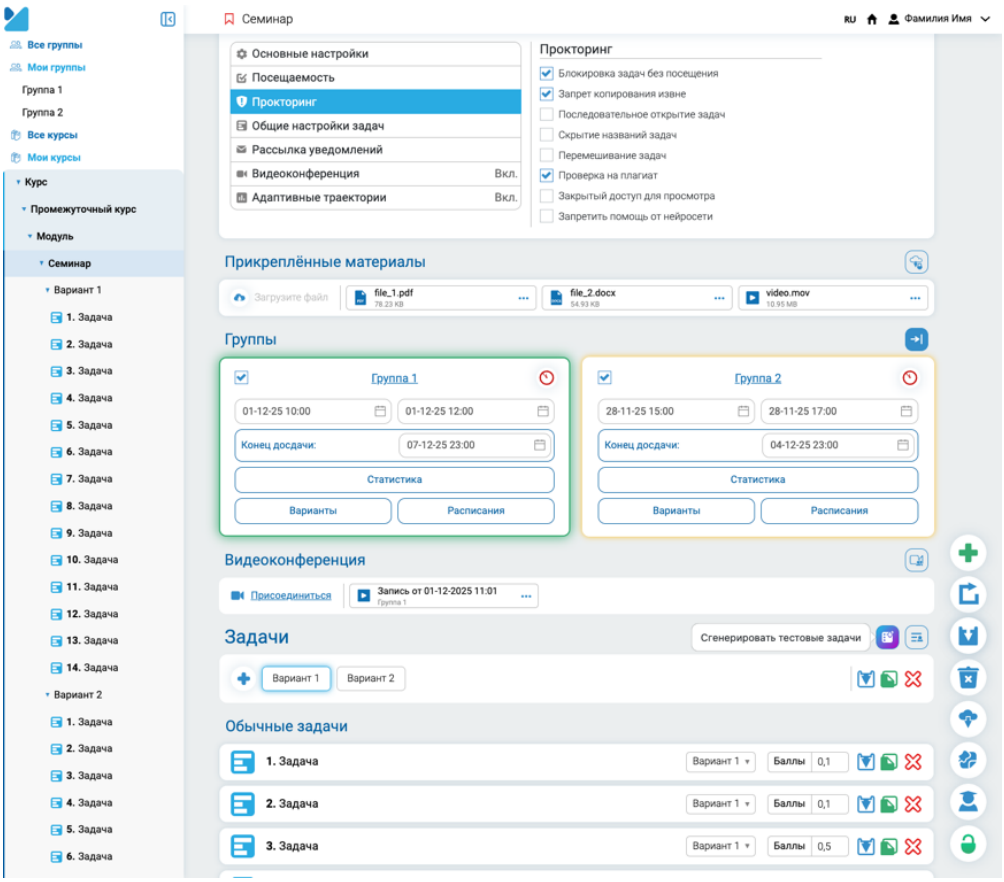


Рис. 1. Страница занятия (контеста) у преподавателя, настройка расписания независимо для всех групп.

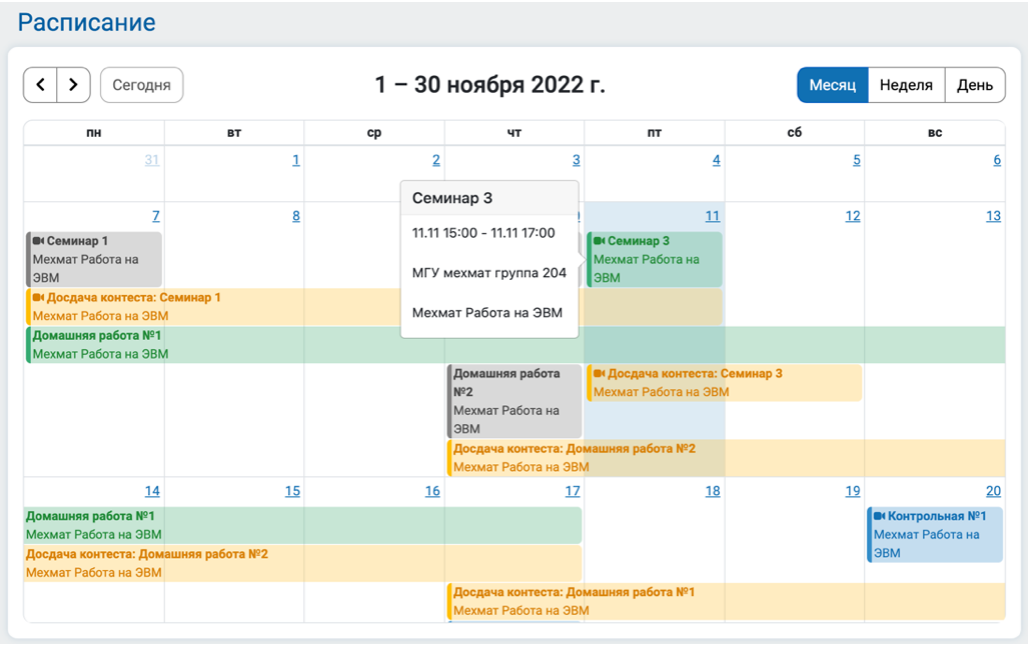


Рис. 2. Календарь со всеми занятиями по всем курсам.

Кроме основного расписания, для каждой группы и каждого студента независимо предусмотрена возможность настройки сроков досдачи учебных материалов, при этом автоматически активируется понижение баллов за задания, которые сдаются после основного дедлайна. Механизм понижения баллов может осуществляться как посредством фиксированного коэффициента, так и в виде постепенного уменьшения баллов во времени. При изменении настроек расписания все баллы, привязанные к просроченным дедлайнам, пересчитываются автоматически. Преподаватель в любой момент может вручную скорректировать данный процесс и отменить понижение баллов для конкретного студента в исключительных случаях. Реализована также возможность выдачи контестов с ограниченным временем выполнения, при этом студент имеет свободу выбора момента начала решения, что повышает гибкость организации контроля знаний.

Для обеспечения коллективной работы преподавателей над одним учебным курсом в платформе реализован набор инструментов, позволяющих гибко управлять правами доступа. Преподавателям могут предоставляться различные уровни прав доступа – от приглашенного лектора с ограниченными возможностями редактирования учебного контента до редактора курса с полными правами. Кроме того, реализован аппарат «наследования» курсов в виде передачи курса от одного преподавателя другому для проведения занятий с ограничением на редактирование материалов, а также широкий спектр иных настроек прав доступа, адаптируемых под разнообразные организационные и методические сценарии.

Благодаря модульной структуре организации курсов и описанной системе тонкой настройки прав доступа обеспечивается эффективная и оперативная совместная разработка учебных программ по смежным дисциплинам. Это достигается посредством практически мгновенной интеграции уже существующих модулей, созданных для отдельных предметов, в новые междисциплинарные и полидисциплинарные курсы.

На платформе Мирера обеспечено хранение всего цифрового следа образовательной деятельности преподавателей и студентов на территории Российской Федерации, что гарантирует соответствие требованиям национального законодательства по вопросам безопасности и конфиденциальности данных. В состав платформы включены специализированные модули, предназначенные для

выполнения типичных аналитических запросов, связанных с обработкой и интерпретацией накопленной информации.

Хранение данных организовано с обеспечением быстрого и удобного доступа к разнообразным материалам всем участникам образовательного процесса: учебным материалам курсов, а также к сведениям о всех успешных и неуспешных попытках решения студентами заданий, реализуемых в рамках курса, данным о посещаемости, записям видеоконференций, обучающим материалам и видеолекциям и другим релевантным данным. На основе этих данных формируются комплексные аналитические отчеты, позволяющие оценить, насколько эффективно студенты усваивают материал курса, что, в свою очередь, служит основой для выработки рекомендаций по корректировке и адаптации индивидуальных траекторий обучения.

ЦОП Мирера обеспечивает автоматический подбор заданий на основе анализа результатов студентов, создавая тем самым адаптивные траектории обучения. Система может автоматически назначать студентам простые задания для повторения и преодоления трудностей, а также более сложные задания, отмеченные «звездочкой», для студентов, демонстрирующих высокие достижения. Аналогично на зачетах предусмотрена автоматическая выдача заданий по темам, которые студент не освоил полностью.

Платформа поддерживает автоматическую генерацию уникальных вариантов заданий для каждого студента, в том числе с учетом тематического распределения. Преподаватель задает перечень тем, указывает количество задач по каждой теме и баллы, которые может получить студент за каждую задачу по этой теме. Студент получает индивидуальный вариант, состоящий из выбранных задач по каждой теме в соответствии с заданными параметрами.

Следуя модели МИИС, в платформе реализованы нейросетевые ассистенты обучающихся и преподавателей, работающие в контролируемом учебном контексте и автоматически использующие данные образовательной среды. Ассистент обучающегося предоставляет методически корректные подсказки без выдачи готовых решений [17], а ассистент преподавателя автоматизирует создание заданий и комментариев на основе учебных материалов, что снижает риск галлюцинаций модели и поддерживает принципы академической честности.

Платформа Мирера дает возможность педагогу проводить мультиформатные интерактивные занятия с большим количеством интегрированных практических заданий, которые проверяются полностью автоматически. Реализованы инструменты для коллаборации в реальном времени преподавателя со студентами. Механизмы визуализации и контроля прогресса студентов позволяют преподавателю отслеживать весь процесс выполнения задания, включая отображение конкретного места в решении, которое в текущий момент времени редактирует студент. Все эти механизмы реализованы по технологии WebSocket, без применения видеонаблюдения или средств прокторинга. Преподаватель имеет возможность оперативно оставлять комментарии, мгновенно отображающиеся студенту, а при необходимости непосредственно подключаться к выполнению задания, при этом у обучающегося блокируется доступ к рабочей области, и он в реальном времени наблюдает действия преподавателя. Технология WebSocket использована в ЦОП Мирера для отображения в реальном времени результатов студентов в статистике, а также других аспектов взаимодействия преподавателя и студентов, требующих актуальной информации у всех участников образовательного процесса без необходимости обновления страницы.

В платформу Мирера интегрировано модифицированное open-source-решение – система видеоконференций BigBlueButton с возможностью записи лекций и их последующей публикацией. Из сервиса видеоконференций в Мireру экспортируются данные об индивидуальной активности участников видеоконференций, что позволяет использовать эту информацию, в частности, для учета посещаемости.

Автоматизация учета посещаемости в условиях смешанного обучения реализована на основе комплекса методов для очного, дистанционного и гибридного форматов. Для очных занятий использованы геолокация участников и динамические QR-коды с JWT-токенами, а для дистанционного обучения – мониторинг активности в видеоконференциях, просмотрах записей и работе с учебными материалами с начислением баллов пропорционально фактическому времени участия [18].

В статистике отображаются набранные баллы, журнал посещаемости, активность студентов и пр., см. рис. 3.

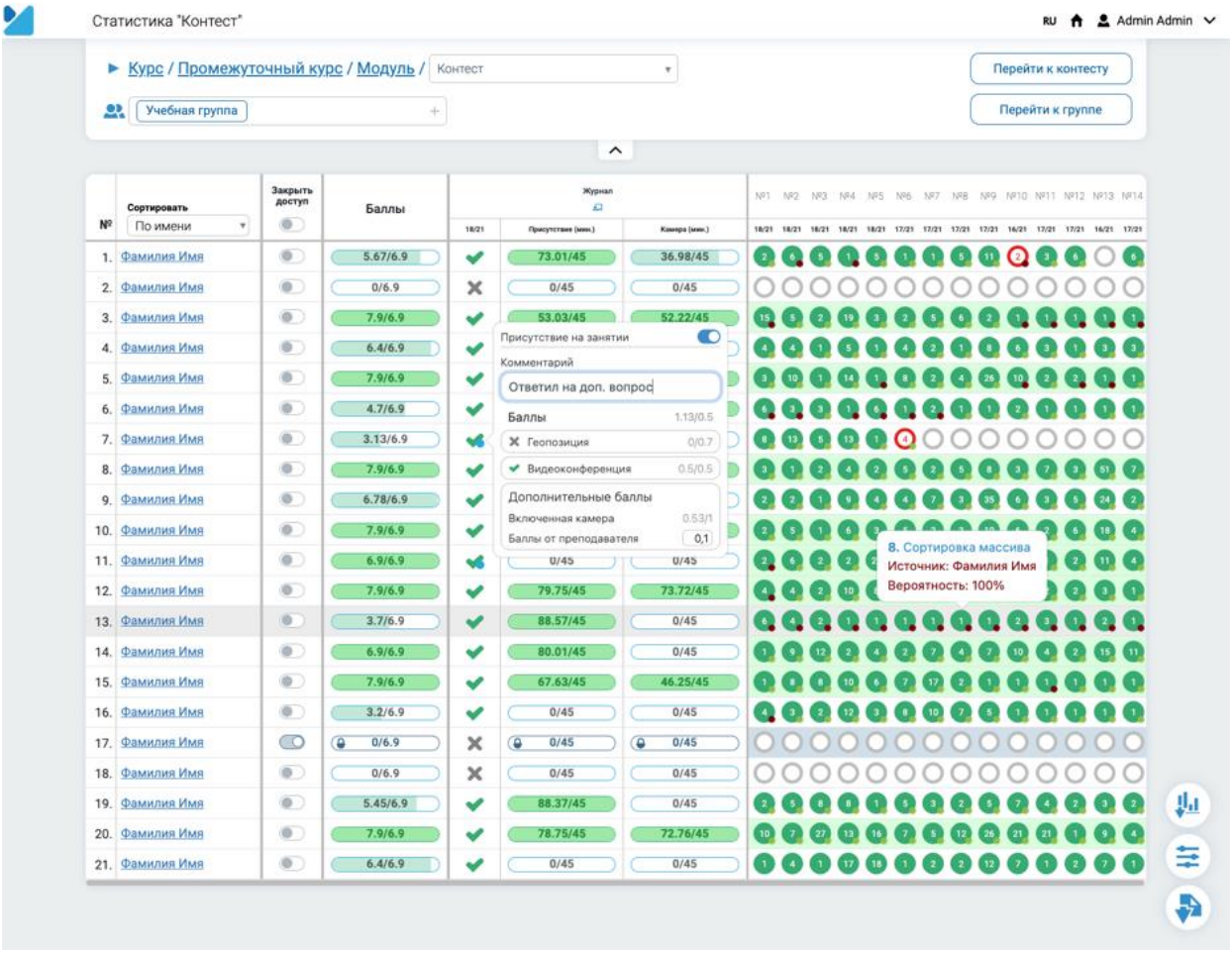


Рис. 3. Интерфейс отображения статистики по контесту.

Цифровая образовательная платформа Мирера спроектирована как многоканальное и доступное с широкого спектра устройств цифровое образовательное пространство, где реализованы различные инструменты коммуникации, совместной работы и доступа к учебным материалам всех участников образовательного процесса. Пользователи платформы могут входить в систему с персональных компьютеров, ноутбуков, планшетов и смартфонов, см. рис. 4 и 5, независимо от их текущего местоположения – будь то учебная аудитория, дом или транспорт.

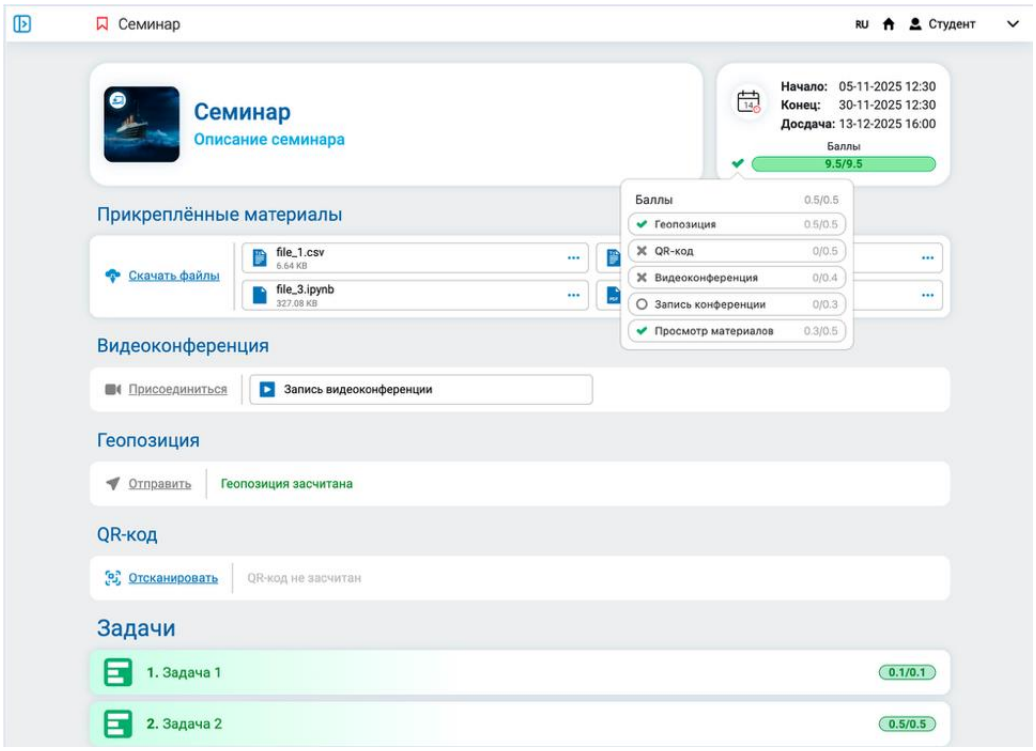


Рис. 4. Страница занятия (контеста) у студента на компьютере.

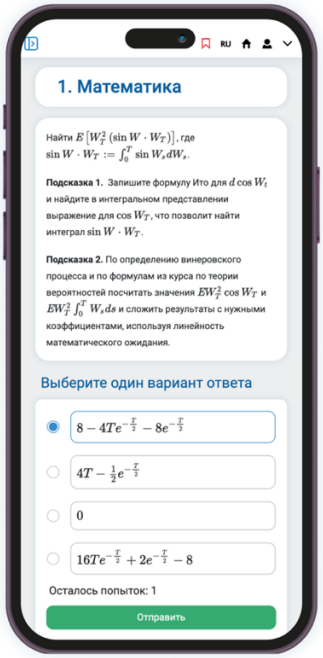


Рис. 5. Страница задачи у студента на мобильном устройстве.

Все ресурсы курсов – лекционные материалы, практические задания, записи прошедших занятий и иные обучающие компоненты – могут быть систематизированы по различным структурным единицам платформы и календарному

плану, что позволяет студентам оперативно находить и использовать необходимые материалы. Интегрированный календарь с функцией напоминаний о предстоящих мероприятиях, выполненный в интуитивно понятном интерфейсе, способствует организации образовательного процесса и формированию устойчивых навыков самостоятельного планирования учебной деятельности.

Особенностью платформы, реализующий принцип LXP+BYOD-learning (Bring Your Own Device), является полное программно-техническое обеспечение образовательного процесса. Так, студенты могут участвовать в видеоконференциях, проходить тестирования, решать задания по программированию и выполнять другие типы формализованных учебных заданий непосредственно в рамках единой образовательной платформы, в том числе интегрирующей различные цифровые образовательные среды. Это позволяет избежать необходимости установки сторонних программ, настройки компиляторов или отдельных сред разработки для каждой дисциплины, тем самым снижая технический барьер и обеспечивая единообразие пользовательского опыта в процессе освоения различных курсов.

В цифровой образовательной платформе Мирера реализован специализированный информационный чат-бот, интегрированный с мессенджером ВКонтакте (VK) и почтовым сервисом. Пользователь платформы может связать свой аккаунт с VK, активировав тем самым получение уведомлений от имени ботов в привычной для себя цифровой социальной среде.

Помимо коммуникационных функций задачами чат-бота стали, в том числе, оперативное предоставление данных о текущей статистике результатов студентов, сроках проведения занятий и домашних заданий, результатах проверки и другой учебной информации. Уведомления, генерируемые интеллектуальными ботами платформы, содержат информацию об учебных и иных событиях и адресованы как студентам, так и преподавателям. В рассылках используются механизмы контекстно-чувствительных триггеров, которые позволяет инициировать уведомления из различных сегментов платформы. Так, со страницы учебной группы может быть отправлена рассылка, адресованная всем ее участникам, а со страницы курса, в свою очередь, могут отправляться уведомления во все группы, прикрепленные к данному курсу. Практические сценарии использо-

вания системы охватывают широкий спектр образовательных и административных событий. Так, запрос на вступление в группу, отправленный студентом, инициирует передачу преподавателю сообщения, содержащего интерфейс для моментального принятия или отклонения заявки. Другим примером является использование встроенной клавиатуры бота с доступом к часто задаваемым вопросам, таким как статистика результатов и вопросы о ближайших занятиях.

Контест в ЦОП Мирера – это дидактическая единица организации образовательного процесса, задающая временные границы, формат (лекция, семинар, практическое занятие, домашняя работа) и условия проведения учебного занятия, а также выступающая узлом генерации автоматизированных коммуникаций, доставки учебных материалов и сбора данных об активности. Для обучающихся аппарат контестов обеспечивает управление временными рамками (уведомления о старте/завершении), доставку материалов в логике «перевернутого класса», персонализированные оповещения для студентов и регулярную агрегированную статистику для преподавателей (активность и вовлеченность в больших группах). Временные характеристики реакции обучаемого на полученные от системы уведомления, активность при взаимодействии с интерфейсом бота, скорость обратной связи и многое другое (цифровой след) служат исходными данными для аналитических моделей и, в частности, используются для индицирования вовлеченности студентов в учебный процесс.

Цифровая образовательная платформа Мирера содержит интегрированную среду разработки программ, поддерживающую полноценный многокомпонентный (мультифайловый) режим, позволяющий работать с совокупностью взаимосвязанных файлов в рамках единого проекта. Преподаватель может задавать участки кода, недоступные студентам для редактирования, а также скрытые файлы, предназначенные, например, для автоматической проверки решений или инкапсуляции вспомогательных функций, см. рис. 6.

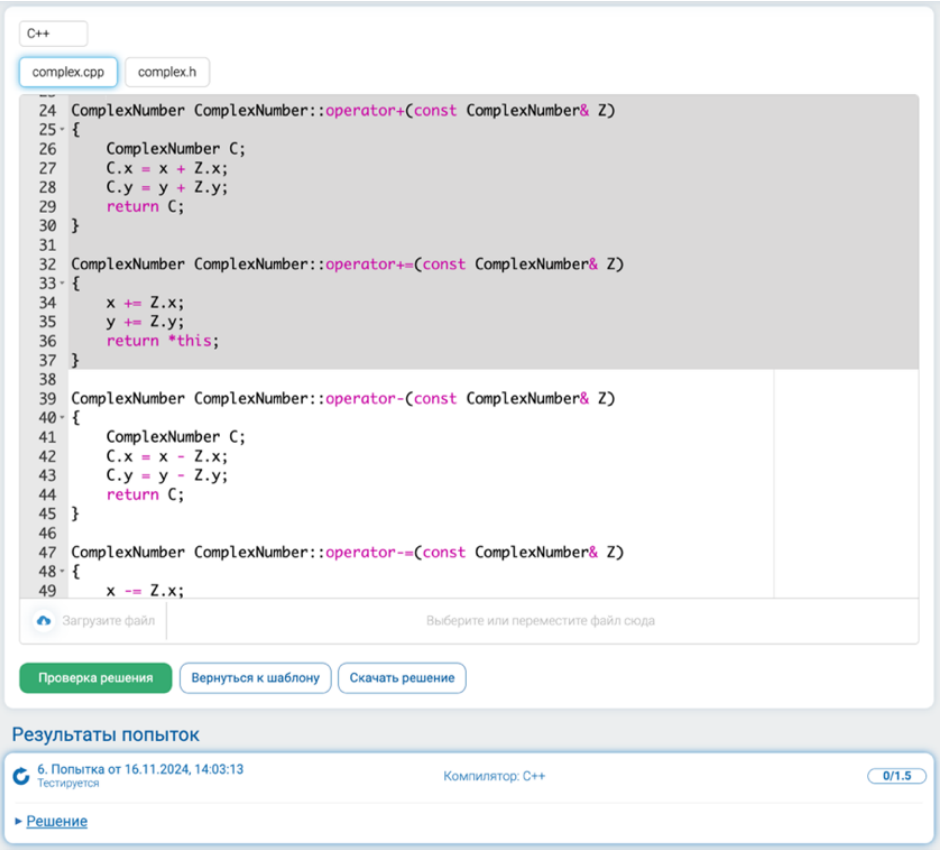


Рис. 6. Встроенный редактор кода с поддержкой многофайлового режима и мгновенной обратной связью.

Платформа Мирера предоставляет преподавателю гибкость при настройке задач по программированию. Преподаватель выбирает компиляторы для себя и студентов, указывает входные и выходные данные, способ проверки, ограничения по времени и памяти с помощью простого и понятного интерфейса, см. рис. 7. Система позволяет выбрать один из множества способов сравнения результатов, включая, например, методы визуального сопоставления изображений, что расширяет спектр возможных формулировок задач.

В частности, платформа Мирера позволяет автоматизировать процесс формирования выходных данных тестов на основе эталонного решения, предварительно заданного преподавателем. В процессе генерации входные данные подаются на вход эталонной программе, после чего результаты ее выполнения автоматически фиксируются как эталонные выходные данные. Эти данные впоследствии используются для проверки решений, предоставленных студентами.

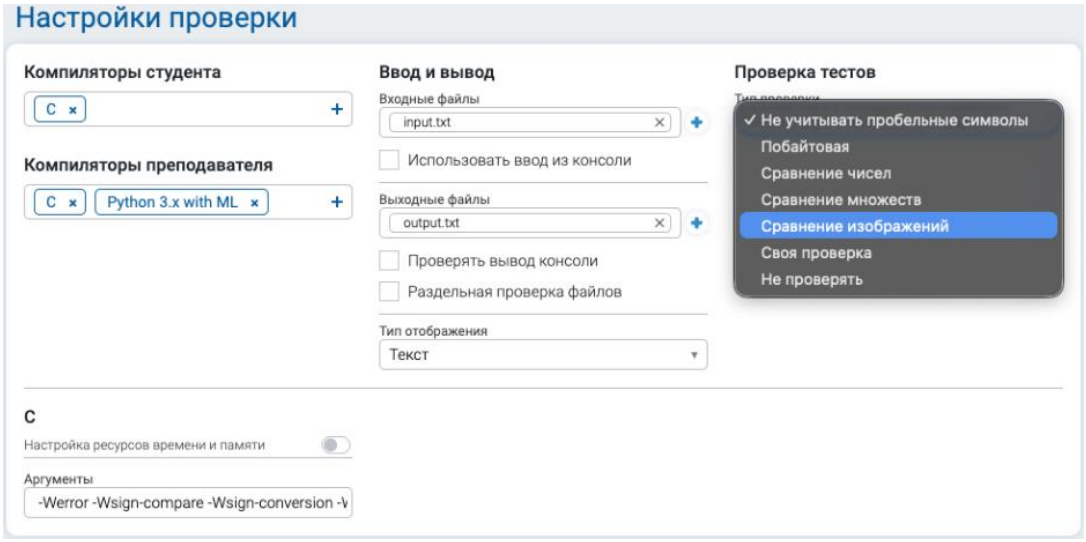


Рис. 7. Настройки проверки задачи по программированию.

Платформа поддерживает возможность автоматизированной генерации входных данных тестов с использованием авторского программного кода, написанного преподавателем. Так, для заданий на декодирование строк входные данные могут формироваться с помощью программы, выполняющей обратную операцию – кодирование строки. Или в случае задач, связанных с выборками из баз данных, возможно автоматическое создание необходимых таблиц посредством специально написанного кода. В платформе Мирера реализован механизм автоматизированной проверки корректности, позволяющий инициировать проверку всех шаблонов и выходных данных тестов всех задач контеста. Функционал производит автоматическую загрузку эталонных решений в соответствующие шаблоны заданий и сверку выходных данных с результатами выполнения этих решений. Совместное использование автоматической генерации входных и выходных данных предоставляет педагогу возможность быстрого масштабируемого создания учебных заданий, см. рис. 8.

Цифровая образовательная платформа Мирера поддерживает широкий мультипарадигмальный набор встроенных компиляторов и языков программирования со встроенной автоматической проверкой. В их число входят Python, C/C++, R, Pascal, Java, QT с поддержкой OpenGL, MPI, eGrep, Flex/Bison, NASM/ARM, SQLite, Prolog, LISP и другие. Кроме того, в платформу интегрирован язык объектного моделирования UML, предназначенный для построения диаграмм бизнес-процессов, календарных планов и иных структур с последующей

автоматизированной верификацией построенных моделей и графов. Архитектура платформы позволяет оперативно расширять данный перечень за счет подключения новых языков или версий, существующих посредством сборки и внедрения дополнительных Docker-контейнеров.

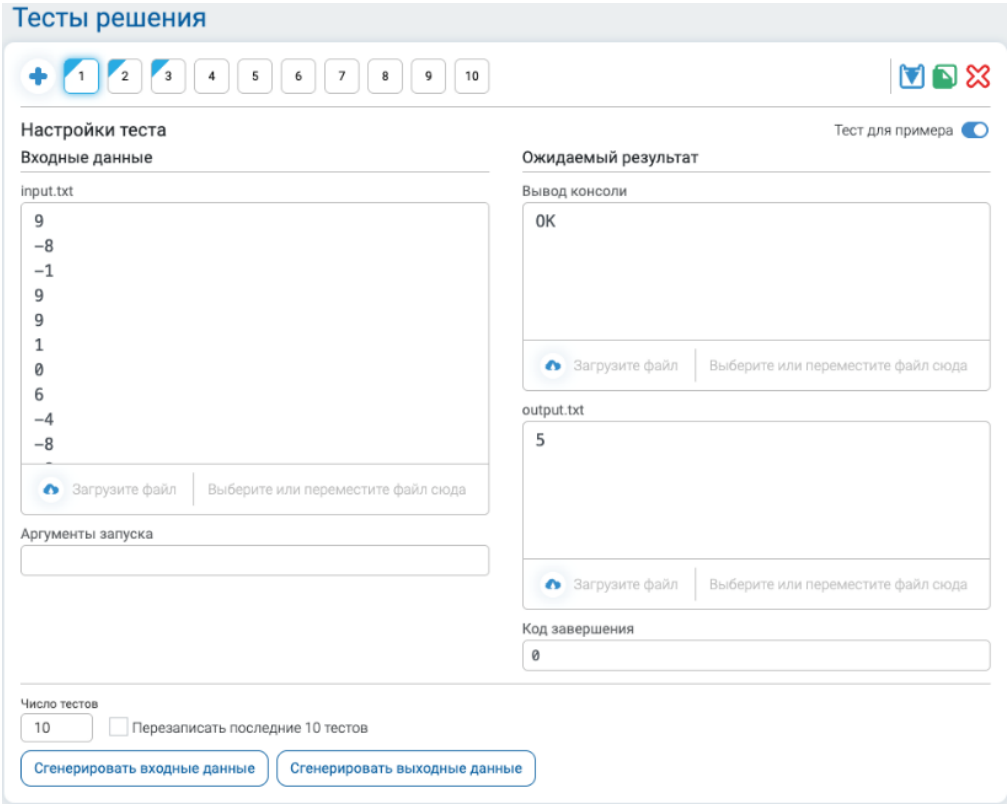


Рис. 8. Настройки тестов в задаче по программированию с возможностью автоматической генерации входных и выходных данных.

Для обеспечения академической честности разработан модуль антиплагиата программного кода, основанный на сопоставлении синтаксически нормализованных фрагментов и выявлении структурных заимствований, устойчивый к переименованию переменных и другим поверхностным модификациям. см. рис. 9. Интеграция модуля с системой мониторинга учебной активности позволяет выявлять как прямые заимствования, так и косвенные признаки академических нарушений на основе анализа поведения обучающихся [19].

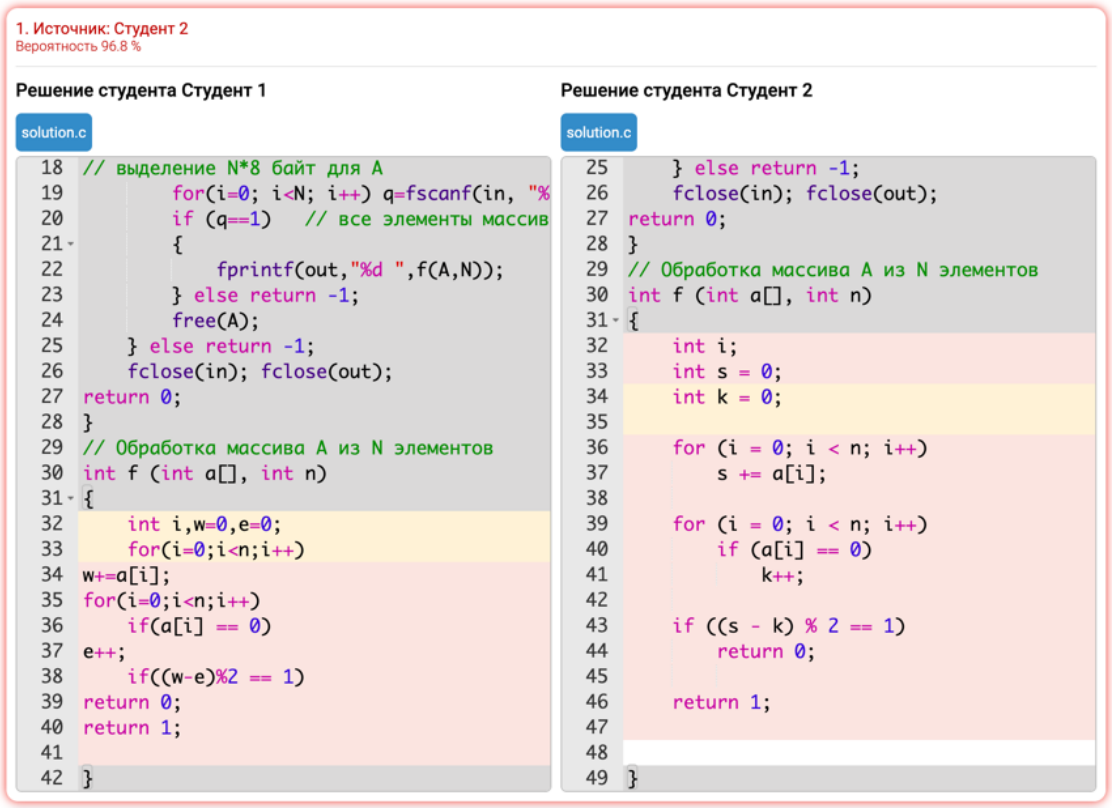


Рис. 9. Пример работы системы плагиата.

Цифровая образовательная платформа Мирера включает разнообразные типы тестовых заданий, которые охватывают как задачи с математическим содержанием, так и задания с открытыми текстовыми ответами. Среди доступных форматов представлены тесты с выбором одного или нескольких вариантов ответа из предложенного списка с возможностью настройки допустимого числа ошибок. Важной особенностью является возможность использования в условиях задач и вариантах ответов не только текстового формата или изображений, но и формул LaTeX – для написания формулы достаточно окружить содержащий ее кусок текста знаками «$», см. рис. 10. Это значительно расширяет возможности составления и проверки разнообразных заданий по математике и другим дисциплинам STEM, см. рис. 11.

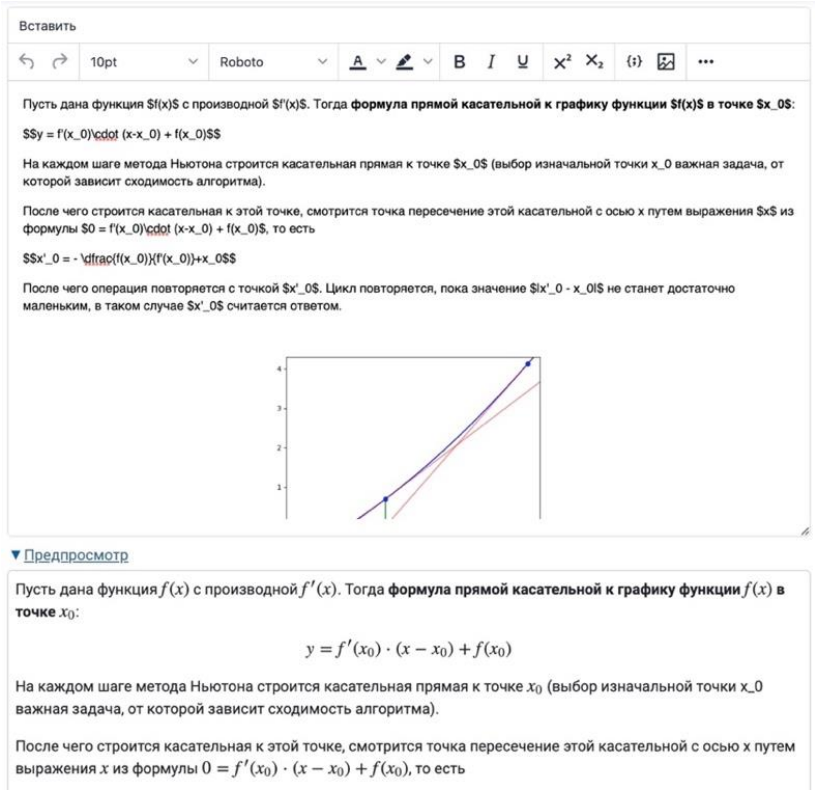


Рис. 10. Встроенный редактор для создания математического контента с возможностью написания формул LaTex.

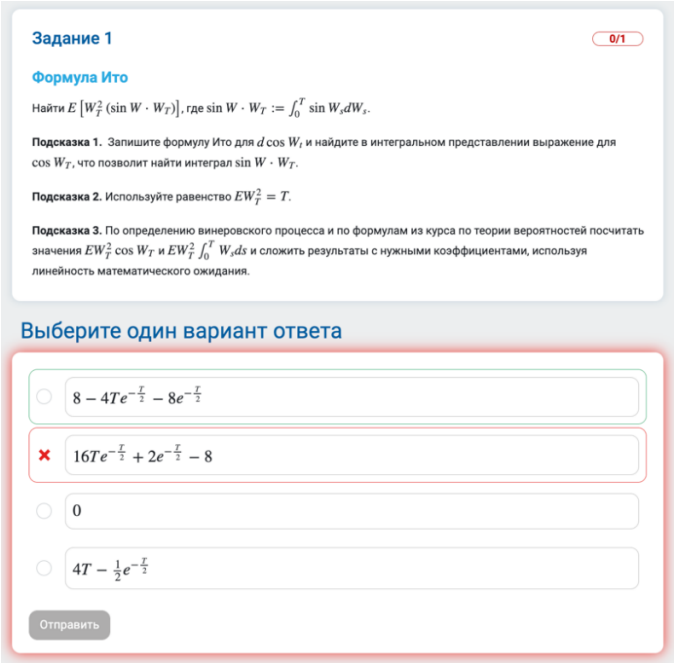


Рис. 11. Пример математической задачи со встроенными LaTex формулами в условии и вариантах ответа.

Формат тестовых заданий на установление правильной последовательности элементов, также реализованный на платформе Мирера, благодаря возможности использования формул LaTex может быть использован для первичной проверки знаний студентов о структуре доказательств теорем или алгоритмов решения типовых задач, где доказательство или решение разбиваются на отдельные шаги, к которым добавляются лишние элементы, и обучаемый должен выбрать корректный порядок этапов из предложенного множества, см. рис. 12. Благодаря включению дополнительных элементов, схожих с корректными, обучаемому необходимо не только уметь устанавливать последовательность логических шагов доказательства теоремы или решения задачи, но и глубоко понимать содержание каждого этапа, включая знание правильных формулировок и соответствующих математических выражений для каждого шага решения или обоснования.

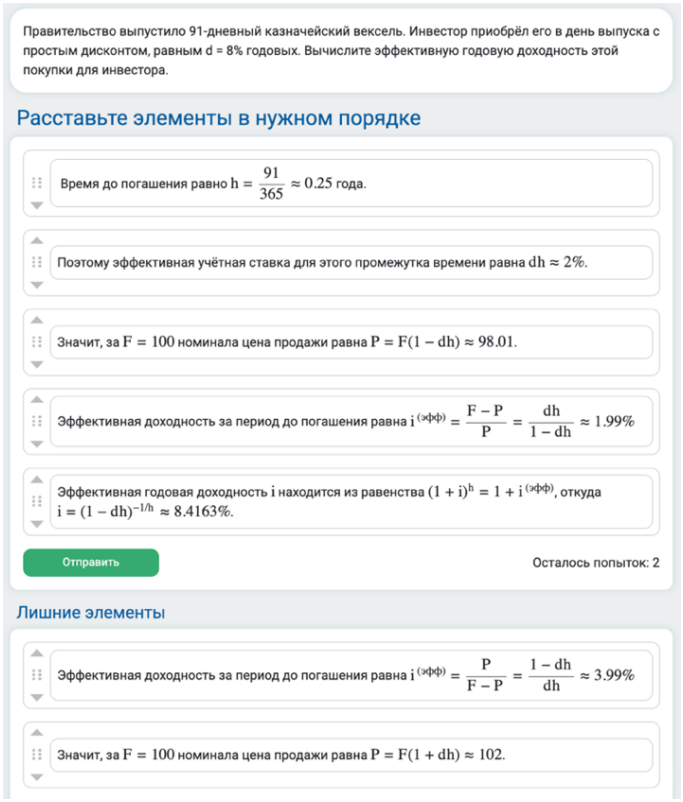


Рис. 12. Пример проверки решения математической задачи через установление правильной последовательности и исключение лишних элементов.

На платформе реализована возможность проверки задач на установление связей между элементами, в том числе сопоставления нескольких элементов одному объекту. При этом как в левой, так и в правой колонках могут находиться

текстовые фрагменты, изображения или формулы LaTeX, что расширяет диапазон проверяемых задач по математике и другим STEM дисциплинам.

Для проверки заданий с открытым вводом ответов Мирера предлагает разнообразные инструменты, применимые как для простых, так и для сложных заданий. Для простых заданий с односложными текстовыми ответами предусмотрена возможность проверки ответов по наличию заданных словосочетаний с учетом регистра и пробелов, числовых значений с заданной погрешностью, изображений и других критериев. Следует отметить, что данный подход не обеспечивает полноценной смысловой оценки больших текстов произвольного содержания. Для решения этой задачи реализован механизм проверки текстов студентов на наличие и семантическое совпадение опорных предложений, заданных преподавателем [20].

РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТАЛЬНЫХ ИССЛЕДОВАНИЙ

Апробация цифровой образовательной платформы Мирера в условиях реального учебного процесса подтвердила эффективность предложенного подхода к организации и сопровождению обучения. Интеграция средств автоматизации, образовательной аналитики и интеллектуальной поддержки в единую цифровую среду позволяет существенно повысить эффективность управления образовательным процессом как для преподавателей, так и для обучающихся.

На платформе зарегистрировано более 30000 пользователей, из них более 250 преподавателей. Создано более 300 цифровых курсов, 8000 контестов и 110000 задач, из них более 60000 по программированию. Автоматически проверено более 2000000 решений студентов, из них более 1200000 по программированию. Более 20000 раз студенты обращались за контекстной помощью к нейропомощнику.

Использование платформы способствовало снижению трудозатрат преподавателей на выполнение рутинных операций, связанных с контролем посещаемости, проверкой заданий, подготовкой учебных материалов и мониторингом образовательных результатов. Освобождение преподавателей от значительной части технической и организационной нагрузки позволило им уделять больше внимания содержательной работе со студентами, совершенствованию учебных материалов и оказанию индивидуальной поддержки обучающихся.

Для студентов внедрение платформы обеспечило оперативную обратную связь, улучшило доступ к академической поддержке и расширило возможности для самостоятельного обучения. Наличие встроенных интеллектуальных инструментов способствовало более активному вовлечению в образовательный процесс, повышению активности при выполнении учебных заданий и более стабильному достижению результатов обучения.

Одним из ключевых результатов внедрения стало создание единого цифрового образовательного пространства, объединяющего данные по различным аспектам образовательной деятельности. Это позволило перейти от анализа отдельных показателей к комплексному мониторингу образовательного процесса, выявлению закономерностей в поведении обучающихся и принятию более обоснованных педагогических решений на основе данных.

Полученные результаты свидетельствуют о том, что применение цифровой образовательной платформы нового поколения, объединяющей инструменты автоматизации, искусственного интеллекта и образовательной аналитики, позволяет повысить качество организации обучения, эффективность использования ресурсов образовательной организации и уровень персонализации образовательного процесса. Проведенная апробация подтверждает перспективность дальнейшего развития подобных платформ как основы цифровой трансформации высшего образования.

ЗАКЛЮЧЕНИЕ

Цифровая образовательная платформа Мирера представляет собой модель МИИС – реализации комплексного подхода к цифровой трансформации высшего образования, при котором технологии искусственного интеллекта и автоматизации не заменяют преподавателя, а органично встраиваются в образовательный процесс, беря на себя рутинные задачи и высвобождая ресурсы для содержательного педагогического взаимодействия. Результаты педагогического эксперимента показали, что интеграция в единую цифровую среду инструментов автоматизированной проверки заданий, образовательной аналитики, учета посещаемости, адаптивного обучения и интеллектуальной поддержки позволяет повысить эффективность управления образовательным процессом, уровень вовлеченности обучающихся и качество принимаемых педагогических решений,

что подтверждает перспективность дальнейшего развития и внедрения подобных платформ в качестве одного из ключевых направлений цифровой трансформации системы высшего образования.

Благодарности

Работа выполнена в рамках темы государственного задания НИЦ «Курчатовский институт» – НИИСИ по теме № FNEF-2024-0001 (1023032100070-3-1.2.1).

СПИСОК ЛИТЕРАТУРЫ

1. *Леонов А.Г., Эпиктетов М.Г.* Практикум Е-87 // Аннотированный библиографический указатель «Депонированные научные работы» ВИНИТИ РАН. 1988. № 4279-В88.

2. *Бесшапошников Н.О., Кушниренко А.Г., Леонов А.Г., Райко М.В., Собакинских О.В.* Цифровая образовательная среда ПиктоМир: опыт разработки и массового внедрения годового курса программирования для дошкольников // Информатика и образование. 2020. № 10. С. 28–40. https://doi.org/10.32517/0234-0453-2020-35-10-28-40

3. *Смолянинова О.Г., Иванов Н.А.* Обзор практик обеспечения электронной поддержки образовательного процесса средствами LMS Moodle: опыт российских вузов // Азимут научных исследований: педагогика и психология. 2019. Т. 8. № 2(27). С. 228–232. https://doi.org/10.26140/anip-2019-0802-0054

4. *Тканов В.* От первых курсов по выбору к массовой индивидуализации: как платформа Modeus помогает российским вузам перейти к новой образовательной модели // Honors College Conference: Материалы II Международной конференции, Красноярск, 27–28 ноября 2020 года. Красноярск: Сибирский федеральный университет, 2020, С. 30–34.

5. *Горчаков Л.В., Стась А.Н., Карташов Д.В.* Обучение программированию с использованием системы EJUDGE // Вестник Томского государственного педагогического университета. 2017. № 9(186). С. 109–112. https://doi.org/10.23951/1609-624X-2017-9-109-112

6. *Ларина Ю.А., Лавровская О.Б.* Особенности использования онлайн-платформы Яндекс.контест для проведения занятий по программированию //

Теоретические и прикладные аспекты развития современной науки и образования: Материалы IV Всероссийской научно-практической конференции, Чебоксары, 26 февраля 2021 года. Чебоксары: Негосударственное образовательное частное учреждение дополнительного профессионального образования «Экспертно-методический центр», 2021. С. 102–105.

7. *Лакотко Э.* COURSERA как альтернативное высшее образование // На путях к новой школе. 2016. № 2. С. 49–54.

8. *Shamsudin N. M., Teoh S.H.* Exploring the Synergy of Learning Experience Platforms (LXP) with Artificial Intelligence for Enhanced Educational Outcomes // Proceedings of the International Conference on Innovation & Entrepreneurship in Computing, Engineering & Science Education (InvENT 2024), Shah Alam, Malaysia, 20–22 August 2024. P. 30–40. https://doi.org/10.2991/978-94-6463-589-8_5

9. *Григорьев С.Г.* Искусственный интеллект в образовании // Информатизация образования и методика электронного обучения: цифровые технологии в образовании: Материалы VII Международной научной конференции, Красноярск, 19–22 сентября 2023 года. Красноярск: Красноярский государственный педагогический университет им. В.П. Астафьева, 2023. С. 1050–1055.

10. *Григорьев С.Г., Аникьева М.А.* Повышение эффективности применения технологий генеративного искусственного интеллекта в образовательной деятельности // Информатика и образование. 2024. Т. 39. № 3. С. 5–15. https://doi.org/10.32517/0234-0453-2024-39-3-5-15

11. *Carbonell J.* AI in CAI: An Artificial-Intelligence Approach to Computer-Assisted Instruction // IEEE Transactions on Man-Machine Systems. 1970. Vol. 11, No. 4. P. 190–202. https://doi.org/10.1109/TMMS.1970.299942

12. *Nye B.D., Graesser A.C., Hu X.* AutoTutor and family: A review of 17 years of natural language tutoring // International Journal of Artificial Intelligence in Education. 2014. Vol. 24, No. 4. P. 427–469. https://doi.org/10.1007/s40593-014-0029-5

13. *Afzal S., Dempsey B., D'Helon C., Mukhi N., Pribic M., Sickler A., Strong P., Vanchiswar M., Wilde L.* The Personality of AI Systems in Education: Experiences with the Watson Tutor, a One-on-One Virtual Tutoring System // Childhood Education. 2019. Vol. 95, No. 1. P. 44–52. https://doi.org/10.1080/00094056.2019.1565809

14. *Alden Rivers B., Whitelock D., Richardson J.T.E., Field D., Pulman S*. Functional, Frustrating and Full of Potential: Learners' Experiences of a Prototype for Automated Essay Feedback. Proceedings of the International Computer Assisted Assessment Conference – Research into E-Assessment, Zeist, The Netherlands, 30 June – 1 July 2014. Communications in Computer and Information Science, Vol 439. P. 40–52. https://doi.org/10.1007/978-3-319-08657-6_4

15. *Arnold K.E., Pistilli M.D.* Course Signals at Purdue: Using Learning Analytics to Increase Student Success // Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (LAK '12), Vancouver, Canada, 29 April – 2 May 2012. P. 267–270. https://doi.org/10.1145/2330601.2330666

16. *Kustitskaya T.A., Esin R.V., Vainshtein Y.V., Noskov M.V.* Hybrid Approach to Predicting Learning Success Based on Digital Educational History for Timely Identification of At-Risk Students // Education Sciences. 2024. Vol. 14, No. 6. P. 657–676. https://doi.org/10.3390/educsci14060657

17. *Кушниренко А.Г., Леонов А.Г., Мащенко К.А., Мартынов Н.С., Пчелин К.К., Шляхов А.В*. Интеграция нейросетевого ассистента обучающихся по программированию в цифровую образовательную платформу // Информатика и образование. 2026. Т. 41, № 1. С. 23–36.
https://doi.org/10.32517/0234-0453-2026-41-1-23-36

18. *Гриднев Н.В., Караваева А.С., Леонов А.Г., Мащенко К.А., Пчелин К.К., Тарасюк Е.Д.* Гибридная модель обучения: новации в учете посещаемости в цифровой образовательной платформе Мирера // Труды НИИСИ. 2025. Т. 15, № 1. С. 58–64. https://doi.org/10.25682/NIISI.2025.1.0008

19. *Кадина Д.И., Леонов А.Г., Мартынов Н.С., Мащенко К.А., Орлов Э.А., Стрекалова А.И.* Автоматизация проверки на плагиат: новый подход к анализу кода в цифровой образовательной платформе Мирера // Труды НИИСИ. 2025. Т. 15, № 1. С. 52–57. https://doi.org/10.25682/NIISI.2025.1.0007

20. *Леонов А.Г., Мартынов Н.С., Мащенко, К. А., Холькина А.А., Шляхов А.В.* Автоматизация проверки семантической составляющей текстовых ответов обучающихся в цифровой образовательной платформе // Программные продукты и системы. 2024. Т. 37, № 3. С. 440–452.
https://doi.org/10.15827/0236-235X.147.440-452

MIRERA DIGITAL EDUCATIONAL PLATFORM

A. G. Leonov[1] [0000-0001-9622-1526], K. A. Mashchenko[2] [0000-0003-0355-6699]

[1, 2]*NRC «Kurchatov Institute» – SRISA, Moscow, Russia*

[1, 2]*Lomonosov Moscow State University, Moscow, Russia*

[1]*Moscow Pedagogical State University, Moscow, Russia*

[1]*State University of Management, Moscow, Russia*

[1]dr.l@vip.niisi.ru, [2]kirill.mashchenko@niisi.ru

Abstract

This paper presents the Mirera digital educational platform designed for comprehensive support of the educational process. The platform integrates learning management tools, automated assessment, learning analytics, attendance monitoring, and AI-based support for both instructors and students. The paper describes the key functionalities of the system, including automated assessment across various disciplines, adaptive learning pathways, analysis of students' digital footprints, integrated communication tools, and AI-powered assistants. Particular attention is paid to the automation of routine educational processes and data-driven pedagogical decision-making. The results of experimental deployment demonstrate that the platform reduces instructors' workload, increases student engagement, and improves the efficiency of educational process management. The findings confirm the potential of integrated digital educational platforms as an effective tool for the digital transformation of higher education.

Keywords*: digital educational platform, artificial intelligence in education, learning analytics, educational automation, adaptive learning, digital transformation of education.*

REFERENCES

1. *Leonov A.G., Ehpiktetov M.G.* Praktikum E-87 // Annotirovannyj bibliograficheskij ukazatel' «Deponirovannye nauchnye raboty» VINITI RAN. 1988. No. 4279-B88.

2. *Besshaposhnikov N.O., Kushnirenko A.G., Leonov A.G., Rajko M.V., Sobakinskih O.V.* Cifrovaya obrazovatel'naya sreda PiktoMir: opyt razrabotki i massovogo

vnedreniya godovogo kursa programmirovaniya dlya doshkol'nikov // Informatika i obrazovanie. 2020. No. 10. P. 28–40.
https://doi.org/10.32517/0234-0453-2020-35-10-28-40

3. *Smolyaninova O.G., Ivanov N.A.* Obzor praktik obespecheniya ehlektronnoj podderzhki obrazovatel'nogo processa sredstvami LMS Moodle: opyt rossijskih vuzov // Azimut nauchnyh issledovanij: pedagogika i psihologiya 2019. Vol. 8, No. 2(27). P. 228–232. https://doi.org/10.26140/anip-2019-0802-0054

4. *Tkanov V.* Ot pervyh kursov po vyboru k massovoj individualizacii: kak platforma Modeus pomogaet rossijskim vuzam perejti k novoj obrazovatel'noj modeli // Honors College Conference: Materialy II Mezhdunarodnoj konferencii, Krasnoyarsk, 27–28 noyabrya 2020 goda. Krasnoyarsk: Sibirskij federal'nyj universitet, 2020, P. 30–34.

5. *Gorchakov L.V., Stas A.N., Kartashov D.V.* Obuchenie programmirovaniyu s ispol'zovaniem sistemy EJUDGE // Vestnik Tomskogo gosudarstvennogo pedagogicheskogo universiteta. 2017. No. 9 (186). P. 109–112.
https://doi.org/10.23951/1609-624X-2017-9-109-112.

6. *Larina Y.A., Lavrovskaya O.B.* Osobennosti ispol'zovaniya onlajn-platformy Yandeks.kontest dlya provedeniya zanyatij po programmirovaniyu // Teoreticheskie i prikladnye aspekty razvitiya sovremennoj nauki i obrazovaniya: Materialy IV Vserossijskoj nauchno-prakticheskoj konferencii, Cheboksary, 26 fevralya 2021 goda. Cheboksary: Negosudarstvennoe obrazovatel'noe chastnoe uchrezhdenie dopolnitel'nogo professional'nogo obrazovaniya "Ehkspertno-metodicheskij centr", 2021. P. 102–105.

7. *Lakotko E.* COURSERA kak al'ternativnoe vysshee obrazovanie // Na putyah k novoj shkole 2016. No. 2. P. 49–54.

8. *Shamsudin N.M., Teoh S.H.* Exploring the Synergy of Learning Experience Platforms (LXP) with Artificial Intelligence for Enhanced Educational Outcomes // Proceedings of the International Conference on Innovation & Entrepreneurship in Computing, Engineering & Science Education (InvENT 2024), Shah Alam, Malaysia, 20–22 August 2024. P. 30–40. https://doi.org/10.2991/978-94-6463-589-8_5

9. *Grigor'ev S.G.* Iskusstvennyj intellekt v obrazovanii // Informatizaciya obra-zovaniya i metodika ehlektronnogo obucheniya: cifrovye tehnologii v obrazovanii : Materialy VII Mezhdunarodnoj nauchnoj konferencii, Krasnoyarsk, 19–22 сентября

2023 года. Krasnoyarsk: Krasnoyarskij gosudarstvennyj pedagogi-cheskij uni-versitet im. V.P. Astaf'eva, 2023. P. 1050–1055.

10. *Grigor'ev S.G., Anik'eva M.A.* Povyshenie ehffektivnosti primeneniya tehnologij generativnogo iskusstvennogo intellekta v obrazovatel'noj deyatel'nosti // Informatika i obrazovanie. 2024. Vol. 39. No. 3. P. 5–15. https://doi.org/10.32517/0234-0453-2024-39-3-5-15

11. *Carbonell J.* AI in CAI: An Artificial-Intelligence Approach to Computer-Assisted Instruction // IEEE Transactions on Man-Machine Systems. 1970. Vol. 11, No. 4. P. 190–202. https://doi.org/10.1109/TMMS.1970.299942

12. *Nye B.D., Graesser A.C., Hu X.* AutoTutor and family: A review of 17 years of natural language tutoring // International Journal of Artificial Intelligence in Education. 2014. Vol. 24, No. 4. P. 427–469. https://doi.org/10.1007/s40593-014-0029-5

13. *Afzal S., Dempsey B., D'Helon C., Mukhi N., Pribic M., Sickler A., Strong P., Vanchiswar M., Wilde L.* The Personality of AI Systems in Education: Experiences with the Watson Tutor, a One-on-One Virtual Tutoring System // Childhood Education. 2019. Vol. 95, No. 1. P. 44–52. https://doi.org/10.1080/00094056.2019.1565809

14. *Alden Rivers B., Whitelock D., Richardson J.T.E., Field D., Pulman S.* Functional, Frustrating and Full of Potential: Learners' Experiences of a Prototype for Automated Essay Feedback. Proceedings of the International Computer Assisted Assessment Conference – Research into E-Assessment, Zeist, The Netherlands, 30 June – 1 July 2014. Communications in Computer and Information Science, Vol. 439. P. 40–52. https://doi.org/10.1007/978-3-319-08657-6_4

15. *Arnold K.E., Pistilli M.D.* Course Signals at Purdue: Using Learning Analytics to Increase Student Success // Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (LAK '12), Vancouver, Canada, 29 April – 2 May 2012. P. 267–270. https://doi.org/10.1145/2330601.2330666

16. *Kustitskaya T.A., Esin R.V., Vainshtein Y.V., Noskov M.V.* Hybrid Approach to Predicting Learning Success Based on Digital Educational History for Timely Identification of At-Risk Students // Education Sciences. 2024. Vol. 14, No. 6. P. 657–676. https://doi.org/10.3390/educsci14060657

17. *Kushnirenko A.G., Leonov A.G., Mashchenko K.A., Martynov N.S., Pchelin K.K., Shlyahov A.V.* Integraciya nejrosetevogo assistenta obuchayushchihsya po

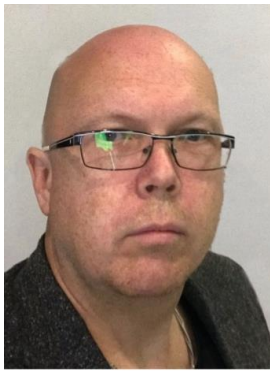
programmirovaniyu v cifrovuyu obrazovatel'nuyu platformu // Informatika i obrazovanie. 2026. Vol. 41, No. 1. P. 23–36.
https://doi.org/10.32517/0234-0453-2026-41-1-23-36

18. *Gridnev N.V., Karavaeva A.S., Leonov A.G., Mashchenko K.A., Pchelin K.K., Tarasyuk E.D.* Gibridnaya model' obucheniya: novacii v uchete poseshchaemosti v cifrovoj obrazovatel'noj platforme Mirera // Trudy NIISI. 2025. Vol. 15, No. 1. P. 58–64. https://doi.org/10.25682/NIISI.2025.1.0008

19. *Kadina D.I., Leonov A.G., Martynov N.S., Mashchenko K.A., Orlov E.A., Strekalova A.I.* Avtomatizaciya proverki na plagiat: novyj podhod k analizu koda v cifrovoj obrazovatel'noj platforme Mirera // Trudy NIISI. 2025. Vol. 15, No. 1. P. 52–57. https://doi.org/10.25682/NIISI.2025.1.0007

20. *Leonov A.G., Martynov N.S., Mashchenko, K. A., Hol'kina A.A., Shlyahov A.V.* Avtomatizaciya proverki semanticheskoj sostavlyayushchej tekstovyh otvetov obuchayushchihsya v cifrovoj obrazovatel'noj platforme // Programmnye produkty i sistemy. 2024. Vol. 37, No. 3. P. 440–452.
https://doi.org/10.15827/0236-235X.147.440-452

СВЕДЕНИЯ ОБ АВТОРАХ



ЛЕОНОВ Александр Георгиевич – учился на механико-математическом факультете МГУ имени М.В. Ломоносова. Кандидат физ.-мат. наук, в 2024 году защитил диссертацию на соискание ученой степени доктора педагогических наук по специальности 5.8.2 «Теория и методика обучения и воспитания (информатика, информатика и вычислительная техника)». В настоящее время работает руководителем цифровой трансформации, заведующим кафедрой дополнительного профессионального образования НИЦ «Курчатовский институт» – НИИСИ, ведущим научным сотрудником Лаборатории вычислительных методов механико-математического факультета МГУ имени М.В. Ломоносова, руководителем программы магистратуры института информационных технологий ГУУ, а также профессором кафедры математики и информатики в начальной школе МПГУ.

Область научных интересов: методология и методики обучения, интеллектуальные информационные системы, цифровые образовательные среды и платформы, нейронные сети и искусственный интеллект в образовании.

Alexander Georgievich LEONOV – studied at the Faculty of Mechanics and Mathematics of Lomonosov Moscow State University. Candidate of Physical and Mathematical Sciences, in 2024 he defended his dissertation for the degree of Doctor of Pedagogical Sciences in specialty 5.8.2. Theory and methodology of teaching and upbringing (computer science, computer science and computer engineering). Currently, he is CDTO (Chief Digital Transformation Officer), head of the Department of Continuing Professional Education at the NRC «Kurchatov Institute» – SRISA, leading researcher at the Laboratory of Computational Methods of the Faculty of Mechanics and Mathematics of Lomonosov Moscow State University, associate professor at the Department of Mathematical Methods in Economics and Management of the Institute of Information Systems of the State University of Management, as well as professor at the Department of Mathematics and Informatics in Elementary School of the Faculty of Elementary Education of the Institute of Childhood of Moscow Pedagogical State University. Research interests: methodology and teaching methods, intelligent information systems, digital educational environments and platforms, as well as neural networks and artificial intelligence in education.

email: dr.l@vip.niisi.ru
ORCID: 0000-0001-9622-1526



МАЩЕНКО Кирилл Алексеевич – учился на механико-математическом факультете МГУ имени М.В. Ломоносова. Закончил аспирантуру по специальности 2.3.8 «Информатика и информационные процессы». В настоящее время является младшим научным сотрудником отдела учебной информатики НИЦ «Курчатовский институт» - НИИСИ, а также младшим научным сотрудником Лаборатории вычислительных методов Механико-математического факультета МГУ имени М.В. Ломоносова.

Область научных интересов: автоматизация и цифровизация образовательного процесса, в том числе с применением методов искусственного интеллекта.

Kirill Alekseevich MASHCHENKO – studied at the Faculty of Mechanics and Mathematics of Lomonosov Moscow State University. He completed postgraduate studies in specialty 2.3.8. Informatics and Information Processes. Currently, he is a junior researcher at the Department of Educational Informatics of the NRC «Kurchatov Institute» – SRISA, as well as a Junior Researcher at the Laboratory of Computational Methods of the Faculty of Mechanics and Mathematics of Lomonosov Moscow State University. Research interests: automation and digital transformation of the educational process, including the application of artificial intelligence methods.

email: kirill.mashchenko@niisi.ru

ORCID: 0000-0003-0355-6699

Материал поступил в редакцию 20 июня 2026 года

УДК 550.388

МЕТОДИКА, РЕЗУЛЬТАТЫ И СРАВНИТЕЛЬНЫЙ АНАЛИЗ ПРИМЕНЕНИЯ НЕЙРОСЕТЕВЫХ МОДЕЛЕЙ MASK-RCNN И YOLO ДЛЯ АВТОМАТИЧЕСКОГО ВЫДЕЛЕНИЯ ТРЕКОВ МОД НА ИОНОГРАММАХ НАКЛОННОГО РАДИОЗОНДИРОВАНИЯ ИОНОСФЕРЫ

С. В. Масюченко[1] [0009-0003-7177-2614], А. О. Щирый[2] [0000-0001-7312-4424]

[1]*НИУ Московский энергетический институт, г. Москва, Россия*

[2]*Институт земного магнетизма, ионосферы и распространения радиоволн им. Н. В. Пушкова РАН, г. Москва, Россия*

[1]stiwwin@gmail.com, [2]andreyschiriy@gmail.com

Аннотация

Автоматическая обработка ионограмм радиозондирования ионосферы необходима как для оперативного мониторинга ионосферы, так и при анализе массивов архивных данных радиозондирования. Основным этапом для достижения этой цели является автоматическое обнаружение и классификация треков – следов отраженных мод радиосигнала на ионограммах. В работе представлены методика, результаты и выполнен сравнительный анализ применения нейросетевых моделей (нейросетевых архитектур) для инстанс-сегментации Mask RCNN и YOLO. Наиболее успешными оказались модели семейства YOLO, которые обеспечивают как высокую точность, так и вычислительную эффективность. При этом точное попиксельное совпадение масок не является критичным – ключевым требованием является корректная локализация полезного сигнала. Использование указанных моделей для выделения треков на ионограммах продемонстрировало умеренную, но достаточную эффективность: основные структуры треков надежно идентифицируются. В будущем планируется расширение набора применяемых моделей, прежде всего за счет визуальных трансформеров.

***Ключевые слова**: ионограмма, выделение сигнала, сегментация изображений, автоматизация радиозондирования ионосферы.*

© С. В. Масюченко, А. О. Щирый, 2026.
Данная статья распространяется на условиях международной лицензии Creative Commons License Attribution 4.0 International (CC BY 4.0).

ВВЕДЕНИЕ

Работа радиотехнических систем (РТС) декаметрового диапазона, таких как системы коротковолновой радиосвязи и загоризонтной (ЗГ) радиолокации, основана на способности коротких волн многократно отражаться от ионосферы и земной поверхности. Важнейшей задачей для обеспечения корректного функционирования таких РТС является адаптация к ионосферным условиям, для чего в составе этих РТС содержится РТС частотного обеспечения, включающая аппаратуру радиозондирования ионосферы и анализатор загруженности радиоканала. При этом возможна адаптация РТС декаметрового диапазона по рабочей частоте, мощности излучения и скорости передачи информации. Аппаратура радиозондирования ионосферы, как правило, обеспечивает диагностику в режимах наклонного зондирования ионосферы (НЗИ) и вертикального зондирования ионосферы (ВЗИ). В случае загоризонтных радиолокационных станций обязателен также режим возвратно-наклонного зондирования [1]. Адаптация РТС требует автоматической обработки данных радиозондирования. При этом важнейшей задачей при такой автоматической обработке является автоматическое выделение треков мод радиосигнала на ионограмме [2–5]. Ионограмма НЗИ (см. рис. 1) представляет собой зависимость времени группового запаздывания и амплитуды каждой моды распространения радиосигнала от частоты излучения [6].

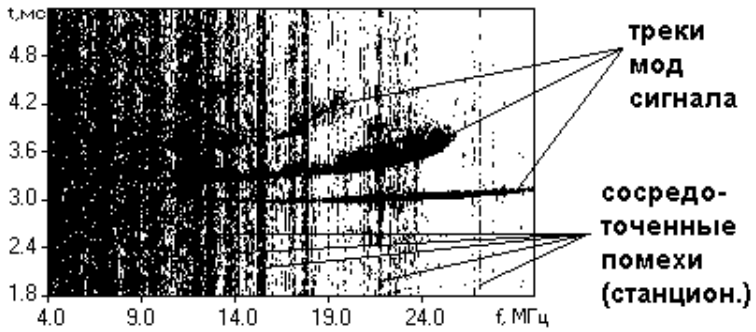


Рис. 1. Пример ионограммы НЗИ.

В настоящей работе использованы только данные НЗИ, хотя задача выделения треков мод радиосигнала актуальна и для ионограмм других видов, но интерпретация результатов существенно отличается. Дадим краткий обзор существующих методов автоматической обработки ионограмм.

Для обработки ионограмм может использоваться метод трафаретных масок [7, 8], когда для выделения сигнальных компонент на ионограммах применяют маски, полученные путем синтеза искусственных ионограмм на основе математических моделей распространения декаметровых радиоволн в ионосферной плазме, далее подгоняемые под исследуемую экспериментальную ионограмму какими-либо алгоритмами. Такой метод обладает рядом недостатков: применяемые математические модели зачастую не обеспечивают приемлемую точность, а также в условиях геомагнитных и солнечных возмущений, особенно в верхних широтах; при таком подходе невозможно выделение сигнальных компонент аномальных мод распространения.

Более примитивные пороговые методы [9], хотя и не подвержены ограничениям моделей распространения радиоволн, но существенно зависят от отношения сигнал/шум (или «качества ионограммы», другими словами).

Значительно более удачными стоит признать методы, основанные на статистических критериях обнаружения резко выделяющихся значений выборки (так называемых «грубых промахов») [10, 11]; однако статистические характеристики сигнала в случаях пересечения треков сигнала с сосредоточенными помехами, диффузности (размытости) сигнала существенно отличаются от характеристик сигнала сильных мод, что может приводить к ошибкам. При этом также не используется информация о протяженной по частоте зондирования структуре трека (из-за обработки ионограммы блоками, равными одной вертикальной линии ионограммы).

Невысокое качество алгоритмов автоматического выделения даже для более простых ионограмм ВЗИ отмечается многими авторами, например, [12]: «существующие на данный момент алгоритмы имеют значительно меньшую эффективность, например, точность автоматической обработки, применяемой дигизондом Рейниша (DPS-4) в летний период, не превышает 70%, а в зимний 50%». Хотя процитированные оценки [12] делались для ионограмм ВЗИ верхних широт, а для ионограмм средних широт эффективность автоматической обработки ощутимо выше, для многих практических применений такая обработка остается недостаточной, тем более, если речь идет об обработке ионограмм НЗИ, а не ВЗИ.

Ввиду недостатков автоматических способов выделения треков (особенно для ионограмм НЗИ) приходилось применять полуавтоматический способ выделения сигнала [2], заключающийся в том, что на ионограмме оператором вручную выделялись прямоугольные области, содержащие следы треков мод полезного сигнала, затем в выбранных областях спектральные отсчеты, превышающие заданный порог, относились к сигналу, далее находились аппроксимирующие полиномы для получения треков мод.

Конечно, делаются попытки применения нейронных сетей для решения задач выделения треков на ионограммах, правда, в основном ВЗИ и с не вполне ясной точностью и качеством [12–16].

С учетом вышесказанного одним из авторов ранее была предложена методика [17] применения нейросетей для задачи выделения треков мод радиосигнала на ионограммах, где, в отличие от перечисленных выше работ, при построении обучающей выборки размеченных ионограмм отбирают ионограммы радиозондирования ионосферы для нормальных условий распространения радиоволн в ионосфере (при отсутствии солнечных и геомагнитных возмущений) и при разметке помечают их бинарным признаком «неаномальна». Дополнительно отбирают и добавляют в выборку ионограммы для аномальных условий распространения и на этапе разметки помечают их бинарным признаком «аномальна», а на стадии обучения модели в качестве признаков используют не только битовые маски выделенных треков, но и метаданные ионограмм (дату–время зондирования, координаты передающего и приемного пунктов).

Для практической реализации предлагаемой методики необходимо разметить, вручную выделяя треки, весьма большие массивы ионограмм (десятки тысяч). Требование больших обучающих выборок обусловлено большим числом внутренних параметров моделей глубокого обучения.

Отдельно необходимо решить задачу выбора базовой архитектуры нейросети. Поэтому целью настоящей работы было сравнение нейросетевых моделей (нейросетевых архитектур) для инстанс-сегментации Mask RCNN и YOLO, как наиболее эффективных моделей для сегментации, во всяком случае до появления визуальных трансформеров ViT (добавление в сравнительный анализ моделей ViT является задачей будущих работ).

ДАННЫЕ ДЛЯ ОБУЧЕНИЯ МОДЕЛЕЙ

В качестве обучающего набора было использовано 300 ионограмм, полученных при наклонном зондировании по шести радиолиниям: Хабаровск – Йошкар-Ола, Магадан – Йошкар-Ола, Норильск – Йошкар-Ола, Усолье – Йошкар-Ола, Кипр – Йошкар-Ола и Кипр – Москва. Указанное число ионограмм отобрано (вручную) из нескольких тысяч ионограмм как наиболее «непохожих» друг на друга (большие массивы слишком похожих ионограмм отрицательно скажутся на качестве сегментации). Проблему похожести ионограмм соседних сеансов зондирования иллюстрирует рис. 2.

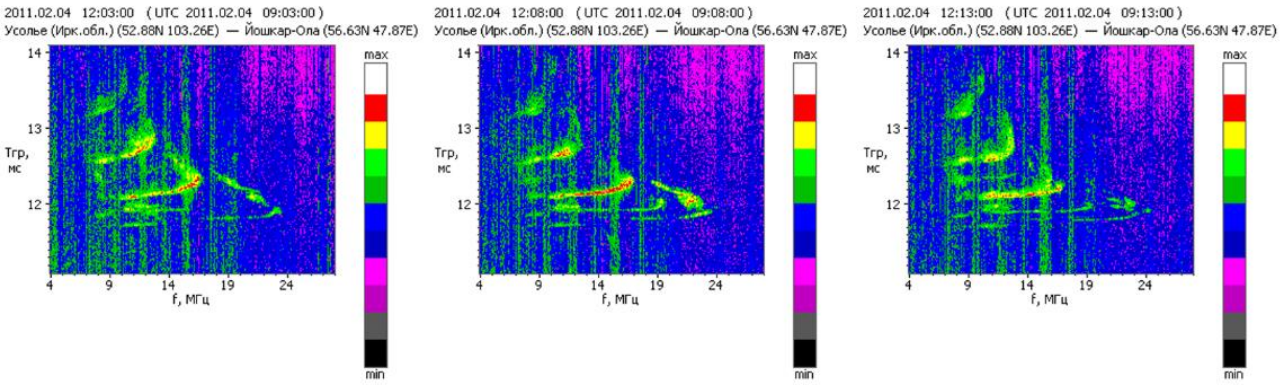


Рис. 2. Ионограммы соседних сеансов могут быть очень похожими.

Вручную была выполнена попиксельная разметка треков с выделением трех классов (цвет, указанный в скобках, использован для маркировки рис. 3 и 4, приведенных далее): L – нижний трек (красный), U – верхний трек (зеленый), N – неопределенные трек (фиолетовый). Пример разметки показан на рис. 3. В результате была получена аннотированная выборка из размеченных ионограмм в формате сегментационных масок, готовая к обучению моделей инстанс-сегментации.

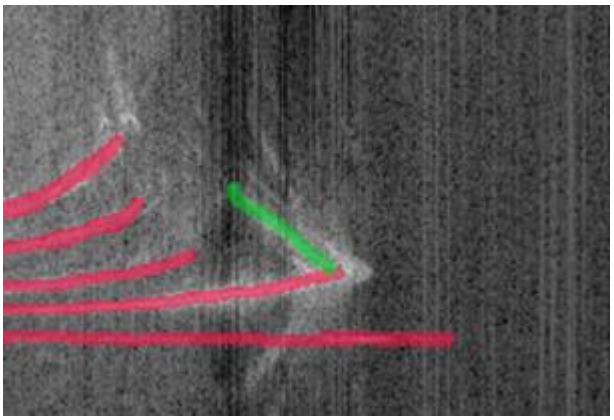


Рис. 3. Пример разметки треков на ионограмме.

Была также предпринята попытка применения аугментации, однако для данной задачи недопустимы трансформации, нарушающие ориентацию трека (например, повороты), поскольку она несет ключевую информацию. В итоге были использованы только следующие преобразования:

- масштабирование по осям X и Y (±50%);
- сдвиг по осям X и Y (±15%);
- случайное изменение контраста;
- добавление размытия (blur).

Подчеркнем, что даже такая ограниченная аугментация применима только для решаемой задачи сегментации. Если же говорить о задачах будущих исследований, таких как построение предиктивных моделей ионосферы на основе ионограмм или моделей интерпретации ионограмм, то там аугментации типа сдвига и масштабирования невозможны, поскольку положение каждого пиксела относительно осей имеет смысл.

МЕТОДИКА ПРИМЕНЕНИЯ МОДЕЛЕЙ

Было проведено сравнительное тестирование ряда передовых моделей, включая классическую Mask R-CNN (с backbone ResNet-50 и Feature Pyramid Network), а также более современные подходы – YOLOv11-Seg, разработанные в рамках семейства YOLO (You Only Look Once) и адаптированные для задач сегментации. Эти модели отличаются высокой скоростью инференса (применения обученной модели) и улучшенной архитектурой feature extraction, что делает их перспективными для обработки больших объемов данных.

Mask RCNN – это архитектура нейронной сети для одновременного обнаружения объектов, их классификации и построения попиксельно точных масок сегментации каждого отдельного экземпляра.

Перечислим этапы работы модели.

1. Извлечение признаков (Feature Extraction)

Сначала изображение проходит через сверточную нейросеть. Она «анализирует» картинку и создает карту признаков, своего рода «описание» содержимого изображения на разных уровнях детализации.

2. Предложение регионов (Region Proposal)

Специальный блок просматривает карту признаков и предлагает возможные места, где могут находиться объекты. Он выдает прямоугольники-«кандидаты» (регионы интереса).

3. Классификация и уточнение рамок (RoI Head)

Каждый предложенный регион обрабатывается отдельно:

- сеть уточняет границы прямоугольника (чтобы он точнее обводил объект);
- определяет класс объекта;.
- отбрасывает регионы, где ничего нет.

4. Генерация маски (Mask Branch)

Параллельно с классификацией работает отдельная ветвь сети – ветвь масок. Для каждого найденного объекта она создает точную бинарную маску, которая показывает, какие именно пиксели принадлежат объекту.

YOLO11-seg – это Instance Segmentation на основе YOLO11, модель выполняет одновременное обнаружение объектов, их классификацию и пиксельно-точную сегментацию. Архитектура включает усовершенствованный backbone (для эффективного извлечения признаков) и оптимизированный neck (для агрегации мультимасштабных представлений), что обеспечивает высокую точность при меньшем количестве параметров по сравнению с предыдущими версиями.

Нами были использованы три конфигурации: nano (2.6 млн параметров), small (9.4 млн параметров) и medium (20.1 млн параметров).

Этапы обучения YOLO11-seg отличаются от приведенных выше, за исключением этапа извлечения признаков. Перечислим их.

1. Извлечение признаков.

2. Сеть генерирует набор «прототипов масок» (Prototypical Masks) на все изображение сразу, это набор абстрактных пятен.

3. Для каждого конкретного объекта сеть предсказывает коэффициенты маски.

4. Итоговая маска объекта получается путем сложения этих «прототипов» с нужными весами (коэффициентами) и обрезкой по границе прямоугольника.

РЕЗУЛЬТАТЫ И СРАВНЕНИЕ

Пример результатов сегментации для модели Mask RCNN представлен на рис. 4.

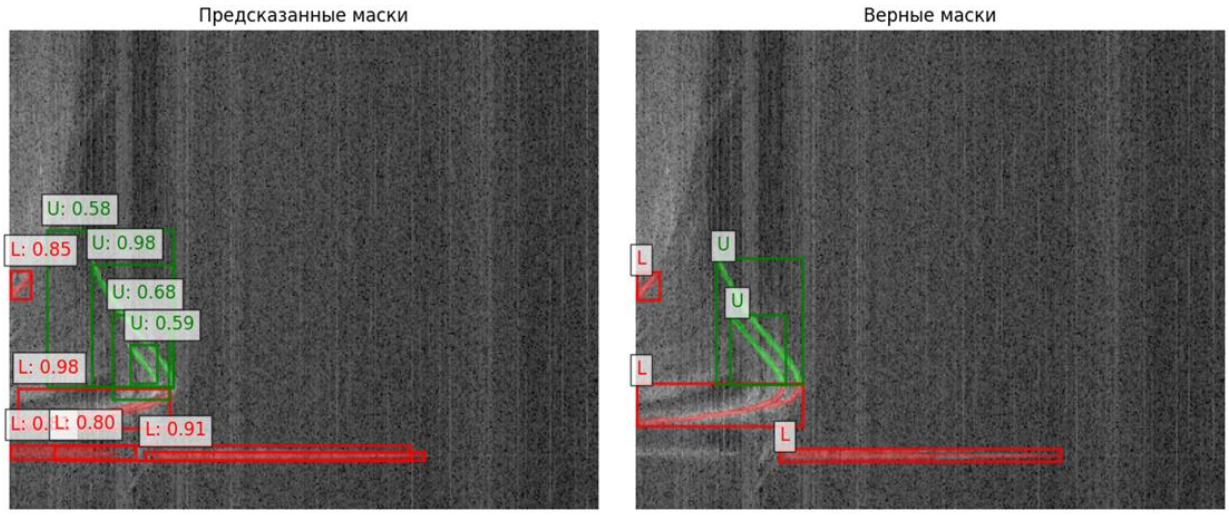


Рис. 4. Пример результатов для Mask RCNN.

Для сравнительного анализа используемых моделей и оценки их качества были использованы следующие метрики, основанные на mAP (mean Average Precision), она учитывает как точность классификации, так и корректность локализации (насколько точно предсказанный bounding box или маска совпадают с разметкой – были проанализированы оба варианта). Это следующие метрики:

- mAP – средняя точность по всем классам, усредненная по диапазону порогов IoU (Intersection over Union) от 0.5 до 0.95 с шагом 0.05;
- mAP@50 – значение mAP при фиксированном пороге IoU = 0.5; учитываются только те предсказания, для которых пересечение с истинной областью составляет не менее 50%;
- mAP per class – значения mAP для каждого класса в отдельности.

В табл. 1 представлены значения указанных метрик.

Табл. 1. Результаты тестирования моделей сегментации.

Модель сегмента-ции	map (box)	map_ 50 (box)	map (L-class) (box)	map (U-class) (box)	map (N-class) (box)	map (mask)	map_50 (mask)	map (L-class) (mask)	map (U-class) (mask)	map (N-class) (mask)
Mask RCNN	0.29	0.61	0.36	0.31	0.20	0.15	0.44	0.26	0.01	0.18
YOLO_v11 nano	0.31	0.57	0.32	0.43	0.18	0.16	0.48	0.23	0.12	0.14
YOLO_v11 small	0.31	0.67	0.33	0.46	0.14	0.16	0.52	0.24	0.11	0.13
YOLO_v11 medium	0.32	0.58	0.37	0.34	0.14	0.17	0.55	0.23	0.15	0.13
YOLO_v11 small	0.36	0.66	0.72	0.78	0.47	0.21	0.60	0.27	0.20	0.17
Mask RCNN	0.35	0.67	0.36	0.34	0.34	0.18	0.49	0.26	0.01	0.28

Из табл. 1 видно, что основные структуры треков надежно идентифицируются. При этом точное попиксельное совпадение масок не является критичным – ключевым требованием является корректная локализация полезного сигнала. Таким образом, подходы на основе ограничивающих прямоугольников (bounding boxes) показали высокие результаты по стандартным метрикам детекции.

ЗАКЛЮЧЕНИЕ

Применение моделей сегментации изображений Mask RCNN и YOLO для выделения треков мод распространения радиосигнала на ионограммах продемонстрировало умеренную, но достаточную эффективность: основные структуры треков надежно идентифицируются. Наиболее успешными оказались модели семейства YOLO, которые обеспечивают как высокую точность, так и вычислительную эффективность.

Текущие стратегии аугментации данных приводят к генерации изображений с чрезмерно высокой визуальной схожестью, что ограничивает разнообразие обучающей выборки. В дальнейшем целесообразно рассмотреть расширен-

ный набор методов аугментации, способных вносить более значимые и физически обоснованные вариации в структуру ионограмм, тем самым повысив обобщающую способность модели.

В дальнейшем набор применяемых моделей будет расширен прежде всего за счет визуальных трансформеров.

Благодарности

Работа выполнена в рамках государственного задания Минобрнауки России для ИЗМИРАН (тема № FFUM-2026-0010).

СПИСОК ЛИТЕРАТУРЫ

1. *Fabrizio G.A.* High frequency over-the-horizon radar: fundamental principles, signal processing, and practical applications. McGraw-Hill Education, 2013.

2. *Щирый А.О.* Разработка и моделирование алгоритмов автоматического измерения характеристик ионосферных коротковолновых радиолиний: Автореф. дис. … канд. техн. наук [Место защиты: С.-Петерб. гос. ун-т телекоммуникаций им. М.А. Бонч-Бруевича]. СПб., 2007. 19 с. EDN: QMISDO

3. *Щирый А.О.* Развитие средств автоматизации наземного радиозондирования ионосферы // Фундаментальные проблемы радиоэлектронного приборостроения. 2014. Т. 14, № 5. С. 170–173.

4. *Щирый А.О.* Архитектура программной части аппаратно-программного комплекса дистанционного наземного радиозондирования ионосферы // Новые информационные технологии в автоматизированных системах. 2015. №18. С. 144–152.

5. *Щирый А.О.* Алгоритмы и программное обеспечение автоматизации процессов измерений и обработки данных оперативной диагностики ионосферы и ионосферных радиолиний // Журнал радиоэлектроники. 2022. №10. https://doi.org/10.30898/1684-1719.2022.10.4

6. *Филипп Н.Д., Блаунштейн Н.Ш., Ерухимов Л.М., Иванов В.А., Урядов В.П.* Современные методы исследования динамических процессов в ионосфере. Кишинев: Штиинца, 1991. 286 с.

7. *Пономарчук С.Н., Грозов В.П., Котович Г.В., Михайлов С.Я.* Обработка и интерпретация ионограмм вертикального и наклонного зондирования для диагностики ионосферы на базе ЛЧМ-ионозонда // Вестник Сибирского государственного аэрокосмического университета им. акад. М.Ф. Решетнева. 2013. № 5(51). С. 163–166.

8. Патент РФ №2697433. *Цыбуля К.Г.* Способ автоматического определения параметров ионосферных слоев по ионограммам. Дата заявки: 26.10.2018. Дата публикации: 14.08.2019.

9. *Егошин А.Б.* Автоматизированная система адаптивной обработки сигналов со сверхбольшой базой для радиозондирования ионосферных радиолиний. Автореф. дис. … канд. техн. наук. Марийский гос. техн. ун-т. Йошкар-Ола, 2003. 25 с.

10. *Колчев А.А., Щирый А.О.* Алгоритм автоматического выделения спектральных компонентов сигнала на ионограмме // Материалы десятого научно-практического семинара «Новые информационные технологии». М.: Московский гос. ин-т электроники и математики, 2007. С. 102–107.

11. *Недопекин А.Е.* Метод обнаружения сигнала ЛЧМ-ионозонда в частотной области с учетом уширения принимаемых мод ионосферного распространения // Журнал Радиоэлектроники: электронный журнал. 2015. №10. URL: http://jre.cplire.ru/jre/oct15/5/text.pdf.

12. *Долгачева С.А., Макарова Л.Н., Николаев А.В.* Обработка ионограмм высокоширотных станций вертикального зондирования с использованием нейронных сетей: Es и F2 слои // Physics of Auroral Phenomena. 2020. Т. 43, № 1. С. 105–108.

13. *Guo L., Xiong J.* Multi-Scale Attention-Enhanced Deep Learning Model for Ionogram Automatic Scaling // Radio Science. 2023. Vol. 58, No. 3. p.e2022RS007566. https://doi.org/10.1029/2022RS007566

14. *Lu Z., Hua C., Wei N., Feng J., Lou P., Liu W.* Research on classification of vertical ionogram based on deep convolution neural network // Progress in Geophysics. 2022. Vol. 37. No. 5. P. 1834–1839. https://doi.org/10.6038/pg2022GG0073

15. *Xiao Z., Wang J., Li J., Zhao B., Hu L., Liu L.* Deep-learning for ionogram automatic scaling // Advances in Space Research. 2020. Vol. 66. No. 4. P. 942–950. https://doi.org/10.1016/j.asr.2020.05.009

16. *De la Jara C., Olivares C.* Ionospheric echo detection in digital ionograms using convolutional neural networks // Radio Science. 2021. Vol. 56. No. 8. P. 1–15.

17. Патент РФ на изобретение № 2859478. *Щирый А.О.* Способ автоматического выделения полезного сигнала на ионограмме радиозондирования ионосферы, основанный на машинном обучении / ФГБУН Институт земного магнетизма, ионосферы и распространения радиоволн им. Н.В. Пушкова Российской академии наук (ИЗМИРАН); № 2023135724; Заявл. 26.12.2023; Опубл. 03.04.2026; Бюл. № 10.

METHODOLOGY, RESULTS, AND COMPARATIVE ANALYSIS OF THE USE OF MASK R-CNN AND YOLO NEURAL NETWORK MODELS FOR THE TASK OF AUTOMATICALLY SELECTING MODE TRACKS ON IONOGRAMS OF OBLIQUE IONOSPHERIC RADIOSONDE

S. V. Masyuchenko[1] [0009-0003-7177-2614], A. O. Schiriy[2] [0000-0001-7312-4424]

[1]*National Research University "Moscow Power Engineering Institute", Moscow, Russia*

[2]*Pushkov Institute of Terrestrial Magnetism, Ionosphere and Radio Wave Propagation of the Russian Academy of Sciences, Moscow, Russia*

[1] stiwwin@gmail.com, [2]andreyschiriy@gmail.com

Abstract

Automatic processing of ionospheric radiosonde ionograms is necessary both for operational monitoring of the ionosphere and for analyzing archived radiosonde data arrays. The central place in this task is occupied by automatic detection and classification of tracks – traces of reflected radio signal modes on ionograms. This paper presents the methodology, results, and comparative analysis of the use of neural network models (neural network architectures) for Mask R-CNN and YOLO instance segmentation. The most successful models were the YOLO family, which provide both high accuracy and computational efficiency. At the same time, the exact pixel–by-pixel matching of the masks is not critical – the key requirement is the correct localization of the useful signal. The use of these models to identify tracks on ionograms

has demonstrated moderate but sufficient efficiency: the main structures of the tracks are reliably identified. The most successful models were the YOLO family, which provide both high accuracy and computational efficiency. In the future, the set of models used will be expanded, primarily due to visual transformers.

***Keywords**: ionogram, signal detection, image segmentation, ionogram processing.*

REFERENCES

1. *Fabrizio G.A.* High frequency over-the-horizon radar: fundamental principles, signal processing, and practical applications. McGraw-Hill Education, 2013.

2. *Shchiry A.O.* Development and simulation of algorithms for automatic measurement of characteristics of ionospheric short-wave radio links. PhD Thesis Abstract. St. Petersburg State University of Telecommunications, St. Petersburg, 2007 (In Russian).

3. *Shchiry A.O.* Development of automation tools for ground-based ionospheric radio sounding // Fundamentalnye Problemy Radioelektronnogo Priborostroeniya [Fundamental Problems of Radioengineering and Device Engineering]. 2014. Vol. 14, No. 5. P. 170–173 (In Russian).

4. *Shchiry A.O.* Software architecture of the hardware-software complex for remote ground-based ionospheric radio sounding // Novye Informatsionnye Tekhnologii v Avtomatizirovannykh Sistemakh [New Information Technologies in Automated Systems]. 2015. No. 18. P. 144–152 (In Russian).

5. *Shchiry A.O.* Algorithms and software for automating measurement and data processing for operational diagnostics of the ionosphere and ionospheric radio links // Zhurnal Radioelektroniki [Journal of Radio Electronics]. 2022. No. 10. https://doi.org/10.30898/1684-1719.2022.10.4 (In Russian).

6. *Filipp N.D., Blaunshtein N.S., Eryukhimov L.M., Ivanov V.A., Uryadov V.P.* Modern methods for studying dynamic processes in the ionosphere. Kishinev: Shtiintsa, 1991 (In Russian).

7. *Ponomarchuk S.N., Grozov V.P., Kotovich G.V., Mikhailov S.Ya.* Processing and interpretation of vertical and oblique sounding ionograms for ionospheric diag-

nostics based on an FMCW ionosonde // Vestnik Sibirskogo gosudarstvennogo aerokosmicheskogo universiteta im. ak. M.F. Reshetneva [Vestnik SibGAU]. 2013. No. 5(51). P. 163–166 (In Russian).

8. *Tsybulya K.G.* Method for automatic determination of ionospheric layer parameters by ionograms. Patent RU 2697433, published August 14, 2019 (In Russian).

9. *Egoshin A.B.* Automated system for adaptive signal processing with a super-large base for radio sounding of ionospheric radio links // PhD Thesis Abstract. Mari State Technical University, Yoshkar-Ola, 2003 (In Russian).

10. *Kolchev A.A., Shchiry A.O.* Algorithm for automatic extraction of signal spectral components on an ionogram // Proc. 10th Sci.-Prac. Seminar "New Information Technologies". Moscow: MIEM, 2007. P. 102–107 (In Russian).

11. *Nedopekin A.E.* Method for detecting the FMCW ionosonde signal in the frequency domain, considering the broadening of received ionospheric propagation modes // Zhurnal Radioelektroniki [Journal of Radio Electronics]. 2015. No. 10. Available at: http://jre.cplire.ru/jre/oct15/5/text.pdf (In Russian).

12. *Dolgacheva S.A., Makarova L.N., Nikolaev A.V.* Processing ionograms of high-latitude vertical sounding stations using neural networks: Es and F2 layers // Physics of Auroral Phenomena. 2020. Vol. 43, No. 1. P. 105–108 (In Russian).

13. *Guo L., Xiong J.* Multi-Scale Attention-Enhanced Deep Learning Model for Ionogram Automatic Scaling // Radio Science. 2023. Vol. 58, No. 3. e2022RS007566. https://doi.org/10.1029/2022RS007566

14. *Lu Z., Hua C., Wei N., Feng J., Lou P., Liu W.* Research on classification of vertical ionogram based on deep convolution neural network // Progress in Geophysics. 2022. Vol. 37, No. 5. P. 1834–1839. https://doi.org/10.6038/pg2022GG0073

15. *Xiao Z., Wang J., Li J., Zhao B., Hu L., Liu L.* Deep-learning for ionogram automatic scaling // Advances in Space Research. 2020. Vol. 66, No. 4. P. 942–950. https://doi.org/10.1016/j.asr.2020.05.009

16. *De la Jara C., Olivares C.* Ionospheric echo detection in digital ionograms using convolutional neural networks // Radio Science. 2021. Vol. 56, No. 8. P. 1–15.

17. *Shchiry A.O.* Method for automatic extraction of a useful signal on an ionospheric radio sounding ionogram based on machine learning. Patent RU 2859478, published April 03, 2026 (In Russian).

СВЕДЕНИЯ ОБ АВТОРАХ



МАСЮЧЕНКО Степан Вадимович – магистрант кафедры прикладной математики и искусственного интеллекта Национального исследовательского университета «МЭИ».

Stepan Vadimovich MASYUCHENKO – student at the Department of “Applied Mathematics and Artificial Intelligence”, National Research University “MPEI”.

email: stiwwin@gmail.com

ORCID: 0009-0003-7177-2614



ЩИРЫЙ Андрей Олегович – кандидат технических наук, старший научный сотрудник Института земного магнетизма, ионосферы и распространения радиоволн РАН им. Н.В. Пушкова (ИЗМИРАН), Москва.

Andrey Olegovich SCHIRIY – PhD (Candidate of Technical Sciences), senior researcher at the Pushkov Institute of Terrestrial Magnetism, Ionosphere and Radio Wave Propagation of the Russian Academy of Sciences (IZMIRAN), Moscow.

email: andreyschiriy@gmail.com

ORCID: 0000-0001-7312-4424

Материал поступил в редакцию 20 июня 2026 года

УДК 004.42

ОБРАБОТКА СКАНИРОВАННЫХ МАТЕМАТИЧЕСКИХ PDF-ДОКУМЕНТОВ В СЕРВИСЕ СЕМАНТИЧЕСКОГО ПОИСКА ФОРМУЛ

К. С. Николаев[0000-0003-3204-238X]

Национальный исследовательский центр «Курчатовский институт», г. Москва, Россия

konnikolaeff@yandex.ru

Аннотация

Рассмотрена задача извлечения семантических связей между математическими формулами и концептами из неструктурированных PDF-документов, включая сканированные. Предложен сквозной конвейер обработки, объединяющий оптическое распознавание, восстановление структуры параграфов, выделение формул и переменных, а также их связывание с концептами на основе синтаксических шаблонов русского языка. Разработаны формальные модели для двух типов связей: «переменная – концепт» и «переменная – главная формула». Проведена экспериментальная оценка качества на коллекции из 150 математических статей; достигнуто повышение точности связывания на 18% по сравнению с предшествующей версией. Описаны архитектурные решения прототипа веб-сервиса, обеспечивающие асинхронную обработку и масштабируемость.

Ключевые слова*: семантический поиск формул, PDF-документы, научные библиотеки, онтологии, веб-сервис.*

ВВЕДЕНИЕ

Обработка научных документов в формате PDF остается актуальной задачей для цифровых библиотек, особенно для сканированных материалов, где отсутствует текстовый слой. Такие документы требуют распознавания символов (OCR) [1] и восстановления логической структуры. Проблема усугубляется для математических текстов, где формулы содержат сложную двумерную структуру, а их семантическая интерпретация зависит от контекста.

© К. С. Николаев, 2026.
Данная статья распространяется на условиях международной лицензии Creative Commons License Attribution 4.0 International (CC BY 4.0).

Существующие подходы к преобразованию PDF (например, PDFX [2], CiTalO [3]) ориентированы на текстовые файлы. Для сканированных документов применяются нейросетевые детекторы [4] и OCR-системы [1], однако они не обеспечивают связывание формул с концептами. Гибридные методы семантической обработки [5] используют комбинацию текстового и структурного анализа, но также требуют качественного текстового слоя. Методы извлечения информации из PDF на основе правил [6] показывают хорошие результаты для структурированных документов, но при этом не адаптированы к математическому содержанию. Работы по обработке математических выражений [7–11] фокусируются на распознавании структуры, они не решают задачу установления смысловых связей между переменными и концептами в тексте. Методы извлечения изображений из научных статей [12] и поисковые системы, такие как Semantic Scholar [13], ориентированы на графические объекты, а не на семантику формул. Оценка качества OCR-систем для научных текстов [14] показывает, что даже современные решения допускают ошибки в математических нотациях. Методы семантической аннотации [15, 16] используют онтологии, но требуют структурированного входного формата.

В работе [17] был представлен прототип сервиса семантического поиска формул, имевший ограничения: поддержка только цифровых PDF, использование эвристик на основе расстояния для связывания, отсутствие асинхронности и авторизации. Настоящее исследование направлено на устранение этих недостатков, а также на формализацию и экспериментальную оценку методов связывания переменных с математическими концептами и главными формулами.

Научная задача состоит в разработке и оценке методов автоматического извлечения семантических связей между математическими объектами (переменными, формулами) и концептами из предметной области в неструктурированных документах, включая сканированные, с использованием лингвистических и структурных признаков.

Введем следующие определения:

- концепт – это единица знания из формальной онтологии (OntoMathPRO [18] или MSC [19]), имеющая уникальный идентификатор и название;
- математический концепт – концепт, относящийся к математической области (например, «группа», «производная»);

- главная формула – формула, выделенная в отдельный абзац, часто нумерованная, содержащая основное утверждение;
- локальная переменная – символ или выражение, встречающееся внутри текстового абзаца и используемое для обозначения параметра или вспомогательной величины;
- семантическая связь – это отношение между переменной и концептом или между переменной и главной формулой, устанавливаемое на основе контекстных и структурных правил.

Далее представлены архитектура прототипа сервиса, модуль оцифровки, формальные модели связывания, модуль визуализации и дана экспериментальная оценка.

АРХИТЕКТУРА ПРОТОТИПА СЕРВИСА

Разработанный прототип веб-сервиса построен как приложение на Flask с использованием асинхронной очереди задач Celery [20] и брокера Redis [21] для хранения состояния задач. Архитектура (рис. 1) включает следующие модули:

- авторизация – регистрация и аутентификация пользователей с подтверждением по электронной почте. Все операции, кроме поиска, требуют авторизации. Форма авторизации приведена на рис. 2 (слева);
- управление задачами – каждый пользователь имеет собственный набор задач с возможностью отслеживания статуса («выполняется», «завершена», «ошибка»). Страница управления задачами показана на рис. 2 (справа);
- обработка PDF – определяет тип документа (цифровой или сканированный) и запускает соответствующий конвейер;
- оцифровка – для сканированных документов добавляет текстовый слой и восстанавливает структуру параграфов;
- извлечение формул и переменных – выделяет главные формулы и локальные переменные;
- связывание – устанавливает связи между переменными и концептами, а также между переменными и главными формулами;
- визуализация – отображает семантическую структуру документа с возможностью фильтрации элементов.

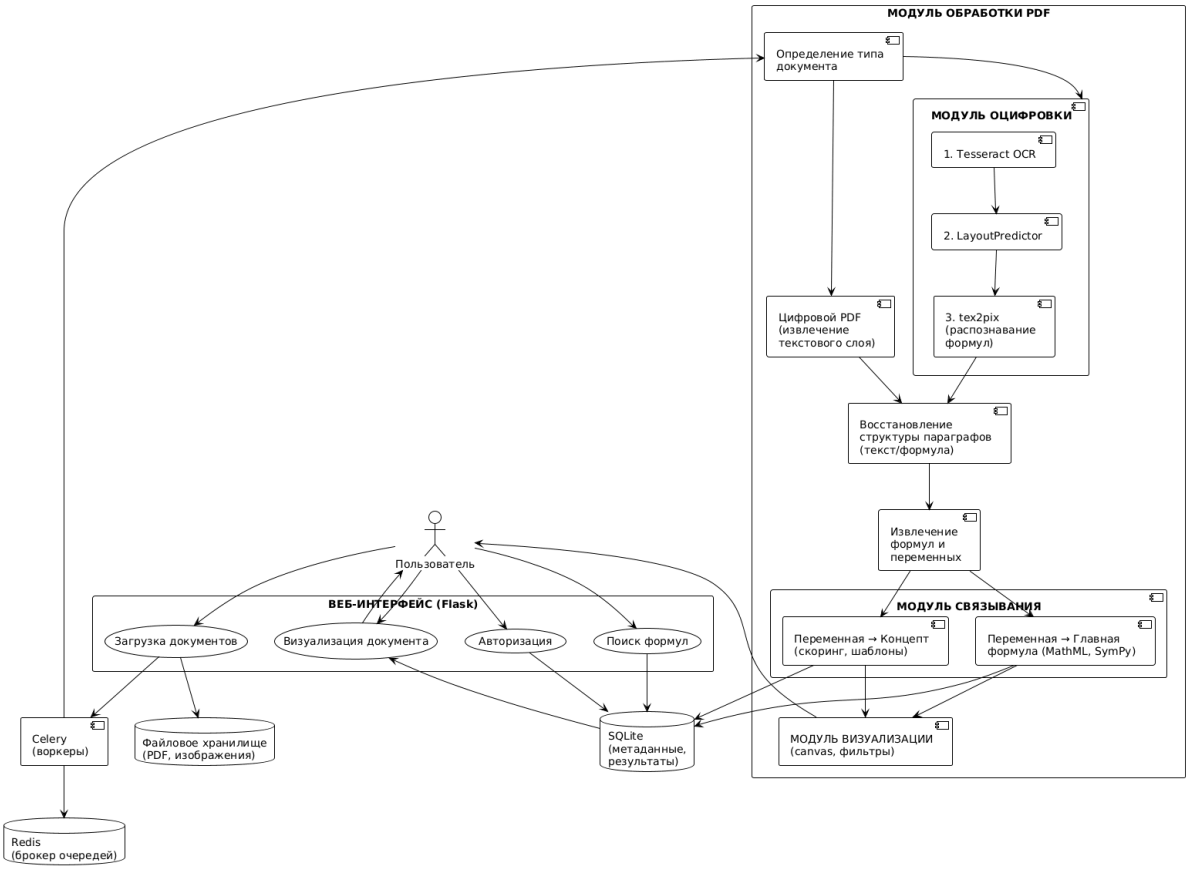


Рис. 1. Архитектура сервиса семантического поиска формул.

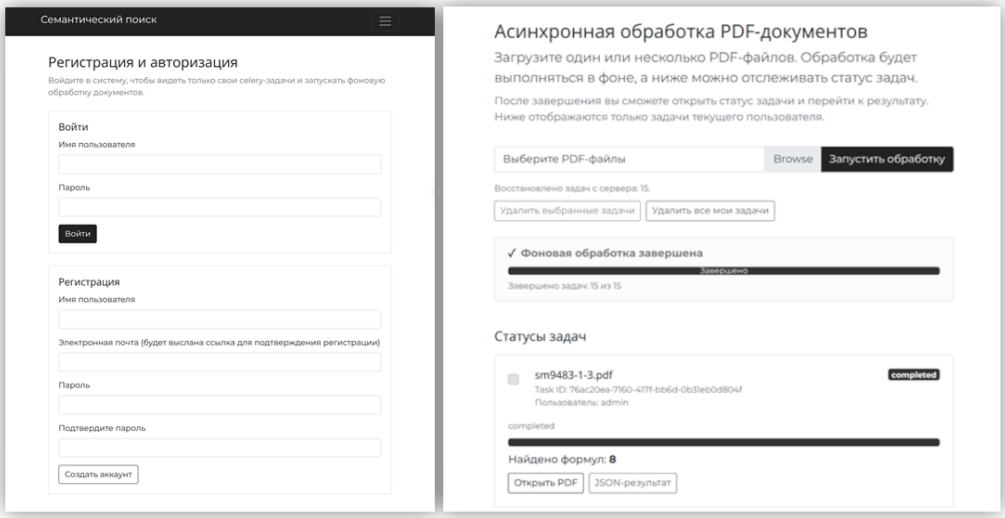


Рис. 2. Форма авторизации и регистрации с подтверждением через электронную почту (слева) и форма загрузки PDF-документов и отслеживания прогресса запущенных задач (справа).

Связь между модулями обработки и визуализации осуществляется через общую базу данных, где хранятся результаты аннотации.

Целевой аудиторией являются исследователи в области математики и компьютерных наук, сотрудники библиотек, издательства, заинтересованные в автоматической индексации и поиске по математическому содержанию.

Перечислим технические требования к вычислительной инфраструктуре: сервер с объемом оперативной памяти не менее 16 ГБ, хранилище на 500 ГБ; поддержка параллельной обработки до 10 документов одновременно; среднее время обработки одного документа – не более 5 мин для 20-страничной статьи.

Инфраструктура прототипа включает: веб-сервер на языке Python, базу данных SQLite для хранения метаданных и результатов, очередь задач Redis. Контейнеризация не используется для упрощения эксплуатации на этапе разработки.

Модуль поиска формул содержит поле поиска на основе коллекции математических концептов (рис. 3).

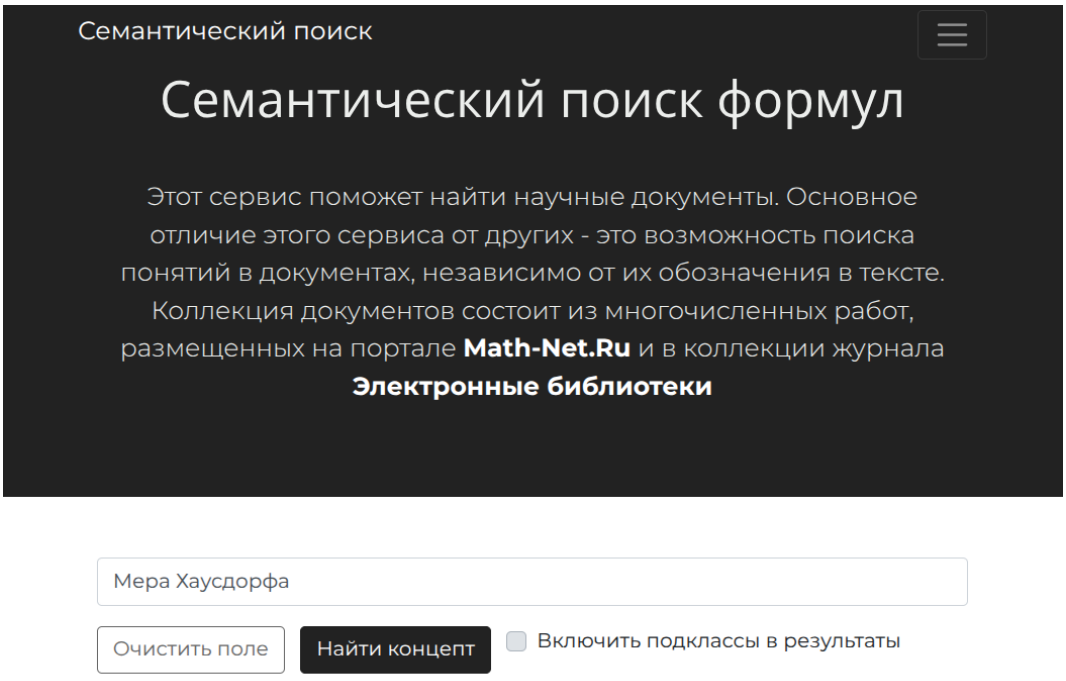


Рис. 3. Форма поиска формул по математическим концептам.

МОДУЛЬ ОЦИФРОВКИ СКАНИРОВАННЫХ ДОКУМЕНТОВ

Модуль предназначен для преобразования сканированных PDF в структурированный текстовый формат. Процесс включает три этапа.

1. Распознавание текста – с помощью ocrmypdf (обертка над Tesseract OCR [1]) создается текстовый слой с координатами строк. Используются русский и английский языки.

2. Восстановление структуры параграфов – применена модель LayoutPredictor из библиотеки surya-ocr, обученная на размеченных научных статьях. Модель выделяет прямоугольники, соответствующие текстовым блокам, формулам, колонтитулам и т. д. Точность определения границ параграфов составляет 0.92 по метрике IoU на тестовой выборке. Визуализация работы LayoutPredictor представлена на рис. 4.

3. Группировка строк – координаты строк из результата OCR сопоставляются с прямоугольниками параграфов; строки, попавшие в один прямоугольник, объединяются в параграф. Каждому параграфу присваивается тип: «текст», «формула», «другое».

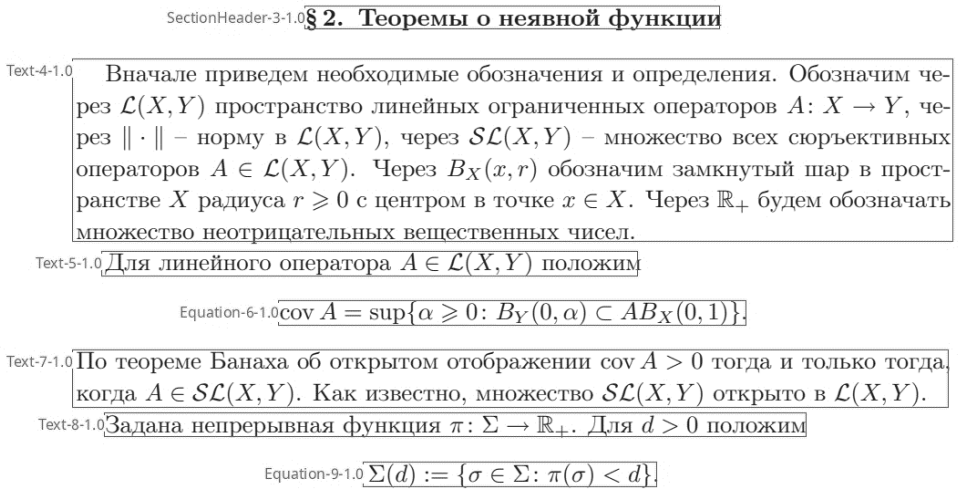


Рис. 4. Пример работы LayoutPredictor для определения границ параграфов и формул.

Для распознавания математических формул, выделенных на этапе 2, применяется библиотека tex2pix, которая принимает на вход изображение формулы и возвращает LaTeX-представление формулы с высокой точностью. Библиотека tex2pix основана на нейросетевой модели, обученной на размеченных наборах

изображений формул и соответствующих им LaTeX-последовательностей. Использование tex2pix позволяет значительно повысить качество распознавания сложных математических выражений по сравнению со стандартными OCR-системами, которые не специализированы на математической нотации. В рамках разработанного конвейера tex2pix вызывается для каждого блока, классифицированного как «формула»; полученное LaTeX-представление сохраняется в структурированном виде для дальнейшей семантической обработки, включая связывание переменных с концептами и главными формулами.

Ограничение метода состоит в том, что Tesseract не распознает сложные многоэтажные формулы (например, цепные дроби) с достаточной точностью. По результатам тестирования на 50 сканированных страницах с формулами среднего уровня сложности точность распознавания символов составила 0.78 для формул против 0.95 для обычного текста. Использование tex2pix позволяет повысить точность распознавания формул до 0.89 на том же наборе данных, что существенно улучшает качество последующего семантического связывания. Это, в свою очередь, накладывает ограничения на дальнейшее связывание: для документов с большим числом сложных формул качество работы снижается.

МОДУЛЬ ФОРМИРОВАНИЯ СЕМАНТИЧЕСКОЙ СТРУКТУРЫ

Формальная постановка задачи связывания переменных и концептов

Пусть имеются множество локальных переменных $V = \{v_1, \dots, v_n\}$, обнаруженных в тексте, и множество концептов $C = \{c_1, \dots, c_m\}$ из онтологии. Для каждой переменной v_i заданы ее левый и правый контексты – последовательности слов L_i и R_i (до и после переменной). Требуется установить бинарное отношение $S \subseteq V \times C$ такое, что $(v_i, c_j) \in S$, если концепт c_j семантически связан с переменной v_i.

Для решения использована функция $f(v_i, c_j)$, агрегирующая следующие признаки:

- вхождение названия концепта (в исходной или лемматизированной формах) в контекст: +1 балл;
- наличие маркеров определения в левом контексте («обозначим», «назовем», «введем обозначение», «через» и др.): +3 балла;

- наличие маркеров определения в правом контексте («это», «есть», «обозначает», «называется», «представляет собой»): +3 балла;
- одновременное присутствие концепта в правом контексте и маркера в левом («где», «здесь», «пусть»): +2 балла;
- присутствие концепта в контекстном окне (6 слов до и 6 после переменной): +2 балла.

Порог принятия решения $\tau = 3$: если $f(v_i, c_j) \geq 3$, то связь считается установленной.

Обоснование порога выбрано эмпирически на валидационной выборке из 30 документов как значение, максимизирующее F-меру.

Формальная постановка задачи связывания переменных и главных формул

Пусть имеется множество главных формул $F = \{f_1, \ldots, f_k\}$, каждая из формул представлена в виде LaTeX-выражения. Для каждой локальной переменной v_i также известно ее LaTeX-представление. Требуется определить отображение $g: V \to F \cup \{\emptyset\}$, где $g(v_i) = f_j$, если переменная входит в состав формулы f_j как параметр или составная часть.

Используем три критерия для установления связи между главной формулой и переменной.

1. Строковое вхождение – LaTeX-код переменной является подстрокой кода формулы.
2. Сравнение символов в MathML – оба выражения конвертируются в MathML согласно спецификации [8]; извлекаются неделимые символы (идентификаторы, числа, операторы); если множество символов переменной есть подмножество символов формулы – то кандидат принимается.
3. Семантический разбор через SymPy – выражения парсятся с помощью parse_latex из библиотеки sympy; сравниваются не только символы, но и структура операторов и функции.

Связывание главной формулы и переменной производится, если выполняется хотя бы один из критериев. Приоритет отдается критерию 3 как наиболее полному.

МОДУЛЬ ВИЗУАЛИЗАЦИИ СЕМАНТИЧЕСКОЙ СТРУКТУРЫ

Для представления результатов разработан модуль визуализации на основе веб-шаблонов Flask, использующих элементы canvas. На странице документа (рис. 5) отображаются:

- локальные переменные;
- главные формулы;
- связи переменных с главными формулами.

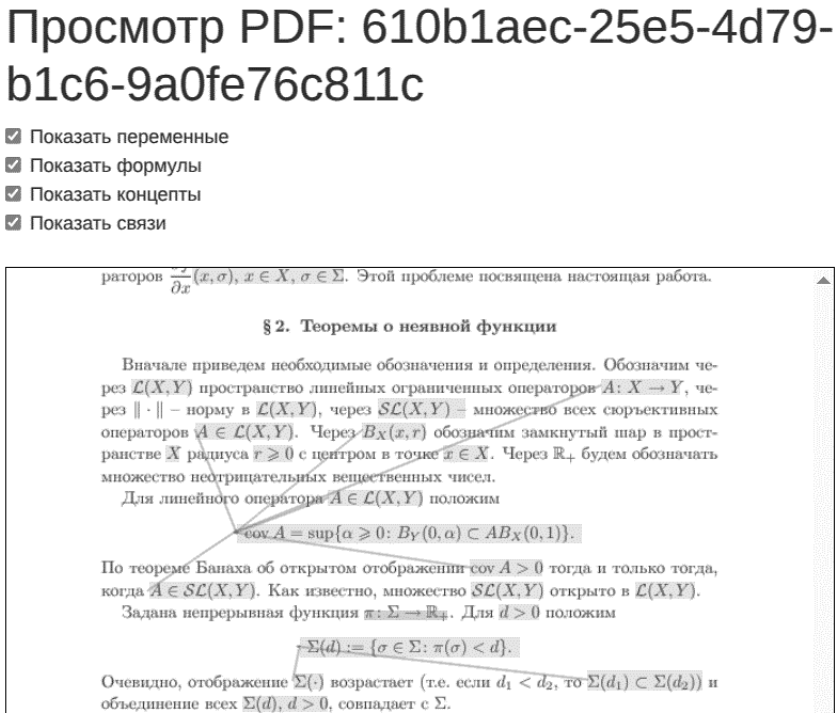


Рис. 5. Фрагмент страницы визуализации семантической структуры документа.

Пользователь может включать/отключать отображение каждой группы элементов через панель управления. Такой подход обеспечивает единообразие для всех браузеров, в отличие от встроенных PDF-аннотаций, поддержка которых различается в разных программах просмотра.

ЭКСПЕРИМЕНТАЛЬНАЯ ОЦЕНКА

Датасет и метрики

Для оценки качества была использована коллекция из 150 математических статей из открытых архивов, из них 50 – сканированные (без текстового слоя).

Все статьи размечены вручную по трем классам связей: переменная – концепт, переменная – главная формула.

Перечислим выбранные метрики:

- точность (Precision) – доля верно установленных связей среди всех установленных системой;
- полнота (Recall) – доля верно установленных связей среди всех истинных связей в разметке;
- F-мера – гармоническое среднее точности и полноты.

Результаты связывания переменных и концептов

В таблице 1 представлено сравнение качества связывания переменных с математическими концептами для двух подходов: эвристики на основе расстояния, использовавшейся в предыдущей версии прототипа, и предложенного скорингового метода, основанного на синтаксических шаблонах русского языка. Оценка выполнена на всей коллекции из 150 документов.

Табл. 1. Значения метрик для различных методов.

Метод	Precision	Recall	F-мера
Расстояние (по [17])	0.62	0.58	0.60
Предложенный скоринговый метод	0.81	0.79	0.80

Прирост F-меры составил 20 процентных пунктов. Для сканированных документов точность ниже (0.72) из-за ошибок OCR, однако метод сохраняет работоспособность благодаря использованию лемматизации и множественных контекстных признаков.

Результаты связывания переменных и главных формул

В таблице 2 приведено сравнение двух подходов к связыванию локальных переменных с главными формулами: с использованием только сравнения символов в MathML и с добавлением семантического разбора через SymPy. Использование SymPy позволяет учитывать не только символьный состав формулы, но и её структуру, что особенно важно для различения семантически разных выражений с одинаковыми символами (например, $f(x)$ и $f \cdot x$).

Табл. 2. Сравнение значений метрик точности, полноты и F-меры для различных подходов к связыванию переменных и главных формул.

Подход	Precision	Recall	F-мера
Только сравнение MathML	0.68	0.65	0.66
С преобразованием в SymPy	0.85	0.82	0.83

Использование SymPy существенно улучшает распознавание параметров формул, обеспечивая прирост F-меры на 17 процентных пунктов.

Границы применимости

На основе анализа ошибок определены следующие границы применимости разработанных методов.

Система эффективна (F-мера ≥ 0.70) для документов, где

- OCR-качество текста ≥ 0.90 (по символьной точности);
- формулы имеют стандартную структуру (не содержат редких нотаций, например нотации Кнута или специальных операторов);
- язык документа – русский.

Для документов с большим количеством сложных формул (например, многоуровневые индексы, матрицы, длинные цепные дроби) качество падает до 0.55–0.60. Это связано с ограничениями Tesseract при распознавании двумерных структур и ограничениями парсера LaTeX в SymPy. Использование tex2pix для распознавания формул позволяет повысить точность распознавания до 0.89 для формул среднего уровня сложности, однако для крайне сложных выражений (например, многострочные системы уравнений или матрицы большого размера) точность tex2pix также снижается до 0.72–0.78, что приводит к ошибкам в извлечении LaTeX-представления и последующем семантическом связывании. Кроме того, tex2pix, как и любая нейросетевая модель, чувствителен к качеству входного изображения: низкое разрешение, шумы, артефакты сканирования или наклон страницы могут существенно ухудшить результат распознавания. Перспективным решением видится использование специализированных моделей распознавания формул на основе графов [11] или нейросетевых подходов с увеличенным объемом обучающих данных для сложных нотаций.

Система также не применима для документов, содержащих формулы исключительно в виде изображений без текстового контекста, поскольку для связывания необходим лингвистический контекст. Для таких документов требуется дополнительный этап извлечения контекстных предложений из окружающего текста, что не всегда возможно в случае отсутствия связного текстового описания формулы.

ЗАКЛЮЧЕНИЕ

Решена задача создания сквозного конвейера для семантической обработки математических PDF-документов, включая сканированные. Предложены формальные модели и методы связывания переменных с математическими концептами и главными формулами. Разработанный конвейер объединяет несколько ключевых компонентов: оптическое распознавание текста на основе Tesseract, восстановление логической структуры документа с помощью LayoutPredictor, распознавание математических формул с использованием tex2pix, а также лингвистический анализ для установления семантических связей на основе синтаксических шаблонов русского языка.

Экспериментально показано значительное повышение качества по сравнению с предыдущей версией прототипа: прирост F-меры для связывания переменной и концепта составил 20 п.п. (с 0.60 до 0.80), для связывания переменных с главными формулами – 17 п.п. (с 0.66 до 0.83). Внедрение tex2pix позволило повысить точность распознавания формул в сканированных документах с 0.78 до 0.89 для выражений среднего уровня сложности, что непосредственно улучшило качество последующего семантического связывания. Определены границы применимости: система эффективна для документов со стандартной структурой формул и качеством OCR не ниже 0.90; для сложных двумерных формул (многострочные системы, матрицы большого размера) качество снижается, что требует дальнейших исследований.

Архитектурные решения (асинхронная очередь задач на базе Celery и Redis, модульная структура, система авторизации) обеспечивают масштабируемость и готовность прототипа к промышленному развертыванию. Разработан-

ный модуль визуализации с использованием canvas предоставляет интерактивный инструмент для анализа семантической структуры документов, что может быть полезно для исследователей и сотрудников библиотек.

Благодарности

Работа выполнена в рамках государственного задания НИЦ «Курчатовский институт».

СПИСОК ЛИТЕРАТУРЫ

1. *Smith R.* An overview of the tesseract OCR engine // Proc. Int. Conf. Doc. Anal. Recognition, ICDAR. 2007. Vol. 2. P. 629–633. https://doi.org/10.1109/ICDAR.2007.4376991

2. *Constantin A., Pettifer S., Voronkov A.* PDFX: Fully-automated PDF-to-XML conversion of scientific literature // DocEng 2013 – Proc. 2013 ACM Symp. Doc. Eng. 2013. P. 177–180. https://doi.org/10.1145/2494266.2494271

3. *Ciancarini P., Iorio A. Di, Nuzzolese A. G., Peroni S., Vitali F.* Semantic annotation of scholarly documents and citations // Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 2013. Vol. 8249 LNAI. P. 336–347. https://doi.org/10.1007/978-3-319-03524-6_29

4. *Amethiya Y., Bajwa G.* Automatic Table Detection and Tabular Data Extraction from Scanned Documents // Smart Innovation, Systems and Technologies. 2025. P. 75–87. https://doi.org/10.1007/978-981-96-0143-1_7

5. *Bertin M., Atanassova I.* Hybrid Approach for the Semantic Processing of Scientific Papers // Semant. Publ. Chall. Track 11 th Eur. Semant. Web Conf. (ESWC 2014). 2014. https://doi.org/10.1007/978-981-96-0143-1_7

6. *Ahmad R., Afzal M. T., Qadir M. A.* Information extraction from PDF sources based on rule-based system using integrated formats // Commun. Comput. Inf. Sci. 2016. Vol. 641. P. 293–308. https://doi.org/10.1007/978-3-319-46565-4_23

7. *Schubotz M., Greiner-Petter A., Scharpf P., Meuschke N., Cohl H. S., Gipp B.* Improving the Representation and Conversion of Mathematical Formulae by Considering their Textual Context // Proc. ACM/IEEE Jt. Conf. Digit. Libr. 2018. P. 233–242. https://doi.org/10.1145/3197026.3197058

8. Mathematical Markup Language (MathML) Version 3.0 2nd Edition. W3C Recommendation, 10 April 2014. URL: https://www.w3.org/TR/MathML3/

9. *Greiner-Petter A., Youssef A., Ruas T., Miller B. R., Schubotz M., Aizawa A., Gipp B.* Math-word embedding in math search and semantic extraction // Scientometrics. 2020. Vol. 125, No. 3. P. 3017–3046. https://doi.org/10.1007/s11192-020-03502-9

10. *Nevzorova O., Kirillovich A., Nevzorov V., Nikolaev K.* The semantic context models of mathematical formulas in scientific papers // CEUR Workshop Proc. 2018. Vol. 2277. P. 33–40.

11. *Shah A. K., Dey A., Zanibbi R.* A Math Formula Extraction and Evaluation Framework for PDF Documents // Lecture Notes in Computer Science. 2021. P. 19–34. https://doi.org/10.1007/978-3-030-86331-9_2

12. *Siegel N., Lourie N., Power R., Ammar W.* Extracting Scientific Figures with Distantly Supervised Neural Networks // Proc. ACM/IEEE Jt. Conf. Digit. Libr. 2018. P. 223–232. https://doi.org/10.1145/3197026.3197040

13. Semantic Scholar. URL: https://www.semanticscholar.org/

14. *Neudecker C., Baierer K., Gerber M., Clausner C., Antonacopoulos A., Pletschacher S.* A survey of OCR evaluation tools and metrics // The 6th International Workshop on Historical Document Imaging and Processing. New York, NY, USA: ACM, 2021. P. 13–18. https://doi.org/10.1145/3476887.3476888

15. *Jaradeh M. Y., Oelen A., Farfar K. E., Prinz M., D'Souza J., Kismihók G., Stocker M., Auer S.* Open research knowledge graph: Next generation infrastructure for semantic scholarly knowledge // K-CAP 2019 – Proceedings of the 10th International Conference on Knowledge Capture. New York, NY, USA: ACM, 2019. P. 243–246. https://doi.org/10.1145/3360901.3364435

16. *Peroni S., Shotton D.* FaBiO and CiTO: Ontologies for describing bibliographic resources and citations // J. Web Semant. 2012. Vol. 17. P. 33–43. https://doi.org/10.1016/j.websem.2012.08.001

17. *Николаев К.С.* Сервис семантического поиска формул по коллекции математических PDF-документов // Информационные технологии и вычислительные системы. 2025. № 3. С. 34–43. https://doi.org/10.14357/20718632250304

18. *Kirillovich A. V., Nevzorova O. A., Lipachev E. K.* OntoMathPRO 2.0 Ontology: Updates of Formal Model // Lobachevskii J. Math. 2022. Vol. 43, No. 12. P. 3504–3514. https://doi.org/10.1134/S1995080222150136

19. MSC2020 –Mathematics Subject Classification System. URL: https://msc2020.org/.

20. Celery –Distributed Task Queue. URL: https://docs.celeryq.dev/.

21. Redis – The real-time context engine for AI apps. URL: https://redis.io/.

PROCESSING OF SCANNED MATHEMATICAL PDF DOCUMENTS IN THE SEMANTIC FORMULA SEARCH SERVICE

K. S. Nikolaev[0000-0003-3204-238X]

National Research Centre "Kurchatov Institute", Moscow, Russia

konnikolaeff@yandex.ru

Abstract

The article addresses the problem of extracting semantic relations between mathematical formulas and concepts from unstructured PDF documents, including scanned ones. A complete processing pipeline is proposed that combines optical character recognition, paragraph structure recovery, formula and variable extraction, and their linking to concepts using syntactic patterns of the Russian language. Formal models are developed for two types of relations: "variable-concept" and "variable-main formula". Experimental evaluation on a collection of 150 mathematical articles shows a significant improvement in linking accuracy, with an increase of 18 percentage points in F-measure compared to the previous version. Architectural improvements of the web service prototype are described, providing asynchronous processing and scalability.

Keywords*: semantic formula search, PDF documents, scientific libraries, ontologies, web service.*

REFERENCES

1. *Smith R.* An overview of the tesseract OCR engine // Proc. Int. Conf. Doc. Anal. Recognition, ICDAR. 2007. Vol. 2. P. 629–633. https://doi.org/10.1109/ICDAR.2007.4376991

2. *Constantin A., Pettifer S., Voronkov A.* PDFX: Fully-automated PDF-to-XML conversion of scientific literature // DocEng 2013 – Proc. 2013 ACM Symp. Doc. Eng. 2013. P. 177–180. https://doi.org/10.1145/2494266.2494271

3. *Ciancarini P., Iorio A. Di, Nuzzolese A. G., Peroni S., Vitali F.* Semantic annotation of scholarly documents and citations // Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 2013. Vol. 8249 LNAI. P. 336–347. https://doi.org/10.1007/978-3-319-03524-6_29

4. *Amethiya Y., Bajwa G.* Automatic Table Detection and Tabular Data Extraction from Scanned Documents // Smart Innovation, Systems and Technologies. 2025. P. 75–87. https://doi.org/10.1007/978-981-96-0143-1_7

5. *Bertin M., Atanassova I.* Hybrid Approach for the Semantic Processing of Scientific Papers // Semant. Publ. Chall. Track 11 th Eur. Semant. Web Conf. (ESWC 2014). 2014. https://doi.org/10.1007/978-981-96-0143-1_7

6. *Ahmad R., Afzal M. T., Qadir M. A.* Information extraction from PDF sources based on rule-based system using integrated formats // Commun. Comput. Inf. Sci. 2016. Vol. 641. P. 293–308. https://doi.org/10.1007/978-3-319-46565-4_23

7. *Schubotz M., Greiner-Petter A., Scharpf P., Meuschke N., Cohl H. S., Gipp B.* Improving the Representation and Conversion of Mathematical Formulae by Considering their Textual Context // Proc. ACM/IEEE Jt. Conf. Digit. Libr. 2018. P. 233–242. https://doi.org/10.1145/3197026.3197058

8. Mathematical Markup Language (MathML) Version 3.0 2nd Edition. W3C Recommendation, 10 April 2014. URL: https://www.w3.org/TR/MathML3/

9. *Greiner-Petter A., Youssef A., Ruas T., Miller B. R., Schubotz M., Aizawa A., Gipp B.* Math-word embedding in math search and semantic extraction // Scientometrics. 2020. Vol. 125, No. 3. P. 3017–3046. https://doi.org/10.1007/s11192-020-03502-9

10. *Nevzorova O., Kirillovich A., Nevzorov V., Nikolaev K.* The semantic context models of mathematical formulas in scientific papers // CEUR Workshop Proc. 2018. Vol. 2277. P. 33–40.

11. *Shah A. K., Dey A., Zanibbi R.* A Math Formula Extraction and Evaluation Framework for PDF Documents // Lecture Notes in Computer Science. 2021. P. 19–34. https://doi.org/10.1007/978-3-030-86331-9_2

12. *Siegel N., Lourie N., Power R., Ammar W.* Extracting Scientific Figures with Distantly Supervised Neural Networks // Proc. ACM/IEEE Jt. Conf. Digit. Libr. 2018. P. 223–232. https://doi.org/10.1145/3197026.3197040

13. Semantic Scholar. URL: https://www.semanticscholar.org/

14. *Neudecker C., Baierer K., Gerber M., Clausner C., Antonacopoulos A., Pletschacher S.* A survey of OCR evaluation tools and metrics // The 6th International Workshop on Historical Document Imaging and Processing. New York, NY, USA: ACM, 2021. P. 13–18. https://doi.org/10.1145/3476887.3476888

15. *Jaradeh M. Y., Oelen A., Farfar K. E., Prinz M., D'Souza J., Kismihók G., Stocker M., Auer S.* Open research knowledge graph: Next generation infrastructure for semantic scholarly knowledge // K-CAP 2019 – Proceedings of the 10th International Conference on Knowledge Capture. New York, NY, USA: ACM, 2019. P. 243–246. https://doi.org/10.1145/3360901.3364435

16. *Peroni S., Shotton D.* FaBiO and CiTO: Ontologies for describing bibliographic resources and citations // J. Web Semant. 2012. Vol. 17. P. 33–43. https://doi.org/10.1016/j.websem.2012.08.001

17. *Nikolaev K.S.* Servis semanticheskogo poiska formul po kollekcii matematicheskih PDF-dokumentov // Informacionnye tehnologii i vychisli-tel'nye sistemy. 2025. № 3. S. 34–43. https://doi.org/10.14357/20718632250304

18. *Kirillovich A. V., Nevzorova O. A., Lipachev E. K.* OntoMathPRO 2.0 Ontology: Updates of Formal Model // Lobachevskii J. Math. 2022. Vol. 43, No. 12. P. 3504–3514. https://doi.org/10.1134/S1995080222150136

19. MSC2020 –Mathematics Subject Classification System. URL: https://msc2020.org/.

20. Celery –Distributed Task Queue. URL: https://docs.celeryq.dev/.

21. Redis – The real-time context engine for AI apps. URL: https://redis.io/.

СВЕДЕНИЯ ОБ АВТОРЕ



НИКОЛАЕВ Константин Сергеевич – младший научный сотрудник Лаборатории суперкомпьютерного моделирования Отделения суперкомпьютерных систем и параллельных вычислений ФГБУ «Национальный исследовательский центр «Курчатовский институт». Основные направления научных исследований: обработка естественного языка, искусственный интеллект.

NIKOLAEV Konstantin Sergeevich – junior researcher at the Supercomputer Modeling Laboratory of the Department of Supercomputer Systems and Parallel Computing of the Federal State Budgetary Institution "National Research Center "Kurchatov Institute". Major fields of scientific research are natural language processing, artificial intelligence.

email: konnikolaeff@yandex.ru
ORCID: 0000-0003-3204-238X

Материал поступил в редакцию 30 июня 2026 года

УДК 004.912+004.8+004.42+004.414.3

МЕТОД СЕМАНТИЧЕСКОГО АНАЛИЗА ИЗМЕНЕНИЙ ТРЕБОВАНИЙ К ПРОГРАММНОМУ ОБЕСПЕЧЕНИЮ В ВЕРСИЯХ ТЕХНИЧЕСКОЙ ДОКУМЕНТАЦИИ

А. С. Савинчева[1] [0009-0005-0756-6001], **З. А. Еникеева**[2] [0000-0002-2468-7455],

К. Ш. Еникеев[3] [0009-0000-6020-9912]

[1–3]*Казанский (Приволжский) федеральный университет, г. Казань, Россия*

[1]asanvlit@gmail.com, [2]ZAEnikeeva@kpfu.ru, [3]KSEnikeev@kpfu.ru

Аннотация

Представлен формализованный метод автоматического обнаружения изменений требований к программному обеспечению в процессе эволюции версий технической документации. Исходными данными служат множества требований, извлеченных и классифицированных на функциональные и нефункциональные, с дополнительной детализацией типов нефункциональных требований (производительность, безопасность, доступность и др.) для каждой версии документации. Задача сопоставления требований между версиями сведена к построению частичного отображения между множествами требований и присвоению каждому требованию статуса: неизмененное (Unchanged), измененное (Modified), новое (New) или удаленное (Deleted). Предложена композитная функция сходства пар требований, объединяющая семантическое сходство на основе векторных представлений (эмбеддингов), лексическое сходство, совпадение класса требования и структурный контекст документа. Представлены результаты применения метода на открытом наборе пар требований. Разработан прототип системы анализа изменений требований в технических документациях.

Ключевые слова*: требования к программному обеспечению, анализ изменений требований, сопоставление версий документации, прослеживаемость требований, семантическое сходство, метод взаимных ближайших соседей, метод Отсу.*

© А. С. Савинчева, З. А. Еникеева, К. Ш. Еникеев, 2026.
Данная статья распространяется на условиях международной лицензии Creative Commons License Attribution 4.0 International (CC BY 4.0).

ВВЕДЕНИЕ

Эволюция технической документации программных систем сопровождается итеративными правками: уточнением формулировок, добавлением новых ограничений, исключением устаревших требований, а также изменением числовых параметров. Для поддержки процессов управления конфигурацией, верификации и сопровождения необходимо не только зафиксировать факт изменения текста, но и интерпретировать его в терминах изменений требований к программному обеспечению (ПО) [1, 2]. Под требованием к ПО будем понимать формализуемое утверждение на естественном языке о свойствах программной системы, ее функциях, ограничениях или условиях эксплуатации, изменение которого влияет на стадии жизненного цикла системы (проектирование, тестирование, приемо-сдаточные испытания и др.).

В большинстве современных систем работы с документацией механизмы анализа изменений опираются на посимвольное или построчное сравнение текста [3, 4]. Такой подход фиксирует все правки, но не обеспечивает их семантическую интерпретацию. Исправление опечатки, перефразирование предложения и изменение критического числового параметра (например, времени отклика сервера) выглядят для подобного классического алгоритма как равнозначные правки. В результате специалист сталкивается с большим объемом изменений, где множество редакторских правок затрудняет процесс выделения именно тех, которые затрагивают содержание требований к ПО.

Альтернативным подходом является анализ изменений через призму уже выделенных требований. Если для каждой версии документации получены множества требований с присвоенными классами (функциональные, нефункциональные по категориям), то задача сводится к сопоставлению этих множеств и определению для каждого требования статуса по отношению к другой версии – неизмененное, измененное, новое или удаленное. Такой подход согласуется с практиками прослеживаемости требований (requirements traceability) и управления изменениями в инженерии требований [1, 5, 6].

В рамках отдельного исследования были разработаны и экспериментально оценены методы извлечения требований из технической документации и их последующей классификации [7]. Такое извлечение было реализовано на основе двунаправленной нейронной сети с долговременной краткосрочной памятью

(Bidirectional Long Short-Term Memory, BiLSTM) и слоя условных случайных полей (Conditional Random Field, CRF), что в совокупности обеспечивает учет контекстуальных зависимостей в последовательности предложений. Классификация требований устроена по таксономии из 13 классов (функциональные, 11 типов нефункциональных требований и второстепенный текст, не содержащий требований) на основе полносвязной нейронной сети. Для задач извлечения и последующей классификации достигнуты значения F1-score macro порядка 0.85. В настоящей работе указанные алгоритмы были использованы в готовом виде для получения множеств требований в различных версиях документации. Цель настоящего исследования состоит в проектировании и разработке метода автоматического обнаружения изменений требований к ПО между версиями технической документации, с присвоенным статусом изменения.

1. ФОРМАЛИЗАЦИЯ ЗАДАЧИ СОПОСТАВЛЕНИЯ ТРЕБОВАНИЙ МЕЖДУ ВЕРСИЯМИ

Пусть имеются две версии технической документации – предыдущая и новая, в общем случае не обязательно последовательные. В результате применения методов извлечения и классификации требований для каждой версии получено множество требований. Каждое требование задано текстовым фрагментом (формулировкой требования) и меткой из фиксированной таксономии классификации C ={Functional, Availability, Legal, Look & Feel, Maintainability, Operational, Performance, Scalability, Security, Usability, Fault tolerance, Portability, Other}. Обозначим множество требований предыдущей версии как $R_{\text{old}} = \{r_1, r_2, \ldots, r_n\}$, а новой версии – как $R_{\text{new}} = \{r'_1, r'_2, \ldots, r'_m\}$. Под сопоставлением требований между версиями будем понимать установление соответствий между элементами R_{old} и R_{new} с целью выявления того, какое требование новой версии (если оно есть) соответствует данному требованию старой версии, и наоборот.

В такой постановке задача обнаружения изменений формулируется следующим образом. Для каждого требования r'_j из R_{new} необходимо определить его статус по отношению к R_{old}:

- неизмененное (Unchanged): в R_{old} существует требование r_i того же класса и с семантически эквивалентным содержанием, при этом допустимы различия в формулировке (например, пунктуация, перефразирование);
- измененное (Modified): в R_{old} существует требование-прообраз r_i, однако содержание требования претерпело изменения (например, изменились числовой параметр или тип требования);
- новое (New): соответствующего прообраза в R_{old} нет, требование добавлено в новую версию.

Для каждого требования r_i из R_{old} определим его актуальность в новой версии. Требование считается удаленным (Deleted), если в R_{new} не найдено требования, сопоставленного с r_i, что интерпретируется как исключение данной функциональности или ограничения из спецификации.

Таким образом, выходом метода является частичное отображение $M \subseteq R_{\text{old}} \times R_{\text{new}}$ – множество сопоставленных пар (r_i, r_j') между версиями, с присвоенным каждому элементу статусом из множества

$$\{\text{Unchanged}, \text{Modified}, \text{New}, \text{Deleted}\}.$$

2. КОМПОЗИТНАЯ ФУНКЦИЯ СХОДСТВА ТРЕБОВАНИЙ

Ключевым элементом метода в предлагаемом подходе является композитный балл сходства $S(i,j)$, вычисляемый для каждой пары требований (r_i, r_j'). Нами использован композитный балл, объединяющий несколько различных признаков, что позволяет учитывать смысловую близость формулировок, лексическое совпадение, совпадение типа требования и структурный контекст документа. Согласно проведенным исследованиям, наиболее эффективными признаются гибридные подходы, сочетающие в себе признаки различных уровней – от строкового сопоставления до учета структурной организации информации [8].

Для каждого требования вычислим векторное представление $e(r) = \text{SBERT}(text(r))$ с помощью модели SBERT, позволяющей получать семантические контекстуализированные эмбеддинги предложений [9]. Семантическое сходство двух требований определим как косинусное сходство их эмбеддингов: $s_{\text{sem}}(i,j) = \cos(e(r_i), e(r_j'))$. Данный показатель отражает близость смыслового содержания требований и устойчив к перефразированию текста [9].

Для учета лексических изменений в формулировках введем коэффициент Жаккара по множествам токенов требований. Пусть *W*(*r*) – множество токенов, полученных из текста требования после предобработки (приведение к нижнему регистру, удаление знаков пунктуации). Тогда показатель лексического сходства s_{lex} определяется как:

$$s_{\text{lex}}(i,j) = \frac{|W(r_i) \cap W(r_j')|}{|W(r_i) \cup W(r_j')|}.$$

Этот показатель отражает степень совпадения токенов в формулировках требований и позволяет различать требования со схожей семантикой, но с различной терминологией [10].

Дополнительным признаком является совпадение классов требований. Пусть $\text{class}(r)$ – метка класса требования из таксономии C. Определим $s_{\text{class}}(i,j) = 1$, если $\text{class}(r_i) = \text{class}(r_j')$, и 0 в противном случае. Совпадение класса усиливает вероятность того, что два требования являются одной и той же сущностью в двух версиях.

Учтем также структурный контекст документа. В технической документации требования, как правило, группируются по разделам (функциональным областям или качественным характеристикам). Учет структурного контекста важен в случаях кратких или универсальных формулировок, которые могут повторяться в различных разделах. Например, требование вида «сервер должен отвечать за 1 секунду» в новых версиях документа может быть помещено в различные разделы (например, в раздел, описывающий скорость поиска, или раздел, посвященный созданию функциональных элементов приложения). Тогда пусть h_i и h_j – заголовки разделов, в которых расположены требования r_i и r_j' соответственно. Сходство заголовков s_{hdr} аналогично определяется по эмбеддингам SBERT как

$$s_{\text{hdr}}(i,j) = \cos(e(h_i), e(h_j)).$$

Нормализованную позицию требования в разделе зададим как $\text{pos}(r) = \frac{k_r}{N_r}$, где k_r – порядковый номер требования в разделе, а N_r – общее число требований в этом разделе. Значение $\text{pos}(r)$ таким образом лежит в отрезке [0, 1]. Близость расположения требований s_{pos} в соответствующих разделах равна

$$s_{\text{pos}}(i,j) = 1 - |\ \text{pos}(r_i) - \text{pos}(r_j')\ |.$$

Чем ближе позиции, тем больше значение s_{pos}. Структурный контекст объединим в один признак:

$$s_{\text{ctx}}(i,j) = \lambda_1 s_{\text{hdr}}(i,j) + \lambda_2 s_{\text{pos}}(i,j).$$

В реализованном варианте были установлены значения $\lambda_1 = 0.7$ и $\lambda_2 = 0.3$. Более высокий вес показателя s_{hdr} позволяет снизить влияние случайных перестановок требований внутри раздела, которые могут возникнуть при редактировании документации.

Итоговую функцию сходства зададим взвешенной суммой с понижающим коэффициентом penalty при несовпадении числовых параметров:

$$S(i,j) = (\alpha s_{\text{sem}} + \beta s_{\text{lex}} + \gamma s_{\text{class}} + \delta s_{\text{ctx}}) \cdot \text{penalty}.$$

Использование penalty обусловлено тем, что изменение численного параметра (например, времени отклика с 3 до 1 с) изменяет суть требования, даже если семантический вектор остается почти неизменным. Веса зададим как $\alpha = 0.7$, $\beta = 0.1$, $\gamma = 0.1, \delta = 0.1$ из определения иерархии признаков, семантика рассмотрена в качестве базиса сопоставления [8].

3. ПОСТРОЕНИЕ ОТОБРАЖЕНИЯ И ПРИСВОЕНИЕ СТАТУСА СОПОСТАВЛЕННЫМ ТРЕБОВАНИЯМ

После вычисления матрицы сходства $S(i,j)$ для всех пар (r_i, r_j') необходимо построить отображение M и присвоить статусы его элементам. Задача сопоставления допускает неполное покрытие: не каждое требование из R_{old} обязано иметь пару в R_{new} (статус Deleted), и не каждое из R_{new}– в R_{old} (статус New). Венгерский алгоритм решения задачи стремится к полному покрытию при создании пар, что в данном контексте может привести к сопоставлению семантически не связанных требований [11]. В связи с этим нами использован метод взаимных ближайших соседей (Mutual Nearest Neighbors, MNN). Пара (r_i, r_j') включается в M, только если r_j' является ближайшим соседом r_i в R_{new} по баллу S, и одновременно r_i является ближайшим соседом r_j' в R_{old}, при этом $S(i,j) \geq \theta_{\text{match}}$. Под порогом сопоставления θ_{match} в дальнейшем понимаем минимальный балл сход-

ства, при котором пара требований допускается в отображение M. Таким образом, в M попадают исключительно взаимно наилучшие совпадения с достаточным уровнем сходства. Остальные требования не образуют пары и получают статус New (для R_{new}) или Deleted (для R_{old}).

Для разграничения неизмененных и измененных требований введем порог неизменности θ_{unch}. Это минимальный балл $S(i,j)$ для пар, уже включенных в M, при превышении которого сопоставленной паре присваивается статус Unchanged. При $S(i,j) < \theta_{\text{unch}}$ пара получает статус Modified.

Оба порога (θ_{match} и θ_{unch}) подбираем автоматически по распределению вычисленных баллов сходства. При наличии как новых, так и измененных/неименных пар требований распределение баллов $S(i,j)$, как правило, имеет бимодальный характер: одна мода сосредоточена в области высоких значений (сопоставленные требования), другая – в области низких (требования, не имеющие пары). Для разделения мод применим метод Отсу, позволяющий определить оптимальный порог бинарного разделения, максимизируя межклассовую дисперсию (и минимизируя внутриклассовую) [12]. Этот метод не требует априорных предположений о виде распределения и адаптируется к конкретным данным, что необходимо для учета разнообразия стилей документации [12]. Для θ_{match} анализируем распределение максимальных баллов сходства для каждого требования из R_{new} (максимум по r_i). Для θ_{unch} проанализируем распределение баллов по парам, уже включенным в M.

Итоговую классификацию по статусам выполним по следующим правилам:

- New: требование r_j' из R_{new} не имеет пары в M;
- Deleted: требование r_i из R_{old} не имеет пары в M;
- Unchanged: пара $(r_i, r_j') \in M$ и $S(i,j) \geq \theta_{\text{unch}}$;
- Modified: пара $(r_i, r_j') \in M$ и $\theta_{\text{match}} \leq S(i,j) < \theta_{\text{unch}}$.

4. ПРИМЕНЕНИЕ МЕТОДА К ТЕХНИЧЕСКОЙ ДОКУМЕНТАЦИИ

Для иллюстрации работы метода рассмотрим два сконструированных фрагмента технической документации в двух версиях. Первая версия имеет вид

```
# Authentication
The system shall authenticate users using a username and password.
The system should process user requests within 3 seconds.
```

The system must log all authentication attempts.

Вторая версия содержит изменения по сравнению с первой:

User Authentication

The system shall authenticate users using a username and password.

The system should process user requests within 2 seconds.

The system should lock a user account after five failed login attempts.

Первая версия содержит раздел Authentication с тремя требованиями: 1) аутентификация по логину и паролю; 2) обработка запросов пользователя в течение 3 с; 3) логирование попыток аутентификации. Новая версия содержит раздел User Authentication с тремя требованиями: аутентификация по логину и паролю (формулировка без изменений); обработка запросов в течение 2 с (изменен числовой параметр); блокировка учетной записи после пяти неудачных попыток входа (новое требование). Третье требование предыдущей версии в новой версии отсутствует.

После извлечения и классификации требований для каждой версии алгоритм вычисляет компоненты $s_{\text{sem}}, s_{\text{lex}}, s_{\text{class}}, s_{\text{ctx}}$ и итоговый $S(i,j)$ для всех пар (табл. 1).

- Для пары (r_2, r_2') семантическое и лексическое сходства высоки, класс совпадает (нефункциональное требование, Performance), структурный контекст близок. Из-за различия числовых значений применен penalty. Итоговый балл оказался ниже порога неизменности, и паре присвоен статус Modified.
- Пара (r_1, r_1') получила высокий балл $S(i,j)$ и соответствующий статус Unchanged благодаря семантическому сходству названий разделов.
- Требование r_3 не нашло прообраза в R_{old}, ему присвоен статус New. Аналогично требование r_3' не нашло пары в R_{new}, его статусом стал Deleted.

Табл. 1. Вычисленные алгоритмом значения $s_{\text{sem}}, s_{\text{lex}}, s_{\text{class}}, s_{\text{ctx}}$ в примере.

Номер требования из R_{new}	Номер требования из R_{old}		
	1	2	3
1	1.00; 1.00; 1.0; 0.95	0.31; 0.12; 0.0; 0.85	0.46; 0.16; 1.0; 0.75
2	0.30; 0.12; 0.0; 0.85	0.98; 0.80; 1.0; 0.95	0.32; 0.24; 0.0; 0.85
3	0.57; 0.13; 1.0; 0.75	0.27; 0.14; 0.0; 0.85	0.47; 0.19; 1.0; 0.95

На рис. 1 представлена вычисленная итоговая оценка сходства $S(i,j)$ для каждой пары требований рассмотренных версий. Таким образом, метод корректно выделил все четыре типа сопоставления из рассмотренного примера.

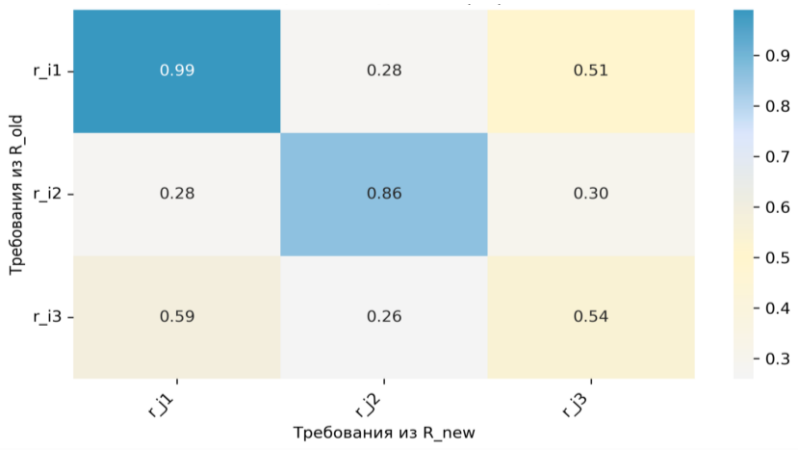


Рис. 1. Итоговая оценка сходства пар требований из примера.

Дополнительно был реализован прототип системы, объединяющий этапы обработки документации от извлечения требований до их сопоставления между версиями. На рис. 2 представлен фрагмент результата работы прототипа на приведенном примере с визуальным выделением статусов требований, а также измененных в них атрибутах между версиями.

Статус	Тип (V1 → V2)	Текст V1	Текст V2	Раздел	Сходство
UNCHANGED	Security (SE) → Security (SE)	The system shall authenticate users using a username and password.	The system shall authenticate users using a username and password.	User Authentication	0.99
MODIFIED	Performance (PE) → Performance (PE)	The system should process user requests within ~~3~~ seconds.	The system should process user requests within 2 seconds.	User Authentication	0.86
DELETED	Security (SE) → -	The system must log all authentication attempts.		Authentication	0.00
NEW	- → Security (SE)		The system should lock a user account after five failed login attempts.	User Authentication	0.00

Рис. 2. Фрагмент результата работы прототипа системы на примере.

Дополнительно для оценки точности метода был использован открытый набор TF_CDN, содержащий пары требований к ПО с экспертными метками о наличии или отсутствии связи между требованиями (категории дубликатов

(Duplicate), конфликтующих (Conflict) и пар, нейтральных по отношению друг к другу (Neutral)) [13]. В рамках эксперимента категории Duplicate и Conflict интерпретировались как наличие связи (Should Match), а Neutral – как отсутствие связи (No Relation). В табл. 2 представлены полученные значения метрик для предложенного метода.

Табл. 2. Результаты применения метода на датасете TF_CDN.

Наименование класса	Значение метрики		
	Precision	Recall	F1-score
No relation	0.97	0.84	0.90
Should Match	0.93	0.99	0.96

Для класса Should Match достигнуты показатели Precision порядка 0.93 и Recall порядка 0.99. Для No Relation метрика Precision составила около 0.97, Recall – 0.84. Общая точность (Accuracy) классификации пар равна 0.94. Высокое значение Recall для класса Should Match указывает на малое число пропущенных связей между требованиями, что является особенно важным показателем для задачи поиска пар изменений. Относительно высокое значение Precision (0.93) свидетельствует о низкой вероятности ложного сопоставления семантически несвязанных требований. Таким образом, полученные результаты подтверждают, что предложенная композитная функция сходства и процедура выбора порогов на основе метода Отсу адекватны для работы с парами требований из промышленных спецификаций.

ЗАКЛЮЧЕНИЕ

Предложен формализованный метод автоматического обнаружения изменений требований к ПО в версиях технической документации. Исходными данными служат множества требований, извлеченных и классифицированных для каждой версии. Задача сведена к построению частичного отображения между множествами требований и присвоению каждому требованию статуса: неизменное, измененное, новое или удаленное.

Ключевыми элементами метода является композитная функция сходства, объединяющая семантическое сходство (с помощью эмбеддингов SBERT), лексическое сходство (коэффициент Жаккара), совпадение класса и структурный контекст (заголовки разделов и относительная позиция требований). Введен понижающий коэффициент при несовпадении числовых параметров в содержании требований. Построение отображения осуществлено по правилу взаимных ближайших соседей (MNN) с вычисляемыми с помощью метода Отсу динамическими порогами по распределениям баллов сходства.

Применение метода на открытом наборе пар требований показало достижение точности идентификации пар на уровне порядка 0.94 при высокой полноте (Recall = 0.99) по классу наличия связей между требованиями, что подтверждает применимость подхода для анализа версий технической документации в терминах изменений требований к ПО.

СПИСОК ЛИТЕРАТУРЫ

1. ISO/IEC/IEEE 29148:2018. Systems and software engineering – Life cycle processes – Requirements engineering.
URL: https://www.iso.org/standard/72089.html.

2. *Sommerville I., Sawyer P.* Requirements Engineering: A Good Practice Guide. Chichester: Wiley, 1997. 404 p.

3. *Hunt J.W., MacIlroy M.D.* An algorithm for differential file comparison. Murray Hill: Bell Laboratories, 1976. 9 p.

4. *Myers E.W.* An O (ND) difference algorithm and its variations // Algorithmica. 1986. No. 1. P. 251–266.

5. *Gotel O.C.Z., Finkelstein C.W.* An analysis of the requirements traceability problem // Труды Международной научной конференции «IEEE international conference on requirements engineering», Колорадо-Спрингс, США, 18–22 апреля 1994. IEEE: 1994, С. 94–101.

6. *Cleland-Huang J., Gotel O. C., Huffman Hayes J., Mäder P., Zisman A.* Software traceability: trends and future directions // Труды Международной научной конференции «Future of Software Engineering, FOSE», Хайдарабад, Индия, 31 мая – 7 июня 2014. ACM: 2014, С. 55–69.

7. Савинчева А.С. Разработка интеллектуальной системы семантического анализа изменений в версиях технической документации программного обеспечения: выпускная квалификационная работа магистра. Казань: КФУ, 2026. С.19-40. URL: https://www.kpfu.ru/student_diplom/10.160.178.20_RIDB8JMIV54UV013SE2J41C1TVBNFK-SUB2Z5YB281KKY836_2X_F_Savincheva.pdf.

8. *Kumar L., Upadhye N.S., Piedy K.* Advances and Challenges in Semantic Textual Similarity: A Comprehensive Survey. URL: https://arxiv.org/abs/2601.03270.

9. *Reimers N., Gurevych I.* Sentence-bert: Sentence embeddings using siamese bert-networks // Труды международной научной конференции «2019 Conference on empirical methods in natural language processing and the 9th International joint conference on natural language processing (EMNLP-IJCNLP)», Гонконг, Китай, 3–7 ноября 2019. ACL: 2019, С. 3982–3992.

10. *Gomaa W.H., Fahmy A.A.* A survey of text similarity approaches // International Journal of Computer Applications. 2013. No. 13. P. 13–18.

11. *Kuhn H.W.* The Hungarian method for the assignment problem // Naval research logistics quarterly. 1955. No. 1-2. P. 83–97.

12. *Otsu N. et al.* A threshold selection method from gray-level histograms // Automatica. 1979. No. 11. P. 285–296.

13. TF-CDN: Initial Release. URL: https://zenodo.org/records/10830023.

METHOD FOR SEMANTIC ANALYSIS OF SOFTWARE REQUIREMENTS CHANGES IN TECHNICAL DOCUMENTATION VERSIONS

A. S. Savincheva[1] [0009-0005-0756-6001], Z. A. Enikeeva[2] [0000-0002-2468-7455], K. S. Enikeev[3] [0009-0000-6020-9912]

[1–3]*Institute of Information Technologies and Intelligent Systems, Kazan (Volga Region) Federal University, Kazan, Russia*

[1]asanvlit@gmail.com, [2]ZAEnikeeva@kpfu.ru, [3]KSEnikeev@kpfu.ru

Abstract

A formalized method for the automated detection of software requirement changes during the evolution of technical documentation versions is described. The input data consist of sets of requirements extracted and classified into functional and non-functional categories, with further granularity regarding non-functional requirement types (e.g., performance, security, availability, etc.) for each documentation version. The task of matching requirements between versions is reduced to constructing a partial mapping between the requirement sets and assigning one of the following statuses to each requirement: Unchanged, Modified, New, or Deleted.

A composite similarity function for requirement pairs is proposed, integrating semantic similarity based on vector representations (embeddings), lexical similarity, requirement class consistency, and the structural context of the document. The results of applying the proposed method to an open dataset of requirement pairs are presented. A prototype system for requirement change analysis in technical documentation has been developed.

Keywords*: software requirements, requirements change analysis, documentation version mapping, requirements traceability, semantic similarity, mutual nearest neighbors' method, Otsu's method.*

REFERENCES

1. ISO/IEC/IEEE 29148:2018. Systems and software engineering – Life cycle processes – Requirements engineering.
URL: https://www.iso.org/standard/72089.html.

2. *Sommerville I., Sawyer P.* Requirements Engineering: A Good Practice Guide. Chichester: Wiley, 1997. 404 p.

3. *Hunt J.W., MacIlroy M.D.* An algorithm for differential file comparison. Murray Hill: Bell Laboratories, 1976. 9 p.

4. *Myers E.W.* An O (ND) difference algorithm and its variations // Algorithmica. 1986. No. 1. P. 251–266.

5. *Gotel O.C.Z., Finkelstein C.W.* An analysis of the requirements traceability problem // Proceedings of the IEEE International Conference on Requirements Engineering, Colorado Springs, USA, April 18–22, 1994. IEEE: 1994, pp. 94–101.

6. *Cleland-Huang J., Gotel O. C., Huffman Hayes J., Mäder P., Zisman A.* Software traceability: trends and future directions // Proceedings of the Future of Software Engineering, FOSE, Hyderabad, India, May 31 – June 7, 2014. ACM: 2014, pp. 55–69.

7. Savincheva A.S. Development of an intelligent system for semantic analysis of changes between versions of software technical documentation: Master's thesis. Kazan: KFU, 2026. pp. 19-40. URL: https://www.kpfu.ru/student_diplom/10.160.178.20_RIDB8JMIV54UV013SE2J41C1TVBNFK-SUB2Z5YB281KKY836_2X_F_Savincheva.pdf.

8. *Kumar L., Upadhye N.S., Piedy K.* Advances and Challenges in Semantic Textual Similarity: A Comprehensive Survey. URL: https://arxiv.org/abs/2601.03270.

9. *Reimers N., Gurevych I.* Sentence-bert: Sentence embeddings using siamese bert-networks // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, November 3–7, 2019. ACL: 2019, pp. 3982–3992.

10. *Gomaa W.H., Fahmy A.A.* A survey of text similarity approaches // International Journal of Computer Applications. 2013. No. 13. P. 13–18.

11. *Kuhn H.W.* The Hungarian method for the assignment problem // Naval research logistics quarterly. 1955. No. 1-2. P. 83–97.

12. *Otsu N. et al.* A threshold selection method from gray-level histograms // Automatica. 1979. No. 11. P. 285–296.

13. TF-CDN: Initial Release. URL: https://zenodo.org/records/10830023.

СВЕДЕНИЯ ОБ АВТОРАХ



САВИНЧЕВА Ангелина Сергеевна – магистрант Института информационных технологий и интеллектуальных систем Казанского (Приволжского) федерального университета.

Angelina Sergeevna SAVINCHEVA – Master's student of the Institute of Information Technologies and Intelligent Systems, Kazan (Volga region) Federal University.

email: asanvlit@gmail.com

ORCID: 0009-0005-0756-6001



ЕНИКЕЕВА Зульфира Аснафовна – к. ф.-м. н., доцент Института информационных технологий и интеллектуальных систем Казанского (Приволжского) федерального университета.

Zulfira Asnafovna ENIKEEVA – Candidate of Physical and Mathematical Sciences, Docent at the Institute of Information Technologies and Intelligent Systems, Kazan (Volga Region) Federal University.

email: ZAEnikeeva@kpfu.ru

ORCID: 0000-0002-2468-7455



ЕНИКЕЕВ Камиль Шамилевич – старший преподаватель кафедры программной инженерии Института информационных технологий и интеллектуальных систем Казанского (Приволжского) федерального университета.

Kamil Shamilevich ENIKEEV – Senior lecturer of the department of software engineering at the Institute of Information Technologies and Intelligent Systems, Kazan (Volga Region) Federal University.

email: KSEnikeev@kpfu.ru

ORCID: 0009-0000-6020-9912

Материал поступил в редакцию 17 мая 2026 года

УДК 551.510.535+004.8

МЕТОДИКА И ПРЕДВАРИТЕЛЬНЫЕ РЕЗУЛЬТАТЫ СИНТЕЗА ИОНОГРАММ ВЕРТИКАЛЬНОГО РАДИОЗОНДИРОВАНИЯ ИОНОСФЕРЫ ПОСРЕДСТВОМ ФИЗИЧЕСКИ ИНФОРМИРОВАННЫХ НЕЙРОСЕТЕЙ

А. М. Симаков[1] [0009-0009-7365-720X], А. О. Щирый[2] [0000-0001-7312-4424]

[1]*НИУ Московский энергетический институт, г. Москва, Россия*

[2]*Институт земного магнетизма, ионосферы и распространения радиоволн им. Н. В. Пушкова РАН, г. Москва, Россия*

[1]nickhaos9@gmail.com, [2]andreyschiriy@gmail.com

Аннотация.

Исследован синтез ионограмм вертикального радиозондирования ионосферы с использованием физически информированных нейронных сетей. В предложенном подходе нейронная сеть применена не для непосредственного формирования изображения, а для аппроксимации высотного профиля электронной концентрации. Обучение сети выполнено с учетом стационарного одномерного уравнения баланса, включающего источник ионизации, рекомбинационные потери и эффективный перенос заряженных частиц. В функцию потерь дополнительно введены граничные условия, ограничение на гладкость профиля и диагностические параметры слоя F_2. Полученный профиль электронной концентрации преобразован в распределение плазменной частоты по высоте. Затем без учета магнитного поля Земли и столкновений электронов с другими частицами рассчитаны точки отражения, групповой показатель преломления и зависимость действующей высоты от частоты зондирования. На основе этой зависимости сформировано синтетическое изображение ионограммы с фоновым шумом и вертикальными помеховыми структурами. В пилотном эксперименте получены физически разумные масштабы максимума электронной концентрации и критической частоты слоя F_2. Результаты подтверждают возможность построения единой физически согласованной цепочки от высотного профиля электронной концентрации до изображения ионограммы. Разработанный подход

© А. М. Симаков, А. О. Щирый, 2026.
Данная статья распространяется на условиях международной лицензии Creative Commons License Attribution 4.0 International (CC BY 4.0).

может применяться для генерации контролируемых синтетических данных, исследования влияния параметров ионосферы на форму ионограммных следов и последующей разработки алгоритмов автоматической обработки результатов радиозондирования.

Ключевые слова: *ионосфера, ионограмма, вертикальное радиозондирование, электронная концентрация, плазменная частота, действующая высота, физически информированная нейронная сеть, PINN, синтетические данные, машинное обучение, радиофизическая модель.*

ВВЕДЕНИЕ

Исследования ионосферы востребованы как для фундаментальной науки (для построения и уточнения моделей геофизических процессов), так и для прикладных целей. Радиозондирование ионосферы является одним из наиболее информативных способов исследования состояния ионосферной плазмы и ионосферных радиолиний [1–6]. Результаты работы ионозонда представляют в виде ионограммы [1, 2]. Существуют различные виды радиозондирования ионосферы. Ионограмма вертикального радиозондирования ионосферы представляет собой зависимость амплитуды принятого сигнала от частоты f и действующей высоты отражения h'. Однако в процессе обработки и интерпретации реальных ионограмм возникает потребность в создании синтетических ионограмм, позволяющих идентифицировать треки мод распространения сигнала, отличать полезный сигнал от фонового шума и сосредоточенных помех, а также отлаживать алгоритмы автоматической обработки [3–5]. Традиционно синтез ионограмм базируется на решении краевой задачи распространения радиоволн в двумерно-неоднородной ионосфере [7–10]. Однако появление физически информированных нейронных сетей (physics-informed neural network, PINN) открывает новые возможности в синтезе ионограмм [11].

Цель настоящей работы – разработка и апробация нового метода синтеза ионограмм, в котором профиль электронной концентрации не задан непосредственно аналитической кривой, а получен как решение дифференциальной задачи, аппроксимируемое физически информированной нейросетью.

МОДЕЛЬ ВЫСОТНОГО ПРОФИЛЯ

Физической основой проведенного исследования служит уравнение (1) баланса для электронной концентрации [12]:

$$\frac{\partial N_e}{\partial t} = P - L - \frac{\partial J}{\partial h}, \quad (1)$$

где N_e – концентрация электронов, P – скорость ионообразования, L – потери, J – вертикальный поток заряженных частиц.

В реализованном прототипе рассмотрены нормированные высота $x \in [0, 1]$ и концентрация $n(x)$. Для синтеза высотного профиля используем следующую стационарную постановку:

$$\frac{\mathrm{d}}{\mathrm{d}x}\left[D(x)\,\frac{\mathrm{d}n}{\mathrm{d}x}\right] + q(x) - \alpha(x)n^2 = 0. \quad (2)$$

Здесь $D(x)$ – эффективный коэффициент транспорта, возрастающий в верхней части расчетной области; $\alpha(x)$ – эффективный коэффициент потерь, более высокий в нижней ионосфере; $q(x)$ – заданный источник ионизации. Для основного вклада области F применена пик-нормированная чепменовская форма

$$q_F(x) = A_F \exp\{1 - z - \exp(-z)\}, \quad z = \frac{x - x_m}{H_x},$$

дополненная слабыми вкладами областей E и F1.

Переход от безразмерной переменной x к физической высоте осуществим с помощью линейного отображения $h = h_0 + (h_1 - h_0)x$, где в пилотном расчете $h_0 = 90$ км и $h_1 = 520$ км. Такая нормировка удобна для PINN-оптимизации: величины в уравнении (2) имеют сопоставимый масштаб, а коэффициенты $D(x)$, $\alpha(x)$ и $q(x)$ легче варьировать без численной жесткости. Нижняя граница соответствует области ниже основного максимума слоя F, тогда как верхняя граница выбрана так, чтобы сохранить заметную электронную концентрацию над максимумом и не обрезать профиль искусственно.

PINN-ПОСТАНОВКА

Нейросеть $\hat{n}_\theta(x)$ принимает высоту x и возвращает положительную оценку концентрации. Производные по высоте вычисляются автоматическим дифференцированием. В функционал обучения включена невязка уравнения (2),

граничные ограничения, слабое априорное ограничение на чепменовскую форму и якоря на диагностические параметры

$foF2$ и $hmF2$:

$$L = \lambda_R\langle R_\theta^2(x)\rangle + \lambda_b[\hat{n}_\theta(0) - n_b]^2 + \lambda_t[\hat{n}_\theta(1) - n_t]^2 + \lambda_f[\hat{n}_\theta(x_m) - n_m]^2 + \lambda_h\left[\frac{\mathrm{d}\hat{n}_\theta}{\mathrm{d}x}(x_m)\right]^2 + \lambda_p\left\langle\left[\hat{n}_\theta(x) - n_p(x)\right]^2\right\rangle.$$

где R_θ – левая часть уравнения (2), x_m соответствует целевой высоте максимума слоя F_2, n_m – концентрация, соответствующая целевой частоте $foF2$, $n_p(x)$ – гладкий профиль-prior.

Такой функционал не сводится к простому «подгону» двух параметров. Слагаемое невязки уравнения задает физическую согласованность профиля во всей области, граничные штрафы стабилизируют решение на краях интервала, а prior ограничивает класс допустимых форм разумными гладкими профилями. Якорь на $hmF2$ фиксирует положение главного максимума, а якорь на $foF2$ определяет его масштаб по концентрации. В результате PINN получает достаточную свободу для перестройки формы $N_e(h)$, но не уходит в нефизические осциллирующие решения, которые часто возникают в плохо обусловленных обратных задачах.

Важно подчеркнуть ограничение текущей постановки: целевые значения $foF2$ и $hmF2$ использованы как управляющие параметры синтеза. Реальные ионограммы в обучение на данном этапе не включались.

ПЕРЕХОД ОТ ПРОФИЛЯ К ИОНОГРАММЕ

После обучения профиль переведем в физические единицы. Плазменную частоту вычислим по формуле

$$f_p = 8.98 \cdot 10^{-6}\sqrt{N_e(\text{м}^{-3})}.$$

Для формирования следа ионограммы используем приближение обыкновенной компоненты без магнитного поля и столкновений. Групповой показатель зададим в виде

$$n_g(f,h) = \frac{1}{\sqrt{1 - \left(\frac{f_p(h)}{f}\right)^2}}. \quad (3)$$

Действующую высоту $h'(f)$ найдем с помощью численного интегрирования n_g от нижней границы модели до точки поворота луча с добавлением участка свободного распространения от Земли до нижней границы. Полученную кривую преобразуем в полутоновое изображение; дополнительно вводим фоновый шум и вертикальные помеховые структуры, характерные для представления ионограмм как изображений [2, 3].

В выбранном приближении отражение обыкновенной компоненты происходит в окрестности высоты, где $f_p(h)$ достигает частоты зондирования. Поэтому форма трассы $h'(f)$ определяется не только положением максимума $N_e(h)$, но и тем, насколько быстро возрастает профиль на подходе к слою F_2. Более пологий рост дает большую действующую высоту на промежуточных частотах, тогда как резкий фронт приводит к более компактной трассе. Это важно для синтетических данных: даже при одинаковом $foF2$ разные формы профиля дают различимую геометрию следа на ионограмме.

ВЫЧИСЛИТЕЛЬНЫЙ ЭКСПЕРИМЕНТ

В вычислительном эксперименте целью синтеза были выбраны $foF2 = 9.4$ МГц и $hmF2 = 340$ км. В обучении были использованы 10000 эпох, 1536 случайных коллокационных точек на эпоху и шаг оптимизатора Adam $5 \cdot 10^{-4}$. Расчетная область по высоте составляла 90–520 км. В табл. 1 представлены основные численные результаты пилотного расчета.

Контрольные значения действующей высоты составили $h'(3\ \text{МГц}) = 186.1$ км, $h'(5\ \text{МГц}) = 287.6$ км, $h'(7\ \text{МГц}) = 359.8$ км, $h'(9\ \text{МГц}) = 457.1$ км. Полученные значения показывают, что расчет воспроизводит не только заданное положение максимума, но и разумный ход групповой задержки в широком диапазоне частот. В частности, $f_p(520\ \text{км}) = 5.15$ МГц указывает на сохранение заметной концентрации на верхней границе расчетной области; это означает, что максимум слоя F_2 не сформирован за счет искусствен-

ного прижатия решения к нулю на верхнем крае интервала. Рост $h'(f)$ при приближении к $foF2$ согласуется с физическим смыслом действующей высоты: по мере уменьшения знаменателя в формуле (3) возрастает вклад группового замедления волны.

Табл. 1. Основные численные результаты пилотного расчета.

Параметр	Значение
Финальная невязка уравнения L_R	$3.08 \cdot 10^{-3}$
Оценка $foF2$	9.36 МГц
Оценка $hmF2$	340.0 км
Максимум N_e	$1.09 \cdot 10^{12}$ м$^{-3}$
f_p на верхней границе 520 км	5.15 МГц
Диапазон частот с отражением	1.23–9.34 МГц
Диапазон рассчитанных h'	97.8–604.8 км

На рис. 1 показаны использованный высотный профиль источника и соответствующий ему профиль плазменной частоты. Видно, что основной вклад источника сосредоточен в области F, а максимум $f_p(h)$ формируется вблизи целевого значения $hmF2$. На рис. 2 приведены восстановленный профиль электронной концентрации и рассчитанная высотно-частотная характеристика. Эти зависимости демонстрируют согласованность между формой $N_e(h)$ и поведением $h'(f)$ в рабочем диапазоне частот. На рис. 3 показано соответствующее синтетическое изображение ионограммы в инвертированной палитре.

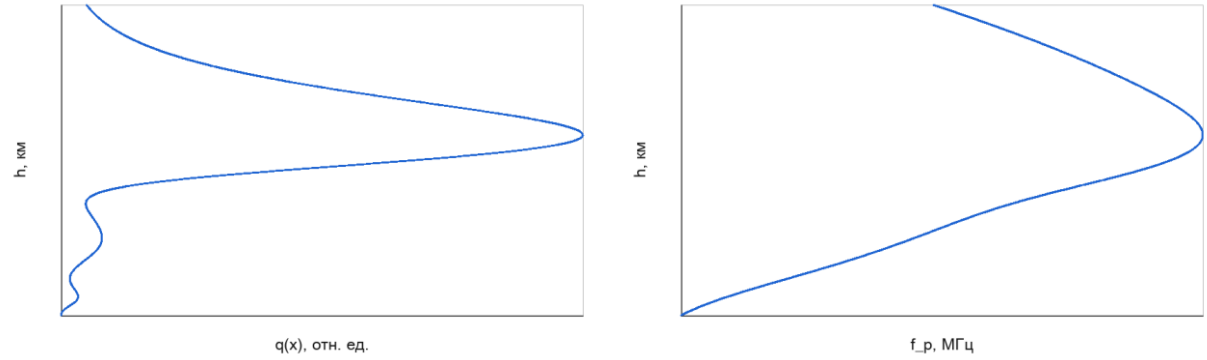


Рис. 1. Слева: высотный профиль источника $\mathrm{q}(x)$, справа: полученный профиль плазменной частоты $\mathrm{f_p}(h)$.

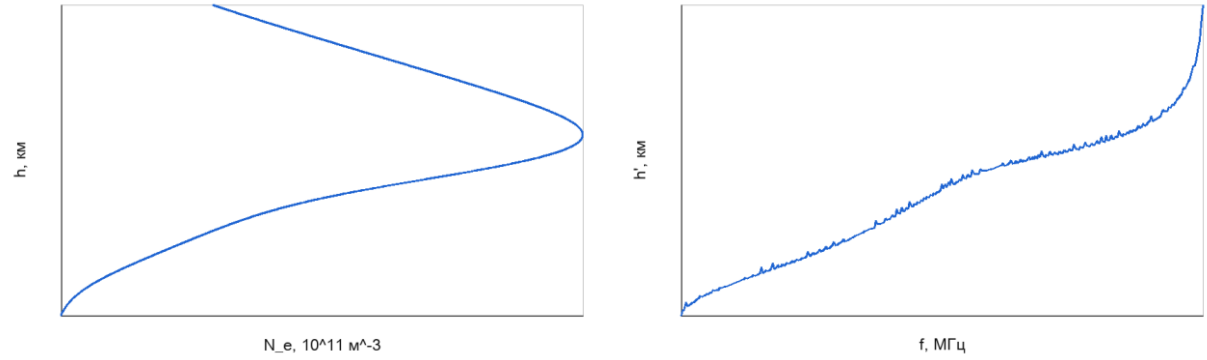


Рис. 2. Слева: профиль электронной концентрации $\mathrm{N_e}(h)$, справа: рассчитанная высотно-частотная характеристика $\mathrm{h}'(f)$.

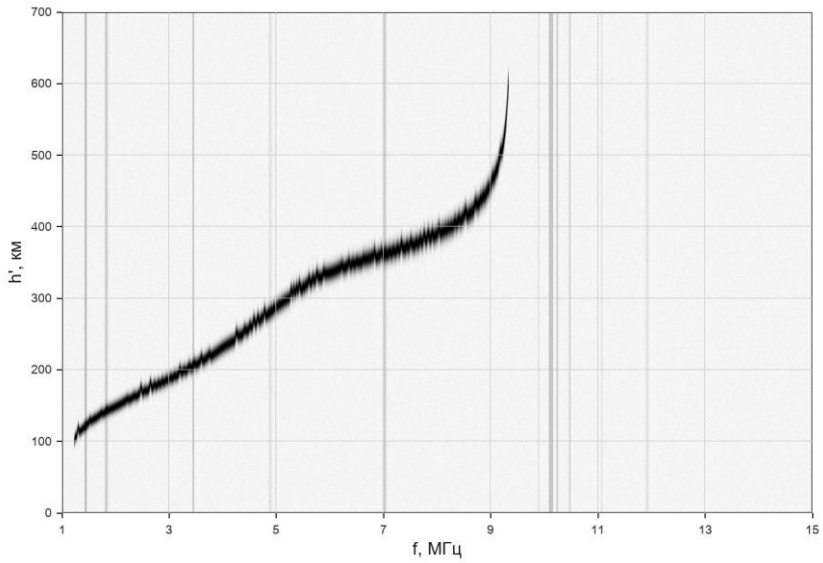


Рис. 3. Синтетическое изображение ионограммы, полученное из рассчитанной кривой $\mathrm{h}'(f)$.

ОБСУЖДЕНИЕ

Предложенная методика демонстрирует возможность получить согласованную цепочку

$$\hat{n}_\theta(x) \;\rightarrow\; N_e(h) \;\rightarrow\; f_p(h) \;\rightarrow\; h'(f) \;\rightarrow\; \text{изображение ионограммы.}$$

По сравнению с прямым заданием ионограммного следа такой подход сохраняет явную связь изображения с высотным профилем электронной концентрации.

Вместе с тем модель остается демонстрационной. В ней нет нестационарности, горизонтальной неоднородности, полного магнитоионного выражения показателя преломления, столкновительного поглощения, отдельной X-компоненты и самостоятельного следа спорадического слоя Es. Поэтому полученную ионограмму следует интерпретировать как контролируемый синтетический образец, а не как результат расчета для конкретного сеанса наблюдений.

Тем не менее именно такая упрощенная постановка полезна как промежуточный этап. Она позволяет связать параметры профиля, диагностические характеристики слоя и итоговое изображение одной вычислительной схемой. Для задач автоматического распознавания это особенно полезно: можно генерировать семейства ионограмм с известной физической подложкой, варьировать уровень шума и помех, а затем проверять устойчивость алгоритмов выделения следа и оценивания $foF2$, $hmF2$ и других параметров. Кроме того, дифференцируемость всей вычислительной схемы делает естественным следующий шаг – включение в функционал наблюдательного слагаемого, зависящего уже не только от профиля, но и от положения следа на самой ионограмме.

На текущем этапе наиболее информативными результатами являются не итоговые полутоновые изображения, а профили $N_e(h)$, $f_p(h)$ и высотно-частотная характеристика $h'(f)$. Именно эти зависимости непосредственно отражают качество физической постановки и позволяют локализовать расхождения между моделью и ожидаемым поведением следа. Поэтому в настоящей работе акцент сделан на анализ промежуточных радиофизических представлений, которые затем могут быть использованы как основа для дальнейшего совершенствования синтеза изображений ионограмм.

ЗАКЛЮЧЕНИЕ

Разработана и проверена методика синтеза ионограмм на основе PINN. Нейросеть использована не как черный ящик для рисования изображения, а как аппроксимация высотного профиля, ограниченного дифференциальным уравнением баланса. В пилотном эксперименте получены физически разумные масштабы максимума электронной концентрации и критической частоты слоя F_2: $foF2 = 9.36$ МГц, $hmF2 = 340.0$ км. В дальнейшем планируется включить наблюдательное слагаемое в функционал PINN и расширить радиофизическую модель до нескольких магнитоионных компонент. Подчеркнем, настоящая постановка задачи синтеза ионограмм с помощью PINN рассмотрена как промежуточный этап между процедурной генерацией изображений и полноценным восстановлением состояния ионосферы по данным наблюдений.

Благодарности

Работа выполнена в рамках государственного задания Минобрнауки России для ИЗМИРАН (тема № FFUM-2026-0010).

СПИСОК ЛИТЕРАТУРЫ

1. Радиозондирование ионосферы спутниковыми и наземными ионозондами / Под ред. С.И. Авдюшина. Труды института прикладной геофизики им. академика Е.К. Федорова. М.: ИПГ, 2008.
URL: http://ipg.geospace.ru/publications/book-2008.pdf

2. РД 52.26.817–2023. Руководство по ионосферным, магнитным и гелиогеофизическим наблюдениям. Часть 1. Ионосферные наблюдения. Обнинск: ФГБУ «ВНИИГМИ-МЦД», 2023.

3. *Shiriy A.O.* HF channel transmit function module measurement // Proceedings of the 5th International Conference on Actual Problems of Electron Devices Engineering, APEDE 2002. Vol. 5. 2002. P. 365–369.
https://doi.org/10.1109/apede.2002.10449645

4. *Щирый А.О.* Разработка и моделирование алгоритмов автоматического измерения характеристик ионосферных коротковолновых радиолиний: Автореф. дис. … канд. техн. наук [Место защиты: С.-Петерб. гос. ун-т телекоммуникаций им. М.А. Бонч-Бруевича]. СПб., 2007. 19 с. EDN: QMISDO

5. *Щирый А.О.* Алгоритмы и программное обеспечение автоматизации процессов измерений и обработки данных оперативной диагностики ионосферы и ионосферных радиолиний // Журнал радиоэлектроники. 2022. № 10. https://doi.org/10.30898/1684-1719.2022.10.4

6. *Щирый А.О.* Две методики измерения помех декаметрового диапазона // Научное приборостроение. 2023. Т. 33, № 4. С. 72–83.

7. Программы расчета траекторных характеристик распространения коротких радиоволн: Сборник / АН СССР, Ин-т земного магнетизма, ионосферы и распространения радиоволн; Отв. ред. канд. физ.-мат. наук Е.М. Жулина. Москва: ИЗМИРАН, 1975. 58 с.

8. *Ларюнин О.А.* Численный синтез ионограмм в горизонтально-неоднородной ионосфере на основе модели комбинированного параболического слоя // Солнечно-земная физика. 2016. №. 3. С. 52–58. https://doi.org/10.12737/18656

9. *Settimi A., Pezzopane M., Pietrella M., Bianchi C., Scotto C., Zuccheretti E., and Makris J.* Testing the IONORT-ISP system: A comparison between synthesized and measured oblique ionograms // Radio Science, March 2013. Vol. 48, No. 2. P. 167–179. https://doi.org/10.1002/rds.20018

10. *Носиков И.А., Падохин А.М., Крашенинников И.В., Клименко М.В., Бессараб П.Ф.* Особенности расчета минимальных частот мод в задаче прогнозирования условий ионосферной радиосвязи // Известия высших учебных заведений. Радиофизика. 2021. Т. 64, № 8-9. С. 672–685.

11. *Raissi M., Perdikaris P., Karniadakis G.E.* Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations // Journal of Computational Physics. 2019. Vol. 378. P. 686–707.

12. *Брюнелли Б.Е., Намгаладзе А.А.* Физика ионосферы. М.: Наука, 1988. 526 с.

METHODOLOGY AND FIRST RESULTS OF IONOGRAM SYNTHESIS BY PHYSICS-INFORMED NEURAL NETWORKS

A. M. Simakov[1] [0009-0009-7365-720X], A. O. Schiriy[2] [0000-0001-7312-4424]

[1]*National Research University "Moscow Power Engineering Institute", Moscow, Russia*

[2]*Pushkov Institute of Terrestrial Magnetism, Ionosphere and Radio Wave Propagation of the Russian Academy of Sciences, Moscow, Russia*

[1]nickhaos9@gmail.com; [2]andreyschiriy@gmail.com

Abstract. The problem of synthesizing ionograms of vertical ionosphere radiosonde using physics-informed neural networks is solved. In the proposed approach, the neural network is used not for direct image generation, but for approximating the high-altitude profile of the electron concentration. Network training is performed taking into account a stationary one-dimensional balance equation, including an ionization source, recombination losses, and effective transfer of charged particles. The loss function additionally includes boundary conditions, a restriction on the smoothness of the profile, and diagnostic parameters of the F2 layer. The resulting electron concentration profile is transformed into a height distribution of the plasma frequency. Therefore, without taking into account the Earth's magnetic field and collisions of electrons with other particles, the reflection points, the group refractive index and the dependence of the effective height on the sounding frequency are calculated. Based on this dependence, a synthetic image of an ionogram with background noise and vertical interference structures is formed. The results confirm the possibility of constructing a single physically consistent chain from the height profile of the electron concentration to the image of the ionogram. The developed approach can be used to generate controlled synthetic data, study the effect of ionospheric parameters on the shape of ionogram trails, and subsequently develop algorithms for automatic processing of radiosonde results.

Keywords: *ionosphere, ionogram, vertical radio sounding, electron density, plasma frequency, virtual height, physics-informed neural network, PINN, synthetic data, machine learning, radiophysical model.*

REFERENCES

1. Satellite and Ground-Based Ionospheric Sounding / Ed. by S.I. Avdyushin. Proceedings of the Academician E.K. Fedorov Institute of Applied Geophysics. Moscow: IPG, 2008. URL: http://ipg.geospace.ru/publications/book-2008.pdf

2. RD 52.26.817–2023. Guidelines for ionospheric, magnetic and heliogeophysical observations. Part 1: Ionospheric observations // VNIIGMI-MCD. Obninsk, 2023.

3. *Shiriy A.O.* HF channel transmit function module measurement // Proceedings of the 5th International Conference on Actual Problems of Electron Devices Engineering, APEDE 2002. Vol. 5. 2002. P. 365–369.
https://doi.org/10.1109/apede.2002.10449645

4. *Schiriy A.O*. Development and Modeling of Algorithms for Automatic Measurement of Characteristics of Ionospheric HF Radio Links*:* Extended Abstract of Cand. Sci. (Eng.) Dissertation. Saint Petersburg State University of Telecommunications named after Prof. M.A. Bonch-Bruevich. St. Petersburg, 2007. 19 p.

5. *Schiriy A.O*. Algorithms and Software for Automation of Measurement and Data Processing in Operational Diagnostics of the Ionosphere and Ionospheric Radio Links // Journal of Radio Electronics. 2022. No. 10.
https://doi.org/10.30898/1684-1719.2022.10.4

6. *Schiriy A.O. Two methods for measuring decameter-band interference //* Nauchnoe Priborostroenie [Scientific Instrumentation]. 2023. Vol. 33, No. 4. P. 72–83.

7. Programs for Calculating Trajectory Characteristics of Shortwave Radio Propagation: Collection / USSR Academy of Sciences, Institute of Terrestrial Magnetism, Ionosphere and Radio Wave Propagation; Ed. by E.M. Zhulina. Moscow: IZMIRAN, 1975. 58 p.

8. *Laryunin O.A.* Numerical Synthesis of Ionograms in a Horizontally Inhomogeneous Ionosphere Based on the Combined Quasi-Parabolic Layer Model // *S*olar-Terrestrial Physics. 2016. No. 3. P. 52–58. https://doi.org/10.12737/18656

9. *Settimi A., Pezzopane M., Pietrella M., Bianchi C., Scotto C., Zuccheretti E., and Makris J.* Testing the IONORT-ISP system: A comparison between synthesized and measured oblique ionograms // Radio Science, March 2013. Vol. 48, No. 2. P. 167–179. https://doi.org/10.1002/rds.20018

10. *Nosikov I.A., Padokhin A.M., Krasheninnikov I.V., Klimenko M.V., Bessarab P.F.* Features of Calculating Minimum Mode Frequencies in the Problem of Forecasting Ionospheric Radio Communication Conditions // Radiophysics and Quantum Electronics. 2021. Vol. 64, No. 8–9. P. 672–685.

11. *Raissi M., Perdikaris P., Karniadakis G.E.* Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations // Journal of Computational Physics. 2019. Vol. 378. P. 686–707.

12. *Brunelli B.E., Namgaladze A.A.* Fizika ionosfery. Moscow: Nauka, 1988. 526 p.

СВЕДЕНИЯ ОБ АВТОРАХ



СИМАКОВ Александр Максимович – магистрант Национального исследовательского университета «МЭИ», Москва, Россия.

Alexander Maximovich SIMAKOV – master's student at the National Research University “Moscow Power Engineering Institute”, Moscow, Russia.

email: nickhaos9@gmail.com

ORCID: 0009-0009-7365-720X



ЩИРЫЙ Андрей Олегович – кандидат технических наук, старший научный сотрудник Института земного магнетизма, ионосферы и распространения радиоволн РАН им. Н.В.Пушкова (ИЗМИРАН), Москва.

Andrey Olegovich SCHIRIY – PhD (Candidate of Technical Sciences), senior researcher at the Pushkov Institute of Terrestrial Magnetism, Ionosphere and Radio Wave Propagation of the Russian Academy of Sciences (IZMIRAN), Moscow.

email: andreyschiriy@gmail.com

ORCID: 0000-0001-7312-4424

Материал поступил в редакцию 16 июня 2026 года

УДК 004.5:159.9

КОГНИТИВНОЕ МОДЕЛИРОВАНИЕ ПОЛЬЗОВАТЕЛЬСКОГО ВЗАИМОДЕЙСТВИЯ В МУЛЬТИРОЛЕВЫХ ВЕБ-ПРИЛОЖЕНИЯХ С 3D-ФУНКЦИОНАЛОМ

М. В. Шматко[1] [0000-0002-7255-8885], Е. В. Евдущенко[2] [0000-0003-3692-2587]

[1]*Омский государственный технический университет, г. Омск, Россия*
[2]*Военный институт (инженерно-технический) Военной академии материально-технического обеспечения, г. Санкт-Петербург, Россия*

[1]marin298@gmail.com, [2]elena.online_ktilp@mail.ru

Аннотация

Исследована когнитивная нагрузка пользователей мультиролевой 3D-платформы. В эксперименте (71 участник: 34 дизайнера, 37 покупателей; 639 наблюдений) объективными метриками и субъективными шкалами выявлены разные профили нагрузки: дизайнеров замедляют формы ввода, покупателей – 3D-контент. На основе 34 проблемных зон сформулированы семь принципов проектирования интерфейса и классификация задач по уровню когнитивной сложности с учетом ролевой специфики. Повторное тестирование (21 участник) подтвердило эффективность: время выполнения задач сократилось на 21–28%, ошибки – на 33–44%, фрустрация – на 25–35%, использование 3D-функций выросло с 40.5% до 90.9%. Разработанная трехкомпонентная модель (таксономический, диагностический, проектный уровни) рекомендована для проектирования аналогичных мультиролевых платформ с 3D-функционалом.

Ключевые слова: *когнитивная нагрузка, мультиролевые веб-приложения, пользовательские 3D-интерфейсы, NASA-TLX, SMEQ, когнитивное моделирование, UX-исследования, человеко-компьютерное взаимодействие.*

ВВЕДЕНИЕ

Мультиролевые веб-приложения – от маркетплейсов (Amazon, AliExpress, Ozon, Wildberries) и платформ для трудоустройства (Upwork, SuperJob, Rabota.ru) до образовательных порталов (Coursera, Учи.ру, InternetUrok) и госу-

© М. В. Шматко, Е. В. Евдущенко, 2026.
Данная статья распространяется на условиях международной лицензии Creative Commons License Attribution 4.0 International (CC BY 4.0).

дарственных цифровых сервисов (Госуслуги, Gov.uk) – все чаще становятся многофункциональными экосистемами, объединяющими разные категории пользователей: производителей, продавцов, покупателей, граждан, представителей бизнеса и госорганов (B2B, B2C, B2G). Проектирование таких экосистем сталкивается с фундаментальным вызовом: как создать интерфейс, одинаково эффективный для пользователей с различными задачами и компетенциями? Внедрение интерактивных технологий: 3D-конфигураторов, AR-примерок, инструментов кастомизации и геймификации – с одной стороны, повышает вовлеченность, с другой – увеличивает когнитивную нагрузку. Исследования показали [1, 2], что высокая когнитивная нагрузка ведет к росту числа ошибок, увеличению времени выполнения задач, снижению конверсии и отказу от использования.

На стыке указанных областей наблюдается дефицит систематических исследований, а среди разработчиков мультиролевых веб-приложений с 3D-функционалом – потребность в практических рекомендациях по оптимизации пользовательского взаимодействия.

Ранее нами была разработана архитектура, определен технологический стек и создан прототип мультибрендовой цифровой платформы со встроенным 3D-конфигуратором для кастомизации одежды [3]. Платформа ориентирована на две ключевые категории пользователей: 3D-дизайнеров (профессионалов, создающих цифровые модели одежды и управляющих ассортиментом и заказами – 37 сценариев) и покупателей (конечных потребителей, использующих платформу для поиска, кастомизации и заказа одежды – 25 сценариев). Предварительный анализ этих сценариев позволил предположить, что уровень когнитивной нагрузки будет существенно различаться не только между ролями, но и внутри каждой роли – в зависимости от сложности задачи. При этом реальный уровень нагрузки может не совпадать с ожидаемым. Именно расхождение между моделью и реальностью, а также между предполагаемой и фактической когнитивными нагрузками представляет для нас исследовательский интерес. В связи с этим в качестве методологической базы исследования было выбрано когнитивное моделирование, которое рассматривается как процесс построения модели взаимодействия пользователей в процессе восприятия и обработки информации. Основами этого процесса выступают три взаимосвя-

занных уровня: таксономический, диагностический и проектный. Такой подход позволяет не только диагностировать проблемы взаимодействия, но и предсказывать эффекты проектных решений, обеспечивая переход от описательной диагностики к конструктивному проектированию.

Цель настоящего исследования – разработать, эмпирически обосновать и валидировать принципы проектирования интерфейса мультиролевого веб-приложения с 3D-функционалом, направленные на снижение когнитивной нагрузки пользователей.

Научная новизна исследования представлена следующими результатами:

– разработана и апробирована методика комплексной оценки когнитивной нагрузки пользователей, сочетающая объективные метрики (время выполнения, количество ошибок, число кликов) с субъективными шкалами (NASA-TLX, SMEQ) и качественным анализом применительно к взаимодействию с 3D-контентом в мультиролевой среде;

– сформулированы и валидированы принципы проектирования интерфейсов для мультиролевых платформ с 3D-функционалом, направленные на снижение когнитивной нагрузки пользователей;

– предложена модель когнитивного взаимодействия пользователей с мультиролевой 3D-платформой.

ТЕОРЕТИЧЕСКИЕ ОСНОВЫ ИССЛЕДОВАНИЯ КОГНИТИВНОЙ НАГРУЗКИ В МУЛЬТИРОЛЕВЫХ 3D-ИНТЕРФЕЙСАХ

Вопрос о том, почему одни интерфейсы воспринимаются как интуитивно понятные, а другие вызывают усталость и раздражение, лежит в плоскости когнитивной нагрузки – меры умственных усилий, требуемых для выполнения задачи.

Теория когнитивной нагрузки (Cognitive Load Theory), разработанная Джоном Свеллером в конце 1980 г., изначально применялась в педагогической психологии, но сегодня составляет фундамент исследований в области UX и HCI. Согласно этой теории, наша рабочая память имеет ограниченный объем. Когда интерфейс требует выполнения множества умственных операций, наступает когнитивная перегрузка, ведущая к ошибкам и замедлению работы [4].

Особую актуальность эта проблема приобретает в контексте современных цифровых платформ. Длительная работа с интерфейсами, создающими высокую когнитивную нагрузку, может приводить не только к ситуативным ошибкам, но и к хроническому стрессу, снижению удовлетворенности и отказу от использования цифрового продукта [5, 6].

Для оценки когнитивной нагрузки исследователями разработан целый ряд методик. Наиболее признанным инструментом остается NASA-TLX (Task Load Index), предложенный Сандрой Харт и Лоуэллом Ставелендом [7]. Этот многомерный опросник оценивает нагрузку по шести шкалам: ментальные, физические, временные требования, усилие, фрустрация и оценка собственной успешности.

Альтернативой NASA-TLX выступает шкала SWAT (Subjective Workload Assessment Technique), оценивающая нагрузку по трем параметрам: временная нагрузка, психологический стресс и умственные усилия. В ситуациях, требующих быстрой повторяющейся оценки, исследователи обращаются к одномерным шкалам – Паaca (Paas Cognitive Load Scale) или SMEQ (Subjective Mental Effort Questionnaire). Эти инструменты занимают 10–15 с, но хорошо коррелируют с более сложными многомерными методиками [8, 9].

Цифровая платформа со встроенным 3D-конфигуратором для кастомизации одежды относится к классу систем, которые в западной традиции обозначают как multi-sided platforms или B2B2C-платформы. Их основная особенность состоит в одновременном обслуживании пользователей с разными ролями, целями, компетенциями и паттернами поведения. Как констатируют эксперты, такие платформы требуют уникального сочетания корпоративных продуктовых и дизайнерских возможностей, тесно интегрированных с потребительским опытом высшего уровня. Это более сложная специализация, чем разработка только для корпоративного или только для потребительского сегмента [6, 10].

Трехмерная визуализация добавляет еще один уровень сложности в проектирование интерфейса B2B2C-платформ. С одной стороны, она позволяет получить более полное и реалистичное представление о продукте, с другой – взаимодействие с 3D-объектом требует освоения новых моторных навыков и создает дополнительную когнитивную нагрузку. Исследования последних лет

подтверждают, что «пространственные 3D-цифровые среды часто рассматриваются как превосходящие 2D-интерфейсы для обмена информацией и знаниями» [11]. Ключевой нюанс: такие улучшения не возникают автоматически – если 3D-представление возможно с меньшими перемещениями, то это будет приводить к более эффективному снижению когнитивной нагрузки [11].

Таким образом, теоретический анализ показал, что когнитивная нагрузка не является однородной величиной, она зависит от сложности задачи, способа представления информации и индивидуальных особенностей пользователя. Для мультиролевой платформы с 3D-функционалом, объединяющей профессиональных и непрофессиональных пользователей, важно дифференцировать задачи по уровню ожидаемой нагрузки, чтобы затем сопоставить эти ожидания с эмпирическими данными.

МЕТОДОЛОГИЯ ЭМПИРИЧЕСКОГО ИССЛЕДОВАНИЯ

Эмпирическое исследование проводилось в лабораторных условиях для минимизации отвлекающих факторов, критически влияющих на когнитивную нагрузку. При этом следует учитывать ограничение внешней валидности: для профессиональных дизайнеров лабораторная среда может создавать дополнительную искусственную нагрузку [12].

Процедура сбора данных включала три последовательных этапа, общая продолжительность которых составила 11 недель.

Этап 1. Организация эмпирического исследования

На данном этапе были определены 62 пользовательские задачи (37 для 3D-дизайнеров и 25 для покупателей), которые были классифицированы по трем уровням когнитивной нагрузки [13]. Классификация осуществлялась методом экспертной оценки с привлечением трех специалистов в области UX-исследований и проектирования интерфейсов (с опытом работы более 7 лет).

Сначала эксперты произвели оценку 62 задач независимо друг от друга, а затем обсудили расхождения и выработали консенсус относительно их распределения по трем уровням когнитивной нагрузки (см. табл. 1). В результате был достигнут высокое значение коэффициента согласованности, или конкордации Кендалла, $W = 0.82$, найденное по формуле [14]

$$W = \frac{12\,S}{m^2(n^3 - n)},$$

где S – сумма квадратов отклонений рангов от их среднего; m – количество экспертов; n – количество параметров.

Табл. 1. Распределение пользовательских задач по уровням когнитивной нагрузки.

Уровень когнитивной нагрузки	Количество задач (доля от общего числа, %)	
	для дизайнеров	для покупателей
Низкий	12 задач (32.4%)	14 задач (56.0%)
Средний	16 задач (43.3%)	8 задач (32.0%)
Высокий	9 задач (24.3%)	3 задачи (12.0%)

На основе распределения, представленного в табл. 2, для каждой роли пользователей было отобрано по 9 ключевых задач, проверенных на репрезентативность и важность для бизнес-процессов платформы. По завершению этапа были сформированы две группы участников (34 дизайнера и 37 покупателей), проведены инструктаж и знакомство с интерфейсом и процедурой «думай вслух».

Этап 2. Проведение лабораторного эксперимента

Участники под наблюдением модераторов выполняли по 9 задач в рандомизированном порядке. Для контроля эффекта научения применялась частичная контрбалансировка по схеме латинского квадрата, порядковый номер задачи использовался как ковариата. Сразу после решения каждой задачи участники заполняли SMEQ для минимизации ретроспективного искажения. Сессии длились 45–60 мин, проводились в первой половине дня для контроля циркадных ритмов [15].

Этап 3. Качественная оценка результатов

На этом этапе совместно с участниками были заполнены полные версии NASA-TLX и проведено посттестовое интервью для сбора качественных комментариев о трудностях работы с интерфейсом.

Методы и инструменты сбора эмпирических данных: в соответствии с задачами исследования выбран комплексный подход, основанный на применении объективных метрик (табл. 2) и субъективных шкал.

Табл. 2. Объективные метрики для сбора эмпирических данных.

№	Метрики	Описание
1	Время выполнения задачи	От момента окончания инструкции до момента достижения результата (в случае тайм-аута фиксировалось максимальное значение – 300 с)
2	Количество ошибок	Фиксация следующих случаев при выполнении задачи: – неверный выбор опции; – попытка выполнить действие меню или разделе, не предназначенном для него; – неправильное заполнение полей формы; – отмена или прерывание действия
3	Количество кликов	Общее число нажатий на интерактивные элементы тестируемого интерфейса, включая ошибочные и повторные; дополнительно фиксировались «неэффективные клики» (действия, не приближающие к результату)

Помимо объективных метрик (см. табл. 2) в исследовании были задействованы субъективные шкалы – одномерная SMEQ для оперативных замеров после каждой задачи (вертикальная линия 150 мм, от «совсем нет усилий» до «чрезвычайно высокие усилия») [16] и многомерная NASA-TLX для итоговой оценки с процедурой попарного взвешивания (для минимизации вариативности NASA-TLX применялись бумажные бланки) [17].

Качественные методы включали протокол «думай вслух» [18] для выявления логики действий пользователей и полуструктурированное посттестовое интервью для определения сложных и понятных элементов интерфейса.

Выборка исследования была сформирована методом критериального отбора с использованием доступных источников – региональных профессиональных сообществ и студенческих групп. Ее суммарный объем составил 71 человек (n = 71), что превышает рекомендованные для юзабилити-тестирования 10–15 человек на роль пользователей [19] и обеспечивает достаточную статистическую мощность для обнаружения эффектов среднего размера при однофакторном дисперсионном анализе [20]. В группу дизайнеров вошли 34 человека со средним стажем 3D-моделирования 4.7 года, почти все они имели опыт работы с онлайн-платформами для продажи одежды (91%). В группу покупателей вошли 37 человек: больше половины из них совершают онлайн-покупки 2–3 раза в месяц, 43% имеют опыт кастомизации одежды, а 27% – опыт использования 3D-конфигураторов.

Статистический анализ: для обработки данных использовался комплекс статистических методов [21]. Информация о них представлена в табл. 3.

Табл. 3. Статистические методы и цели их применения.

Цель анализа	Используемый метод	Критерии/уточнения
Описание данных	Средние (*M*), стандартные отклонения (*SD*), медианы (*Mdn*), интерквартильные размахи (*IQR*)	Для всех количественных переменных
Проверка нормальности	Критерий Шапиро – Уилка	Фиксация величин асимметрии и эксцесса
Сравнение двух групп респондентов	*t*-критерий Стьюдента (при нормальности)/*U*-критерий Манна – Уитни (при отсутствии)	При множественных сравнениях – поправка Бонферрони [22]
Сравнение внутри группы (разные уровни сложности)	Однофакторный ANOVA с повторными измерениями	Апостериорное сравнение – критерий Тьюки; при нарушении сферичности – поправка Гринхауса – Гейссера; контроль эффекта научения – порядковый номер задачи как ковариата
Совместное влияние факторов (роль и уровень сложности)	Многофакторный MANOVA с повторными измерениями	Метод оценки максимального правдоподобия

Выявление предикторов высокой нагрузки	Множественный линейный регрессионный анализ (пошаговое включение)	Зависимые переменные: NASA-TLX, SMEQ
Структура связей между показателями	Корреляционный анализ (Пирсон / Спирмен) и факторный анализ (метод главных компонент с варимакс-вращением)	Проверка соответствия теоретической модели трех типов нагрузки
Качественный анализ протоколов	Тематический анализ по процедуре Браун и Кларк [23]	Выделенные темы использовались для интерпретации количественных результатов
Размер эффекта	η^2 (эта-квадрат) для ANOVA, *d*-Коэна для *t*-тестов, r_s для корреляций	Интерпретация по рекомендациям Коэна [24]

Уровень статистической значимости во всех тестах установлен на уровне $p < 0.05$. Размер эффекта был рассчитан по формулам

$$\eta^2 = \frac{SS_{\text{effect}}}{SS_{\text{total}}}, \quad d = \frac{M_1 - M_2}{SD}$$

для η^2 (эта-квадрат) и *d*-Коэна соответственно. Здесь где SS_{effect} – сумма квадратов, обусловленная влиянием изучаемого фактора; SS_{total} – общая сумма квадратов, отражающая общую вариабельность данных; M_1 и M_2 – средние двух групп; *SD* – общее стандартное отклонение. Интерпретация размеров эффекта проведена в соответствии с рекомендациями Коэна: малый ($\eta^2 \approx 0.01$, $d \approx 0.2$), средний ($\eta^2 \approx 0.06$, $d \approx 0.5$) и большой ($\eta^2 \approx 0.14$, $d \approx 0.8$) [24].

Исследование было проведено в соответствии с установленными этическими принципами. Конфиденциальность была обеспечена присвоением каждому участнику уникального идентификационного кода.

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ ПОЛЬЗОВАТЕЛЬСКОГО ВЗАИМОДЕЙСТВИЯ В МУЛЬТИРОЛЕВОЙ 3D-ПЛАТФОРМЕ

В ходе исследования проведено 71 тестирование (34 дизайнера, 37 покупателей), каждый участник выполнил по 9 задач. Все 639 наблюдений включены в анализ. Проверка распределений критерием Шапиро – Уилка показала отклонение от нормальности для большинства показателей ($p < 0.05$), поэтому

наряду с параметрическими были использованы непараметрические методы и бутстреп.

Объективные метрики когнитивной нагрузки

Время выполнения задач: среднее время по всем задачам составило 87.3 с (*SD* = 42.6). Стандартные отклонения рассчитаны по формуле

$$SD = \sqrt{\frac{(x_i - M)^2}{n-1}},$$

где x_i – каждое значение в наборе данных; M — среднее арифметическое всей совокупности; n — размер выборки.

Дизайнеры выполняли задачи быстрее (*M* = 74.2 с, *SD* = 31.8), чем покупатели (*M* = 99.1 с, *SD* = 48.3), различия статистически значимы (*U* = 34.5, $p < 0.001$, *d* = 0.62 – средний эффект), расчет *U*-критерия выполнен по формуле

$$U = (n_1 n_2) + \frac{[n_x(n_x + 1)]}{2} - R_x,$$

где n_x – размер выборки с большей суммой рангов; n_1 – размер первой выборки (дизайнеры); n_2 – размер второй выборки (покупатели); R_x – сумма рангов этой выборки.

Наибольший разрыв между группами наблюдался на задачах высокого уровня сложности (121.3 с у дизайнеров против 162.0 с у покупателей, *d* = 0.84) (см. табл. 4).

Табл. 4. Среднее время выполнения задач по ролям пользователей и уровням сложности (в с).

Уровень сложности задач	Дизайнеры (*M ± SD*)	Покупатели (*M ± SD*)	Различие	*p*	Размер эффекта (*d*)
Низкий	32.4 (12.7)	41.8 (18.3)	+9.4	0.012	0.43
Средний	68.9 (24.1)	93.5 (36.2)	+24.6	< 0.001	0.71
Высокий	121.3 (41.2)	162.0 (57.4)	+40.7	< 0.001	0.84

Количество ошибок: всего было зафиксировано 347 ошибок: среднее количество на задачу – 0.54 (*SD* = 0.31). У покупателей ошибок в 2.2 раза больше,

чем у дизайнеров (0.89 против 0.41 на задачах высокого уровня сложности, d = 0.92) (см. табл. 5).

Табл. 5. Среднее количество ошибок на задачу по ролям пользователей и уровням сложности задач.

Уровень сложности задач	Дизайнеры ($M \pm SD$)	Покупатели ($M \pm SD$)	Различие	p	Размер эффекта (d)
Низкий	0.08 (0.03)	0.14 (0.06)	+0.06	0.045	0.31
Средний	0.23 (0.11)	0.47 (0.19)	+0.24	< 0.001	0.67
Высокий	0.41 (0.18)	0.89 (0.27)	+0.48	< 0.001	0.92

Качественный анализ показал, что наиболее частые ошибки покупателей связаны со следующими моментами:

– непониманием применения ткани к модели (23% всех ошибок) – отсутствие визуальной обратной связи при кастомизации;

– потерей ориентации в многошаговом процессе создания заказа (18%);

– неудачными попытками вращения или масштабирования 3D-модели (15%) – отсутствие явных элементов управления.

У дизайнеров основные ошибки обусловлены заполнением форм: пропуск обязательных полей (28% ошибок) и загрузка файлов в неподдерживаемом формате (22% ошибок) из-за отсутствия явных указаний.

Неэффективные клики: дизайнеры совершают меньше неэффективных кликов (M = 1.8 на задачу, SD = 0.9), чем покупатели (M = 3.4, SD = 1.7), $p < 0.001$ (U-критерий Манна – Уитни, r_s = 0.46, средний эффект). Коэффициент корреляции Спирмена r_s был рассчитан по формуле

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)},$$

где d_i – разность между рангами соответствующих признаков; n – количество наблюдений (пар значений).

Полученные результаты свидетельствуют о более «извилистом» пути выполнения задач покупателями.

Субъективные оценки когнитивной нагрузки

SMEQ (оперативные замеры): оценки SMEQ высоко согласованы с объективными метриками: корреляция Спирмена с временем выполнения – $r_s = 0.72$ ($p < 0.001$) и с количеством ошибок – $r_s = 0.68$ ($p < 0.001$). Для покупателей задачи высокого уровня сложности оценены в 104.8 балла из 150 (69.9% от максимума), что указывает на высокую, но не экстремальную нагрузку (см. табл. 6). Качественный анализ протоколов подтверждает дискомфорт: «Я уже третий раз выбираю эту ткань, но не понимаю, применилась она или нет».

Табл. 6. Средние баллы SMEQ по ролям пользователей и уровням сложности задач (шкала 0–150 мм).

Уровень сложности задач	Дизайнеры (*M ± SD*)	Покупатели (*M ± SD*)	Различие	*p*	Размер эффекта (*d*)
Низкий	28.4 (12.1)	34.7 (15.3)	+6.3	0.038	0.35
Средний	52.7 (18.4)	71.2 (22.8)	+18.5	< 0.001	0.73
Высокий	81.3 (24.6)	104.8 (28.4)	+23.5	< 0.001	0.89

NASA-TLX (интегральная оценка): профили когнитивной нагрузки выявили принципиально разные паттерны для двух групп (см. табл. 7):

– ментальные требования значительно выше у покупателей (78.3 против 62.4 у дизайнеров, $d = 0.81$). Покупатели прикладывают больше усилий для понимания интерфейса, что согласуется с объективными данными (большее время и количество ошибок).

– временные требования выше у дизайнеров (54.2 против 43.8 у покупателей, $d = 0.58$). Несмотря на более быстрое выполнение задач (74.2 с против 99.1 с), дизайнеры воспринимают время как более ограниченное, вероятно, из-за профессиональных ожиданий.

– фрустрация – самый большой разрыв между группами (63.7 у покупателей против 42.3 у дизайнеров, $d = 0.93$). Основные источники фрустрации покупателей – это невозможность понять, применилось ли действие, и неочевидность сброса фильтров.

– субъективная успешность (обратная шкала) у покупателей значительно ниже (52.6 против 35.1, d = 0.89), что может снизить готовность повторно использовать платформу.

Табл. 7. Профили когнитивной нагрузки по шкалам NASA-TLX.

Шкала	Дизайнеры (*M ± SD*)	Покупатели (*M ± SD*)	Различие	*p*	Размер эффекта (*d*)
Ментальные требования	62.4 (14.2)	78.3 (16.7)	+15.9	< 0.001	0.81
Физические требования	18.7 (8.3)	24.1 (11.2)	+5.4	0.024	0.42
Временные требования	54.2 (15.6)	43.8 (14.1)	–10.4	0.003	0.58
Усилие	58.9 (13.8)	71.4 (15.9)	+12.5	< 0.001	0.74
Фрустрация	42.3 (17.2)	63.7 (19.4)	+21.4	< 0.001	0.93
Успешность (обратная*)	35.1 (12.8)	52.6 (16.3)	+17.5	< 0.001	0.89

**Примечание:* по шкале успешности более высокий балл означает более низкую оценку собственной успешности пользователя.

Совместное влияние факторов (MANOVA): для оценки совместного влияния роли пользователя (межгрупповой фактор) и уровня сложности задач (внутригрупповой фактор) был проведен двухфакторный MANOVA с повторными измерениями. Результаты представлены в табл. 8.

Табл. 8. Результаты двухфакторного MANOVA с повторными измерениями.

Эффект	Wilks' λ	*F*	df_1	df_2	*p*	η^2_p
Роль пользователя	0.67	18.4	6	64	< 0.001	0.12
Уровень сложности задач	0.52	24.7	12	140	< 0.001	0.15
Взаимодействие факторов (Роль × Уровень)	0.81	7.3	12	140	< 0.001	0.05

Лямбда Уилкса (Wilks' λ) представляет собой многомерный статистический тест для оценки значимости различий между факторами. Для расчета показателя размера эффекта η^2_p (частная эта-квадрат) применяют формулу

$$\eta_p^2 = \frac{df_1 \cdot F}{df_1 \cdot F + df_2},$$

где *F*-статистика описывает отношение межгрупповой дисперсии к внутригрупповой; df_1, df_2 – это степени свободы, которые используются в процессе анализа; *p* – достигнутый уровень статистической значимости: если $p < 0.05$, то результат считают статистически значимым [24].

Получены значимые главные эффекты роли пользователя ($p < 0.001$, $\eta^2_p = 0.12$ – средний эффект) и уровня сложности задач ($p < 0.001$, $\eta^2_p = 0.15$ – средний/большой эффект), а также значимый эффект их взаимодействия ($p < 0.001$, $\eta^2_p = 0.05$ – малый эффект). Это указывает на то, что влияние уровня сложности на когнитивную нагрузку неодинаково для двух групп: нагрузка покупателей в большей степени чувствительна к усложнению задач.

Качественный анализ проблемных зон: на основе анализа объективных метрик, протоколов «думай вслух» и посттестовых интервью выявлены 34 проблемные зоны, сгруппированные в три категории: навигация (№ 1), управление 3D-контентом (№ 2) и формы ввода данных (№ 3). Их распределение подтвердило ролевую специфику: у дизайнеров преобладают трудности с формами ввода данных (6 из 11 проблем), у покупателей – с 3D-контентом (7 из 15).

Приведем примеры проблемных зон интерфейса из категории № 1: **«Проблемы навигации».**

1. Неочевидность перехода между разделами «Мои модели» и «Заказы»: 8 из 34 пользователей-дизайнеров (23.5%) при выполнении задачи «Найти заказ по конкретной модели» начинали поиск в разделе «Мои модели» и затрачивали в среднем 45 с на ошибочный путь, прежде чем перейти в нужный раздел.

Среднее время выполнения задачи у пользователей, столкнувшихся с этой проблемой, составило 112 с ($SD = 24.3$) против 67 с ($SD = 18.7$) у пользователей, не столкнувшихся с этой проблемой (различия статистически значимы: $U = 34.5$, $p = 0.008$, $d = 0.91$), также при выполнении данной задачи было зафиксировано в среднем 3.4 неэффективных клика.

2. Потеря фильтров при переходе между страницами каталога: 15 из 37 пользователей-покупателей (40.5%) применяли фильтры, а затем переходили на следующую страницу результатов и обнаруживали, что фильтры сбросились.

Это вызывало у них сильную фрустрацию, поэтому 11 пользователей (29.7%) начали процесс фильтрации заново, а 4 (10.8%) не завершили выполнение задачи (зафиксирован тайм-аут 300 с).

Далее приведены примеры проблемных зон интерфейса из категории № 2: **«Проблемы управления 3D-контентом».**

1. Необнаруживаемость функций вращения модели одежды в 3D-окне: 22 из 37 пользователей-покупателей (59.5%) не использовали функцию вращения 3D-модели, хотя она была им доступна. 17 из 22 участников, не использовавших вращение (77.2%), что составляет 46% от всей группы покупателей, в посттестовом интервью сказали, что не знали о такой возможности.

Среди использовавших вращение (15 пользователей) среднее время изучения кастомизированной 3D-модели составило 47 с (SD = 12.3), а среди не использовавших вращение – только 23 с (SD = 8.7). Различия статистически значимы ($t(35) = 3.8$, $p < 0.001$, $d = 1.24$ – большой эффект). Расчет t-критерия Стьюдента выполнен по формуле

$$t = \frac{M_1 - M_2}{\sqrt{\frac{SD_1^2}{n_1} + \frac{SD_2^2}{n_2}}},$$

где M_1 и M_2 – средние двух групп; SD_1 и SD_2 – стандартные отклонения групп, n_1 и n_2 – размеры выборок.

Однако интерпретация этого различия требует осторожности: участники, использовавшие вращение, могли просто дольше рассматривать модель, а не «изучать ее полнее».

2. Отсутствие обратной связи при применении ткани или цвета: как пользователи-покупатели, так и пользователи-дизайнеры не всегда понимали, применилось ли изменение к 3D-модели одежды. Особенно остро проблема проявлялась, когда выбранные ткань или цвет были близки к исходным вариантам либо модель загружалась с задержкой.

Так, например, 14 из 37 пользователей-покупателей (37.8%) повторно нажимали на уже выбранную опцию, думая, что первое нажатие «не сработало», а 9 пользователей-дизайнеров (26.5%) в посттестовом интервью отметили, что «иногда непонятно, обновилась 3D-модель или нет».

3. Затруднения с масштабированием модели: функция масштабирования не была интуитивно понятна для 51.4% пользователей-покупателей. Они пытались применять на компьютере жесты, привычные для мобильных устройств, либо не могли найти соответствующие кнопки.

В ходе тестирования 19 из 37 пользователей-покупателей (51.4%) не смогли приблизить 3D-модель для рассмотрения ее деталей самостоятельно. Этим участникам модератор предоставлял вербальную подсказку после фиксации факта затруднения. Среднее время, затраченное на попытки масштабирования до получения подсказки, составило 28 с (*SD* = 9.4).

Теперь приведем примеры проблемных зон интерфейса из категории № 3: **«Проблемы форм ввода данных».**

1. Непонятные требования к форматам загружаемых файлов: при загрузке 3D-моделей дизайнеры получали сообщение об ошибке только после попытки загрузить файл. Сообщение было общим – «Ошибка загрузки файла» без указания причины.

В среднем пользователи-дизайнеры тратили 2.3 попытки (*SD* = 1.1, диапазон от 1 до 5) на успешную загрузку файла, 15 из 34 дизайнеров (44.1%) совершили хотя бы одну ошибку при выполнении данной задачи, затратив на нее в среднем на 52 с(*SD* = 18.7) больше, чем дизайнеры, выполнившие задачу с первой попытки.

2. Отсутствие индикации обязательных полей ввода данных: в формах создания карточки товара, а также в формах оформления заказа пользователи пропускали некоторые поля, обязательные для заполнения, и получали сообщение об ошибке при отправке формы.

Так, в ходе тестирования 21 участник из 71 (29.6%) столкнулся с этой проблемой, а 8 из них (38.1%) после получения сообщения об ошибке не смогли самостоятельно определить, какое конкретно поле пропущено в форме.

3. Слишком большое количество полей на одной странице: форма создания карточки товара в тестируемом варианте интерфейса содержала 14 полей ввода данных без логической группировки. Это вызвало у пользователей-дизайнеров когнитивную перегрузку и привело к пропуску некоторых полей.

Среднее время заполнения пользователями-дизайнерами данной формы составило 4.2 мин (*SD* = 1.3) против ожидаемых 2.5 мин по оценке экспертов (различия статистически значимы: $t(33) = 4.6$, $p < 0.001$, $d = 1.12$); 12 из них (35.3%) пропустили хотя бы одно поле при первом заполнении формы и обнаружили это только при неудачной попытке опубликовать карточку товара.

В ходе тестирования некоторые проблемные зоны были выявлены в обеих группах, но проявлялись они по-разному. Например, проблема с непониманием статуса загрузки 3D-модели: пользователи-дизайнеры беспокоились о технических аспектах (цитата из протокола «думай вслух»: «Долго грузится, наверное, модель тяжелая. Может, надо ее оптимизировать?»), а пользователи-покупатели интерпретировали это как поломку (цитата из протокола «думай вслух»: «Зависло. Надо обновить страницу»).

В табл. 9 представим результаты сопоставления предварительной экспертной классификации задач с полученными эмпирическими данными.

Табл. 9. Соответствие экспертной классификации задач по уровню когнитивной нагрузки эмпирическим данным.

Роль	Экспертная оценка уровня сложности задач	Совпадение с эмпирическими данными	Завышение уровня сложности задач	Занижение уровня сложности задач
Пользователи-дизайнеры	Низкий	100%	0%	0%
	Средний	75%	12.5%	12.5%
	Высокий	89%	11%	0%
Пользователи-покупатели	Низкий	93%	7%	0%
	Средний	62.5%	0%	37.5%
	Высокий	33%	0%	67%

Так, для покупателей 37.5% задач, отнесенных экспертами к среднему уровню, оказались высокосложными, а 67% задач, отнесенных экспертами к высокосложным, – значительно сложнее из-за необходимости одновременно управлять 3D-объектом, отсутствия обратной связи, потери ориентации в последовательности действий и отсутствия индикации прогресса. Для дизайнеров задачи, связанные с фильтрацией и сортировкой, оказались проще, чем ожидали эксперты, благодаря знакомым им паттернам интерфейса.

Выявленные отличия обусловлены тремя факторами, недостаточно учтенными в экспертной классификации: (1) новизной 3D-парадигмы для непрофессиональных пользователей, (2) профессиональной деформацией экспертов и (3) кумулятивным эффектом от сочетания нескольких мелких проблем, создающих «когнитивный шум» для непрофессиональных пользователей.

РАЗРАБОТКА И ВАЛИДАЦИЯ ПРИНЦИПОВ ПРОЕКТИРОВАНИЯ ИНТЕРФЕЙСА МУЛЬТИРОЛЕВОЙ 3D-ПЛАТФОРМЫ

На основе анализа 34 проблемных зон и их интерпретации были сформулированы принципы проектирования интерфейса новой цифровой платформы с 3D-функционалом, направленные на снижение когнитивной нагрузки для обеих ролей пользователей – дизайнеров и покупателей.

Принципы проектирования интерфейса

1. Принцип явного управления 3D-контентом: все функции для управления 3D-контентолжны быть представлены визуально различимыми элементами интерфейса, а также дублироваться через альтернативные способы взаимодействия.

2. Принцип многоуровневой обратной связи: каждое действие пользователя должно получать немедленную, явную и информативную обратную связь на нескольких уровнях: визуальном, текстовом и, при возможности, тактильном.

3. Принцип ролевой адаптации интерфейса: интерфейс должен динамически адаптироваться под текущую роль пользователя – «дизайнер» или «покупатель», изменяя плотность информации, сложность элементов управления и доступные функции.

4. Принцип превентивного информирования: вся информация, необходимая для корректного выполнения действия, должна быть доступна пользователю до начала действия, а не в момент ошибки.

5. Принцип сохранения контекста: состояния интерфейса (фильтры, сортировка, выбранные опции) должны сохраняться при навигации в пределах одной сессии, если иное не выбрано пользователем.

6. Принцип структурирования сложности: сложные многошаговые процессы должны быть разбиты на логические этапы с явной индикацией прогресса, большое количество полей – сгруппировано в визуально различимые блоки.

7. Принцип постепенного раскрытия функциональных возможностей (progressive disclosure): интерфейс должен показывать пользователю только те функции и опции, которые необходимы на текущем этапе работы, раскрывая их сложность по мере освоения.

На основе сформулированных принципов в прототип были внесены изменения. Наиболее значительные модификации произведены для экранной формы кастомизации для покупателей и формы создания карточки товара для дизайнеров.

Так, в модифицированной версии интерфейса для кастомизации (см. рис. 1) реализованы: явное управление 3D-контентом (кнопки вращения, слайдер масштабирования, анимация при загрузке, подсказки при наведении); многоуровневая обратная связь (подсветка выбранной ткани, плавное изменение текстуры, уведомление «Ткань применена»); ролевая адаптация (упрощенный режим для новых покупателей, кнопка «Больше опций»); «Мастер» кастомизации с прогресс-баром (шаг 1: выбор ткани, шаг 2: выбор фасона, шаг 3: выбор деталей).

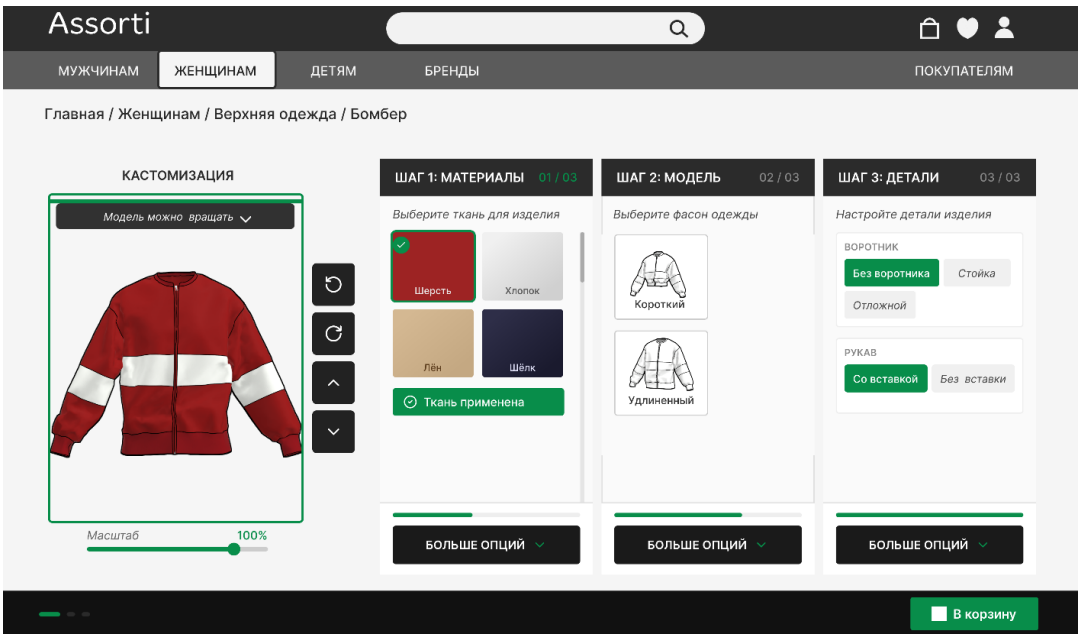


Рис. 1. Макет модифицированной экранной формы интерфейса для кастомизации модели одежды покупателем.

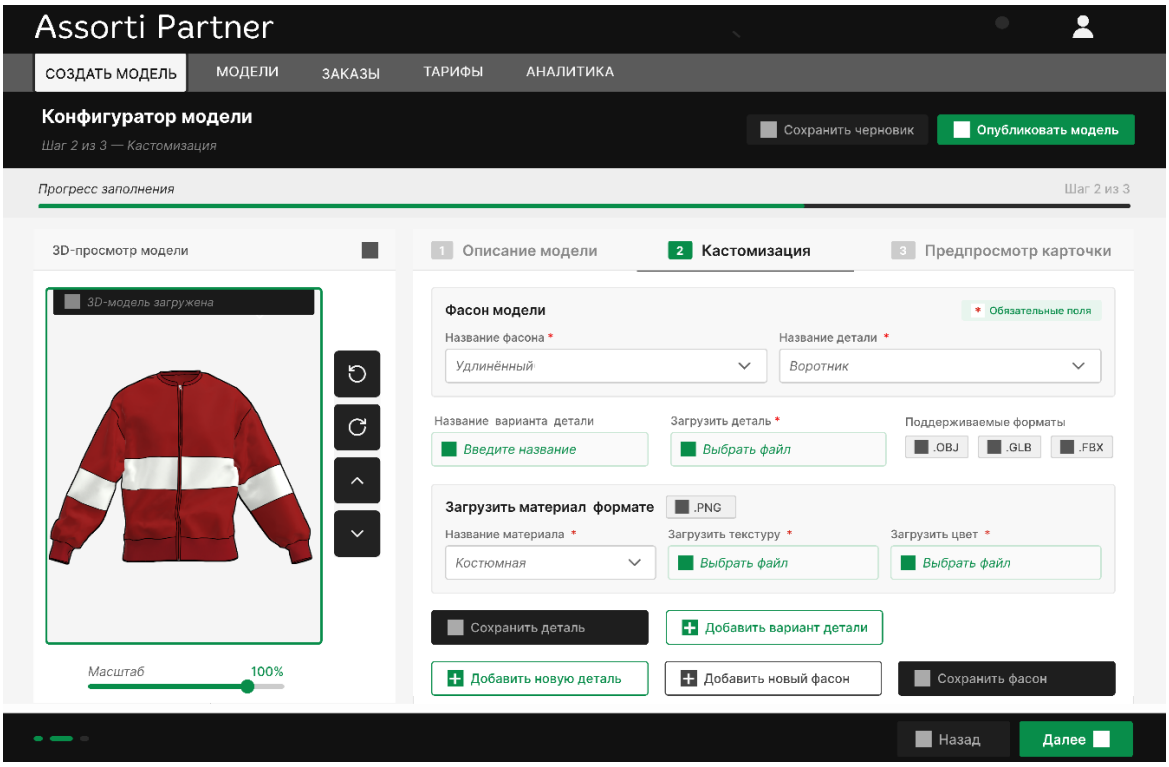


Рис. 2. Макет одной из экранных форм модифицированного интерфейса для создания карточки товара дизайнером.

На экранной форме для создания карточки товара (см. рис. 2) реализованы: структурирование сложности (форма разделена на три логических блока с аккордеоном, прогресс-бар); превентивное информирование (обязательные поля помечены звездочкой, указаны допустимые форматы с иконками, предпросмотр карточки до публикации, ссылка «Скачать модель» в режиме превью); сохранение контекста (сохранение введенных данных при переходах между блоками, возможность сохранить черновик).

Кроме того, были произведены и другие модификации: навигация (индикация связи между 3D-моделями и заказами и пр.); фильтры (сохранение состояния фильтров в URL и при пагинации и пр.); поиск (расширенный поиск для дизайнеров, упрощенный с автодополнением для покупателей).

Для оценки эффективности предложенных изменений проведено повторное тестирование с 21 новым участником (10 дизайнеров и 11 покупателей, не принимавших участия в первом этапе). Использовались те же задачи.

Сравнительный анализ метрик до и после модификации представлен в табл. 10 и 11. Наибольший эффект для дизайнеров дали превентивное ин-

формирование и структурирование сложности: снижение количества попыток загрузки на 48% (*d* = 0.68), снижение времени на 21% (*d* = 0.51). Для покупателей наибольший эффект дали явное управление 3D-контентом и многоуровневая обратная связь: рост использования 3D-функций на 50.4 процентных пункта (с 40.5% до 90.9%), снижение фрустрации на 35% (*d* = 0.69), снижение ошибок на 44% (*d* = 0.64). Важно отметить, что изменения интерфейса, направленные на снижение когнитивной нагрузки в одной группе, не ухудшили показатели в другой.

Табл. 10. Динамика показателей в группе пользователей-дизайнеров после модификации в интерфейсе.

Показатель	До модификации	После модификации	Изменение	*p*	Размер эффекта (*d*)
Среднее время выполнения задач, сек.	74.2	58.6	–15.6 (–21%)	< 0.01	0.51
Среднее количество ошибок на задачу	0.27	0.18	–0.09 (–33%)	< 0.05	0.42
Средний балл SMEQ	54.1	42.3	–11.8 (–22%)	< 0.01	0.47
Фрустрация (NASA-TLX)	42.3	31.7	–10.6 (–25%)	< 0.01	0.44
Количество попыток загрузки файла	2.3	1.2	–1.1 (–48%)	< 0.01	0.68

Табл. 11. Динамика показателей в группе пользователей-покупателей после модификации в интерфейсе.

Показатель	До модификации	После модификации	Изменение	*p*	Размер эффекта (*d*)
Среднее время выполнения задач, сек.	99.1	71.3	–27.8 (–28%)	< 0.01	0.58
Среднее количество ошибок на задачу	0.52	0.29	–0.23 (–44%)	< 0.01	0.64
Средний балл SMEQ	70.2	48.6	–21.6 (–31%)	< 0.01	0.63
Ментальные требования (NASA-TLX)	78.3	58.4	–19.9 (–25%)	< 0.01	0.52

Фрустрация (NASA-TLX)	63.7	41.2	–22.5 (–35%)	< 0.01	0.69
Использование вращения 3D-модели	40.5%	90.9%	+50.4%	< 0.01 (χ^2)	–

В результате исследования предложена модель когнитивного взаимодействия пользователей с мультиролевой 3D-платформой. Ее ключевой особенностью является замкнутость: три компонента опираются друг на друга и во взаимосвязи обеспечивают эффективность подхода.

1. Таксономический компонент – классификация задач для дизайнеров и покупателей с учетом когнитивной нагрузки (коэффициент конкордации Кендалла W = 0.82) и эмпирическая валидация, выявившая систематические расхождения между экспертной оценкой и практикой (для покупателей 67% задач, отнесенных к высокому уровню, продемонстрировали нагрузку, превышающую медианные значения).

2. Диагностический компонент – построение профилей когнитивной нагрузки на основе NASA-TLX (у дизайнеров – временное давление 54.2, у покупателей – фрустрация 63.7) и идентификация 34 проблемных зон (навигация – 10, управление 3D-контентом – 10, формы ввода данных – 14).

3. Проектный компонент – формулировка 7 принципов проектирования, их реализация в прототипе и эмпирическая валидация (снижение времени на 21–28%, ошибок на 33–44%, фрустрации на 25–35%; размеры эффекта d от 0.42 до 0.69).

Разработанные принципы проектирования интерфейса, а также модель когнитивного взаимодействия разных групп пользователей с 3D-платформой прошли эмпирическую валидацию. С учетом ограничений, связанных с междугрупповым дизайном и меньшим объемом выборки в повторном тестировании (n = 21), они могут быть рекомендованы для применения в аналогичных мультиролевых веб-приложениях, в том числе со сложным 3D-функционалом.

ЗАКЛЮЧЕНИЕ

Разработана, эмпирически обоснована и проверена на практике модель когнитивного взаимодействия пользователей с мультиролевой 3D-платформой

для кастомизации одежды. В эксперименте участвовали 34 дизайнера и 37 покупателей. Всего осуществлено 639 наблюдений.

Объективными метриками (время, ошибки, неэффективные клики) и субъективными шкалами (SMEQ, NASA-TLX) выявлены разные профили когнитивной нагрузки: дизайнеров замедляют формы (временное давление 54.2), покупателей – 3D-контент (фрустрация 63.7). На основе анализа 34 проблемных зон сформулированы 7 принципов проектирования интерфейса: явное управление 3D, многоуровневая обратная связь, ролевая адаптация, превентивное информирование, сохранение контекста, структурирование сложности и progressive disclosure. Повторное тестирование прототипа подтвердило их эффективность: время выполнения задач сократилось на 21–28%, количество ошибок – на 33–44%, фрустрация снизилась на 25–35%, использование 3D-функций выросло с 40.5% до 90.9%.

В качестве направлений для дальнейших исследований можно выделить проверку прогностического потенциала разработанной когнитивной модели, а также ее адаптацию для других предметных областей – конфигураторов обуви, интерьеров и пр. Для разрабатываемой платформы перспективна адаптация предложенных принципов под мобильные устройства, чтобы изучить сенсорные паттерны взаимодействия с 3D-контентом. Отдельный интерес представляет интеграция с VR и анализ источников когнитивной нагрузки в иммерсивных средах.

СПИСОК ИСТОЧНИКОВ

1. *Guo J., Zhao Y., Wenqian Z., Lu K., Feng X., Xia T.* The influence of the information richness of interfaces on consumers' purchase intention: the sequential mediating effects of cognitive load, mental imagery, and flow experience // Behavioral Sciences. 2025. Vol. 15, No 5. P. 673. https://doi.org/10.3390/bs15050673?urlappend=%3Futm_source%3Dresearchgate.net%26utm_medium%3Darticle

2. *Kumari A.* The Future of Commerce: How Customer Expectations Are Redefining B2B and B2C. Acorel Blog, 2025. URL: https://blog.acorel.nl/2025/12/the-future-of-commerce-how-customer-expectations-are-redefining-b2b-and-b2c/ (дата обращения 02.04.2026).

3. *Евдущенко Е.В., Шматко М.В.* Разработка цифровой платформы со встроенным 3D-конфигуратором для кастомизации одежды // Электронные библиотеки. Т. 29. Вып. 1. С. 207–239.
https://doi.org/10.26907/1562-5419-2026-29-1-207-239

4. *Paas F., Renkl A., Sweller J.* Cognitive Load Theory and Instructional Design: Recent Developments // Educational Psychologist. 2003. Vol. 38, No 1. P. 1–4. https://doi.org/10.1207/S15326985EP3801_1

5. *Raees M., Khan V.J., Papangelis K.* UX Challenges in Implementing an Interactive B2B Customer Segmentation Tool. In Companion Publication of the 2025 Conference on Human Factors in Computing Systems (CHI'25), 2025.
https://doi.org/10.1145/3706599.3706687

6. *Wasilewski A., Kolaczek G.* One size does not fit all: multivariant user interface personalization in e-commerce // IEEE Access. 2024. Vol. 12. P. 65570–65582. https://doi.org/10.1109/ACCESS.2024.3398192

7. *Hart S.G., Staveland L.E.* Development of NASA-TLX: Results of Empirical and Theoretical Research // Advances in Psychology. 1988. Vol. 52. P. 139–183.
https://doi.org/10.1016/S0166-4115(08)62386-9

8. *Jin H., Xiao M., Liu L., Kan S., Fu Y, Zhang D.* Relationship between physical fatigue and mental fatigue based on multimodal measurement under different load levels // Ergonomics. 2024. Vol. 67(11). P. 1748–1763.
https://doi.org/10.1080/00140139.2024.2364667

9. *Von Janczewski N., Kraus J., Engeln A., Baumann M.* A subjective one-item measure based on NASA-TLX to assess cognitive workload in driver-vehicle interaction. Transportation Research Part F: Traffic Psychology and Behavior. 2022. Vol. 86. P. 210–225. https://doi.org/10.1016/j.trf.2022.02.012

10. *Qiu Y., Duan H., Xie H., Ding X., Jiao Y.* Design and development of a web-based interactive twin platform for watershed management // Transactions in GIS. 2022. Vol. 26. P. 1299–1317. https://doi.org/10.1111/tgis.12904

11. *Sudár A., Csapó Á.B.* Comparing desktop 3d virtual reality with web 2.0 interfaces: identifying key factors behind enhanced user capabilities // Heliyon. 2024. Vol. 10, No 11. P. e31717. https://doi.org/10.1016/j.heliyon.2024.e31717

12. *Albert B., Tullis T.* Measuring the User Experience (3rd ed.). Morgan Kaufmann, 2022. 384 p.
URL:https://www.educate.elsevier.com/book/details/9780128180808 (дата обращения 10.04.2026).

13. *Wickens C.D., Hollands J.G., Banbury S., Parasuraman R.* Engineering Psychology and Human Performance (4th ed.). Pearson, 2013. 518 p.
URL: https://studentebookhub.com/wp-content/uploads/2024/preview/9781032042145.pdf (дата обращения 20.04.2026).

14. *Kendall M.G., Babington Smith B.* The problem of m rankings // The Annals of Mathematical Statistics. 1939. Vol. 10, No. 3. P. 275–287.
https://doi.org/10.1214/aoms/1177732186

15. *Schmidt C., Collette F., Cajochen C., Peigneux P*. A time to think: Circadian rhythms in human cognition // Cognitive Neuropsychology. 2007. Vol. 24, No. 7. P. 755–789. https://doi.org/10.1080/02643290701754158

16. *Sauro J., Dumas J.S.* Comparison of three one-question, post-task usability questionnaires // In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09). Association for Computing Machinery, New York. NY, USA. 2009. P. 1599–1608. https://doi.org/10.1145/1518701.1518946

17. *Taylor S.S., Bolton M.L.* Design matters: Scale design and rounding ratings impacts NASA-TLX results. Proceedings of the Human Factors and Ergonomics Society Annual Meeting. 2025. Vol. 69, No 1. P. 1192–1194.
https://doi.org/10.1177/10711813251371038

18. *Голова М.* Протокол «Думай вслух»: что это такое и как его использовать. 4brain.ru, 2024. URL: https://4brain.ru/blog/protokol-dumay-vsluh-chto-eto-takoe-i-kak-ego-ispolzovat/ (дата обращения 03.04.2026).

19. *Nielsen J.* How Many Test Users in a Usability Study? Nielsen Norman Group, 2012. URL: https://www.nngroup.com/articles/how-many-test-users/ (дата обращения 04.04.2026).

20. *Faul F., Erdfelder E., Lang A.G., Buchner A.* G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences // Behav. Res. Methods. 2007. Vol. 39, No. 2. P. 175–191. https://doi.org/10.3758/BF03193146

21. *Tabachnick B.G., Fidell L.S.* Using Multivariate Statistics. Seventh edition. New York: Pearson, 2019. 850 p. URL: https://pdf4pro.com/view/seventh-edition-using-multivariate-statistics-692f15.html (дата обращения 04.05.2026).

22. *Armstrong R.A.* When to use the Bonferroni correction // Ophthalmic and Physiological Optics. 2014. Vol. 34, No. 5, P. 502–508. https://doi.org/10.1111/opo.12131

23. *Braun V., Clarke V.* Thematic Analysis: A Practical Guide. SAGE Publications. 2021. 376 p. URL: https://books.google.ru/books/about/Thematic_Analysis.html?id=mToqEAAAQBAJ&redir_esc=y (дата обращения 18.04 .2026).

24. *Cohen J.* Statistical Power Analysis for the Behavioral Sciences (2nd ed.). NY: Routledge.1988. 567 p. https://doi.org/10.4324/9780203771587

COGNITIVE MODELING OF USER INTERACTION IN MULTI-ROLE WEB APPLICATIONS WITH 3D FUNCTIONALITY

M. V. Shmatko[1] [0000-0002-7255-8885], E. V. Evdushenko[2] [0000-0003-3692-2587]

[1]*Omsk State Technical University, Omsk, Russia*

[2]*Military Engineering Institute of the Military Logistics Academy, Saint Petersburg, Russia*

[1]marin298@gmail.com, [2]elena.online_ktilp@mail.ru

Abstract

The article investigates cognitive load in users of a multi-role 3D platform. An experiment (71 participants: 34 designers, 37 customers; 639 observations) using objective metrics and subjective scales revealed different load profiles: designers are slowed down by data entry forms, while customers are hindered by 3D content. Based on 34 problem areas, seven interface design principles and a task classification by cognitive complexity level, accounting for role-specific factors, were formulated. Retesting (21 participants) confirmed their effectiveness: task completion time decreased by 21–28%, errors by 33–44%, frustration by 25–35%, and the use of 3D features increased from 40.5% to 90.9%. The developed three-component model

(taxonomic, diagnostic, and design levels) is recommended for designing similar multi-role platforms with 3D functionality.

Keywords: *cognitive load, multi-role web applications, user 3D interfaces, NASA-TLX, SMEQ, cognitive modeling, UX research, human-computer interaction.*

REFERENCES

1. *Guo J., Zhao Y., Wenqian Z., Lu K., Feng X., Xia T.* The influence of the information richness of interfaces on consumers' purchase intention: the sequential mediating effects of cognitive load, mental imagery, and flow experience // Behavioral Sciences. 2025. Vol. 15, No. 5. P. 673. https://doi.org/10.3390/bs15050673?urlappend=%3Futm_source%3Dresearchgate.net%26utm_medium%3Darticle

2. *Kumari A.* The Future of Commerce: How Customer Expectations Are Redefining B2B and B2C. Acorel Blog, 2025. URL: https://blog.acorel.nl/2025/12/the-future-of-commerce-how-customer-expectations-are-redefining-b2b-and-b2c/.

3. *Evdushchenko E.V., Shmatko M.V.* Razrabotka cifrovoy platformy so vstroennym 3D-konfiguratorom dlya kastomizacii odezhdy [Development of a digital platform with an integrated 3D configurator for clothing customization]. Russian Digital Libraries Journal. 2026. Vol. 29, No. 1. S. 207–239. https://doi.org/10.26907/1562-5419-2026-29-1-207-239

4. *Paas F., Renkl A., Sweller J.* Cognitive Load Theory and Instructional Design: Recent Developments // Educational Psychologist. 2003. Vol. 38, No. 1. P. 1–4. https://doi.org/10.1207/S15326985EP3801_1

5. *Raees M., Khan V.J., Papangelis K.* UX Challenges in Implementing an Interactive B2B Customer Segmentation Tool. In Companion Publication of the 2025 Conference on Human Factors in Computing Systems (CHI'25), 2025. https://doi.org/10.1145/3706599.3706687

6. *Wasilewski A., Kolaczek G.* One size does not fit all: multivariant user interface personalization in e-commerce // IEEE Access. 2024. Vol. 12. P. 65570–65582. https://doi.org/10.1109/ACCESS.2024.3398192

7. *Hart S.G., Staveland L.E.* Development of NASA-TLX: Results of Empirical and Theoretical Research // Advances in Psychology. 1988. Vol. 52. P. 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

8. *Jin H., Xiao M., Liu L., Kan S., Fu Y, Zhang D.* Relationship between physical fatigue and mental fatigue based on multimodal measurement under different load levels // Ergonomics. 2024. Vol. 67(11). P. 1748–1763. https://doi.org/10.1080/00140139.2024.2364667

9. *Von Janczewski N., Kraus J., Engeln A., Baumann M.* A subjective one-item measure based on NASA-TLX to assess cognitive workload in driver-vehicle interaction. Transportation Research Part F: Traffic Psychology and Behavior. 2022. Vol. 86. P. 210–225. https://doi.org/10.1016/j.trf.2022.02.012

10. *Qiu Y., Duan H., Xie H., Ding X., Jiao Y.* Design and development of a web-based interactive twin platform for watershed management // Transactions in GIS. 2022. Vol. 26. P. 1299–1317. https://doi.org/10.1111/tgis.12904

11. *Sudár A., Csapó Á.B.* Comparing desktop 3d virtual reality with web 2.0 interfaces: identifying key factors behind enhanced user capabilities // Heliyon. 2024. Vol. 10, No 11. P. e31717. https://doi.org/10.1016/j.heliyon.2024.e31717

12. *Albert B., Tullis T.* Measuring the User Experience (3rd ed.). Morgan Kaufmann, 2022. 384 p. URL:https://www.educate.elsevier.com/book/details/9780128180808 (accessed: April 10, 2026).

13. *Wickens C.D., Hollands J.G., Banbury S., Parasuraman R.* Engineering Psychology and Human Performance (4th ed.). Pearson, 2013. 518 p. URL: https://studentebookhub.com/wp-content/uploads/2024/preview/9781032042145.pdf (accessed: April 10, 2026).

14. *Kendall M.G., Babington Smith B.* The problem of m rankings // The Annals of Mathematical Statistics. 1939. Vol. 10, No. 3. P. 275–287. https://doi.org/10.1214/aoms/1177732186

15. *Schmidt C., Collette F., Cajochen C., Peigneux P*. A time to think: Circadian rhythms in human cognition // Cognitive Neuropsychology. 2007. Vol. 24, No. 7. P. 755–789. https://doi.org/10.1080/02643290701754158

16. *Sauro J., Dumas J.S.* Comparison of three one-question, post-task usability questionnaires // In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09). Association for Computing Machinery, New York. NY, USA. 2009. P. 1599–1608. https://doi.org/10.1145/1518701.1518946

17. *Taylor S.S., Bolton M.L.* Design matters: Scale design and rounding ratings impacts NASA-TLX results. Proceedings of the Human Factors and Ergonomics Society Annual Meeting. 2025. Vol. 69, No. 1. P. 1192–1194. https://doi.org/10.1177/10711813251371038

18. *Golova M.* Protokol «Dumaj vsluh»: chto ehto takoe i kak ego ispol'zovat'. 4brain.ru. 2024. URL:https://4brain.ru/blog/protokol-dumay-vsluh-chto-eto-takoe-i-kak-ego-ispolzovat/.

19. *Nielsen J.* How Many Test Users in a Usability Study? Nielsen Norman Group, 2012. URL: https://www.nngroup.com/articles/how-many-test-users/ (accessed: April 10, 2026).

20. *Faul F., Erdfelder E., Lang A.G., Buchner A.* G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences // Behav. Res. Methods. 2007. Vol. 39, No. 2. P. 175–191. https://doi.org/10.3758/BF03193146

21. *Tabachnick B.G., Fidell L.S.* Using Multivariate Statistics. Seventh edition. New York: Pearson, 2019. 850 p. URL: https://pdf4pro.com/view/seventh-edition-using-multivariate-statistics-692f15.html (accessed: April 10, 2026).

22. *Armstrong R.A.* When to use the Bonferroni correction // Ophthalmic and Physiological Optics. 2014. Vol. 34, No. 5. P. 502–508. https://doi.org/10.1111/opo.12131

23. *Braun V., Clarke V.* Thematic Analysis: A Practical Guide. SAGE Publications. 2021. 376 p.
URL: https://books.google.ru/books/about/Thematic_Analysis.html?id=mToqEAAAQBAJ&redir_esc=y (accessed: April 10, 2026).

24. *Cohen J.* Statistical Power Analysis for the Behavioral Sciences (2nd ed.). NY: Routledge. 1988. 567 p. https://doi.org/10.4324/9780203771587

СВЕДЕНИЯ ОБ АВТОРАХ



ШМАТКО Марианна Владимировна – к. филос. н, доцент кафедры «Математические методы и информационные технологии в экономике» ОмГТУ. Сфера научных интересов – разработка новых цифровых продуктов, UX-аналитика и юзабилити информационных систем, гейм-дизайн. Сфера практической деятельности – разработка технологических стартап-проектов.

Marianna Vladimirovna SHMATKO – Candidate of Philosophical Sciences, associate professor at the Department "Mathematical Methods and Information Technologies in Economics" of Omsk State Technical University. Research interests: digital product development, UX analytics, information systems usability, and game design. Practical activities: development of technology startup projects.

email: marin298@gmail.com

ORCID: 0000-0002-7255-8885



ЕВДУЩЕНКО Елена Владимировна – к. техн. н., доцент кафедры «Военная архитектура, автоматизированные системы проектирования, естественнонаучные дисциплины» ВИ (ИТ) ВА МТО, г. Санкт-Петербург. Сфера научных интересов – 3D-моделирование одежды, развитие ИТ-продуктов.

Elena Vladimirovna EVDUSHCHENKO – Candidate of Technical Sciences, associate professor at the Department “Military Architecture, Automated Design Systems and Natural Sciences” of the Military Engineering Institute of the Military Logistics Academy, St. Petersburg. Research interests include 3D modeling of clothing and the development of IT products.

email: elena.online_ktilp@mail.ru

ORCID: 0000-0003-3692-2587

Материал проступил в редакцию 22 июня 2026 года