ОГЛАВЛЕНИЕ

М. Ш. Адыгамов, А. О. Голубь, Э. Р. Сайфуллин,
Т. Р. Гимадиев, Н. Ю. Серов
ИНТЕЛЛЕКТУАЛЬНЫЙ РОБОТ-ХИМИК:
НА ПУТИ К АВТОНОМНОЙ ЛАБОРАТОРИИ **997–1014**

Р. А. Бурнашев, Я. В. Сергеев
ПРОЕКТИРОВАНИЕ ДИНАМИЧЕСКОЙ ЭКСПЕРТНОЙ СИСТЕМЫ
ПО АНАЛИЗУ ВЛИЯНИЯ КЛИМАТИЧЕСКИХ ВОЗДЕЙСТВИЙ
НА МАЛЫЕ И СРЕДНИЕ ПРЕДПРИЯТИЯ **1015–1035**

В. К. Вершинин, И. В. Ходненко, С. В. Иванов
НОРМАЛИЗАЦИЯ ТЕКСТА, РАСПОЗНАННОГО ПРИ ПОМОЩИ
ТЕХНОЛОГИИ ОПТИЧЕСКОГО РАСПОЗНАВАНИЯ СИМВОЛОВ,
С ИСПОЛЬЗОВАНИЕМ ЛЕГКОВЕСНЫХ LLM **1036–1056**

А. М. Ганиева
ЦИФРОВОЕ МОДЕЛИРОВАНИЕ ТЕМАТИЧЕСКОГО ПОЛЯ ИЗУЧЕНИЯ
КУЛЬТУРНОЙ КОНГРУЭНТНОСТИ В ПСИХОЛОГИЧЕСКОМ КОНТЕКСТЕ **1057–1069**

Ю. А. Загорулько, Е. А. Сидорова, И. Р. Ахмадеева
АВТОМАТИЧЕСКОЕ ИЗВЛЕЧЕНИЕ АРГУМЕНТАТИВНЫХ ОТНОШЕНИЙ
ИЗ ТЕКСТОВ НАУЧНОЙ КОММУНИКАЦИИ **1070–1084**

М. В. Исангулов, А. М. Елизаров, А. Р. Кунафин,
А. Р. Гатиатуллин, Н. А. Прокопьев
НЕЙРОСИМВОЛИЧЕСКИЙ ПОДХОД К ДОПОЛНЕННОЙ ГЕНЕРАЦИИ
ТЕКСТА НА ОСНОВЕ АВТОМАТИЗИРОВАННОЙ ИНДУКЦИИ
МОРФОТАКТИЧЕСКИХ ПРАВИЛ **1085–1102**

А. А. Красновский
ОЦЕНКА НЕОПРЕДЕЛЕННОСТИ В ТРАНСФОРМЕРНЫХ ЦЕПЯХ
НА ОСНОВЕ ПРИНЦИПА СОГЛАСОВАННОСТИ
ЭФФЕКТИВНОЙ ИНФОРМАЦИИ **1103–1119**

Д. А. Лютова, В. А. Малых
АБСТРАКТИВНАЯ СУММАРИЗАЦИЯ НОВОСТЕЙ ВНЕШНЕЙ ТОРГОВЛИ НА ОСНОВЕ НОВОГО СПЕЦИАЛИЗИРОВАННОГО КОРПУСА ДАННЫХ 1120–1137

Д. Р. Пойманов, М. С. Шутов
ИССЛЕДОВАНИЕ КВАНТОВАНИЯ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ: ОЦЕНКА ЭФФЕКТИВНОСТИ С АКЦЕНТОМ НА РУССКОЯЗЫЧНЫЕ ЗАДАЧИ 1138–1164

О. Ю. Рогов, Д. Е. Инденбом, Д. С. Корж, Д. В. Пугачёва, В. А. Воронов, Е. В. Тутубалина
СОКРЫТИЕ В СМЫСЛЕ: СЕМАНТИЧЕСКОЕ КОДИРОВАНИЕ ДЛЯ ГЕНЕРАТИВНО-ТЕКСТОВОЙ СТЕГАНОГРАФИИ 1165–1185

И. А. Свиридов, К. С. Егоров
УСЛОВНАЯ ГЕНЕРАЦИЯ ЭЛЕКТРОКАРДИОГРАММ С ПОМОЩЬЮ ИЕРАРХИЧЕСКИХ ВАРИАЦИОННЫХ АВТОКОДИРОВЩИКОВ 1186–1206

А. Таха, Р. А. Лукманов
ГДЕ НАХОДЯТСЯ ЛУЧШИЕ ПРИЗНАКИ? ПОСЛОЙНЫЙ АНАЛИЗ СЛОЕВ ТРАНСФОРМЕРА ДЛЯ ЭФФЕКТИВНОЙ КЛАССИФИКАЦИИ ЭНДОСКОПИЧЕСКИХ ИЗОБРАЖЕНИЙ 1207–1229

Ю. В. Трофимов, А. Д. Лебедев, А. С. Ильин, А. Н. Аверкин
ЯДРО ВЕРИФИЦИРУЕМОЙ ОБЪЯСНИМОСТИ: ГИБРИДНАЯ АРХИТЕКТУРА GD-ANFIS/SHAP ДЛЯ XAI 2.0 * 1230–1252

П. А. Филоненко, В. Н. Кох, П. Д. Блинов
ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ В РЕШЕНИИ ПРОБЛЕМЫ ОНКОПРОФИЛАКТИКИ: РЕТРОСПЕКТИВНОЕ ИССЛЕДОВАНИЕ 1253–1266

И. З. Хаялеева, М. М. Абрамский
СТИЛОМЕТРИЧЕСКИЙ АНАЛИЗ В ЗАДАЧЕ ПОИСКА ЗАИМСТВОВАНИЙ ТЕКСТОВ НА ТАТАРСКОМ ЯЗЫКЕ 1267–1278

Тематический выпуск по материалам Междисциплинарной конференции по искусственному интеллекту «ИИ-ЗАМАН», 17 сентября 2025 года

Редакторы-составители: В.С. Белобородов, А.Р. Гатиатуллин, Р.А. Гильмуллин, Л.Р. Гильмутдинова, А.К. Ковалёв, А.В. Кузнецов, О.В. Попова, Е.В. Тутубалина, А.Ф. Хасьянов, А.А. Шпильман

ОТ СОСТАВИТЕЛЕЙ

Настоящий тематический выпуск журнала «Электронные библиотеки» включает статьи, подготовленные на основе докладов, представленных на Междисциплинарной научной конференции «ИИ-ЗАМАН». Конференция прошла 17 сентября 2025 года в рамках международного форума «Kazan Digital Week – 2025» и была посвящена фундаментальным и прикладным исследованиям в области искусственного интеллекта.

Основные направления конференции: компьютерное зрение, обработка естественного языка, воплощённый искусственный интеллект и робототехника, применение искусственного интеллекта в научных исследованиях.

Основная цель проведенной конференции — объединение специалистов, исследователей и студентов для обсуждения современных актуальных задач искусственного интеллекта, обмена результатами и опытом, а также содействие междисциплинарному научному диалогу. Организаторами конференции выступили Академия наук Республики Татарстан, Институт искусственного интеллекта AIRI и Университет Иннополис.

УДК 004.896

ИНТЕЛЛЕКТУАЛЬНЫЙ РОБОТ-ХИМИК: НА ПУТИ К АВТОНОМНОЙ ЛАБОРАТОРИИ

М. Ш. Адыгамов[1] [0009-0006-2364-9867], А. О. Голубь[2] [0009-0004-0090-0292],
Э. Р. Сайфуллин[3] [0000-0003-0823-9051], Т. Р. Гимадиев[4] [0000-0001-5012-0308],
Н. Ю. Серов[5] [0000-0002-5772-8399]

[1,3-5]Федеральный исследовательский центр «Казанский научный центр Российской академии наук», г. Казань, Россия
[1-5]Казанский (Приволжский) федеральный университет, г. Казань, Россия

[1]musa20930@gmail.com, [2]toxa.mix7@gmail.com, [3]mr.emilsr@gmail.com,
[4]Timur.Gimadiev@gmail.com, [5]Serov.Nikita@gmail.com

Аннотация

Представлена программно-аппаратная платформа, которая позволяет проводить химические синтезы в автоматическом режиме, включая приготовление реакционных смесей, их нагрев и перемешивание, а также отбор проб с разбавлением после синтеза и отправку на анализ методом высокоэффективной жидкостной хроматографии с последующей автоматической обработкой результатов. Для управления отдельными элементами роботизированной установки создана собственная библиотека ChemBot на языке Python, а для управления всей системой – клиентский веб-сервер; для просмотра состояния установки и хода выполнения синтезов разработан веб-интерфейс. Работа всей платформы по выполнению экспериментов протестирована при выполнении синтезов по альдольной конденсации, где варьировались соотношение реагентов, катализатор и его количество, температура и время синтеза. Написание собственного кода для контроля и управления всей системой стало важным шагом на пути интеграции роботизированной установки и искусственного интеллекта (ИИ), что в перспективе позволит осуществить переход к автономной лаборатории, когда предсказание целевой молекулы и ее синтеза, экспериментальное осуществление и анализ, а также, при необходимости, уточнение или изменение использованной модели будут осуществляться в автоматическом режиме, без вмешательства человека.

© М. Ш. Адыгамов, А. О. Голубь, Э. Р. Сайфуллин, Т. Р. Гимадиев, Н. Ю. Серов, 2025.
Данная статья распространяется на условиях международной лицензии Creative Commons License Attribution 4.0 International (CC BY 4.0).

***Ключевые слова**: искусственный интеллект, роботизация, химический синтез, автономная лаборатория, хемоинформатика.*

ВВЕДЕНИЕ

Современные химические исследования столкнулись с парадоксом: растущая сложность задач (синтез новых материалов, лекарственных соединений, оптимизация процессов) требует экспоненциального повышения количества и сложности экспериментов, в то время как традиционные лабораторные методы остаются ресурсоемкими и длительными. Решением этой проблемы стали автоматизированные лаборатории (англ. self-driving labs, SDLs [1]), интегрирующие аппаратные платформы, методы машинного обучения (МО), искусственного интеллекта (ИИ) и облачные технологии. Их развитие трансформирует научный процесс, что обеспечивает:

- высокую пропускную способность (параллельное проведение сотен реакций);
- воспроизводимость (исключение «человеческого фактора»);
- автономность (круглосуточная работа без вмешательства оператора);
- интеллектуальную оптимизацию (адаптивное планирование экспериментов на основе данных, полученных на экспериментальной установке).

Однако, несмотря на очевидные преимущества, внедрение автоматизированных систем в область химического синтеза сопряжено с рядом серьезных вызовов [2]. Среди них можно выделить отсутствие аппаратной гибкости некоторых коммерческих роботизированных установок, их высокую стоимость [3, 4], а также сложность разработки и объединения всех процессов: предсказание эксперимента, его проведение и анализ – в автономную систему, в которой все процессы последовательно выполняются без участия человека, а полученные данные используются для дообучения моделей МО и ИИ [5]. Стоит отметить, что некоторые научные группы занимаются разработкой открытых и более доступных решений для автоматизации синтезов [6, 7], но, как правило, подобные системы не являются коммерчески доступными и имеют некоторые конструктивные ограничения (например, установка [6] имеет ограниченное число реакторов, а устройство в [7] создавалось под определенный узкий круг задач).

Для создания полностью автономной платформы автоматизации необходимо интегрировать три программных пакета, работающих в режиме «закрытого цикла»: ПО для поиска целевых молекул и оптимального пути их синтеза, ПО для проведения эксперимента и ПО для обработки экспериментальных данных [5]. В традиционном подходе, в отличие от систем «закрытого цикла», модели планирования синтезов обучаются на ограниченном наборе данных, после чего их предсказания проверяются экспериментально (вручную или автоматизированным способом). В подходе «закрытого цикла» модели продолжают дообучаться за счет новых данных, полученных на автоматизированных установках (при этом автоматизируется и синтез, и анализ). Подобные решения позволяют быстрее получать результаты более высокого качества по сравнению с традиционным подходом к дизайну новых соединений.

В настоящей работе представлена программно-аппаратная платформа для автономного проведения химических синтезов, их анализа и планирования. В текущей версии программный пакет ChemBot включает ПО для проверки возможности приготовления синтезов, приготовления реакционных смесей на описанной далее установке, проведения и планирования параллельных синтезов, отбора образцов на анализ и автоматического запуска анализов на системе высокоэффективной жидкостной хроматографии (ВЭЖХ). Система была использована для проведения реакций альдольной конденсации в присутствии аминокислотных комплексов цинка и хорошо показала себя при осуществлении синтезов с различными условиями такими, как варьирование реагентов и их соотношений, катализатора и его количества, а также условий – температуры, кислотности среды за достаточно короткое время.

УСТРОЙСТВО РОБОТИЗИРОВАННОЙ УСТАНОВКИ

Установка, используемая для проведения синтезов, сделана на базе автоматизированной системы LifeBot для ПЦР-тестов (ООО «Эвотэк-Мирай Геномикс» г. Иннополис, Россия) [8]. С учетом внесенных модификаций робот способен параллельно проводить до 48 химических синтезов при нагревании и перемешивании в реакционной зоне. Роботизированная установка работает в сочетании с

ВЭЖХ-системой Smartline (фирма Knauer, Германия) [9] с возможностью автоматического отбора проб и запуска анализа. Общий вид установки представлен на рис. 1.

Часть исходных элементов установки была заменена на собственные разработки, среди которых можно отметить включение в рабочее поле реакторных зон, приемника для отбора проб, создание нового держателя пипеток. Все модификации робота сделаны с использованием коммерчески доступной электроники и деталей, напечатанных на 3D-принтере, что в перспективе позволит удешевить подобные установки, сделать их более доступными для научного сообщества и популяризировать роботизированные установки в химических лабораториях.

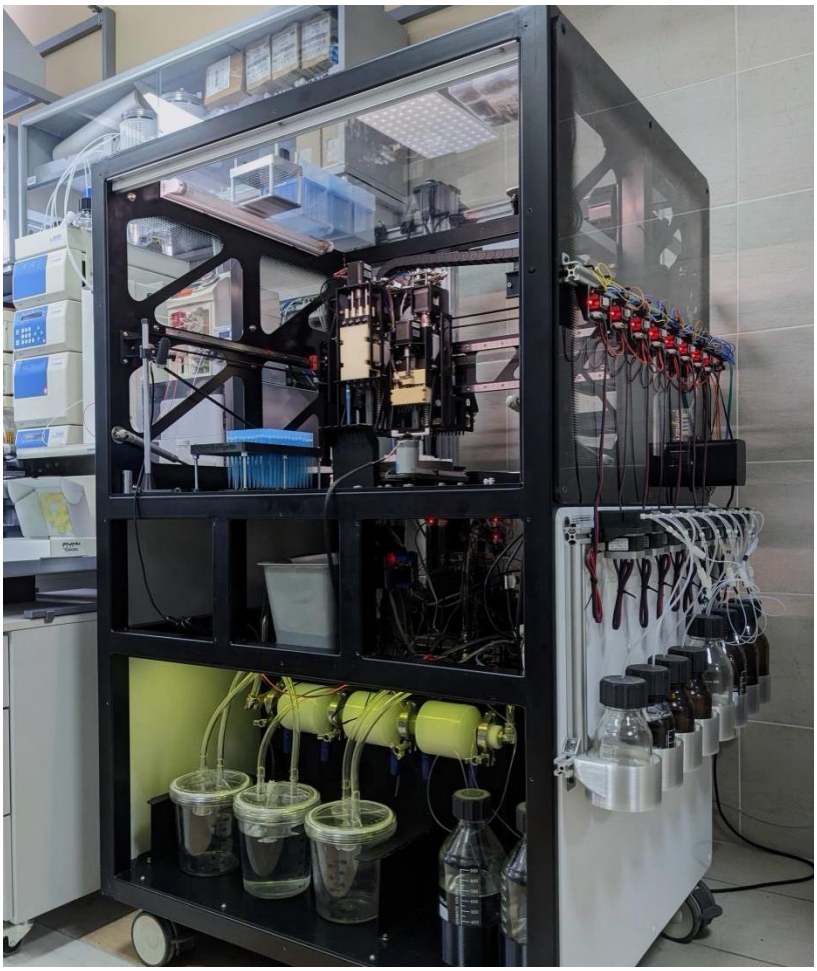


Рис. 1. Общий вид установки. По центру находится рабочее поле с манипулятором. Справа расположены хранилища растворителей с перистальтическими насосами. Слева на заднем плане виден хроматограф.

Для более понятного представления об устройстве установки на рис. 2 приведена ее блок-схема. Центральной частью установки является рабочее поле (обведено зеленым), в котором расположены различные устройства. Внутри поля движется манипулятор, подключенный к линиям растворителей, а также имеющий встроенную пипетку, благодаря чему может переносить жидкие реагенты или их растворы и дозировать растворители (переносы жидкостей показаны синими стрелками на рис. 2). Для исключения загрязнения реагентов и образцов предусмотрена смена наконечников пипетки – чистые наконечники берутся из хранилища наконечников, а для грязных предусмотрен сброс. Дополнительно на манипуляторе имеется устройство для открытия/закрытия реакторов и хранилищ, что необходимо для предотвращения испарения при выполнении синтезов и при длительном хранении.

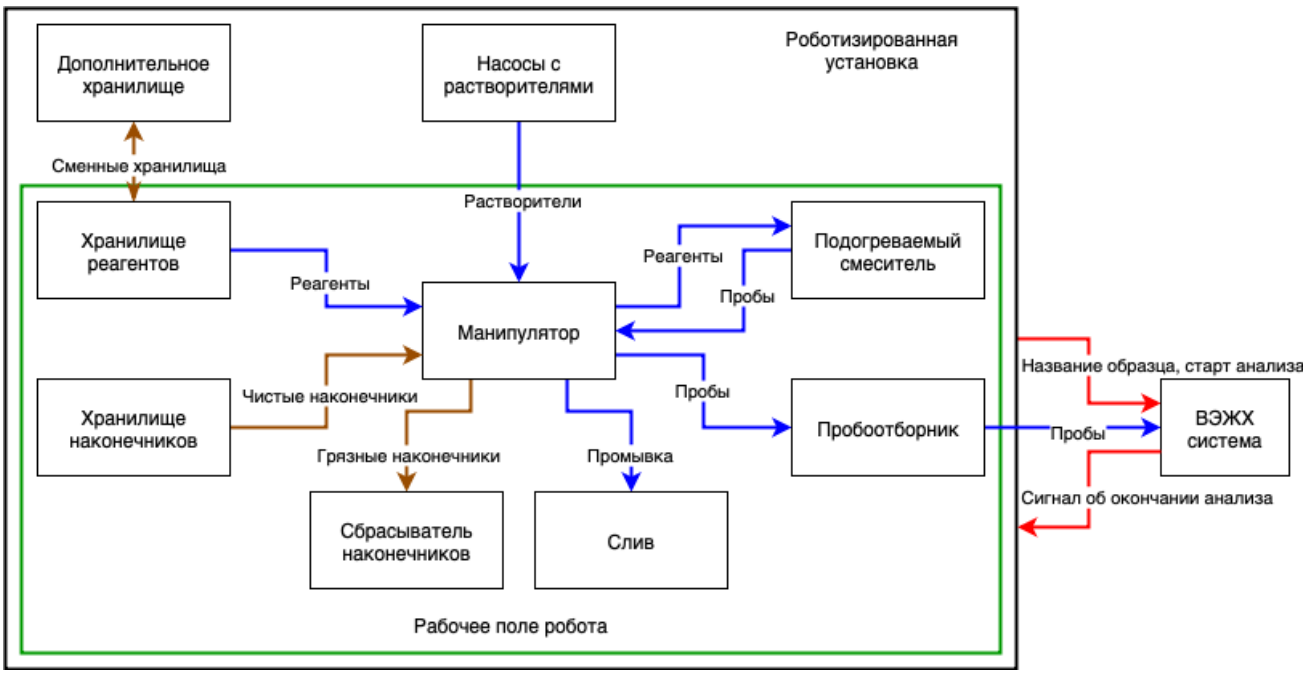


Рис. 2. Блок-схема установки.

Зеленым цветом выделено рабочее поле робота, синим – потоки жидкостей (растворителей, растворов реагентов и проб), коричневым – физически сменяемые части (наконечники, хранилища с реагентами), красным – обмен сигналов с внешним оборудованием (ВЭЖХ-система).

Реагенты хранятся в сменных хранилищах, вместимостью по 96 позиций, замена которых между рабочим полем и стойкой хранилища осуществляется с помощью самодельной роборуки. Для осуществления важного этапа любого синтеза – перемешивания с поддержанием определенной температуры – в рабочем поле находится подогреваемый смеситель. Для выполнения анализа имеется пробоотборник, в который переносится образец после синтеза, разбавляется растворителем и переносится в ВЭЖХ-систему. Кроме переноса самих образцов взаимодействие с хроматографической системой осуществляется и путем отправки названия образца и команды на старт анализа, а также получение обратных сигналов о готовности петли к внесению следующего образца и об окончании анализа.

Очень важной частью работы стало создание собственной библиотеки ChemBot для управления отдельными устройствами роботизированной установки на языке Python, а также клиентского веб-сервера для управления всей системой. Отметим, что контроль за роботизированной установкой и выполнение команд осуществляются с помощью одноплатного компьютера Raspberry Pi.

При «ручной» работе с установкой пользователь только вводит схемы реакций и конфигурацию хранилища робота, после чего робот автоматически пересчитывает объемы всех реагентов с учетом количества искомого вещества в хранилище, проводит приготовление реакционной смеси, проведение реакций и их анализ после завершения времени реакции.

Для просмотра хода выполнения нескольких синтезов и содержимого хранилищ и реакторных зон был создан веб-интерфейс, представленный на рис. 3. Интерфейс включает три колонки: слева представлены реакторные зоны (при нажатии на реакторы открывается карточка с синтезом, назначенным в этот реактор), в центре – информация по синтезам с графиком временной линии их выполнения, справа можно посмотреть расположение и количества веществ в хранилище. Разработанное интерактивное веб-приложение позволяет просматривать состояния смесителей нагревателей, ход выполнения всех синтезов (их запланированное время запуска, время, прошедшее с момента начала синтеза), содержимое хранилищ веществ, растворителей и держателя пипеток.

Программно-аппаратный комплекс также способен автоматически анализировать хроматограммы полученных реакционных смесей и передавать эти данные для обучения ИИ и/или активного обучения по ходу выполнения экспериментов. Однако в настоящей работе использовалось встроенное ПО хроматографа ClarityChrom [14] для обработки результатов хроматограмм выполненных синтезов, т. к. собственный подход автоматического анализа хроматограмм находится на этапе разработки. Преимуществами использования собственного подхода должны стать такие важные аспекты, как обнаружение зашкаливающих пиков и недостаточной интенсивности всех пиков, что позволит дать обратную связь на выполнение повторного анализа с меньшим или бóльшим разбавлением образца для получения корректных результатов.

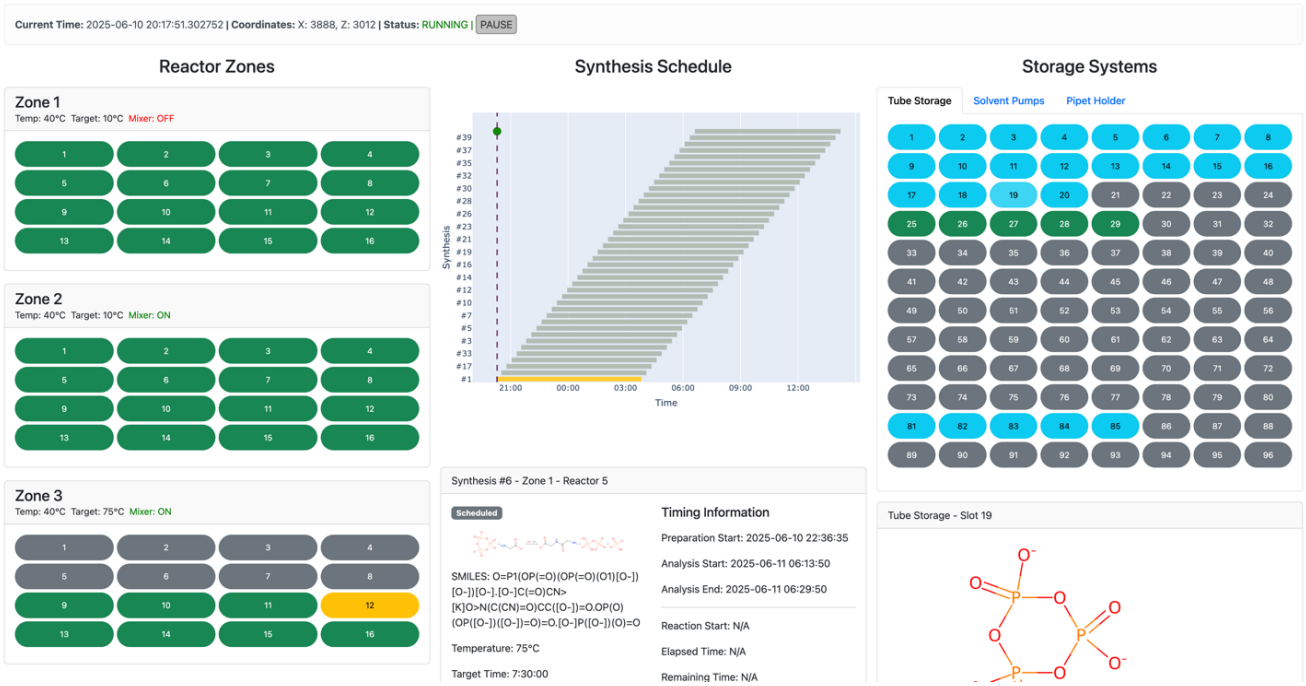


Рис. 3. Веб-приложение для отслеживания хода выполнения синтезов (по центру), просмотра конфигурации хранилищ (справа) и состояний реакторных зон (слева).

В качестве алгоритмов ИИ для предсказания путей синтеза могут быть использованы модели оптимизации отдельных типов реакций [10, 11], модели

предсказания оптимального пути синтеза целевых молекул [12] или модели поиска новых путей синтеза [13]. Кроме того, при наличии подходящих баз данных с помощью машинного обучения могут быть разработаны и новые модели.

ЭКСПЕРИМЕНТАЛЬНЫЕ ИССЛЕДОВАНИЯ

После настройки и контроля правильности функционирования отдельных узлов системы (такие, как работа с крышками и пипетками, правильность дозирования, перемешивание и поддержание температуры и т. д.) точность и воспроизводимость пробоподготовки были проверены путем автоматизированного приготовления и анализа образцов для построения калибровочной зависимости из одного концентрированного исходного раствора. Было приготовлено пять различных концентраций, каждая из которых воспроизводилась трижды. В результате установлено, что созданная установка обеспечивает достаточное для синтезов качество приготовления образцов – погрешность в приготовлении не превосходит 2%, что сопоставимо с точностью использованной для анализа обращенной-фазовой хроматографии с градиентным режимом (буферный раствор в воде – ацетонитрил).

Для проверки полной работоспособности робота-химика были проведены серии пробных синтезов альдольной конденсации (схема реакции представлена на рис. 4) с автоматизированным проведением анализа и ручной обработкой хроматограмм.

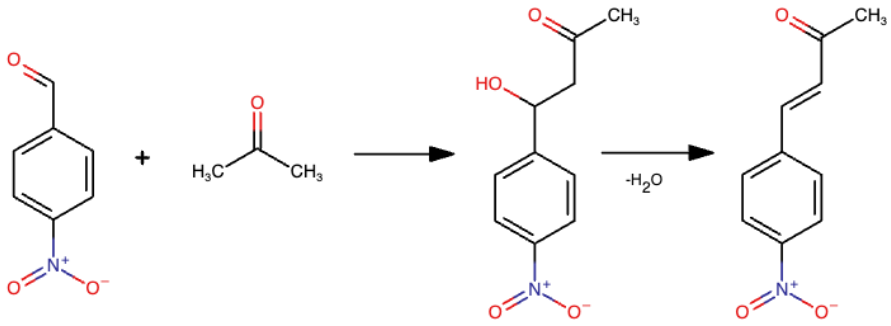


Рис. 4. Реакция альдольной конденсации нитробензальдегида и ацетона с участием аминокислотных комплексов цинка в качестве катализаторов.

В экспериментах варьировались количества катализатора и щелочи, а также температура. В качестве аминокислот в *бис*-комплексе цинка выступали: глицин,

пролин, глутаминовая кислота, аспарагин, глутамин, изолейцин, аргинин и гистидин. Примеры результатов осуществленных синтезов при одном из наборов экспериментальных условий представлены на рис. 5, из которого видно, что наиболее эффективным катализатором является *бис*-пролинат цинка.

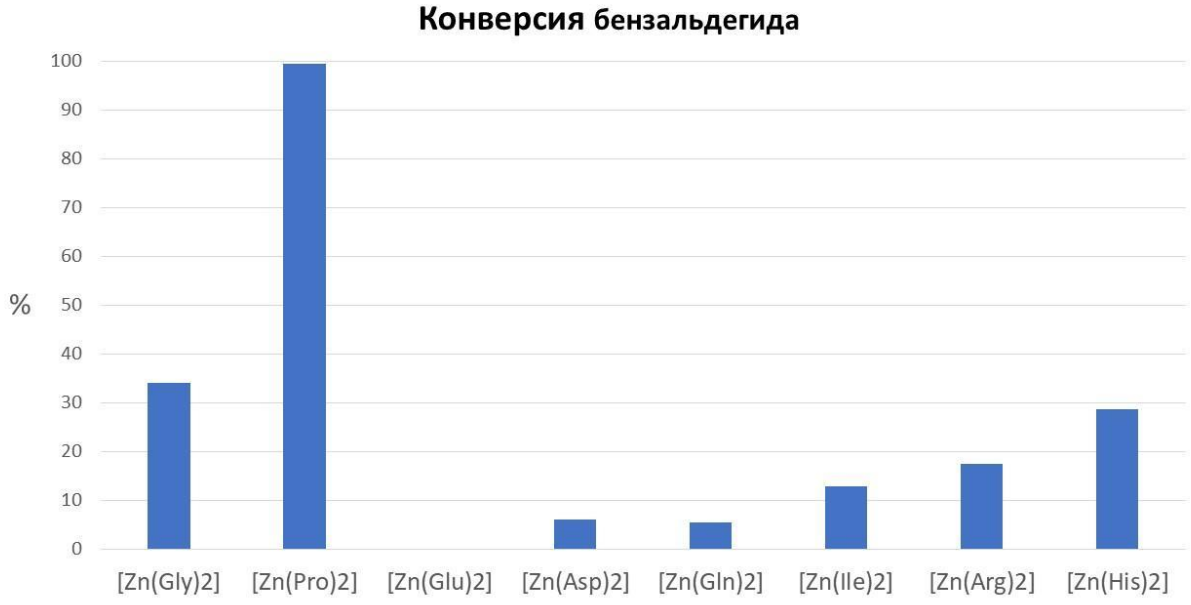


Рис. 5. Влияние *бис*-аминокислотных комплексов цинка на конверсию. Условия: комнатная температура, 0.1 м. % катализатора, 0.05 м. % щелочи.

ЗАКЛЮЧЕНИЕ

Представленная роботизированная установка позволяет осуществлять химические синтезы с последующим отбором проб и их анализом на хроматографе в автоматизированном режиме. Одним из важных ее достоинств является параллельное проведение синтезов, что с учетом времени ВЭЖХ-анализа (около 2–3 образцов в час в зависимости от методики разделения) позволяет осуществлять до 48 синтезов в сутки, что делает ее удобной платформой для проверки путей синтеза, а также для оптимизации условий реакции. В нашей научной группе в настоящее время активно разрабатываются методы предсказания путей синтеза целевых соединений и оптимизации реакционных условий с использованием машинного обучения и анализа больших данных. Следующим этапом исследований станет объединение всех разработок – роботизированной платформы, алгоритмов ИИ и систем адаптивной оптимизации – в единую интегрированную систему.

Такое сочетание роботизации и ИИ в перспективе позволит замкнуть цикл «предсказание – синтез и анализ – подтверждение или изменение модели» и продвинуться в сторону автоматизированных лабораторий.

Благодарности

Работа выполнена в рамках государственного задания ФИЦ КазНЦ РАН.

СПИСОК ЛИТЕРАТУРЫ

1. *Tom G., Schmid S.P., Baird S.G., Cao Y., Darvish K., Hao H., Lo S., Pablo-García S., Rajaonson E.M., Skreta M., Yoshikawa N., Corapi S., Akkoc G.D., Strieth-Kalthoff F., Seifrid M., and Aspuru-Guzik A.* Self-Driving Laboratories for Chemistry and Materials Science // Chemical Reviews. 2024. Vol. 16, No. 124, P. 9633–9732. https://doi.org/10.1021/acs.chemrev.4c00055

2. *Seifrid M., Pollice R., Aguilar-Granda A., Morgan Chan Z., Hotta K., Ser C.T., Vestfrid J., Wu T.C., Aspuru-Guzik A.* Autonomous Chemical Experiments: Challenges and Perspectives on Establishing a Self-Driving Lab // Accounts of Chemical Research. 2022. Vol. 55, I. 17. P. 2454–2466. https://doi.org/10.1021/acs.accounts.2c00220

3. *Burger B., Maffettone P.M., Gusev V.V., Aitchison C.M., Bai Y., Wang X., Li X., Alston B.M., Li B., Clowes R., Rankin N., Harris B., Spick R.S., Cooper A.I.* A mobile robot chemist // Nature. 2020. Vol. 583. P. 237–241. https://doi.org/10.1038/s41586-020-2442-2

4. *Martin K.N., Rubsamen M.S., Kaplan N.P., Hendrick M.P.* Method for Interfacing a Plate Reader Spectrometer Directly with an OT-2 Liquid Handling Robot // ChemRxiv. 2022. https://doi.org/10.26434/chemrxiv-2022-6z4q1

5. *Sanchez-Lengeling B., Aspuru-Guzik A.* Inverse molecular design using machine learning: Generative models for matter engineering // Science. 2018. Vol. 361, I. 6400. P. 360–365. https://doi.org/10.1126/science.aat2663

6. *Manzano J.S., Hou W., Zalesskiy S.S., Frei P., Wang H., Kitson P.J., Cronin L.* An autonomous portable platform for universal chemical synthesis // Nature Chemistry. 2022. Vol. 14. P. 1311–1318. https://doi.org/10.1038/s41557-022-01016-w

7. *Lee E.C., Salley D., Sharma A., Cronin L.* AI-Driven Robotic Crystal Explorer for Rapid Polymorph Identification // arXiv. 2024. https://doi.org/10.48550/arXiv.2409.05196

8. Автоматизированная станция пробоподготовки LifeBot.
URL: https://evotech-mg.com/products/avtomatizirovannaya-stanciya-probopodgotovki

9. Separations. Analytical Instruments. Smartline HPLC Series.
URL: https://separations.nl/en/products/detail/smartline-hplc-series-knauer

10. *Afonina V.A., Mazitov D.A., Nurmukhametova A., Shevelev M.D., Khasanova D.A., Nugmanov R.I., Burilov V.A., Madzhidov T.I., Varnek A.* Prediction of Optimal Conditions of Hydrogenation Reaction Using the Likelihood Ranking Approach // International Journal of Molecular Sciences. 2022. Vol. 23, I. 1. P. 248. https://doi.org/10.3390/ijms23010248

11. *Ahneman D.T., Estrada J.G., Dreher S.D., Doyle A.G.* Predicting reaction performance in C–N cross-coupling using machine learning // Science. 2018. Vol. 360, I. 6385. P. 186–190. https://doi.org/10.1126/science.aar5169

12. K*ashafutdinova I.M., Poyezzhaeva A., Gimadiev T., Matzhidov T.* Active learning approaches in molecule pKi prediction // Molecular Informatics. 2024. Vol. 44, I. 1. Art. e202400154. https://doi.org/10.1002/minf.202400154

13. *Bort W., Baskin I.I., Gimadiev T., Mukanov A., Nugmanov R., Sidorov P., Marcou G., Horvath D., Klimchuk O., Madzhidov T., Varnek A.* Discovery of novel chemical reactions by deep generative recurrent neural network // Scientific Reports. 2021. Vol. 11. P. 3178. https://doi.org/10.1038/s41598-021-81889-y

14. Knauer. ClarityChrom CDS.
URL: https://www.knauer.net/software-claritychrom-cds

INTELLIGENT CHEMIST ROBOT: TOWARDS AN AUTONOMOUS LABORATORY

M. S. Adygamov[1] [0009-0006-2364-9867], A. O. Golub[2] [0009-0004-0090-0292], E. R. Saifullin[3] [0000-0003-0823-9051], T. R. Gimadiev[4] [0000-0001-5012-0308], N. Yu. Serov[5] [0000-0002-5772-8399]

[1,3-5]Federal Research Center "Kazan Scientific Center of Russian Academy of Science", Kazan, Russia

[1-5]Kazan Federal University, A.M. Butlerov Chemistry Institute, Kazan, Russia

[1]musa20930@gmail.com, [2]toxa.mix7@gmail.com, [3]mr.emilsr@gmail.com, [4]Timur.Gimadiev@gmail.com, [5]Serov.Nikita@gmail.com

Abstract

This paper describes a hardware and software platform that enables automated chemical syntheses, including the preparation, heating, and mixing of reaction mixtures, as well as post-synthesis dilution sampling and sending for high-performance liquid chromatography (HPLC) analysis, followed by automated processing of the results. A custom Python library, ChemBot, was developed to control individual robotic devices, and a client web server was created to manage the entire system. A web interface was created to view the system status and the progress of syntheses. The performance of the entire platform for performing experiments was tested by performing aldol condensation syntheses, where the ratio of reagents, the catalyst and its amount, the temperature and time of synthesis were varied. Writing custom code to monitor and control the entire system is an important step toward integrating the robotic system with artificial intelligence (AI), which will ultimately enable the transition to an autonomous laboratory, where target molecule prediction and synthesis, experimental execution and analysis, and, if necessary, refinement or modification of the model will be performed automatically, without the need for human intervention.

Keywords*: artificial intelligence, robotics, chemical synthesis, self-driving lab, chemoinformatics.*

REFERENCES

1. *Tom G., Schmid S.P., Baird S.G., Cao Y., Darvish K., Hao H., Lo S., Pablo-García S., Rajaonson E.M., Skreta M., Yoshikawa N., Corapi S., Akkoc G.D., Strieth-Kalthoff F., Seifrid M., and Aspuru-Guzik A.* Self-Driving Laboratories for Chemistry and Materials Science // Chemical Reviews. 2024. Vol. 16, No. 124, P. 9633–9732. https://doi.org/10.1021/acs.chemrev.4c00055

2. *Seifrid M., Pollice R., Aguilar-Granda A., Morgan Chan Z., Hotta K., Ser C.T., Vestfrid J., Wu T.C., Aspuru-Guzik A.* Autonomous Chemical Experiments: Challenges and Perspectives on Establishing a Self-Driving Lab // Accounts of Chemical Research. 2022. Vol. 55, I. 17. P. 2454–2466. https://doi.org/10.1021/acs.accounts.2c00220

3. *Burger B., Maffettone P.M., Gusev V.V., Aitchison C.M., Bai Y., Wang X., Li X., Alston B.M., Li B., Clowes R., Rankin N., Harris B., Spick R.S., Cooper A.I.* A mobile robot chemist // Nature. 2020. Vol. 583. P. 237–241. https://doi.org/10.1038/s41586-020-2442-2

4. *Martin K.N., Rubsamen M.S., Kaplan N.P., Hendrick M.P.* Method for Interfacing a Plate Reader Spectrometer Directly with an OT-2 Liquid Handling Robot // ChemRxiv. 2022. https://doi.org/10.26434/chemrxiv-2022-6z4q1

5. *Sanchez-Lengeling B., Aspuru-Guzik A.* Inverse molecular design using machine learning: Generative models for matter engineering // Science. 2018. Vol. 361, I. 6400. P. 360–365. https://doi.org/10.1126/science.aat2663

6. *Manzano J.S., Hou W., Zalesskiy S.S., Frei P., Wang H., Kitson P.J., Cronin L.* An autonomous portable platform for universal chemical synthesis // Nature Chemistry. 2022. Vol. 14. P. 1311–1318. https://doi.org/10.1038/s41557-022-01016-w

7. *Lee E.C., Salley D., Sharma A., Cronin L.* AI-Driven Robotic Crystal Explorer for Rapid Polymorph Identification // arXiv. 2024. https://doi.org/10.48550/arXiv.2409.05196

8. Автоматизированная станция пробоподготовки LifeBot. URL: https://evotech-mg.com/products/avtomatizirovannaya-stanciya-probopodgotovki

9. Separations. Analytical Instruments. Smartline HPLC Series. URL: https://separations.nl/en/products/detail/smartline-hplc-series-knauer

10. *Afonina V.A., Mazitov D.A., Nurmukhametova A., Shevelev M.D., Khasanova D.A., Nugmanov R.I., Burilov V.A., Madzhidov T.I., Varnek A.* Prediction of Optimal Conditions of Hydrogenation Reaction Using the Likelihood Ranking Approach // International Journal of Molecular Sciences. 2022. Vol. 23, I. 1. P. 248. https://doi.org/10.3390/ijms23010248

11. *Ahneman D.T., Estrada J.G., Dreher S.D., Doyle A.G.* Predicting reaction performance in C–N cross-coupling using machine learning // Science. 2018. Vol. 360, I. 6385. P. 186–190. https://doi.org/10.1126/science.aar5169

12. K*ashafutdinova I.M., Poyezzhaeva A., Gimadiev T., Matzhidov T.* Active learning approaches in molecule pKi prediction // Molecular Informatics. 2024. Vol. 44, I. 1. Art. e202400154. https://doi.org/10.1002/minf.202400154

13. *Bort W., Baskin I.I., Gimadiev T., Mukanov A., Nugmanov R., Sidorov P., Marcou G., Horvath D., Klimchuk O., Madzhidov T., Varnek A.* Discovery of novel chemical reactions by deep generative recurrent neural network // Scientific Reports. 2021. Vol. 11. P. 3178. https://doi.org/10.1038/s41598-021-81889-y

14. Knauer. ClarityChrom CDS. URL: https://www.knauer.net/software-claritychrom-cds

СВЕДЕНИЯ ОБ АВТОРАХ

АДЫГАМОВ Муса Шамильевич – учится на 1 курсе аспирантуры Казанского федерального университета в Институте геологии и нефтегазовых технологий по специальности 2.8.4 Разработка и эксплуатация нефтяных и газовых месторождений. В настоящее время работает младшим научным сотрудником в Федеральном исследовательском центре «Казанский научный центр Российской академии наук». Область научных интересов: машинное обучение, QSPR моделирование, автоматизация химических/биологических экспериментов.

Musa Shamilevich ADYGAMOV – a 1st-year postgraduate student at the Kazan Federal University, Institute of Geology and Petroleum Technologies, specialty 2.8.4 Development of oil and gas deposits. He currently works as a junior researcher at the Federal Research Center "Kazan Scientific Center of the Russian Academy of Sciences". Research interests: machine learning, QSPR modelling, chemical/biological experiments automation.

email: musa20930@gmail.com

ORCID: 0009-0006-2364-9867

ГОЛУБЬ Антон Олегович – учится на 2 курсе магистратуры Казанского (Приволжского) федерального университета на кафедре органической и медицинской химии по специальности 04.04.01 «Химия» и профилю «Хемоинформатика и молекулярное моделирование». Область научных интересов: машинное обучение, программирование.

Anton Olegovich GOLUB – Second-year Master's student at Kazan Federal University. Specializing in Chemoinformatics and Molecular Modeling (Chemistry degree 04.04.01) at the Department of Organic and Medicinal Chemistry. Research interests: machine learning and programming.

email: toxa.mix7@gmail.com

ORCID: 0009-0004-0090-0292

САЙФУЛЛИН Эмиль Ринатович – закончил Казанский федеральный университет. В 2019 году защитил диссертацию на соискание степени кандидата технических наук по специальности 01.04.14 «Теплофизика и теоретическая теплотехника». В настоящее время работает старшим научным сотрудником в Федеральном исследовательском центре «Казанский научный центр Российской академии наук», а также старшим научным сотрудником Казанского федерального университета. Область научных интересов: теплофизика, физхимия, нефтедобыча, хемоинформатика.

Emil Rinatovich SAIFULLIN – graduated from Kazan Federal University. In 2019, he defended his dissertation for the degree of Candidate of Technical Sciences in the specialty 01.04.14 "Thermal Physics and Theoretical Heat Engineering". He currently works as a senior researcher at the Federal Research Center "Kazan Scientific Center of the Russian Academy of Sciences" and as a senior researcher at Kazan Federal University. His research interests include thermal physics, physical chemistry, oil production, and chemoinformatics.

email: mr.emilsr@gmail.com

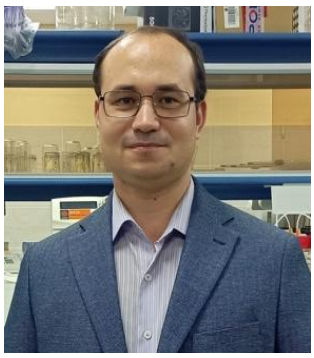
ORCID: 0000-0003-0823-9051

ГИМАДИЕВ Тимур Рустемович – закончил университет Страсбурга, Франция (UniStra, France). В 2018 получил степень доктора философии (PhD) в университете Страсбурга по направлению Хемоинформатика, химические науки. В настоящее время работает старшим научным сотрудником в Федеральном исследовательском центре «Казанский научный центр Российской академии наук», а также доцентом кафедры органической химии Химического института им. А.М. Бутлерова Казанского (Приволжского) федерального университета. Область научных интересов: роботизированные лаборатории, координационная химия, моделирование химических процессов, хемоинформатика, искусственный интеллект в химии и биохимии.

Timur Rustemovich GIMADIEV – graduated from the University of Strasbourg, France (UniStra, France). In 2018, he earned his PhD in Chemoinformatics and Chemical Sciences from the University of Strasbourg. Currently, he serves as a senior researcher at the Federal Research Centre “Kazan Scientific Center of the Russian Academy of Sciences” and as an Associate Professor at the Department of Organic Chemistry, A.M. Butlerov Institute of Chemistry, Kazan (Volga Region) Federal University. His research interests include robotic laboratories, coordination chemistry, chemical process modeling, chemoinformatics, and artificial intelligence in chemistry and biochemistry.

email: Timur.Gimadiev@gmail.com

ORCID: 0000-0001-5012-0308



СЕРОВ Никита Юрьевич – закончил Казанский (Приволжский) федеральный университет (КФУ). В 2021 году защитил диссертацию на соискание степени кандидата химических наук по специальности 02.00.01 «Неорганическая химия». В настоящее время работает ведущим научным сотрудником в Федеральном исследовательском центре «Казанский научный центр Российской академии наук», а также доцентом кафедры неорганической химии Химического института им. А.М. Бутлерова КФУ. Область научных интересов: роботизированные лаборатории, координационная химия, моделирование химических процессов, искусственный интеллект в химии и биохимии.

NIkita Yurievich SEROV – graduated from Kazan Federal University. In 2021, he defended his dissertation for the degree of Candidate of Chemical Sciences in the specialty 02.00.01 "Inorganic Chemistry". He currently works as a leading researcher at the Federal Research Center "Kazan Scientific Center of the Russian Academy of Sciences" and as an associate professor in the Department of Inorganic Chemistry at the A.M. Butlerov Chemical Institute of Kazan Federal University. Research interests: robotic laboratories, coordination chemistry, chemical process modelling, and artificial intelligence in chemistry and biochemistry.

email: Serov.Nikita@gmail.com

ORCID: 0000-0002-5772-8399

Материал поступил в редакцию 11 октября 2025 года

УДК 004.891

ПРОЕКТИРОВАНИЕ ДИНАМИЧЕСКОЙ ЭКСПЕРТНОЙ СИСТЕМЫ ПО АНАЛИЗУ ВЛИЯНИЯ КЛИМАТИЧЕСКИХ ВОЗДЕЙСТВИЙ НА МАЛЫЕ И СРЕДНИЕ ПРЕДПРИЯТИЯ

Р. А. Бурнашев[1] [0000-0002-1057-0328], Я. В. Сергеев[2] [0009-0009-3370-1464]

[1, 2]Казанский (Приволжский) федеральный университет, г. Казань, Россия

[1]r.burnashev@inbox.ru, [2]sergeevyarik7@yandex.ru

Аннотация

Растущая нестабильность климата создает новые вызовы и риски для устойчивости малых и средних предприятий. В работе предложена архитектура прототипа динамической экспертной системы, интегрирующей несколько ключевых модулей: пользовательский интерфейс, базу знаний, серверное приложение и модуль динамического обновления данных с API-интерфейсами реального времени. Особенностью системы является применение аппарата Z^+-чисел, реализованного на основе программной библиотеки scikit-fuzzy, что позволяет учитывать градуированную уверенность в оценках. Этот подход дает более обоснованные и адаптивные оценки рисков, чувствительные к изменению качества исходных данных. Интерактивная визуализация результатов реализована на основе картографической платформы OpenStreetMap. Приведены примеры агрегации экспертных оценок в формате *Z*-чисел, а также описана методика адаптации функций уверенности системы на основе исторических данных.

***Ключевые слова**: Z-числа, нечеткая логика, экспертная система, неопределенность, климатические риски, малые и средние предприятия, визуализация данных, принятие решений.*

ВВЕДЕНИЕ

Задачи экспертного оценивания и поддержки принятия решений в условиях неполной информации требуют адекватного представления неопределенности. Процессы коммуникации между экспертами, формулировка гипотез и оценок часто сводятся к использованию лингвистических выражений естественного языка, таких как «высокая температура», «низкий риск», «я совершенно уверен» и др.

© Р. А. Бурнашев, Я. В. Сергеев, 2025.
Данная статья распространяется на условиях международной лицензии Creative Commons License Attribution 4.0 International (CC BY 4.0).

Для обработки подобных высказываний требуется специальный математический аппарат. Для решения этой проблемы Л. Заде в 2011 г. предложил концепцию *Z*-чисел [1], описывающую неопределенную переменную как упорядоченную пару нечетких чисел (*A*, *B*). Компонента *A* представляет собой нечеткое ограничение на возможные значения переменной, а компонента *B* — нечеткую меру надежности (уверенности) этого ограничения. Данная концепция служит инструментом формализации экспертных знаний.

Однако в динамических системах, обрабатывающих потоки данных из разнородных источников с изменяющимся уровнем достоверности, общая мера уверенности *B* зачастую оказывается недостаточной. Например, уверенность в прогнозной температуре +35°C может быть существенно выше, чем в значении +42°C; оценка влияния засухи на сельхозпредприятие более надежна, чем на IT-компанию. Это обуславливает необходимость перехода к более сложной конструкции — Z^+-числу (*A*, *R*), где *R* является нечетким отношением, задающим уровень уверенности в зависимости от конкретного значения в пределах *A* [2, 3].

Целью настоящей работы является практическая реализация концепции Z^+-чисел в рамках прототипа динамической экспертной системы, предназначенной для оценки климатических рисков для малых и средних предприятий (МСП). В статье рассмотрена методология преобразования экспертных оценок в *Z*-числа, архитектура системы, механизм обработки и визуализации данных с учетом градуированной неопределенности.

Таким образом, изучаемая проблема заключается в неспособности классических *Z*-чисел адекватно моделировать градуированную уверенность в динамических системах с разнородными данными. Гипотеза состоит в том, что использование Z^+-чисел с адаптивным механизмом обучения функции уверенности $R(x)$ позволит повысить обоснованность и точность оценок климатических рисков. Основные результаты работы включают: 1) архитектуру динамической экспертной системы, реализующую Z^+-числа; 2) метод адаптации $R(x)$ на основе исторических данных; 3) экспериментальное подтверждение эффективности подхода.

ТЕОРЕТИЧЕСКИЕ ОСНОВЫ *Z*- И *Z*⁺-ЧИСЕЛ

Лингвистические переменные и нечеткие множества. Концепции лингвистической переменной и нечеткого множества составляют основу для понимания *Z*-чисел. Лингвистическая переменная — это переменная, значениями которой выступают слова или предложения естественного или искусственного языка [4—6]. Например, лингвистическая переменная «Скорость» может принимать значения «очень низкая», «низкая», «средняя», «высокая». Каждое такое значение формализуется с помощью теории нечетких множеств.

Нечеткое множество A в универсуме X характеризуется функцией принадлежности $\mu_A(x)$: $X \rightarrow [0,1]$, которая определяет степень принадлежности каждого элемента $x \in X$ множеству A. Типичным примером служит треугольное нечёткое число, задаваемое тройкой (a_1, a_2, a_3), где a_2— ядро (степень принадлежности равна 1), a_1 и a_3— левая и правая границы носителя.

***Z*-число** представляет собой упорядоченную пару $Z = (A, B)$, где A — нечеткое ограничение на значения некоторой переменной X, а B — нечеткая мера надежности этого ограничения [7]. Например, высказывание эксперта «Завершение проекта займет около 2 недель, я почти уверен» формализуется как $Z = (A, B)$. Здесь A — нечеткое число «около 2 недель» (например, треугольное число (1.5, 2, 2.5)), а B — нечеткое число «почти уверен» (например, (0.7, 0.8, 0.9)).

***Z*⁺-число** развивает концепцию *Z*-чисел, представляя собой пару *Z*⁺= (A,*R*) [2]. Как и прежде, A является нечетким ограничением, а *R* — мера надежности [8]— задается в виде нечеткого отношения или распределения. Ключевое отличие заключается в том, что *R* в *Z*⁺-числе определяет не общий уровень уверенности во всем множестве A, а распределение этой уверенности по его элементам. Это позволяет моделировать ситуации, когда уверенность в разных частях нечеткого интервала различна. Например, для нечеткого числа A = «Высокая температура» (30—45°C) отношение *R* может быть убывающей функцией: уверенность 1.0 при 35°C, 0.6 при 40°C и 0.2 при 45°C, что адекватно отражает меньшую надежность экстремальных прогнозных значений.

Для целей настоящей работы ключевым является свойство Z^+-чисел, позволяющее дифференцировать уровень уверенности внутри нечеткого интервала A.

Это свойство использовано для моделирования ситуаций, когда надежность прогноза климатического параметра (например, температуры) снижается по мере приближения к экстремальным значениям, что и отражается в виде функции $R(x)$.

ПРОЕКТИРОВАНИЕ ПРОТОТИПА ДИНАМИЧЕСКОЙ ЭКСПЕРТНОЙ СИСТЕМЫ

Динамическая экспертная система [9, 10] построена по модульному принципу и включает четыре основных компонента (рис. 1):

- пользовательский интерфейс;
- базу знаний (БЗ);
- серверное приложение;
- модуль динамического обновления данных.

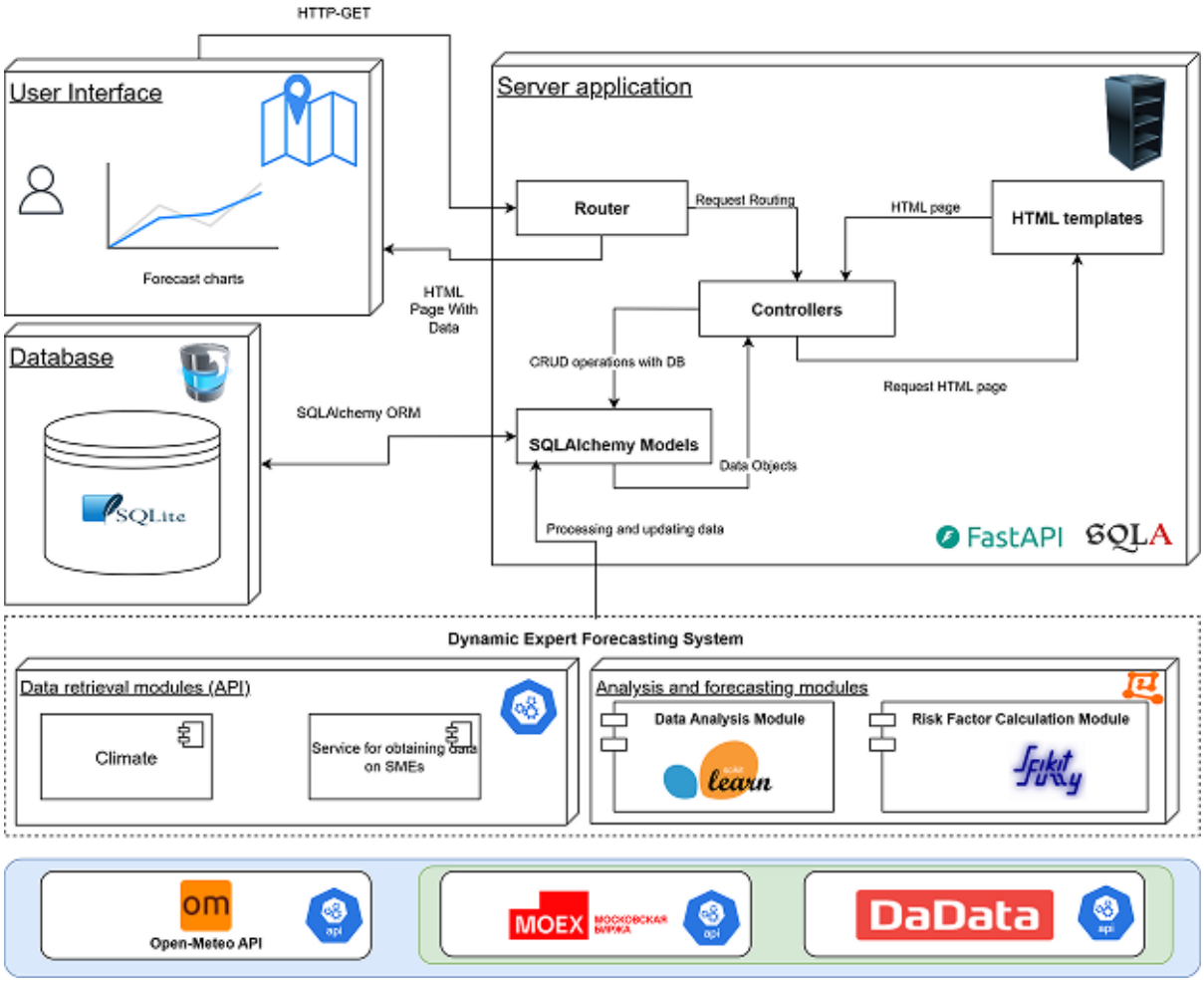


Рис. 1. Архитектура динамической экспертной системы

Пользовательский интерфейс обеспечивает взаимодействие пользователя с системой и реализует динамическую визуализацию пространственных данных. На основе библиотеки OpenStreetMap[1] создана интерактивная карта, отображающая населенные пункты по географическим координатам. На карте, с помощью интерактивных элементов и реестра предприятий, представлены актуальные результаты анализа климатического влияния: степени рисков и значения ключевых индикаторов.

База знаний [11, 12] содержит структурированные наборы данных: исторические климатические данные[2] и реестр МСП[3]. В ходе предварительной обработки исторические данные по климату и предприятиям были очищены от пропусков, приведены к сопоставимой временной шкале и единому формату представления (.csv).

Серверное приложение [13] является вычислительным ядром системы, скрытым от пользователя. Его ключевые функции включают:

- маршрутизацию HTTP-запросов от пользовательского интерфейса;
- обработку запросов с использованием контроллеров;
- управление данными в базе знаний (CRUD-операции);
- генерацию HTML-страниц для пользовательского интерфейса;
- выполнение аналитических моделей и нечеткого вывода на основе библиотеки scikit-fuzzy[4];
- автоматическую актуализацию результатов анализа в БЗ при поступлении новых данных.

[1] OpenStreetMap Foundation. OpenStreetMap. URL: https://www.openstreetmap.org (дата обращения: 10.07.2025).

[2] ВНИИГМИ-МЦД. Температура и осадки: набор данных. URL: http://meteo.ru/data/ (дата обращения: 10.07.2025).

[3] Реестр субъектов МСП за 2016–2024 гг. URL: https://data.rcsi.science/data-catalog/datasets/205/ (дата обращения: 10.07.2025).

[4] Fuzzy logic toolkit for SciPy. https://doi.org/10.5281/zenodo.802396. https://github.com/scikit-fuzzy/scikit-fuzzy?tab=readme-ov-file.

Модуль динамического обновления данных обеспечивает постоянную актуализацию информации в системе. Реализованный как набор специализированных программных компонентов, этот модуль выполняет следующие функции.

1. **Получение данных из внешних источников.** Модуль взаимодействует с открытыми API веб-сервисов в режиме, близком к реальному времени. Основными источниками климатических данных выступают:

- специализированные метеорологические API (Open-Meteo) [5], предоставляющие актуальные погодные условия, долгосрочные прогнозы и исторические данные. При API-запросе передаются параметры широты, долготы, временной зоны и количества дней;
- API центров мониторинга окружающей среды для получения специализированных индексов[6] (солнечная активность, качество воздуха, уровень осадков).

2. **Предварительная обработка и верификация.** Полученные сырые данные (в форматах JSON или XML) проходят этапы очистки от аномалий, преобразования к единому внутреннему формату системы (.csv) и проверки на целостность и непротиворечивость.

3. **Интеграция с базой знаний.** Подготовленные массивы данных передаются в базу знаний через слой объектно-реляционного преобразования (ORM), который абстрагирует работу с системой управления базами данных (СУБД) на уровне объектов предметной области, что обеспечивает структурированное хранение и эффективный доступ к данным для последующего анализа.

4. **Взаимодействие с серверным приложением.** Модуль не только обновляет сырые данные, но и предоставляет серверному приложению актуальную историческую статистику, необходимую для адаптации функций уверенности R в Z^+-числах, обеспечивая динамическую подстройку логики вывода системы под изменяющиеся условия и точность источников.

[5] Open-Meteo. Historical and Forecast Weather Data API. URL: https://open-meteo.com/ (дата обращения: 10.07.2025).

[6] NOAA Space Weather Prediction Center. Daily Solar Indices.
URL: https://services.swpc.noaa.gov/text/daily-solar-indices.txt (дата обращения: 10.07.2025).

Адаптивный механизм обучения уверенности. Ключевой задачей серверного приложения выступает снижение субъективности начального задания функций уверенности *R* в Z^+-числах. Для этого в механизм нечеткого вывода интегрирован модуль адаптивного обучения, который уточняет вид отношений *R* на основе ретроспективных данных.

Процесс адаптации реализуется по следующему алгоритму.

1. **Накопление эталонных данных.** Для каждого типа климатического параметра и класса предприятий в базе знаний сохраняются исторические пары «прогноз — фактическое значение» за продолжительный период.

2. **Сравнение и вычисление ошибки.** Для каждой такой пары вычисляется ошибка прогноза, например для температуры $\Delta = \left|T_{\text{прогноз}} - T_{\text{факт}}\right|$.

3. **Построение эмпирической функции надежности.** На основе накопленной статистики ошибок для каждого значения прогнозного параметра x (например, для каждой прогнозируемой температуры) строится эмпирическая мера уверенности — функция, обратно пропорциональная средней ошибке прогноза для данного x: $R(x) \approx 1/(1 + \alpha\ \mathrm{MSE}(x))$, где $\mathrm{MSE}(x)$ — средняя квадратичная ошибка прогнозов, давших значение x, а α — нормирующий коэффициент.

4. **Аппроксимация и сглаживание.** Полученное точечное отображение $x \rightarrow R(x)$ аппроксимируется [14] с помощью регрессионных моделей для получения гладкой нечеткой функции уверенности *R*, готовой к использованию в механизме нечеткого вывода.

Таким образом, функция *R*, первоначально заданная экспертом, со временем постоянно уточняется и адаптируется к реальной точности источников данных, что снижает субъективность и повышает обоснованность оценок системы.

Механизм нечеткого вывода. Ключевым этапом работы серверного приложения является преобразование Z^+-числа Z^+=(*A*, *R*) в обычное нечеткое число *A*', которое может быть использовано в классическом механизме нечеткого вывода для расчета итогового риска. Это преобразование осуществляется оператором вероятностного ограничения [2], реализованным на основе программной библиотеки scikit-fuzzy.

Пусть *A* задано на универсуме *X* функцией принадлежности $\mu_A(x)$, а *R* — функцией уверенности $R(x)$, отображающей каждое значение $x \in X$ в степень уверенности из интервала [0, 1]. Тогда результирующее нечеткое множество *A'* рассчитывается по формуле $\mu_{A'}(x) = R(x)\mu_A(x)$.

Данная операция реализована программно как поэлементное умножение массива значений функции принадлежности *A* на массив значений функции уверенности *R* для каждого дискретного значения универсума *X*. Таким образом, уверенность *R*(*x*) выступает в роли модулятора, напрямую снижая степень принадлежности тех значений *x*, в достоверности которых система менее уверена.

Пример. Пусть для температуры 45°C исходная степень принадлежности к понятию «Высокая температура» $\mu_A(45) = 0.8$. Если функция уверенности, обученная на исторических данных, дает для этого значения *R*(45) = 0.3, то итоговая степень принадлежности в расчетах будет снижена: $\mu'_A(45)$ = 0.3*0.8 = 0.24.

Итоговое нечеткое множество *A'* агрегируется с другими параметрами в системе нечеткого вывода, и для него вычисляется четкое значение риска (дефаззификация), которое отображается пользователю. Итоговая уверенность для визуализации (например, в виде насыщенности цвета) рассчитывается как среднее значение функции $R(x)$ по носителю нечеткого множества *A*.

Динамический характер экспертной системы [15—18] проявляется также в возможности адаптации нечетких отношений уверенности *R* в Z^+-числах при поступлении новых данных. Например, уверенность в прогнозных значениях температуры (R) автоматически снижается модулем обработки по мере увеличения горизонта прогноза или при обнаружении расхождений между моделями.

Автоматический перерасчет показателей обеспечивается тем, что поступление в базу знаний актуальных климатических данных, таких как текущая погода, прогнозы, индексы автоматически инициирует выполнение аналитических процедур серверным приложением.

Визуализация неопределенности. Реализована интерактивная картографическая система на основе OpenStreetMap, где каждый маркер предприятия визуализирует результаты расчета Z^+-чисел: цвет указывает на уровень риска, а насы-

щенность цвета отражает степень уверенности в оценке. При клике на маркер открывается всплывающее окно с детализацией расчета, включая исходную степень принадлежности, уровень уверенности и итоговый скорректированный риск.

Для эффективного донесения информации об уверенности до пользователя реализованы следующие методы визуализации.

- **Насыщенность цвета:** Цвет маркера МСП на карте (от зеленого к красному) показывает уровень риска (*A*), а насыщенность этого цвета соответствует степени уверенности в этой оценке (*R*). Низкая насыщенность сигнализирует о высокой неопределенности.
- **Всплывающие подсказки** (Tooltips)**:** При наведении курсора на элемент отображается текстовая информация вида: «Риск: Высокий (Уверенность: 75%)».
- **Графики функций:** В «карточке» предприятия может отображаться график функции $R(x)$ для ключевых параметров, позволяя анализировать распределение уверенности по диапазону возможных значений.

Программная реализация и визуализация системы. Для верификации предложенного подхода разработаны программные модули на языке Python, реализующие динамическую экспертную систему оценки климатических рисков на основе Z^+-чисел. Архитектура системы включает следующие ключевые компоненты:

- модуль обработки Z^+-чисел — реализует преобразование $(A, R) \rightarrow A'$;
- интерактивная визуализация МСП на OpenStreetMap (рис. 2);
- аналитический дашборд — сравнительный анализ эффективности методов;
- генератор тестовых данных — создание реалистичных сценариев.

На интерактивной карте (рис. 2) в качестве примера представлены результаты оценки климатических рисков для 10 тестовых предприятий различных отраслей. Маркеры визуализируют применение Z^+-подхода.

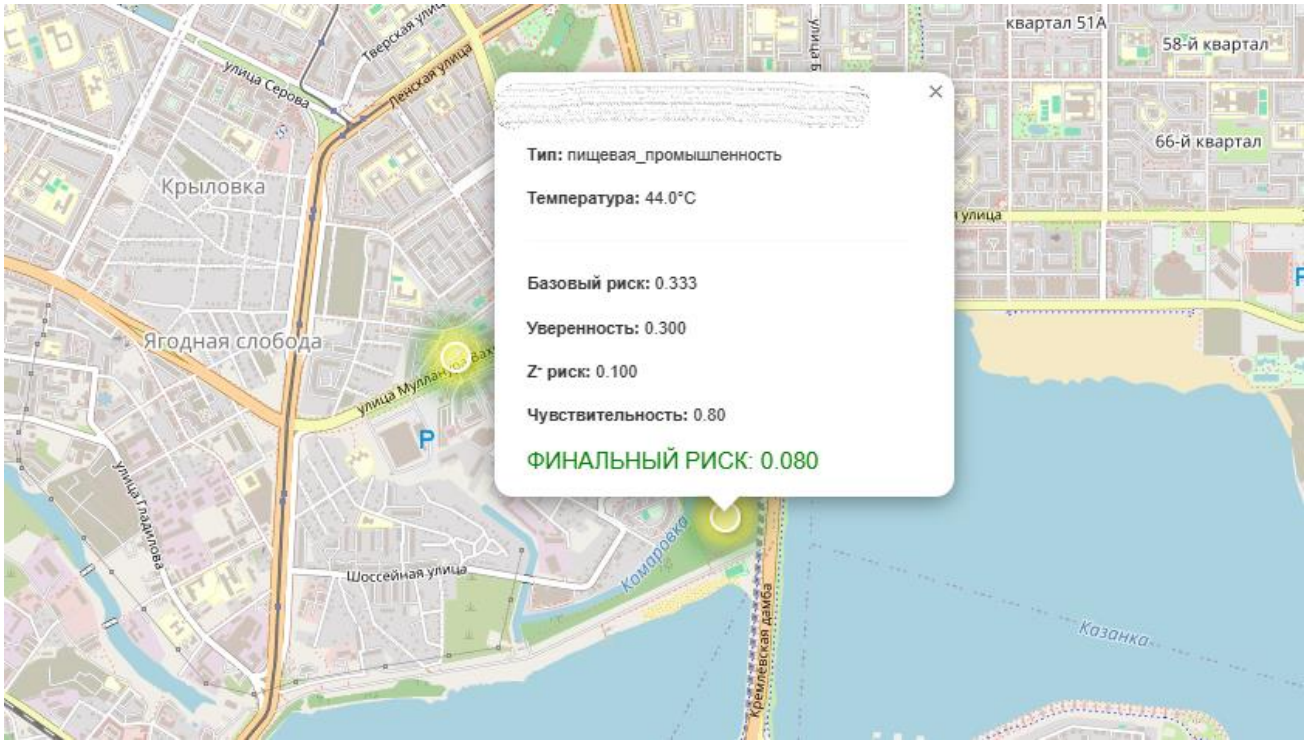


Рис. 2. Пример интерактивной визуализация МСП

ЭКСПЕРИМЕНТАЛЬНАЯ ОЦЕНКА ЭФФЕКТИВНОСТИ

В этом разделе проведено сравнение трех методов – двух базовых, и предложенного в настоящей работе. Это:

- (*Стандартная нечеткая логика*): нечеткая система, оперирующая только лингвистическими переменными (компонента *A*);
- (*Z-числа*): система на основе классических *Z*-чисел с постоянным уровнем уверенности *B* для всего интервала *A*;
- (*Z^+-числа*): описанная в работе система с адаптивными Z^+-числами.

В качестве метрики использовалась калибровка уверенности. Для валидации подхода был создан синтетический набор данных, имитирующий исторические данные за 2023—2024 гг. для 5000 предприятий различных отраслей экономики. Использование синтетического набора данных позволило сгенерировать эталонные метки «ущерб / нет ущерба» с детерминированной зависимостью от климатических событий, что необходимо для объективной валидации способности системы выявлять именно эту зависимость, минуя шумы и неопределенности реальных данных.

Оценка системы считалась ошибочной, если она присваивала высокий риск при эталонной метке «отсутствие ущерба», и наоборот. Результаты оценки точности прогнозов в зависимости от заявленной уверенности системы представлены в табл. 1.

Табл. 1. Зависимость точности прогнозов от заявленной уверенности системы

Уровень уверенности системы	**Стандартная нечеткая логика**	**Z-числа**	**Z^+-числа**
Низкая (0—30%)	—	55%	78%
Средняя (31—70%)	—	72%	85%
Высокая (71—100%)	82%	82%	95%

Для оценки калибровки уверенности мы группировали прогнозы системы по уровням заявленной уверенности (низкий, средний, высокий) и вычисляли точность прогнозирования (долю верных оценок) внутри каждой группы.

Анализ результатов эксперимента свидетельствует о лучшей калибровке предложенного метода на основе Z^+-чисел по сравнению с базовыми подходами. В условиях, когда система присваивает оценкам высокий уровень уверенности, достигается точность прогнозирования 95%, что существенно превышает показатели сравниваемых методов. Это позволяет утверждать, что высокая уверенность системы статистически обоснована и повышает надежность соответствующих предупреждений. Кроме того, система продемонстрировала способность к корректной идентификации ситуаций с низкой предсказуемостью, обеспечивая точность 78% в группе с низкой уверенностью, тогда как классический подход на основе *Z*-чисел показывает склонность к ошибочным оценкам в аналогичных условиях.

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

Внедрение концепции Z^+-чисел позволило преодолеть ключевое ограничение традиционных нечетких систем — предположение о равномерной уверенности в пределах всего нечеткого интервала. На тестовых примерах показано, что система выдает различные оценки для ситуаций, которые классический подход трактует одинаково. Например, для двух предприятий с расчетным риском «Высокий» (*A*) система, учитывая низкую уверенность (*R*) для IT-компании из-за косвенности связи, визуализировала его менее насыщенным цветом, что сигнализировало пользователю о необходимости более осторожной интерпретации результата.

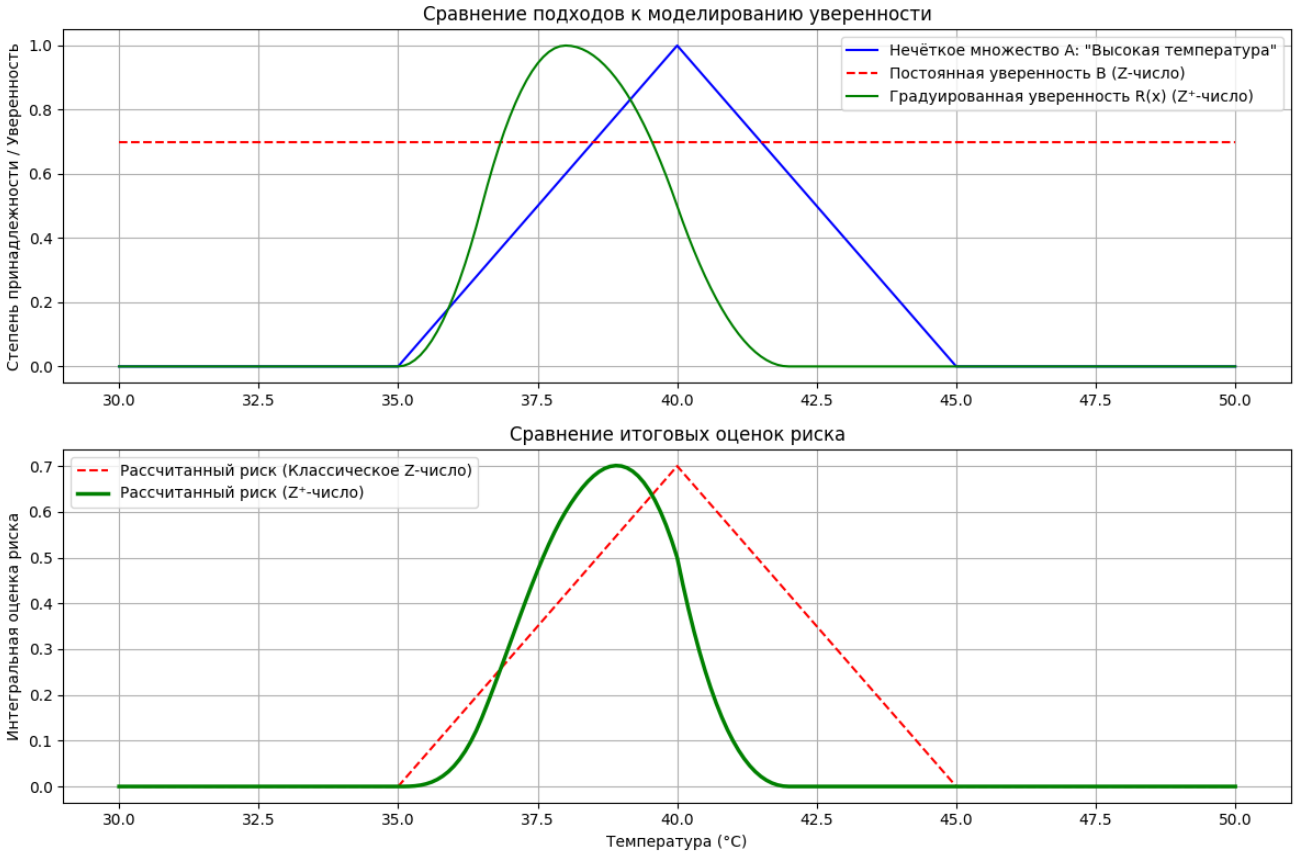


Рис. 3. Сравнение подходов *Z*-системы и Z^+-системы.

На рис. 3 представлены результаты моделирования для сценария оценки риска «Высокая температура», наглядно демонстрирующие различия между подходами. Классический *Z*-подход использует постоянный уровень уверенности (*B*)

на протяжении всего интервала *A*, в то время как Z^+-подход применяет градуированную уверенность $R(x)$, которая максимальна вблизи наиболее вероятного значения (38°C) и снижается к границам интервала. Это точнее отражает онтологию экспертных знаний: уверенность в прогнозе температуры 38°C обычно выше, чем в 45°C.

Полученные оценки риска (рис. 3, нижняя часть) показывают, что классическое *Z*-число дает завышенную и «плоскую» оценку для всех высоких температур. Напротив, Z^+-число обеспечивает более детализированную и контекстуально зависимую оценку: пик риска соответствует зоне наибольшей уверенности, а оценка автоматически снижается на краях интервала, указывая пользователю на высокую неопределенность. Это предотвращает принятие необоснованно радикальных решений на основе экстремальных, но ненадежных прогнозов.

Оценка производительности и масштабируемости системы. Практическое внедрение экспертной системы требует оценки ее производительности и способности обрабатывать растущие объемы данных. Для оценки этих характеристик проведен нагрузочный тест прототипа системы на серверной платформе с конфигурацией: CPU Intel Xeon E-2388G (8 ядер, 3.2 GHz), 32 GB RAM, SSD NVMe.

Методика тестирования включала последовательную загрузку в систему данных о МСП и измерение времени отклика при выполнении стандартного сценария: расчет рисков для всех предприятий при поступлении нового пакета климатических данных. Полученные результаты представлены в табл. 2.

Анализ результатов показал, что система демонстрирует приемлемую линейную сложность $O(n)$ на объемах данных, характерных для регионального сегмента МСП (до 5—7 тыс. предприятий). Основным узким местом, как и предполагалось, является этап агрегации Z^+-чисел, что проявляется в резком росте загрузки CPU при обработке более 5000 записей. Тем не менее, даже для 10000 предприятий время обработки (~2.2 мин.) остается адекватным для задач, не требующих обработки в реальном времени. Для дальнейшего повышения масштабируемости в будущих работах планируется исследование методов параллельных вычислений на GPU и предварительной агрегации данных.

Табл. 2. Результаты нагрузочного тестирования

Количество предприятий	Время обработки, с	Загрузка CPU, %	Использование RAM, MB
100	1.2	15	250
1000	8.5	42	580
5000	48.7	89	1450
10000	132.4	98	2850

ЗАКЛЮЧЕНИЕ

Показаны практическая значимость концепций *Z*- и Z^+-чисел для задач управления в условиях неопределенности и их реализуемость. Разработана архитектура прототипа динамической экспертной системы по анализу влияния климатических воздействий на малые и средние предприятия.

Использование Z^+-чисел и их динамическая адаптация к актуальности данных значительно повышают обоснованность выдаваемых системой рекомендаций для малых и средних предприятий, что способствует их устойчивости к климатическим изменениям.

Основными выявленными ограничениями системы являются: 1) высокая вычислительная сложность агрегации Z^+-чисел, ограничивающая масштабируемость, и 2) субъективность начального задания функций уверенности *R*.

Перспективы дальнейших исследований включают:

- оптимизацию алгоритмов агрегации за счет внедрения методов параллельных вычислений (GPU-ускорение) и создание редуцированных аппроксимирующих алгоритмов;
- разработку методов автоматического обучения функций $R(x)$ на основе анализа исторических данных и временных рядов с применением методов машинного обучения для минимизации субъективности.

СПИСОК ЛИТЕРАТУРЫ

1. *Zadeh L.A.* A Note on Z-numbers // Information Sciences. 2011. Vol. 181. P. 2923–2932.

2. *Aliev R.A., Alizadeh A.V., Huseynov O.H.* The arithmetic of discrete Z-numbers // Information Sciences. 2015. Vol. 290. P. 134–155.

3. *Lala M. Zeinalova.* Choquet aggregation based decision making under Z-information // ICTACT Journal on Soft Computing. 2014. Vol. 4, № 4. P. 819–824.

4. *Бурнашев Р.А., Сергеев Я.В., Назипова А.Ф.* Методы гранулирования нечётких временных рядов для анализа данных // Онтология проектирования. 2025. Т. 15, № 3(57). С. 404–417. https://doi.org/10.18287/2223-9537-2025-15-3-404-417.

5. *Enikeeva A. I., Burnashev R.A., Farahov R.R.* Development of an Expert System Based on Fuzzy Logic for Pneumonia Diagnostics // Automatic Documentation and Mathematical Linguistics. 2024. Vol. 58, No. S4. P. S202–S215. https://doi.org/10.3103/S000510552570027X.

6. *Enikeev A. I., Burnashev R.A., Vakhitov G.Z.* Software tools and techniques for the expert systems building // Advances in Intelligent Systems and Computing. 2020. Vol. 1041. P. 191–199. https://doi.org/10.1007/978-981-15-0637-6_16.

7. *Полещук О. М.,* Поярков Н.Г, Тумор С.В. Принятие решений на основе байесовского подхода и Z-чисел // Лесной вестник. 2019. Т. 23, № 4. С. 112–116. – https://doi.org/10.18698/2542-1468-2019-4-112-116.

8. *Полещук О. М., Чернова Т.В.* Z - числа и их новые возможности для моделирования реального мира // Современные проблемы физико-математического образования : сборник материалов VI Международной заочной научно-практической конференции, Орехово-Зуево, 12–13 декабря 2016 года / Государственный гуманитарно-технологический университет. Орехово-Зуево: Государственный гуманитарно-технологический университет, 2016. С. 33–35.

9. *Kostikova A.V.,* Tereliansky P.V., Shuvaev A.V., [et al.] Expert Fuzzy Modeling of Dynamic Properties of Complex Systems // ARPN Journal of Engineering and Applied Sciences. 2016. Vol. 11, No. 17. P. 10601–10608.

10. *Мозгачев А.В., Рыбина Г.В., Шанцер Д.И., Блохин Ю.М.* Динамические интеллектуальные системы на основе интегрированных экспертных систем// Приборы и системы. Управление, контроль, диагностика. 2012. № 5. С. 13–20.

11. *Титов Н. А., Макрушин С. В.* Технология создания доменной базы знаний вопрос-ответной системы на основе крупномасштабной универсальной базы знаний// Computational Nanotechnology. 2022. Т. 9, № 1. С. 115–124. https://doi.org/0.33693/2313-223X-2022-9-1-115-124.

12. *Davydenko I. T.* Semantic models, method and tools of knowledge bases coordinated development based on reusable components // Открытые семантические технологии проектирования интеллектуальных систем. 2018. No. 8. P. 99–119.

13. *Бочкарев А. М.* Эффективность использования информационных платформ разработки клиент-серверных приложений для информационных систем промышленных предприятий // Финансовый бизнес. 2021. № 4(214). С. 17–19.

14. *Гринюк Д. А., Сухорукова И. Г., Олиферович Н. М.* Использование алгоритмов аппроксимации для сглаживания трендов измерительных преобразователей// Труды БГТУ. Серия 3: Физико-математические науки и информатика. 2017. № 2(200). С. 82–87.

15. Bi L., Cao W., Hu W., Wu M. A Dynamic-Attention-Based Heuristic Fuzzy Expert System for the Tuning of Microwave Cavity Filters // IEEE Transactions on Fuzzy Systems. 2022. Vol. 30, No. 9. P. 3695–3707. https://doi.org/10.1109/tfuzz.2021.3124643.

16. *Livio J., Hodhod R.* AI Cupper: A Fuzzy Expert System for Sensorial Evaluation of Coffee Bean Attributes to Derive Quality Scoring. IEEE Transactions on Fuzzy Systems. 2018. Vol. 26 (6). P. 3418–3427. https://doi.org/10.1109/TFUZZ.2018.2832611.

17. *Samanta S., Pratama M., Sundaram S.* Bayesian Neuro-Fuzzy Inference System for Temporal Dependence Estimation. IEEE Transactions on Fuzzy Systems. 2021. Vol. 29 (9). P. 2479–2490. https://doi.org/10.1109/TFUZZ.2020.3001667.

18. *Giiven M.K., Passino K.M.* Avoiding exponential parameter growth in fuzzy systems. IEEE Transactions on Fuzzy Systems. 2001. Vol. 9 (1). P. 194–199. DOI: 10.1109/91.917125.

DESIGN OF A DYNAMIC EXPERT SYSTEM FOR ANALYZING THE IMPACT OF CLIMATE EFFECTS ON SMALL AND MEDIUM SIZED ENTERPRISES

R. A. Burnashev[1] [0000-0002-1057-0328], Y. V. Sergeev[2] [0009-0009-3370-1464]

[1, 2]*Kazan Federal University, Kazan, Russia*

[1]r.burnashev@inbox.ru, [2]sergeevyarik7@yandex.ru

Abstract

Growing climate instability is creating new challenges and risks for the resilience of small and medium-sized enterprises (SMEs). This article proposes a prototype architecture for a dynamic expert system comprising several key modules: a user interface, a knowledge base, a server application, and a dynamic data update module with real-time APIs. A distinctive feature of the system is the application of Z^+-number calculus, implemented using the scikit-fuzzy library, which allows for accounting of graded confidence in evaluations. This approach provides more robust and adaptive risk assessments that are sensitive to changes in the quality of input data. Interactive visualization of the results is built upon OpenStreetMap. The system's methodology for self-adaptation of confidence measures based on historical data is described.

Keywords*: Z-numbers, fuzzy logic, expert system, uncertainty, climate risks, small and medium enterprises, data visualization, decision making.*

REFERENCES

1. *Zadeh L.A.* A Note on Z-numbers // Information Sciences. 2011. Vol. 181. P. 2923–2932.

2. *Aliev R.A., Alizadeh A.V., Huseynov O.H.* The arithmetic of discrete Z-numbers // Information Sciences. 2015. Vol. 290. P. 134–155.

3. *Zeinalova L.M.* Choquet aggregation based decision making under Z-information // ICTACT Journal on Soft Computing. 2014. Vol. 4, № 4. P. 819–824.

4. Burnashev R.A., Sergeev Y.V., Nazipova A.F. Metody granulyatsii nechetkikh vremennykh ryadov dlya analiza dannykh // Ontologiya proektirovaniya. 2025. T. 15, No. 3(57). P. 404-417. https://doi.org/10.18287/2223-9537-2025-15-3-404-417.

5. Enikeeva A.I., Burnashev R.A., Farahov R.R. Development of an Expert System Based on Fuzzy Logic for Pneumonia Diagnostics // Automatic Documentation and

Mathematical Linguistics. 2024. Vol. 58, No. S4. P. S202–S215. https://doi.org/10.3103/S000510552470027X.

6. Enikeev A.I., Burnashev R.A., Vakhitov G.Z. Software tools and techniques for the expert systems building // Advances in Intelligent Systems and Computing. 2020. Vol. 1041. P. 191-199. https://doi.org/10.1007/978-981-15-0637-6_16

7. *Poleshchuk O.M., Poyarkov N.G., Tumor S.V. Prinyatiye resheniy na osnove bayesovskogo podkhoda i Z-chisel // Lesnoy vestnik. 2019. T. 23, No. 4. S. 112–116. https://doi.org/10.18698/2542-1468-2019-4-112-116.*

8. *Poleshchuk O.M., Chernova T.V.* Z - chisla i ikh novyye vozmozhnosti dlya modelirovaniya real'nogo mira // Sovremennyye problemy fiziko-matematicheskogo obrazovaniya: sbornik materialov VI Mezhdunarodnoy zaochnoy nauchno-prakticheskoy konferentsii, Orekhovo-Zuyevo, 12–13 dekabrya 2016 goda / Gosudarstvennyy gumanitarno-tekhnologicheskiy universitet. Orekhovo-Zuyevo: Gosudarstvennyy gumanitarno-tekhnologicheskiy universitet, 2016. P. 33–35.

9. *Kostikova A.V., Tereliansky P.V., Shuvaev A.V. [i dr.] Expert Fuzzy Modeling of Dynamic Properties of Complex Systems // ARPN Journal of Engineering and Applied Sciences. 2016. Vol. 11, No. 17. P. 10601–10608.*

10. *Mozgachev A.V., Rybina G.V., Shantser D.I., Blokhin Y.M. Dinamicheskiye intellektual'nyye sistemy na osnove integrirovannykh ekspertnykh sistem // Pribory i sistemy. Upravleniye, kontrol', diagnostika. 2012. No. 5. S. 13–20.*

11. *Titov N.A., Makrushin S.V.* Tekhnologiya sozdaniya domennoy bazy znaniy vopros-otvetnoy sistemy na osnove krupnomasshtabnoy universal'noy bazy znaniy // Computational Nanotechnology. 2022. V. 9, № 1. P. 115–124. https://doi.org/10.33693/2313-223X-2022-9-1-115-124.

12. *Davydenko I.T.* Semantic models, method and tools of knowledge bases coordinated development based on reusable components // Otkrytyye semanticheskiye tekhnologii proektirovaniya intellektual'nykh sistem. 2018. No. 8. P. 99–119.

13. *Bochkarev A.M.* Effektivnost' ispol'zovaniya informatsionnykh platform razrabotki klient-servernykh prilozheniy dlya informatsionnykh sistem promyshlennykh predpriyatiy // Finansovyy biznes. 2021. № 4(214). P. 17–19.

14. *Grinyuk D.A., Sukhorukova I.G., Oliferovich N.M.* Ispol'zovaniye algoritmov approksimatsii dlya sglazhivaniya trendov izmeritel'nykh preobrazovateley // Trudy BGTU. Seriya 3: Fiziko-matematicheskiye nauki i informatika. 2017. № 2(200). P. 82–87.

15. Bi L., Cao W., Hu W., Wu M. A Dynamic-Attention-Based Heuristic Fuzzy Expert System for the Tuning of Microwave Cavity Filters // IEEE Transactions on Fuzzy Systems. 2022. Vol. 30, No. 9. P. 3695–3707. https://doi.org/10.1109/tfuzz.2021.3124643.

16. *Livio J., Hodhod R.* AI Cupper: A Fuzzy Expert System for Sensorial Evaluation of Coffee Bean Attributes to Derive Quality Scoring. IEEE Transactions on Fuzzy Systems.2018. Vol. 26 (6). P. 3418–3427. https://doi.org/10.1109/TFUZZ.2018.2832611.

17. *Samanta S., Pratama M., Sundaram S.* Bayesian Neuro-Fuzzy Inference System for Temporal Dependence Estimation. IEEE Transactions on Fuzzy Systems. 2021. Vol. 29 (9). P. 2479–2490. https://doi.org/10.1109/TFUZZ.2020.3001667.

18. *Giiven M.K., Passino K.M.* Avoiding exponential parameter growth in fuzzy systems. IEEE Transactions on Fuzzy Systems. 2001. Vol. 9 (1). P. 194–199 https://doi.org/10.1109/91.917125.

СВЕДЕНИЯ ОБ АВТОРАХ

БУРНАШЕВ Рустам Арифович – окончил Казанский (Приволжский) федеральный университет. В настоящее время работает заведующим кафедрой анализа данных и технологий программирования Института вычислительной математики и информационных технологий Казанского (Приволжского) федерального университета.

Область научных интересов: системы поддержки принятия решений, экспертные системы, машинное обучение, анализ данных, нечёткая логика. Автор более 50 научных работ.

Rustam Arifovich BURNASHEV – graduated from Kazan (Volga Region) Federal University. Since 2025, he has been the Head of the Department of Data Analysis and Programming Technologies at the Institute of Computational Mathematics and Information Technologies of Kazan (Volga Region) Federal University.

Research interests: decision support systems, expert systems, machine learning, data analysis, fuzzy logic. Author of more than 50 scientific papers.

email: r.burnashev@inbox.ru

ORCID: 0000-0002-1057-0328

СЕРГЕЕВ Ярослав Владиславович **–** обучается на 2 курсе магистратуры Казанского (Приволжского) федерального университета на кафедре анализа данных и технологий программирования по специальности 01.04.02 «Прикладная математика и информатика». В настоящее время работает преподавателем на кафедре анализа данных и технологий программирования Института вычислительной математики и информационных технологий Казанского (Приволжского) федерального университета.

Область научных интересов: Системы поддержки принятия решений.

Yaroslav Vladislavovich SERGEEV – is a 2nd-year master's student at the Department of Data Analysis and Programming Technologies of Kazan (Volga Region) Federal University, majoring in 01.04.02 Applied Mathematics and Informatics. He currently works as a teacher at the Department of Data Analysis and Programming Technologies, Institute of Computational Mathematics and Information Technologies at Kazan (Volga Region) Federal University.

Research interests: decision support systems.

email: sergeevyarik7@yandex.ru

ORCID: 0009-0009-3370-1464

Материал поступил в редакцию 11 октября 2025 года

УДК 004.82

НОРМАЛИЗАЦИЯ ТЕКСТА, РАСПОЗНАННОГО ПРИ ПОМОЩИ ТЕХНОЛОГИИ ОПТИЧЕСКОГО РАСПОЗНАВАНИЯ СИМВОЛОВ, С ИСПОЛЬЗОВАНИЕМ ЛЕГКОВЕСНЫХ LLM

В. К. Вершинин[1] [0009-0001-9425-0881], И. В. Ходненко[2] [0009-0003-7787-7126], С. В. Иванов[3] [0000-0002-1128-2942]

[1-3]*Университет ИТМО, г. Санкт-Петербург, Россия*

[1]vershinin@itmo.ru, [2]svivanov@itmo.ru,[3]Ivan.Khodnenko@itmo.ru

Аннотация

Несмотря на значительный прогресс, технологии оптического распознавания символов (OCR) для исторических газет по-прежнему допускают 5–10% ошибок на уровне символов. В работе представлена полностью автоматизированная система нормализации пост-OCR, объединяющая легкие языковые модели (LLM) объемом 7–8 млрд параметров, обученные по инструкциям и квантизованные до 4 бит (INT4), с небольшим набором регулярных выражений. На наборе данных BLN600 (600 страниц британских газет XIX в.) лучшая модель YandexGPT-5-Instruct Q4 снижает Character Error Rate (CER) с 8.4% до 4.0% (–52.5%) и Word Error Rate (WER) с 20.2% до 6.5% (–67.8%), повышая при этом семантическое сходство до 0.962. Система работает на потребительском оборудовании (RTX-4060 Ti, 8 ГБ VRAM) со скоростью около 35 секунд на страницу и не требует дополнительного обучения или параллельных данных. Полученные результаты показывают, что компактные INT4-LLM являются практичной альтернативой крупным моделям для постобработки OCR исторических документов.

Ключевые слова*: оптическое распознавание символов, пост-OCR-коррекция, исторические газеты, большие языковые модели, квантизация, INT4, конвейер нормализации, ошибка на уровне символов, семантическое сходство, регулярные выражения, YandexGPT-5, легкие модели, обработка естественного языка, цифровые гуманитарные науки, оцифровка документов.*

© В. К. Вершинин, И. В. Ходненко, С. В. Иванов, 2025.
Данная статья распространяется на условиях международной лицензии Creative Commons License Attribution 4.0 International (CC BY 4.0).

ВВЕДЕНИЕ

Масштабная оцифровка архивных коллекций опирается на системы оптического распознавания символов (OCR) [1], качество работы которых напрямую определяет полнотекстовый поиск, аналитические исследования и извлечение структурированных данных. Для исторических печатных источников уровень ошибок по-прежнему остается значительным: согласно современным исследованиям, средний показатель Character Error Rate (CER) находится в диапазоне 5–10% [2, 3], что существенно снижает качество поиска и автоматического извлечения фактов.

Метрики и целевые уровни

В работе использованы следующие три стандартные метрики.

Character Error Rate (CER): вычисляется по формуле

$$\mathrm{CER} = \frac{S+I+D}{N},$$

где *S* — количество замен, *I* — вставок, *D* — удалений, *N* — общее количество символов в эталонном тексте.

Word Error Rate (WER) определяется аналогично, но на уровне слов:

$$\mathrm{WER} = \frac{Sw+Iw+Dw}{Nw}.$$

Semantic Similarity (SS) — косинусное сходство между векторными представлениями MiniLM для эталонного текста и гипотезы.

Практические ориентиры из литературы показывают, что для крупномасштабной оцифровки «хорошее» качество OCR соответствует точности 98–99% (т. е. CER ≈ 1–2%), «среднее» — 90–98% (CER ≈ 2–10%), «плохое» — менее 90% (CER > 10%) [4]. Для задач извлечения информации и поиска наблюдается заметное ухудшение результатов при уменьшении точности ниже 70–80% [5]. На основании этого, а также эмпирических наблюдений, согласно которым при CER ≈ 2–3% обычно фиксируется WER ≈ 8–12% [6], в настоящей работе приняты следующие пороговые значения: CER < 0.05, WER < 0.10 и SS > 0.90 – как консервативные критерии «пригодности к использованию».

Исторические газеты XIX в. представляют особую трудность: сложная многоколоночная верстка, смешение шрифтов и износ бумаги приводят к ошибкам сегментации и распознавания, повышая исходные уровни CER и WER по сравнению с современными публикациями. Крупные проекты по оцифровке газет (например, Europeana Newspapers) прямо указывают эти факторы как основные причины снижения качества OCR для исторических периодических изданий, подчеркивая влияние сложной верстки и низкого качества оригиналов [7–10]. Это также отражено в наших экспериментах: для корпуса BLN600 исходные значения составляют CER = 0.084 и WER = 0.202 (см. табл. 1, строка Baseline OCR), которые уменьшаются после нормализации (см. разд. 4).

Определение

Нормализация текста – это процесс преобразования текста в стандартизованную каноническую форму. Она включает исправление орфографических ошибок, раскрытие аббревиатур, устранение сокращений, нормализацию пунктуации, регистра и других языковых вариаций с целью обеспечения согласованного и однородного представления текстовых данных [11].

Последние исследования показывают, что посткоррекция с использованием больших языковых моделей (LLM) позволяет снизить CER ниже 5% и существенно повысить пригодность текста для последующих задач [2, 12, 13]. Однако большинство предложенных решений опирается на модели объемом 13–70 млрд параметров, требующие серверов с объемом видеопамяти ≥ 40 ГБ и/или трудоемкой донастройки на параллельных корпусах [14, 15]. Эти требования делают технологию малодоступной для региональных и институциональных архивов.

Постобучающая квантизация весов (INT8/INT4; GPTQ, AWQ, NF4 и др.) резко снижает потребление памяти и часто сохраняет качество, близкое к полновесным моделям, для задач с коротким и средним контекстом [16]. Однако при сверхдлинных входных данных (> 64 К токенов) агрессивная 4-битная квантизация может заметно ухудшить качество, тогда как 8-битная остается практически без потерь [17]. В нашей задаче (OCR-фрагменты на уровне страниц) длины контекста малы, что позволяет использовать компактные 4-битные модели класса 7–8 В на потребительских видеокартах (≤ 8 ГБ VRAM).

В настоящем исследовании рассмотрено, насколько далеко можно продвинуть нормализацию пост-OCR текстов исторических газет в условиях таких ограничений по ресурсам. Реализован минимальный конвейер: OCR → предобработка → нормализация LLM → постобработка → текст и проведена его оценка на открытом корпусе BLN600 [18]. Показано, что достижение целевых порогов качества возможно при 4-битном выводе на GPU с 8 ГБ видеопамяти, что делает данный подход практически применимым для широкого круга архивных учреждений.

1. ОБЗОР СВЯЗАННЫХ РАБОТ

Vision-LLM, использующийся как прямой OCR, демонстрирует возможности GPT-4V для распознавания текстовых изображений [12]. На рукописном наборе данных IAM они сообщили о значениях CER равных 3.32% и 13.75% на уровне страницы и строки соответственно, в то время как лучшая специализированная CTC-модель достигала 2.89% и 6.52% [12, табл. 4].

На многоязычном наборе данных уличных знаков MLT19 качество резко падает для нелатинских алфавитов: F1-score снижается с 82 (EN) до 1–11 (AR, KO, [12, табл. 2].

Seq2seq-коррекция

LSTM seq2seq-модель применяется после базового OCR и снизили CER с 7–9 % до 4–5 % на корпусе ECCO-TCP [14, разд. 4.2]. Метод требует крупного параллельного корпуса пар «сырой OCR / эталон», что затруднительно для маломасштабных коллекций [15].

Инструкционное дообучение трансформеров

Показано, что Llama-2-7B, дообученная в режиме следования инструкциям, снижает CER с 20% до 9% для газет XIX века (для BART – 15%; [2, табл. 1]), что подчеркивает преимущество моделей, обученных по инструкциям, над ранними seq2seq-трансформерами [2].

Синтетический шум и адаптация при инференсе (TTA)

Добавление марковского шума с последующим дообучением снижает CER с 5–7% до 2–3% [3, разд. 4.3]. Подход SCN-TTA дает сопоставимые 2–3%, начиная с 9% [13, табл. 6].

Многовидовое объединение

Комбинирование нескольких версий OCR одного и того же документа снижает WER с 8–10% до 6–7% [19, табл. 5], что дополняет подходы seq2seq для языков с ограниченной аннотацией [15].

Выводы

Существуют три эффективных направления:

- использование крупных или дообученных LLM;
- генерация синтетического шума и адаптация при инференсе (TTA);
- многовидовое объединение.

Все они демонстрируют CER < 5 %, но требуют либо очень больших моделей (> 13 B) и GPU > 40 ГБ, либо крупных корпусов для дообучения. Наш подход следует идее (i), но использует квантизованные LLM 7–8 B без дополнительного обучения, помещающиеся в 8 ГБ VRAM, устраняя тем самым главный инфраструктурный барьер.

2. МЕТОД

Мощные языковые модели с размером 13–70 млдр параметров успешно уменьшают ошибки OCR, однако требуют ≥ 40 ГБ видеопамяти или трудоемкого дообучения на параллельных корпусах, что недопустимо для региональных архивов с ограниченными ресурсами и небольшими наборами данных.

В настоящей работе показано, что компактные 7–8 B модели, квантизованные до 4 бит и работающие «из коробки», по точности (CER/WER) не уступают крупным моделям, при этом потребляют в несколько раз меньше вычислительных ресурсов.

Как показано на рис. 1, входной поток состоит из сканов газет (в форматах JPEG/PNG/PDF, с разрешением 150–300 dpi). Если исходное OCR-распознавание

не предоставлено, применяется Tesseract 5.3 с языковыми пакетами eng/ru и параметрами oem 3 – psm 4.

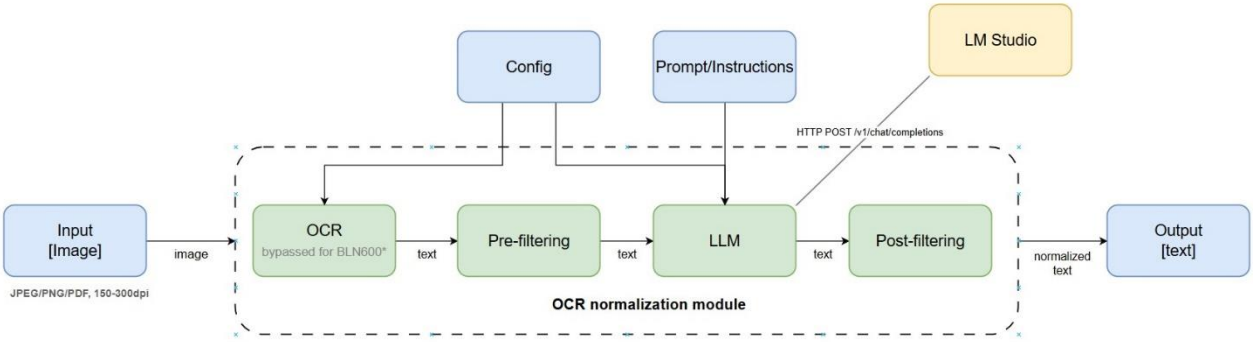


Рис. 1. Схема конвейера нормализации от OCR к LLM-модулю.

Предварительная фильтрация

Наивный слой регулярных выражений (RegEx) удаляет шум, с которым языковая модель (LLM) справляется слабо (см. Листинг 1).

Листинг 1. Правила RegEx, используемые при предварительной и последующей фильтрации

```
def prefilter(text: str) -> str:
    text = text.lower()
    text = re.sub(r"[^\\w\\s]", " ", text)
    text = re.sub(r"\\s+", " ", text)
    return text.strip()

def postfilter(text: str) -> str:
    text = re.sub(r"[()\\[\\]']", "", text)
    text = re.sub(r"\\s+", " ", text)
    return text.strip()
```

Как видно из Листинга 1, набор регулярных выражений, используемых для пред- и постфильтрации, намеренно минимален. Он служит только для удаления шумовых символов, с которыми языковые модели работают плохо, оставляя основную часть нормализации самой модели. Такой подход позволяет избежать переобучения на конкретных артефактах OCR и гарантирует, что улучшения, приведенные в табл. 1, обусловлены именно этапом LLM-нормализации, а не ручной предобработкой.

Нормализация с помощью LLM

Предобработанный текст передается в LM Studio (REST-интерфейс, запрос POST /v1/chat/completions). По умолчанию используется модель YandexGPT-5-Lite-8B-Instruct Q4_K_M (4,9 ГБ); альтернативно — Mistral-7B-Instruct Q4_K_M (4,4 ГБ). Параметры инференса задаются в YAML-конфигурации: temperature = 0.2, top_p = 0.7, top_k=50, max_tokens=4096. Используется промпт под названием «correction» (см. Листинг 2), который явно запрещает галлюцинации и добавление нового содержания.

Листинг 2. Шаблон промпта, используемого для корректировки текста

```
You are an expert text corrector, specialized in fixing OCR error.
1) Fix spelling/grammar.
2) Keep original wording if correct.
3) Do NOT add new sentences.
4) Remove or repair only garbled words.
Text to correct:
"{input_text}"
```

Постфильтрация

Заключительный проход регулярных выражений удаляет оставшиеся лишние символы и двойные пробелы, формируя итоговый текстовый файл .txt.

3. ИНСТРУМЕНТЫ И СРЕДА ВЫПОЛНЕНИЯ

Программное обеспечение: Python 3.12, transformers 4.51.3, sentence-transformers 4.1.0, LM Studio 0.3.17 (build 10).

Аппаратное обеспечение: RTX 4060 Ti (8 ГБ VRAM), Ryzen 5 5600G, 36 ГБ оперативной памяти, CUDA 11.8.

Подсчет семантического сходства выполняется с использованием модели all-MiniLM-L6-v2, при этом вычисляется косинусное сходство после L2-нормализации.

Как показано на рис. 2, YAML-конфигурация описывает используемый набор данных, выбранную языковую модель (LLM) и параметры сэмплирования. Скрипт run_inference.py последовательно обрабатывает пары «сырой OCR / эталонный

текст», сохраняя тексты после каждого этапа (журнал конвейера). Скрипт evaluate_metrics.py вычисляет значения WER, CER и SS для финального результата, а evaluate_aggregates.py усредняет метрики по всему корпусу.

Эксперименты проводились исключительно на корпусе BLN600 (600 страниц британских газет XIX в.), включающем pdf-изображения, результаты «сырого» OCR и эталонный (размеченный) текст. Промежуточные метрики после OCR и обработки LLM используются для анализа вклада каждого этапа.

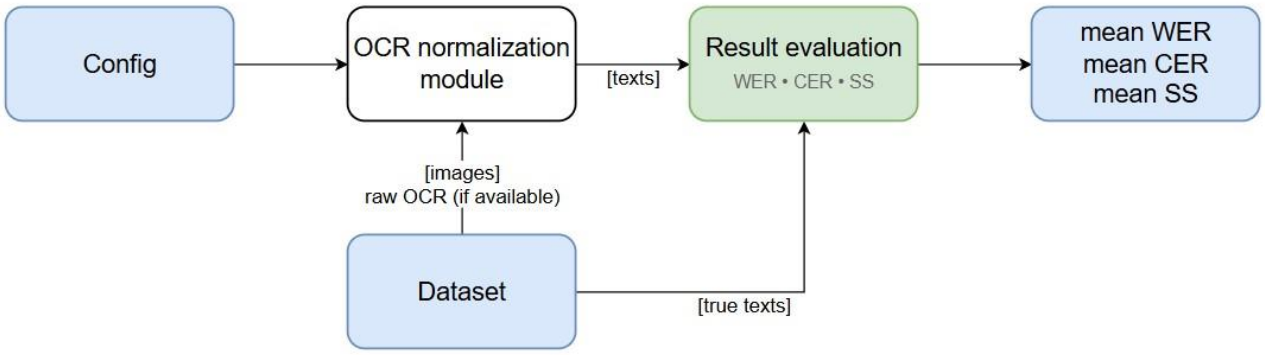


Рис. 2. Схема процесса оценки на наборе данных BLN600

Пример конфигурации

Как показано в Листинге 3, YAML-файл конфигурации определяет набор данных, выбранную языковую модель и параметры сэмплирования.

Листинг 3. YAML-конфигурация для YandexGPT-5 на корпусе BLN600

```
dataset:
  name: BLN600
  path: TMP/bln600
llm:
  model_name: yandexgpt-5-lite-8b-instruct
  host: http://localhost:1234/v1/chat/completions
  temperature: 0.2
  top_p: 0.7
  top_k: 50
  max_tokens: 4096
experiment:
  output_path: results/bln600_yandex.yaml
```

Унифицированный формат результатов

Скрипты для оценки ожидают наличие ключей:

«ground_truth, ocr_text, corrected_text».

Таким образом, предложенный конвейер может выполняться на графических процессорах с объемом памяти ≤ 8 ГБ, не требует дополнительных данных для дообучения и легко переиспользуется благодаря YAML-конфигурациям и API-сервису LM Studio.

4. РЕЗУЛЬТАТЫ

В табл. 1 представлены метрики моделей без дообучения (результаты, полученные нашим подходом).

Табл. 1. Качество на корпусе BLN600 после нормализации (GPU 8 ГБ)

Модель	**CER ↓**	**WER ↓**	**SS ↑**
Базовый OCR	0.0840	0.2020	0.8455
Mistral-7B-Q4	0.0921	0.1240	0.9315
YandexGPT-5-Q4	**0.0399**	**0.0650**	**0.9616**
Llama-2-7B-Q4	0.1490	0.1650	0.9279
Llama-2-13B-Q4	0.4205	0.4060	0.8400

В табл. 2 представлены метрики моделей с дообучением, взятые из источников.

Табл. 2. Данные, приведенные в литературе

Модель (источник)	**CER ↓**	**WER ↓**	**SS ↑**
Llama-2-13B-Q8*	0.038	n/a	n/a
Llama-2-7B-Q8*	0.048	n/a	n/a
Llama-3-8B-F16**	0.080	0.190	n/a

* Модели дообучены на исторических данных OCR [2].

** Данные из [3].

Основные выводы

Модель YandexGPT-5-Q4 снижает Character Error Rate (CER) на 52% и Word Error Rate (WER) на 68% по сравнению с исходным OCR, при этом помещаясь в 4.9 ГБ VRAM.

Модель Mistral-7B-Q4 демонстрирует умеренные улучшения при том же объеме памяти.

Обе модели обеспечивают семантическое сходство выше 0.93, что превышает установленный порог пригодности к использованию (0.90).

Динамика ошибок

На рис. 3–5 приведены значения метрик до и после нормализации на корпусе BLN600 для моделей, используемых в нашем пайплайне, а также для моделей из литературы.

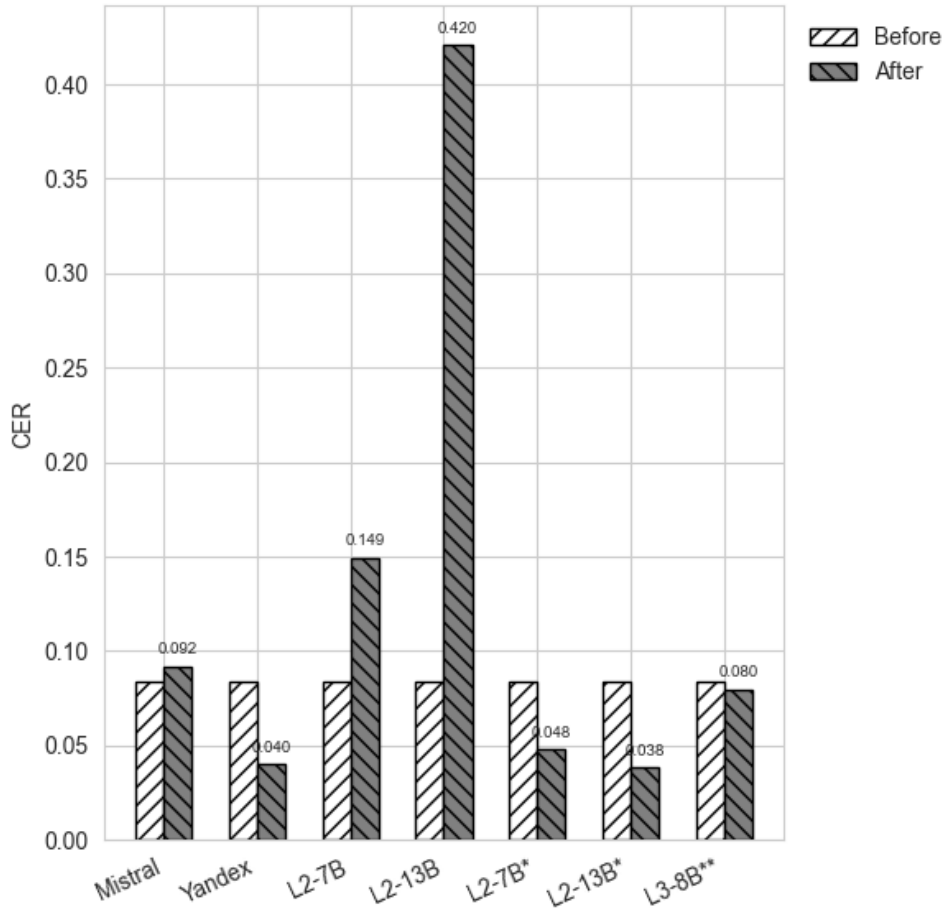


Рис. 3. Ошибка на уровне символов (CER) до и после нормализации на корпусе BLN600.

Нестержневые (без звездочек) столбцы — наши запуски zero-shot; звездочками отмечены результаты, приведенные в литературе. Столбец «До» (CER = 0.084) общий для всех моделей.

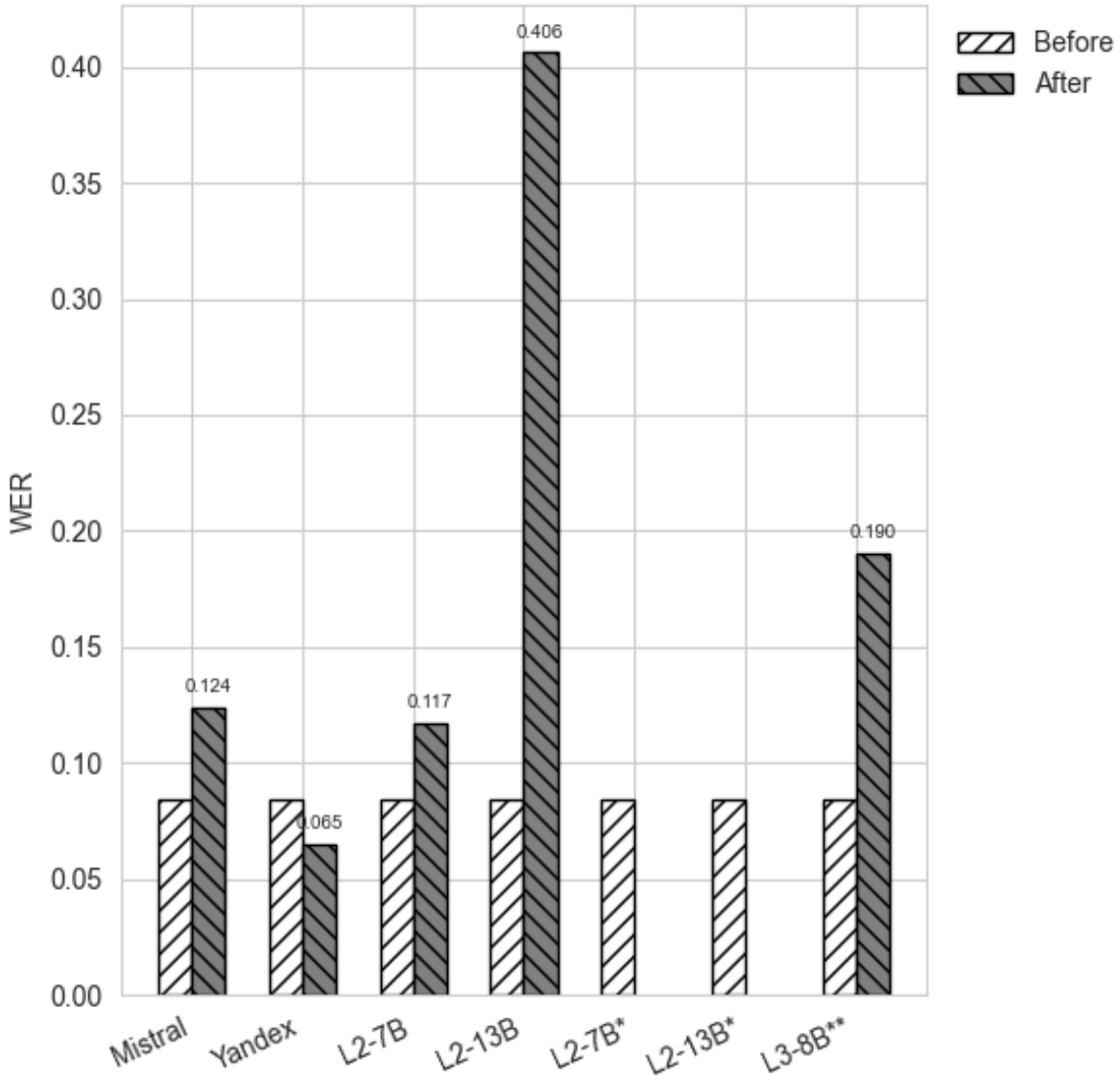


Рис. 4. Ошибка на уровне слов (WER) до и после нормализации на BLN600. Нестержневые столбцы — наши результаты без дообучения; звездочками отмечены литературные базовые модели.

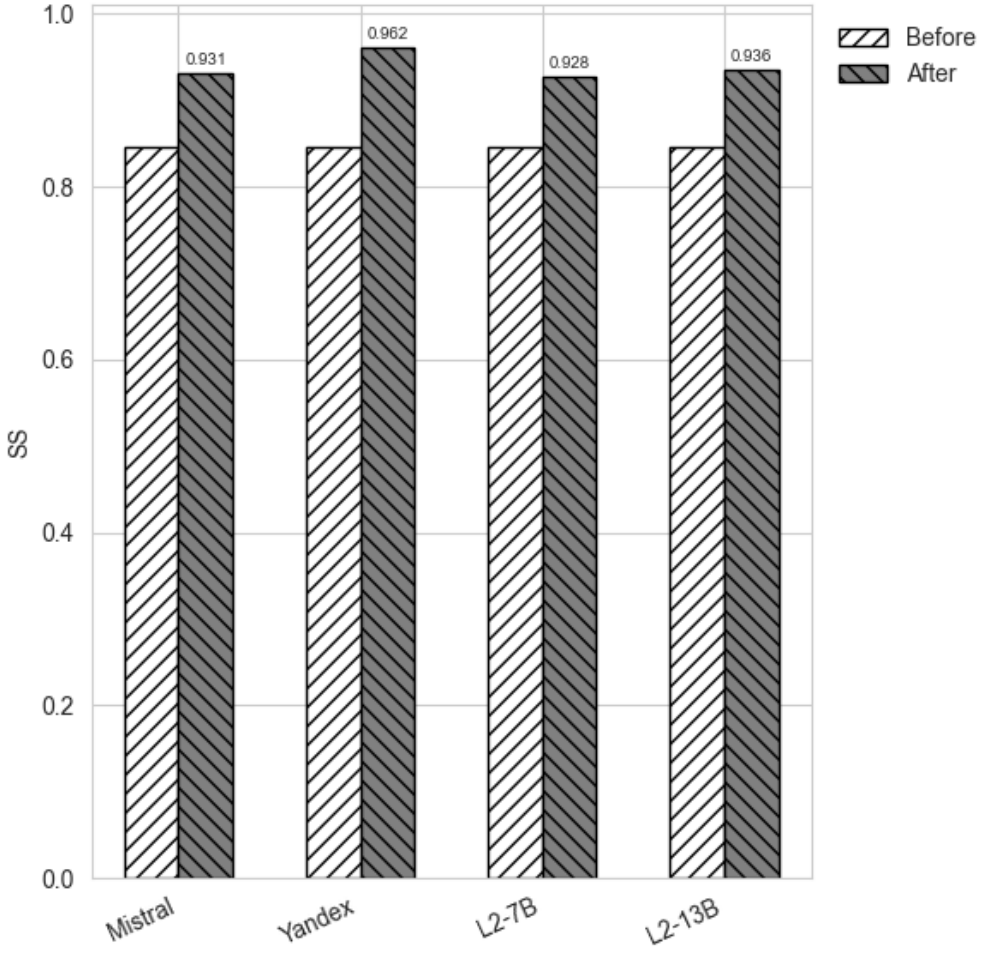


Рис. 5. Семантическое сходство (MiniLM) до и после нормализации для наших zero-shot запусков. Дообученные модели из литературы не предоставляют значения SS и поэтому не показаны.

Время выполнения

В табл. 3 представлено время инференса моделей на указанной видеокарте.

Табл. 3. Задержка и использование памяти на RTX 4060 Ti (8 ГБ)

Модель	VRAM (ГБ)	Время (с/страница)
Mistral-7B-Q4	4.1	32.7
YandexGPT-5-Q4	**4.6**	**35.3**
Llama-2-7B-Q4	3.6	38.3
Llama-2-13B-Q4	6.9	112.5

5. ОБСУЖДЕНИЕ

Важно отметить, что предложенный нами метод не оценивался на таких бенчмарках, как IAM или MLT19, которые ориентированы на OCR-задачи построчного или пословного уровня, а также на распознавание уличных знаков. Эти наборы данных включают очень короткие входные контексты – часто ограниченные одним словом или одной строкой.

В подобных условиях наш подход, основанный на использовании LLM для нормализации фрагментов на уровне целой страницы, вероятно, показал бы худшие результаты, поскольку успешная коррекция в минимальном контексте обычно требует специализированных словарей, лексиконов или дополнительных внешних источников, обеспечивающих недостающий контекст.

В противоположность этому наш конвейер изначально спроектирован для OCR-фрагментов на уровне страницы, где доступен более широкий текстовый контекст. Это позволяет LLM использовать соседние токены, чтобы корректно интерпретировать и исправлять ошибки OCR.

Когда контекстное окно сужается — до длины строки или слова, языковой модели не хватает информации для надежной коррекции, если она опирается только на внутренние вероятностные представления. В таких условиях модели обычно выигрывают от использования доменных словарей, ограниченного декодирования или правил постобработки.

Таким образом, эффективность предложенного нами конвейера нормализации обусловлена наличием достаточного контекстного окружения и не может напрямую распространяться на задачи с крайне коротким входным текстом.

Для полноты и воспроизводимости исследования мы включили наборы IAM и MLT19 в сопровождающий GitHub-репозиторий настоящей работы. Эти датасеты не входят в основную часть оценки, однако предоставлены как вспомогательные ресурсы — для будущих сравнительных исследований и для исследователей, заинтересованных в адаптации нашего подхода к OCR-задачам с коротким контекстом.

Таким образом, эффективность предложенного конвейера нормализации напрямую связана с доступностью достаточного контекста для анализа, и его применение может быть ограничено задачами, где длина входных данных крайне

мала. В дальнейшем работа будет направлена на проверку обобщающей способности подхода при сокращенном контексте и в многоязычных корпусах.

ЗАКЛЮЧЕНИЕ

Представлен zero-shot конвейер с 4-битной квантизацией, способный устранять ошибки OCR при использовании потребительского оборудования.

На бенчмарке BLN600 предложенный конвейер снижает показатели ошибок CER с 8.4% до 4.0% и WER с 20.2% до 6.5% при использовании модели YandexGPT-5-Instruct-Q4 (8B), а также CER до 9.2% и WER до 12.4% при модели Mistral-7B-Instruct-Q4.

Обе модели помещаются в видеопамять объемом ≤ 5 ГБ и обрабатывают одну страницу газеты примерно за 35 секунд на RTX 4060 Ti, демонстрируя результаты, сопоставимые или превосходящие значительно более крупные дообученные модели — без необходимости дополнительного обучения или использования параллельных данных.

Укажем возможные направления дальнейшей работы:

- расширение на многоязычные данные (кириллица + латиница);
- применение легковесных архитектур типа Mixture-of-Experts и chain-of-thought prompting для сложных макетов страниц;
- создание открытого российского бенчмарка с административными формами для стимулирования открытых исследований.

В целом компактные 4-битные LLM представляют собой практичную и экономичную альтернативу крупным моделям для постобработки OCR, открывая возможности масштабного цифрового восстановления исторических и специализированных архивов. Исходный код и скрипты для оценки результатов доступны в открытом репозитории: https://github.com/Kerysfel/OCRNorm

СПИСОК ЛИТЕРАТУРЫ

1. *Memon J., Sami M., Khan R.A.* Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR) // IEEE Access. 2020. Vol. 8. P. 142642–142668. https://doi.org/10.1109/ACCESS.2020.3012542

2. *Thomas A., Gaizauskas R., Lu H.* Leveraging LLMs for Post-OCR Correction of Historical Newspapers // Proceedings of the 2nd Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA). 2024. P. 116–121. https://doi.org/10.18653/v1/2024.lt4hala-1.6

3. *Bourne J.* Scrambled text: training Language Models to correct OCR errors using synthetic data // arXiv preprint. 2024. arXiv:2409.19735. https://doi.org/10.48550/arXiv.2409.19735

4. *Holley R.* How Good Can It Get? Analysing and Improving OCR Accuracy in Large-Scale Historic Newspaper Digitisation Programs // D-Lib Magazine. 2009. Vol. 15, No. 3/4. https://doi.org/10.1045/march2009-holley

5. *van Strien D., Beelen K., Coll Ardanuy M., Hosseini K., McGillivray B., Tolfo G.S.* Assessing the Impact of OCR Quality on Downstream NLP Tasks // Proceedings of the 12th International Conference on Agents and Artificial Intelligence (ICAART 2020). 2020. P. 484–496. https://doi.org/10.5220/0009169004840496

6. *Drobac S., Friberg Heppin K., Wirén M., Lindén K.* Optical Character Recognition with Neural Networks and Post-Correction with Finite State Methods // International Journal on Document Analysis and Recognition (IJDAR). 2020. Vol. 23. P. 279–295. https://doi.org/10.1007/s10032-020-00359-9

7. *Neudecker C., Antonacopoulos A.* Making Europe's Historical Newspapers Searchable // DAS 2016 Workshop / Europeana Newspapers. 2016.(Workshop paper). URL: https://www.primaresearch.org/www/files/das2016/Europeana%20Newspapers.pdf.

8. *Boillet M., Kermorvant C., Paquet T.* Robust text line detection in historical documents: learning and evaluation methods // International Journal on Document Analysis and Recognition (IJDAR). 2022. Vol. 25. P. 95–114. https://doi.org/10.1007/s10032-022-00395-7

9. *Ermakova L., Tolfo G.S., Hosseini K.* On the Impact of OCR Quality on Named Entity Extraction from Historical Newspapers // DH Benelux 2021 (Extended abstracts). 2021. URL: https://dhbenelux.org/wp-content/uploads/booklet2021.pdf#page=66

10. *Kettunen K.* Optical Character Recognition Quality Affects Perceived Usefulness and Trust // arXiv preprint. 2022. arXiv:2209.08222.

11. *Sreelekha S., Sumam A.R., Nair R.R.* Systematic Review on Text Normalization Techniques and Its Approach to Non-Standard Words // Preprint. 2023 (ResearchGate). URL: https://www.researchgate.net/publication/373877004.

12. *Shi Y., Peng D., Liao W., Lin Z., Chen X., Liu C., Zhang Y., Jin L.* Exploring OCR Capabilities of GPT-4V(ision): A Quantitative and In-Depth Evaluation // arXiv preprint. 2023. arXiv:2310.16809.

13. *Guan S., Xu C., Lin M., Greene D.* Effective Synthetic Data and Test-Time Adaptation for OCR Correction // Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024). 2024. P. 15412–15425 (ACL Anthology).

14. *Kanerva J., Ledins C., Käpyaho S., Ginter F.* OCR Error Post-Correction with LLMs in Historical Documents: No Free Lunches // Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025). 2025. Tallinn, Estonia (ACL Anthology).

15. *Rijhwani S., Anastasopoulos A., Neubig G.* OCR Post-Correction for Endangered Language Texts // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020. P. 5931–5942.
https://doi.org/10.18653/v1/2020.emnlp-main.478

16. *Jin R., Du J., Huang W., Liu W., Luan J., Wang B., Xiong D.* A Comprehensive Evaluation of Quantization Strategies for Large Language Models // Findings of ACL 2024 (also arXiv preprint). 2024. arXiv:2402.16775.
https://doi.org/10.48550/arXiv.2402.16775

17. *Mekala A., Atmakuru A., Song Y., Karpinska M., Iyyer M.* Does Quantization Affect Models' Performance on Long-Context Tasks? // arXiv preprint. 2025. arXiv:2505.20276. https://doi.org/10.48550/arXiv.2505.20276.

18. *Booth C.W., Thomas A., Gaizauskas R.* BLN600: A Parallel Corpus of Machine/Human Transcribed Nineteenth-Century Newspaper Texts // Proceedings of LREC-COLING 2024. 2024. P. 2440–2446.
https://doi.org/10.15131/shef.data.25439023.

19. *Gupta H., Del Corro L., Broscheit S., Hoffart J., Brenner E.* Unsupervised Multi-View Post-OCR Error Correction with Language Models // Proceedings of the

2021 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2021. P. 8647–8652. https://doi.org/10.18653/v1/2021.emnlp-main.680

NORMALIZATION OF TEXT RECOGNIZED BY OPTICAL CHARACTER RECOGNITION USING LIGHTWEIGHT LLMs

V. K. Vershinin[1] [0009-0001-9425-0881], I. V. Khodnenko[2] [0009-0003-7787-7126], S. V. Ivanov[3] [0000-0002-1128-2942]

[1–3]*ITMO University, Saint-Petersburg, Russia*

[1]vershinin@itmo.ru, [2]Ivan.Khodnenko@itmo.ru, [3]svivanov@itmo.ru

Abstract

Despite recent progress, Optical Character Recognition (OCR) on historical newspapers still leaves 5–10% character errors. We present a fully automated post-OCR normalization pipeline that combines lightweight 7–8B instruction-tuned LLMs quantized to 4-bit (INT4) with a small set of regex rules. On the BLN600 benchmark (600 pages of 19th-century British newspapers), our best model YandexGPT-5-Instruct Q4 reduces Character Error Rate (CER) from 8.4% to 4.0% (–52.5%) and Word Error Rate (WER) from 20.2% to 6.5% (–67.8%), while raising semantic similarity to 0.962. The system runs on consumer hardware (RTX-4060 Ti, 8 GB VRAM) at about 35 seconds per page and requires no fine-tuning or parallel training data. These results indicate that compact INT4 LLMs are a practical alternative to large checkpoints for post-OCR cleanup of historical documents.

Keywords*: optical character recognition, post-OCR correction, historical newspapers, large language models, quantization, INT4, normalization pipeline, character error rate, semantic similarity, regex rules, YandexGPT-5, lightweight models, natural language processing, digital humanities, document digitization.*

REFERENCES

1. *Memon J., Sami M., Khan R.A.* Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR) // IEEE Access. 2020. Vol. 8. P. 142642–142668. https://doi.org/10.1109/ACCESS.2020.3012542

2. *Thomas A., Gaizauskas R., Lu H.* Leveraging LLMs for Post-OCR Correction of Historical Newspapers // Proceedings of the 2nd Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA). 2024. P. 116–121. https://doi.org/10.18653/v1/2024.lt4hala-1.6

3. *Bourne J.* Scrambled text: training Language Models to correct OCR errors using synthetic data // arXiv preprint. 2024. arXiv:2409.19735. https://doi.org/10.48550/arXiv.2409.19735

4. *Holley R.* How Good Can It Get? Analysing and Improving OCR Accuracy in Large-Scale Historic Newspaper Digitisation Programs // D-Lib Magazine. 2009. Vol. 15, No. 3/4. https://doi.org/10.1045/march2009-holley

5. *van Strien D., Beelen K., Coll Ardanuy M., Hosseini K., McGillivray B., Tolfo G.S.* Assessing the Impact of OCR Quality on Downstream NLP Tasks // Proceedings of the 12th International Conference on Agents and Artificial Intelligence (ICAART 2020). 2020. P. 484–496. https://doi.org/10.5220/0009169004840496

6. *Drobac S., Friberg Heppin K., Wirén M., Lindén K.* Optical Character Recognition with Neural Networks and Post-Correction with Finite State Methods // International Journal on Document Analysis and Recognition (IJDAR). 2020. Vol. 23. P. 279–295. https://doi.org/10.1007/s10032-020-00359-9

7. *Neudecker C., Antonacopoulos A.* Making Europe's Historical Newspapers Searchable // DAS 2016 Workshop / Europeana Newspapers. 2016.(Workshop paper). URL: https://www.primaresearch.org/www/files/das2016/Europeana%20Newspapers.pdf.

8. *Boillet M., Kermorvant C., Paquet T.* Robust text line detection in historical documents: learning and evaluation methods // International Journal on Document Analysis and Recognition (IJDAR). 2022. Vol. 25. P. 95–114. https://doi.org/10.1007/s10032-022-00395-7

9. *Ermakova L., Tolfo G.S., Hosseini K.* On the Impact of OCR Quality on Named Entity Extraction from Historical Newspapers // DH Benelux 2021 (Extended abstracts). 2021.
URL: https://dhbenelux.org/wp-content/uploads/booklet2021.pdf#page=66

10. *Kettunen K.* Optical Character Recognition Quality Affects Perceived Usefulness and Trust // arXiv preprint. 2022. arXiv:2209.08222.

11. *Sreelekha S., Sumam A.R., Nair R.R.* Systematic Review on Text Normalization Techniques and Its Approach to Non-Standard Words // Preprint. 2023 (ResearchGate). URL: https://www.researchgate.net/publication/373877004.

12. *Shi Y., Peng D., Liao W., Lin Z., Chen X., Liu C., Zhang Y., Jin L.* Exploring OCR Capabilities of GPT-4V(ision): A Quantitative and In-Depth Evaluation // arXiv preprint. 2023. arXiv:2310.16809.

13. *Guan S., Xu C., Lin M., Greene D.* Effective Synthetic Data and Test-Time Adaptation for OCR Correction // Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024). 2024. P. 15412–15425 (ACL Anthology).

14. *Kanerva J., Ledins C., Käpyaho S., Ginter F.* OCR Error Post-Correction with LLMs in Historical Documents: No Free Lunches // Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025). 2025. Tallinn, Estonia (ACL Anthology).

15. *Rijhwani S., Anastasopoulos A., Neubig G.* OCR Post-Correction for Endangered Language Texts // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020. P. 5931–5942.
https://doi.org/10.18653/v1/2020.emnlp-main.478

16. *Jin R., Du J., Huang W., Liu W., Luan J., Wang B., Xiong D.* A Comprehensive Evaluation of Quantization Strategies for Large Language Models // Findings of ACL 2024 (also arXiv preprint). 2024. arXiv:2402.16775.
https://doi.org/10.48550/arXiv.2402.16775

17. *Mekala A., Atmakuru A., Song Y., Karpinska M., Iyyer M.* Does Quantization Affect Models' Performance on Long-Context Tasks? // arXiv preprint. 2025. arXiv:2505.20276. https://doi.org/10.48550/arXiv.2505.20276.

18. *Booth C.W., Thomas A., Gaizauskas R.* BLN600: A Parallel Corpus of Machine/Human Transcribed Nineteenth-Century Newspaper Texts // Proceedings of LREC-COLING 2024. 2024. P. 2440–2446. https://doi.org/10.15131/shef.data.25439023.

19. *Gupta H., Del Corro L., Broscheit S., Hoffart J., Brenner E.* Unsupervised Multi-View Post-OCR Error Correction with Language Models // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2021. P. 8647–8652. https://doi.org/10.18653/v1/2021.emnlp-main.680

СВЕДЕНИЯ ОБ АВТОРАХ

ВЕРШИНИН Владислав Константинович – ассистент факультета технологий искусственного интеллекта, инженер лаборатории систем поддержки принятия решений, Университет ИТМО. Окончил образовательную программу «Big Data and Machine Learning» Университета ИТМО.

Область научных интересов: RAG-системы, автоматизация цифровой обработки изображений, машинное обучение, обработка естественного языка.

Vladislav Konstantinovich VERSHININ – Teaching Assistant at the Faculty of Artificial Intelligence Technologies and Engineer at the Laboratory of Decision Support Systems, ITMO University. Graduate of the “Big Data and Machine Learning” program at ITMO University.

Research interests: Retrieval-Augmented Generation (RAG) systems, automated digital image processing, machine learning, natural language processing.

email: vershinin@itmo.ru

ORCID: 0009-0001-9425-0881

ИВАНОВ Сергей Владимирович – доцент факультета технологий искусственного интеллекта, Университет ИТМО; руководитель образовательной программы «Инженерия искусственного интеллекта». Кандидат технических наук (2008).

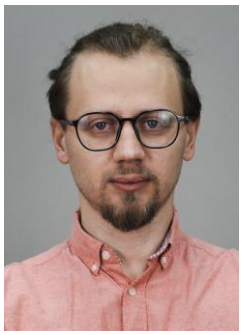
Область научных интересов: машинное обучение, искусственный интеллект, математическое моделирование, оптимизация.

Sergey Vladimirovich IVANOV – Associate Professor at the Faculty of Artificial Intelligence Technologies, ITMO University; Head of the “Artificial Intelligence Engineering” educational program. PhD in Engineering Sciences (2008).

Research interests: machine learning, artificial intelligence, mathematical modeling, optimization.

email: svivanov@itmo.ru

ORCID: 0000-0002-1128-2942



ХОДНЕНКО Иван Владимирович — старший научный сотрудник и старший преподаватель, Национальный центр когнитивных разработок, Университет ИТМО. Окончил Волгоградский государственный технический университет (бакалавр) и Университет ИТМО (магистратура, аспирантура). Кандидат технических наук (2022).

Область научных интересов: машинное обучение, компьютерное зрение, распознавание документов.

Ivan Vladimirovich KHODNENKO – Senior Researcher and Senior Lecturer at the National Center for Cognitive Research, ITMO University. Holds a Bachelor’s degree from Volgograd State Technical University and completed Master’s and PhD studies at ITMO University. PhD in Technical Sciences (2022).

Research interests: machine learning, computer vision, document recognition.

email: Ivan.Khodnenko@itmo.ru

ORCID: 0009-0003-7787-7126

Материал поступил в редакцию 10 октября 2025 года

УДК 159.922.6

ЦИФРОВОЕ МОДЕЛИРОВАНИЕ ТЕМАТИЧЕСКОГО ПОЛЯ ИЗУЧЕНИЯ КУЛЬТУРНОЙ КОНГРУЭНТНОСТИ В ПСИХОЛОГИЧЕСКОМ КОНТЕКСТЕ

А. М. Ганиева[0000-0003-1323-9363]

Казанский (Приволжский) федеральный университет, г. Казань, Россия

ganieva.aisylu@mail.ru

Аннотация

В работе установлены ключевые темы в современных психологических исследованиях культурной конгруэнтности с использованием метода тематического цифрового моделирования массива научных публикаций.

Актуальность и значимость проведенного исследования обусловлены ростом значимости культурной конгруэнтности в условиях цифровой трансформации общества, изменяющей способы социализации и взаимодействия. Современные технологии требуют переосмысления психологических механизмов адаптации индивида к культурной среде, особенно в детском и подростковом возрастах. Несмотря на активное изучение этого феномена, наблюдается очевидный недостаток исследований, посвященных культурной конгруэнтности взрослых. Применение цифрового моделирования и искусственного интеллекта позволяет систематизировать знания и выявить структуру тематического поля с высокой точностью. Полученные данные открывают перспективу для дальнейшего изучения культурной конгруэнтности в ходе онтогенеза.

Конструирование тематического поля исследований культурной конгруэнтности, основанный на анализе цифровых анналов, содержащих коллекцию научных публикаций по данной тематике (112 статей), был выполнен с использованием алгоритма тематического моделирования (topic modeling) на языке программирования Python и с применением цифровых платформ, включая инструменты на основе мультимодальных нейросетей (GigaChat, Qwen, DeepSeek). В результате проведенного анализа возрастных особенностей феномена культурной конгруэнтности выделены четыре возрастные группы: дошкольники, младшие школьники, подростки и взрослые.

© А. М. Ганиева, 2025.
Данная статья распространяется на условиях международной лицензии Creative Commons License Attribution 4.0 International (CC BY 4.0).

***Ключевые слова**: культурная конгруэнтность, психологическое исследование, возрастная психология, общая психология, тематическое моделирование.*

ВВЕДЕНИЕ

В работе представлен тематический обзор, посвященный анализу научных исследований в области культурной конгруэнтности. Обзор предполагал целенаправленное структурирование существующих публикаций по выбранной теме, выявление ключевых концепций, теоретических подходов и пробелов в научной литературе. Методологическая основа исследования построена на строгом соблюдении рекомендаций протокола PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) [1], что обеспечивает прозрачность, воспроизводимость, систематичность и полноту процесса отбора и анализа источников [2]. Применение подхода scoping review позволило не только уточнить ключевые теоретические дефиниции и концептуальные рамки, но и идентифицировать основные феномены, связанные с проявлениями культурной конгруэнтности, а также выявить существующие пробелы в научной литературе и наметить перспективные направления дальнейших исследований.

В настоящей работе рассмотрены следующие задачи с целью выделения ключевых тематических доменов для определения сущности культурной конгруэнтности в психологическом контексте.

1. Какие системообразующие тематические домены, интегрирующие теоретические концепции культурной конгруэнтности в психологическом контексте, доминируют в современном научном дискурсе.

2. Какова структура исследовательского поля культурной конгруэнтности в психологическом контексте, включая основные направления ее изучения, их теоретическую обоснованность, а также связанные с ними методологические и концептуальные ограничения.

ФОРМИРОВАНИЕ ВЫБОРКИ ДЛЯ ОБЗОРА ТЕМАТИЧЕСКОГО ПОЛЯ

Формирование выборки осуществлялось посредством поиска литературы в электронной библиотеке eLIBRARY.RU [3]. Поиск в международных библиографических базах данных Scopus [4] и WoS [5] не был произведен ввиду существующих на сегодняшний день ограничений. Процедура поиска и отбора статей включала три этапа (рис. 1).

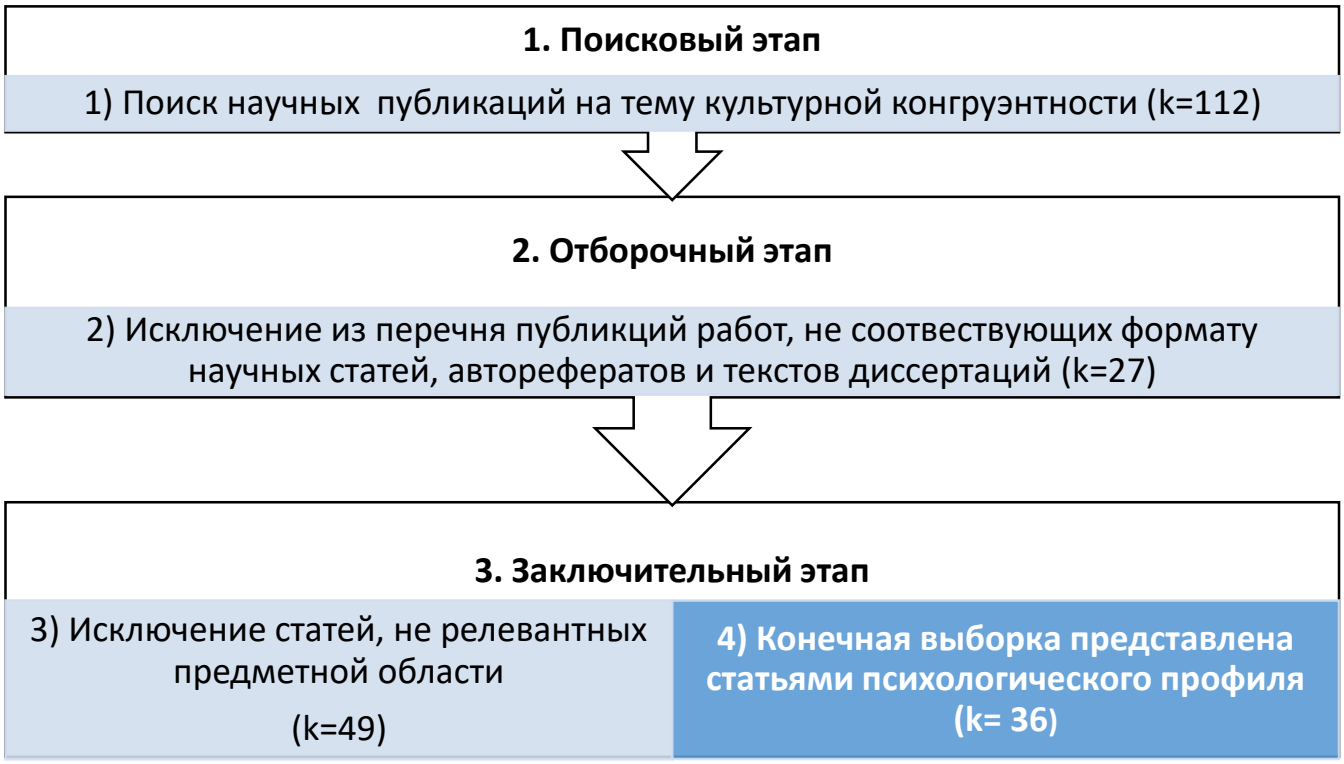


Рис. 1. Процедура формирования репрезентативной выборки

Запрос для базы данных eLIBRARY.RU происходил по следующему алгоритму:

при формировании релевантной выборки исследования на первом этапе путем поисковых запросов было констатировано наличие 112 публикаций на тему культурной конгруэнтности [6]:

```
elibrary_search_query = '''("культурная конгруэнтность" OR "cultural congruence")
AND (дети OR подростки OR взрослые OR children OR adolescents OR adults)'''
print ("Запрос для eLibrary:", elibrary_search_query)
```

Поиск данных выполнен через веб-интерфейс сайта:

• включены заголовки, аннотации, ключевые слова: *культурная конгруэнтность, дети, подростки, взрослые; cultural congruence, children, adolescents, adults*

• применены фильтры: рецензируемые журналы, ВАК, диссертации

• тематика: психологические науки

• язык: русский, английский

• период: без ограничений.

На втором этапе до начала предметного анализа из общей совокупности документов были исключены все материалы, не соответствующие формату полноценной научной статьи, автореферата или диссертации. В этой категории оказались такие типы публикаций, как тезисы конференции, глава в книге, отчеты о НИР, патенты, учебные пособия (всего было исключено 27 документов). В том числе были исключены дублирующиеся публикации. В результате осталось 85 текстов, удовлетворяющих установленным критериям в рамках содержательного анализа.

На третьем этапе из имеющегося списка были исключены 49 статей, которые не были валидны изучаемой тематике, например исследования культурной конгруэнтности в области филологии, сельского хозяйства, в менеджменте и др. В результате тщательного отбора в подборку вошли только 36 статей, которые были релевантны тематике исследования. Отобранные статьи были повторно проанализированы с целью обоснования их актуальности для решения поставленных исследовательских вопросов.

На рис. 2 представлена динамика публикационной активности по теме культурной конгруэнтности в период с 2013 по 2025 годы, отраженная в виде горизонтальной гистограммы. Примечательно, что количество научных работ увеличилось с единичного случая в 2013 году до максимума в 16 публикаций в 2024 году, что свидетельствует о прогрессирующем интересе исследовательского сообщества к данной проблематике. Несмотря на временные снижения объема публикаций в отдельные годы, общая тенденция остается восходящей, что подтверждается линией тренда на графике. Такая динамика указывает на возрастающую значимость темы в современных гуманитарных и социальнонаправленных исследованиях. Таким образом, стоит отметить, что культурная конгруэнтность становится все более значимым объектом для теоретического и эмпирического анализа.

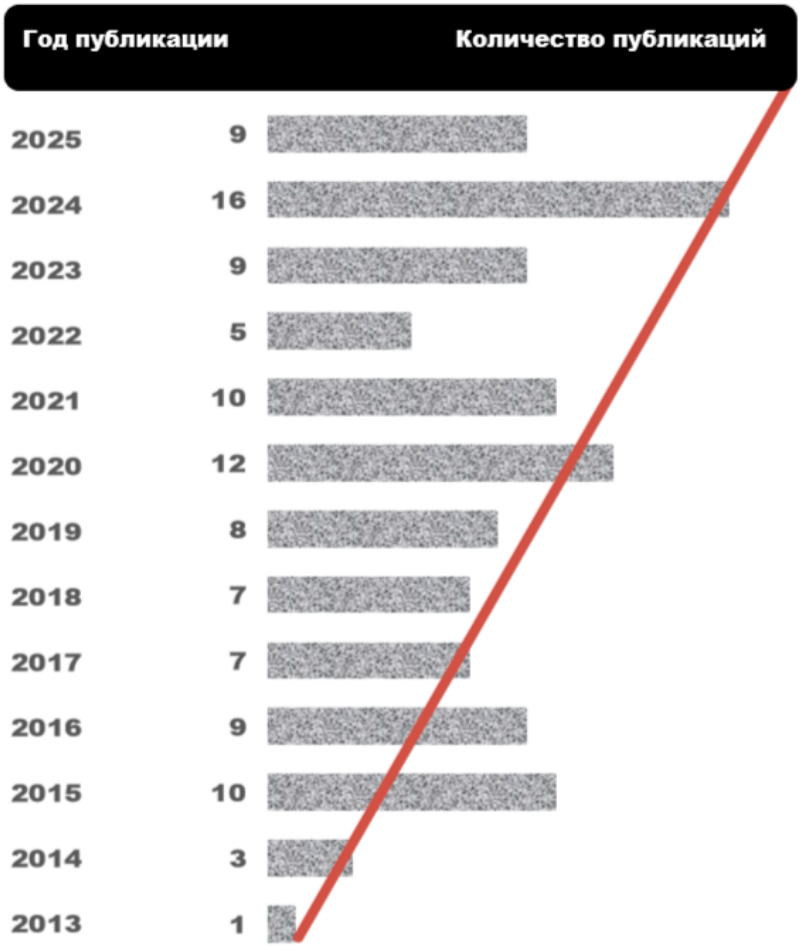


Рис. 2. Динамика научных публикаций, посвященных культурной конгруэнтности, по годам

АНАЛИЗ ТЕМАТИЧЕСКОГО ПОЛЯ: ЭКСТРАКЦИЯ ДАННЫХ

В рамках основной части настоящего исследования из пула собранных аннотаций были экстрагированы данные с применением специализированного языка программирования Python [7]. В соответствии с методологическими рекомендациями ранее опубликованных исследований [8], первичная обработка текстов включала удаление цифр, знаков препинания и прочих символов, последующее разделение текста на отдельные лексические единицы (токены) и приведение всех слов к нижнему регистру. Для сегментации текста была использована униграммная токенизация - подход, который, согласно результатам тематического моделирования, обеспечивает более высокую точность анализа по сравнению с другими методами [9].

На следующем этапе был проведен процесс лемматизации - приведения словоформ к их базовой словарной форме. После этого из конструкта были исключены термины, которые были односложными и включали в свой состав менее трех символов и стоп-слов (из специальной библиотеки NLTK (Natural Language Toolkit)), не представляющих высокой семантической значимости в рамках настоящего исследования.

Далее текстовые данные были представлены в виде матрицы («документы×термины»), где строки соответствовали отобранным аннотациям, а столбцы - выбранным лексическим единицам. Выбор биграмм обусловлен необходимостью более адекватно отразить семантическую структуру текста по сравнению с одиночными словами [10]. Среди биграмм ключевым является основное понятие в рамках настоящего анализа - культурная конгруэнтность.

В процессе обработки текстовых данных была выполнена их векторизация с помощью класса CountVectorizer из библиотеки scikit-learn. В результате анализа было идентифицировано 16 устойчивых биграмм, которые послужили основой для последующего тематического моделирования. С применением алгоритма тематического моделирования, в ходе которого систематически варьировались ключевые параметры, включая число итераций и начальные случайные состояния, удалось выделить четыре устойчивых тематических кластера, проявлявших высокую воспроизводимость при многократных ретестах.

Основная часть работы состояла в анализе полных текстов публикаций с целью интерпретации выделенных тем и их оценки на предмет релевантности с экспертной точки зрения.

На следующем этапе с помощью инструментов, работающих на основе мультимодальной нейросети (Qwen [11], Deepseek [12], GigaChat [13]), были проанализированы аннотации к статьям и выполнены их группировки по ключевым словам и возрастным периодам (табл. 1).

Следует отметить, что в проанализированных публикациях отсутствуют работы, фокусирующиеся на взрослых как отдельной возрастной группе в контексте культурной конгруэнтности. Все выявленные публикации касаются детей

(дошкольного и младшего школьного возраста) и подростков. В связи с этим возрастная группа взрослых в исследовании не представлена ввиду отсутствия соответствующих данных.

Табл. 1. Распределение статей по тематикам

Номер группы	Количество наименований	Сочетания слов
1	14	культурная конгруэнтность дошкольника, культурная конгруэнтность младшего школьника, соответствие нормативной ситуации, поведение детей, уровень культурной конгруэнтности, формирование культурной конгруэнтности, диагностика культурной конгруэнтности, культурная конгруэнтность и когнитивное развитие, культурная конгруэнтность и интеллект, культурная конгруэнтность и регуляторные функции, культурная конгруэнтность и эмоциональное распознавание, культурная конгруэнтность и личностные характеристики, культурная конгруэнтность и диалектическое мышление, культурная конгруэнтность и экранное время
2	7	культурная конгруэнтность подростка, культурная конгруэнтность и креативность, культурная конгруэнтность и дивергентное мышление, культурная конгруэнтность и безопасность, культурная конгруэнтность и самоорганизация, культурная конгруэнтность и социальное взаимодействие, культурная конгруэнтность и учебная деятельность, культурная конгруэнтность и личностные свойства, культурная конгруэнтность и урбанизация, культурная конгруэнтность и экопсихологические условия
3	0	–

4	15	культурная конгруэнтность, нормативная ситуация, соответствие правилам, диалектическое мышление, формально-логическое мышление, теоретическое мышление, регуляторные функции, рабочая память, переключаемость, сдерживающий контроль, интеллектуальное развитие, уровень интеллекта, IQ, личностные характеристики, общительность, возбудимость, тревожность, склонность к риску, социальная смелость, самооценка, распознавание эмоций, понимание эмоций, эмоциональная сфера, время перед монитором, цифровая среда, половая принадлежность, гендерные различия, трудные жизненные ситуации, проблемно-противоречивые ситуации, психические новообразования, рефлексия, внутренний план действия, уровни поведения в нормативной ситуации

Как видно из табл. 1, группа 1 включает аннотации 14 работ, в которых изучается культурная конгруэнтность детей в контексте развития, поведения, когнитивных функций, эмоциональной сферы и влияния внешних факторов. При этом в научных работах сравнительный анализ дошкольников и младших школьников проводится по одним и тем же компонентам. В целях избежания дублирования было принято решение объединить эти работы в одну группу.

Группа 2 включает аннотации 7 работ, посвященных культурной конгруэнтности в подростковом возрасте, ее связи с креативностью, дивергентным мышлением, личностными свойствами, а также ее обусловленности экопсихологическими и социальными условиями.

Группа 3 не содержит публикаций, так как в представленном наборе отсутствуют исследования, посвященные исследованию культурной конгруэнтности взрослых.

Группа 4 содержит научные публикации, которые не вошли ни в один предыдущий кластер (группу).

Таким образом, анализ рассмотренных в настоящей работе публикаций с помощью цифровых технологий показал, что тема культурной конгруэнтности весьма актуальна в психологической науке и изучается достаточно активно, однако на сегодняшний день отсутствуют исследования, посвященные культурной конгруэнтности взрослых, что составляет перспективу настоящего исследования.

ЗАКЛЮЧЕНИЕ

Анализ современного научного дискурса с использованием методов цифрового моделирования и тематического анализа выявил высокую актуальность исследования культурной конгруэнтности в психологическом контексте, особенно в отношении детей и подростков.

Выделены четыре основные тематические группы, отражающие ключевые аспекты проявления культурной конгруэнтности: ее связь с когнитивным, эмоциональным и личностным развитием, регуляторными функциями, стилями мышления и влиянием на ее уровень выраженности социальных и экологических факторов.

Установлено, что большинство исследований посвящено дошкольникам, младшим школьникам и подросткам, что свидетельствует о достаточной изученности культурной конгруэнтности на ранних этапах онтогенеза как в теоретическом, так и в эмпирическом плане.

В ходе анализа выявлен существенный научный пробел - отсутствие исследований, в которых взрослые рассматриваются как отдельная возрастная группа в контексте культурной конгруэнтности, что ограничивает понимание ее динамики на протяжении всей жизни.

Использование цифровых технологий, в частности методов тематического моделирования, основанных на алгоритмах искусственного интеллекта, показало высокую результативность в систематизации и наглядном представлении тематического поля исследования, что способствует повышению объективности и обеспечивает воспроизводимость получаемых выводов.

Полученные данные подчеркивают необходимость дальнейших исследований в области культурной конгруэнтности взрослых, а также свидетельствуют

о перспективности применения цифровых методов для анализа сложных психологических конструктов.

СПИСОК ЛИТЕРАТУРЫ

1. *Кулакова Е.Н., Настаушева Т.Л., Кондратьева И.В*. Систематическое обзорное исследование литературы по методологии scoping review // Вопросы современной педиатрии. 2021. Т. 20. № 3. С. 210–222.
https://doi.org/10.15690/vsp.v20i3/2271

2. *Волкова Н.В., Бордунос А.К., Чикер В.А., Почебут Л.Г., Кораблева С.А.* Цифровое моделирование тематического поля изучения социального капитала поколений в организациях // Социальная психология и общество. 2025. Т. 16. № 1. С. 5–27. URL: https://psyjournals.ru/journals/sps/archive/2025_n1/Volkova_et_al

3. eLIBRARY.ru. URL: https://www.elibrary.ru/defaultx.asp

4. Scopus. URL: https://www.elsevier.com

5. WoS. URL: https://clarivate.com

6. *Баянова Л.Ф., Ганиева А.М.* Креативность и культурная конгруэнтность подростков // Национальный психологический журнал. 2023. Т. 18, № 4. С. 16–24. https://doi.org/10.11621/npj.2023.0402

7. Python. URL: https://www.python.org/

8. *Kobayashi V.B., Mol S.T., Berkers H.A., Kismihók G., Den Hartog D.N.* Text Mining in Organizational Research // Organizational Research Methods, 2018. Vol. 21, No. 3. P. 733—765.

9. *Hagen L.* Content analysis of e-petitions with topic modeling: How to train and evaluate LDA models? // Information Processing & Management. 2018. Vol. 54. No. 6. P. 1292–1307.

10. *Chauhan U., Shah A.* Topic Modeling Using Latent Dirichlet allocation: A Survey // ACM Comput. Surv. 2022. Vol. 54. No 7. P. 1–35.

11. Gigachat. URL: https://giga.chat/

12. Qwen. URL: https://chat.qwen.ai/

13. Deepseek. URL: https://www.deepseek.com

DIGITAL MODELING FOR SCOPING REVIEW IN STUDYING INTERGENERATIONAL CULTURAL CONGRUENCE

A. M. Ganieva[0000-0003-1323-9363]

Kazan (Volga Region) Federal University, Kazan, Russia

ganieva.aisylu@mail

Abstract

The aim of the work is to identify key topics in modern psychological research of cultural congruence using the method of thematic digital modeling of an array of scientific publications.

The modernity and significance of the conducted research is due to the growing importance of cultural congruence in the context of the digital transformation of society, which is changing the ways of socialization and interaction. Modern technologies require rethinking the psychological mechanisms of individual adaptation to the cultural environment, especially in childhood and adolescence. Despite the active study of this phenomenon, there is a noticeable shortage of research on the cultural congruence of adults. The use of digital modeling and artificial intelligence allows us to systematize knowledge and identify the structure of the thematic field with high accuracy. The obtained data opens up the prospect for further study of cultural congruence throughout the entire life cycle.

The thematic field review of cultural congruence research was conducted based on an analysis of digital archives comprising a curated collection of 112 scholarly publications on the topic. The review employed a topic modeling algorithm implemented in the Python programming language and leveraged digital platforms incorporating multimodal neural network–based tools (GigaChat, Qwen, DeepSeek). The data analysis yielded four distinct age groups that reflect the developmental specificity of cultural congruence manifestations: preschoolers, primary school–age children, adolescents, and adults.

Keywords*: cultural congruence, psychological research, developmental psychology, topic modeling, scoping review.*

REFERENCES

1. *Kulakova E.N., Nastausheva T.L., Kondratyeva I.V.* Sistemnoye obzornoye issledovaniye literatury po metodologii scoping review [Systematic literature review on the methodology of scoping review] // Voprosy sovremennoy pediatrii [Current Pediatrics Issues]. 2021. Vol. 20, No. 3. P. 210–222. https://doi.org/10.15690/vsp.v20i3/2271

2. *Volkova N.V., Bordunos A.K., Chiker V.A., Pochepbut L.G., Korableva S.A.* Tsifrovoye modelirovaniye tematicheskogo polya izucheniya sotsial'nogo kapitala pokoleniy v organizatsiyakh [Digital modeling of the thematic field of studying intergenerational social capital in organizations] // Sotsial'naya psikhologiya i obshchestvo [Social Psychology and Society]. 2025. Vol. 16, No. 1. P. 5–27. URL: https://psyjournals.ru/journals/sps/archive/2025_n1/Volkova_et_al

3. eLIBRARY.ru. URL: https://www.elibrary.ru/defaultx.asp

4. Scopus. URL: https://www.elsevier.com

5. WoS. URL: https://clarivate.com

6. *Bayanova L.F., Ganieva A.M.* Kreativnost' i kul'turnaya kongruentnost' podrostkov [Creativity and cultural congruence of adolescents] // Natsional'nyy psikholog-icheskiy zhurnal [National Psychological Journal]. 2023. Vol. 18, No. 4. P. 16–24. https://doi.org/10.11621/npj.2023.0402

7. Python. URL: https://www.python.org/

8. *Kobayashi V.B., Mol S.T., Berkers H.A., Kismihók G., Den Hartog D.N.* Text Mining in Organizational Research // Organizational Research Methods. 2018. Vol. 21, No. 3. P. 733–765.

9. *Hagen L.* Content Analysis of E-Petitions with Topic Modeling: How to Train and Evaluate LDA Models? // Information Processing & Management. 2018. Vol. 54, No. 6. P. 1292–1307.

10. *Chauhan U., Shah A.* Topic Modeling Using Latent Dirichlet Allocation: A Sur-vey // ACM Computing Surveys. 2022. Vol. 54, No. 7. P. 1–35.

11. Gigachat. URL: https://giga.chat/

12. Qwen. URL: https://chat.qwen.ai/

13. Deepseek. URL: https://www.deepseek.com

СВЕДЕНИЯ ОБ АВТОРЕ

ГАНИЕВА Айсылу Мунавировна – закончила Казанский (Приволжский) федеральный университет. В настоящее время является ассистентом кафедры педагогической психологии Института психологии и образования Казанского (Приволжского) федерального университета и педагогом-психологом I категории.

Область научных интересов: психология, возрастная психология, общая психология.

Aisylu Munavirovna GANIEVA – graduated from Kazan (Volga Region) Federal University. She currently works as an assistant lecturer at the Department of Pedagogical Psychology of the Institute of Psychology and Education of Kazan (Volga Region) Federal University and as Educational Psychologist, Grade I.

Research interests: psychology, developmental psychology, general psychology.

email: ganieva.aisylu@mai.ru

ORCID: 0000-0003-1323-9363

Материал поступил в редакцию 14 октября 2025 года

УДК 81'322.2+004.855.5

АВТОМАТИЧЕСКОЕ ИЗВЛЕЧЕНИЕ АРГУМЕНТАТИВНЫХ ОТНОШЕНИЙ ИЗ ТЕКСТОВ НАУЧНОЙ КОММУНИКАЦИИ

Ю. А. Загорулько[1] [0000-0002-7111-6524], Е. А. Сидорова[2] [0000-0001-8731-3058], И. Р. Ахмадеева[3] [0000-0002-7371-1087]

[1–3]Институт систем информатики им. А.П. Ершова СО РАН, г. Новосибирск, Россия

[1]zagor@iis.nsk.su, [2]lsidorova @iis.nsk.su , [3]i.r.akhmadeeva@iis.nsk.su

Аннотация

Сложность задачи извлечения аргументативных структур связана с такими проблемами, как выделение аргументативных сегментов, прогнозирование дальних связей между неконтактными сегментами, обучение на данных, размеченных с низкой степенью согласованности между аннотаторами. В настоящей работе рассмотрен подход к извлечению аргументативных отношений из достаточно больших текстов, относящихся к области научной коммуникации. Проведен сравнительный анализ методов тонкой настройки с использованием предобученной языковой модели типа Longformer, позволяющей учитывать длинные контексты, и двух методов, позволяющих учитывать расхождения аннотаторов в разметке аргументов за счет использования так называемых мягких меток, полученных путем равномерного сглаживания меток и усреднения экспертных оценок. Эксперименты проводились на четырех наборах данных, содержащих положительные и отрицательные примеры пар утверждений (посылка, заключение) и различающихся способами сегментации и средним размером текста. Наилучшие результаты получены на модели с усреднением экспертных оценок. В то же время отмечено, что модель, использующая сглаженные метки, также повышает точность классификаторов, но ухудшает полноту.

***Ключевые слова**: анализ аргументации, извлечение аргументативных отношений, научная коммуникация, проблемы сегментации, мягкая метка, сглаживание меток, языковая модель.*

© Ю. А. Загорулько, Е. А. Сидорова, И. Р. Ахмадеева, 2025.
Данная статья распространяется на условиях международной лицензии Creative Commons License Attribution 4.0 International (CC BY 4.0).

ВВЕДЕНИЕ

Аргументация является одной из конституирующих составляющих научной коммуникации (коммуникации на научные темы), поскольку содержанием научных и научно-популярных текстов является научно обоснованное знание. В научной сфере общения автор должен убедить в правомерности своих идей коллег-ученых, а в научно-популярной – широкую аудиторию.

При анализе аргументации, представленной в тексте, требуется не только извлекать аргументы и цепочки аргументов, подтверждающие или опровергающие некий тезис (абстрактная аргументация), но и исследовать структуру каждого аргумента, ее роль и значимость для всей аргументации в целом (структурная аргументация). Можно выделить два подхода к решению задачи автоматического извлечения структурированной аргументации:

— первый предполагает выполнение последовательности следующих шагов: сегментация текста → аргумент/не аргумент → простая структура аргумента → уточненная структура аргумента;

— второй подход, получивший название «все в одном» (all-in-one) включает всего два шага: сегментация текста → уточненная структура аргумента; он сразу решает задачу связанности выделенных сегментов аргументативными отношениями, и уже на этом основании решается вопрос об аргументативности утверждений и их типе.

Целью настоящей работы было экспериментальное исследование подхода all-in-one на основе нейросетевых моделей. Эксперименты проводились на двух русскоязычных корпусах текстов, относящихся к области научной коммуникации.

1. ОБЗОР РАБОТ

Для задачи интеллектуального анализа аргументации, как и для многих задач анализа естественного языка, характерен высокий уровень неоднозначности в размеченных обучающих данных. От такой неоднозначности сложно избавиться, поскольку она является неотъемлемым свойством самого языка. Поэтому в последнее время появились работы (см., например [1]), указывающие на необходимость учитывать неоднозначность при решении задач понимания естественного языка.

Одним из способов учета такой неоднозначности является сглаживание меток (label smoothing) – техника, широко используемая в глубоком обучении, которая, как было показано, улучшает качество работы модели при обучении на шумных данных [2] и предотвращает излишнюю самоуверенность (overconfidence) модели в ответе [3]. Эта техника вместо стандартного обучения с использованием прямого кодирования (one-hot encoding) предполагает использование сглаженных меток, полученных путем подмешивания равномерного вектора в исходный вектор меток.

Впервые техника сглаживания меток была предложена для решения задачи классификации изображений [4], но впоследствии была продемонстрирована ее применимость для задач анализа текста. Так, в работе [5] сглаживание меток применялось для калибровки модели при решении задачи автоматического определения логической связи между текстами (NLI), для которой характерна большая неоднозначность в разметке. Другой способ сглаживания меток заключается в усреднении меток, полученных от разных аннотаторов.

В настоящей работе исследована возможность применения техники сглаживания меток в новой области – интеллектуальном анализе аргументации.

2. КОРПУС ТЕКСТОВ С АРГУМЕНТАТИВНОЙ РАЗМЕТКОЙ

Используем два русскоязычных аннотированных корпуса текстов, представленных на платформе ArgNetBank Studio (https://uniserv.iis.nsk.su/arg). Научно-популярный корпус включает тексты двух подстилей, характеризующихся как тематической неоднородностью, так и жанровым разнообразием: научные новости и статьи с сайта habr.com/ru (habr-статьи). Корпус научной коммуникации [6] включает краткие научные статьи, научные обзоры и полноформатные научные статьи с комментариями рецензентов. В целом процесс разметки существенно зависит от жанра.

Короткие научные статьи, как правило, имеют один главный тезис, который достаточно легко выявить по позиции (два последних абзаца текста), и не имеют «длинных» связей. Разметка таких статей облегчается благодаря наличию в них элементов содержательной структуры (актуальность, цель работы, новизна) и формально-логической жанровой структуры (введение, заключение, обзор, методы, эксперимент, выводы).

В **рецензиях** главный тезис также находится в конце и сводится к трем вариантам: принять рецензируемую статью к публикации, принять после доработки или отклонить. Аргументативные отношения образуют кустовую структуру, аргументами первого порядка являются достоинства и/или недостатки статьи.

Главный тезис **научных новостей**, напротив, располагается в начале (лиде) новостной заметки; для заметки характерны повторы разной степени общности, имеется четкая жанровая структура с выделением хэдлайна, лида и бэкграунда.

В **статьях с комментариями рецензентов,** с одной стороны, как в обычной научной статье, можно опираться на цель, выделив главный тезис (один или несколько), с другой – комментарии образуют компонентный диалог, представляющий собой обмен стимулирующими и реагирующими репликами.

Habr-статьи могут вообще не иметь главного тезиса. Здесь можно опираться на содержательные разделы, а комментарии к текстам привносят в изложение черты каскадного диалога и полилога.

На основе анализа жанровых особенностей текстов было выделено два подкорпуса:

S-корпус – подкорпус коротких научных статей (в среднем 1053 словоупотребления);

L-корпус – подкорпус длинных (в среднем 3500 слов) текстов научной коммуникации, в котором собраны полноразмерные научные статьи, рецензии, аналитические научно-популярные статьи и новости о событиях в науке.

3. ПОДГОТОВКА НАБОРА ДАННЫХ

Для применения методов машинного обучения к задаче извлечения аргументативных связей необходимо:

а) создать на базе аннотированных текстов наборы данных, содержащие положительные и отрицательные примеры аргументативных отношений;

б) разработать механизм предварительного построения гипотез по заданному тексту, т. е. определить, каким образом текст будет разбиваться на утверждения и какие пары утверждений будут проверяться нашей моделью.

3.1. ПРОБЛЕМА СЕГМЕНТАЦИИ

Аргументативный анализ начинается с сегментации, т. е. разделения текста на осмысленные с точки зрения аргументации фрагменты (argumentative discourse units, ADUs). В общем случае это утверждения, в основе которых лежат пропозиции. Однако анализ сегментации, выполненной аннотаторами-людьми, показывает, что сегментами могут быть как более мелкие, так и более крупные фрагменты текста. При этом между аннотаторами далеко не всегда наблюдается согласие в отношении выполнения сегментации. В связи с этим трудно говорить о «золотом стандарте» в данном вопросе.

(S1) ChatGPT стал вторым чат-ботом, прошедшим широко известный Тест Тьюринга. (S2) Это значит, что во взаимодействии с ним судейской коллегии было невозможно определить общаются ли они с человеком или программой. (S3.1) Вдохновленные таким несомненным успехом, (S3.2) а также свободным доступом к боту предоставленном в OpenAI, (S3.3) многочисленные «уверовавшие в ИИ» начали наперебой предлагать приткнуть бота во все возможные ниши: от программирования до медицинских диагнозов. (S4) Даже поисковые системы забили тревогу в ожидании того, что бот подвинет их в предложении услуг поиска информации. (S5) На самом деле все эти ожидания не имеют под собой абсолютно никаких оснований. (S6) Ниже проиллюстрирую этот факт на конкретных примерах.

Рис. 1. Фрагмент текста из habr-статьи

В табл. 1 представлено распределение единиц рассуждения (ADUs), соответствующее аргументативно насыщенному абзацу из habr-статьи (см. рис. 1), который включили в анализ все аннотаторы (в эксперименте участвовали четыре аннотатора, обозначенных как A-1, A-2, A-3, A-4).

Табл. 1. Различия в сегментации фрагмента текста 4 аннотаторами

A-1	S1	S2	S3.1	S3.2	S3.3	S4		S5	
A-2	S1		S3.1	S3.2	S3.3	S4.1	S4.2	S5-S6	
A-3	S1	S2	S3					S5	S6
A-4	S1	S2	S3			S4		S5	

При этом A-2 исключил(а) из анализа второе предложение (S2), посчитав, что оно является не аргументирующим, а объяснительным, и предназначенным

для тех читателей, которые не знают, что такое тест Тьюринга. Но большинство аннотаторов посчитали S2 важным, обосновывающим S1 в качестве причины (A-4) или знака (A-1 и A-3). A-1 и A-4 удалили из анализа S6, возможно, сочли его метаязыковым, т. е. аргументативно незначимым. Последнее подтверждается включением S6 в комплекс с S5 аннотатором A-2, который посчитал(а) S6 индикатором примеров, представленных в тексте далее. Первые два аннотатора разбили S3, выделив самостоятельные сегменты S3.1 и S3.2 как основания для сегмента S3.3. S4 исключен из анализа аннотатором A-3, а A-2 разбил его на две клаузы, которые использованы в качестве поддерживающих посылок к S3.3.

Данный пример показывает, что, осуществляя аргументативную разметку, человек опирается на формальное выделение сегментов, но не ограничивается им, т. к. этот процесс скорее идет в ногу с процессом рассуждений относительно наличия и типа аргументативной связи. Усредненная оценка согласия между аннотаторами на данном фрагменте текста составляет 57.3%.

Оценка согласия сегментаций вычислялась в соответствии с алгоритмом, предложенным в работе [7]. Этот алгоритм основан на мере сходства множеств и дополнительно учитывает случаи, когда один сегмент пересекается с несколькими другими. Среднее значение согласия сегментаций на всем корпусе равно 61.2%, а коэффициенты каппа Коэна (Cohen’s k) и альфа Криппендорфа (Krippendorff’s α) равны 0.42 и 0.43 соответственно.

Для автоматической сегментации текста помимо базового подхода— сегментации на предложения — использовались еще два подхода для более тонкой сегментации:

Q1: выделение клауз на основе синтаксического дерева предложения. Выделялись глагольные группы, причастные и деепричастные обороты;

Q2: выделение дискурсивных единиц на основе риторического анализа текста.

Сравнение работы сегментаторов с ручной сегментацией выявило существенные различия. Так, процент совпадения сегментов (на фрагментах с аргументацией) составил 49.5% и 46.75% для Q1 и Q2 соответственно. Кроме того, Q1 показал лучшие результаты на S-корпусе (56.2% против 44.9%), а Q2 — на L-корпусе (47% против 46.2%).

Анализ работы сегментаторов выявил следующие типичные ошибки: некорректная обработка разделителей предложений (точки или их отсутствие), неточное обращение с прямой речью (например, прямая речь со вставленной речью говорящего или косвенная речь с предшествующей речью говорящего), ошибочное разделение на клаузы при наличии свернутого предложения, представленного субстантивным предикатом, неверная идентификация разрывных сегментов и т. д.

3.2. ПОСТРОЕНИЕ ПАР УТВЕРЖДЕНИЙ ДЛЯ КЛАССИФИКАЦИИ

Для моделирования процесса построения гипотез все множество аргументативных связей было разделено на ближние (связь между двумя контактными сегментами) и дальние, соотношение которых в корпусе составило примерно 2 к 3. Для предсказания ближних связей рассматривались пары утверждений, получаемые с помощью скользящего окна. В качестве окна брались два подряд идущих сегмента (предложение и/или клауза). Для построения гипотез дальних связей механизм окна не подходил, поэтому были исследованы различные варианты взаиморасположения утверждений, относящихся к одному аргументу (посылка и заключение), друг относительно друга. В итоге для дальних связей отбирались утверждения, расположенные в границах одного предложения и располагающиеся либо в одном, либо в соседних абзацах (гипотеза компактности), что составило 52.9% от общего количества размеченных аннотаторами аргументативных связей.

Для обучения классификаторов было построено четыре набора данных, которые содержат положительные и отрицательные примеры пар утверждений (посылка, заключение). Пара считается положительным примером, если существует аргументативная связь между посылкой и заключением хотя бы в одной экспертной разметке. Отрицательные примеры отбирались следующим образом: для каждой позитивной пары утверждений генерировалась пара, в которой утверждения находятся в тех же абзацах (или в соседних, при отсутствии подходящих утверждений) и между которыми отсутствует путь от посылки к заключению во всех графах аргументации, соответствующих этому тексту.

На основе S-корпуса был создан один датасет, в котором утверждениями выступают предложения. На основе L-корпуса были созданы три датасета: в первом утверждениями также являются предложения, а в двух других — клаузы, полученные двумя различными методами сегментации (Q1 и Q2).

Каждому примеру были сопоставлены два типа меток: жесткие (one-hot encoding) и мягкие (soft labels). Применялись два способа получения мягких меток: а) равномерное сглаживание (label-smoothing), и б) усреднение (label-averaging). В случае равномерного сглаживания бинарные «жесткие» метки 0 и 1 заменялись на значения p и $1-p$ соответственно, где настраиваемый гиперпараметр p (коэффициент сглаживания меток) был подобран равным 0.1. При усреднении меток вес каждого примера определялся пропорционально доле аннотаторов, посчитавших данную пару утверждений связанной аргументативным отношением. Если в процессе усреднения для определенного текста существовала только одна разметка, то к полученным меткам применялся первый способ — равномерное сглаживание.

При построении тестовых датасетов в качестве положительных примеров отбирались только такие пары утверждений, которые посчитали аргументативными два и более аннотаторов. Пары утверждений, которые один из аннотаторов посчитал аргументативно связанными, а другой — нет, считались неоднозначными и в построении тестовых датасетов не участвовали.

4. ЭКСПЕРИМЕНТАЛЬНОЕ ИССЛЕДОВАНИЕ

В эксперименте использовался подход к решению задачи извлечения аргументативных отношений на основе тонкой настройки языковой модели (Fine Tuning). Задача извлечения аргументации all-in-one была сведена к предсказанию наличия или отсутствия связи между двумя утверждениями (предложениями, их частями или группами предложений). Другими словами, модель должна была решать задачу бинарной классификации: положительный ответ — связь есть, отрицательный ответ — связь отсутствует.

Предыдущие эксперименты показали, что в задаче предсказания аргументативной связи между утверждениями существенную роль играет контекст, в которых эти утверждения встречаются [8]. Поэтому были рассмотрены модели типа

Longformer [9], способные обрабатывать длинные фрагменты текста. По результатам предварительных экспериментов модель *kazzand/ru-longformer-large-4096* была выбрана для построения векторных представлений пар утверждений с учетом контекста их появления. Векторные представления утверждений строились путем усреднения векторов соответствующих токенов, а в качестве векторного представления контекста использовался вектор специального токена *"[CLS]".* Конкатенация этих трех векторов подавалась в полносвязный слой для классификации на два класса: наличие или отсутствие аргументативного отношения.

В табл. 2 представлены результаты, полученные тремя моделями:

1) base model: базовая модель, обученная на жестких метках;
2) model-LS: модель, обученная на мягких метках, полученных из жестких меток с использованием сглаживания меток;
3) model-AA: модель, обученная на мягких метках, полученных путем усреднения экспертных оценок.

Все модели были обучены в течение 15 эпох со скоростью обучения $1e^{-5}$, коэффициент сглаживания меток для model-LS составил 0.1.

Табл. 2. Результаты предсказания наличия аргументативной связи

	S-корпус (предложения)			L-корпус (предложения)			L-корпус (клаузы-Q1)			L-корпус (клаузы-Q2)		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
base model	72.9	69.6	71.2	67.5	81	**73.6**	74.7	72.1	**73.3**	67.8	67.8	67.8
model-LS	73.5	80	76.6	72.6	60.5	66	75.9	65.9	70.6	68.4	57.8	62.7
model-AA	77.2	77.8	**77.5**	72.1	65.2	68.5	72.9	66.8	69.7	67.3	69.5	**68.4**

Как можно видеть, использование мягких меток повышает точность классификаторов, но для L-корпуса показатель полноты снижается. Это может быть связано с тем, что L-корпус характеризуется большим количеством текстов, размеченных только одним аннотатором. Для S-корпуса наилучшие результаты дала model-AA.

Более высокие показатели получены на S-корпусе, что может объясняться его гомогенностью, небольшой длиной текстов, относительной простотой аргументации и использованием аннотаторами единой методики разметки, что повысило уровень согласия между ними. Так, согласие аннотаторов на S-корпусе составило 64.2% для утверждений и 44.2% для связей, тогда как на L-корпусе – 58.9% и 33.2% соответственно.

ЗАКЛЮЧЕНИЕ

Мы рассмотрели подход к извлечению аргументативных отношений из русскоязычных текстов, обладающих достаточно сложной (для автоматической обработки) организационной структурой, свободным стилем изложения аргументов и достаточно большим объемом. Модель с мягкими метками повышает точность классификаторов, но ухудшает полноту. Наилучшие результаты были получены на модели с усреднением экспертных оценок.

В целом подход на основе мягких меток не показал значительных улучшений качества извлечения аргументации, поэтому дальнейшие работы мы планируем вести в направлении улучшения качества датасетов путем повышения согласия аннотаторов за счет строгого следования методическим рекомендациям и разработки правил унификации разметки.

Благодарности

Исследование выполнено при финансовой поддержке Российского научного фонда № 23-11-00261, https://rscf.ru/project/23-11-00261/.

СПИСОК ЛИТЕРАТУРЫ

1. *Meissner J.M., Thumwanit N., Sugawara S., Aizawa A.* Embracing Ambiguity: Shifting the Training Target of NLI Models // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Online, August 2021. Association for Computational Linguistics: Vol. 2: Short Papers, P. 862–869.
https://doi.org/10.18653/v1/2021.acl-short.109

2. *Lukasik M., Bhojanapalli S., Menon A., Kumar S.* Does label smoothing mitigate label noise? // Proceedings of the 37th International Conference on Machine Learning, Virtual, 13–18 July 2020, Vol. 119, P. 6448–6458.
URL: https://proceedings.mlr.press/v119/lukasik20a.html

3. *Haque S., Bansal A., McMillan C.* Label smoothing improves neural source code summarization // 2023 IEEE/ACM 31st International Conference on Program Comprehension (ICPC), Melbourne, Australia, 15–16 May 2023. Institute of Electrical and Electronics Engineers: 2023, P. 101–112.
https://doi.org/10.1109/ICPC58990.2023.00025

4. *Szegedy C., Vanhoucke V., Ioffe S., Shlens J., Wojna Z.* Rethinking the Inception Architecture for Computer Vision // *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 27–30 June, 2016. Institute of Electrical and Electronics Engineers: P. 2818–2826. https://doi.org/10.1109/CVPR.2016.308

5. *Wang Y., Wang M., Chen Y., Tao S., Guo J., Su C., Zhang M., Yang H*. Capture Human Disagreement Distributions by Calibrated Networks for Natural Language Inference // Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland. May 2022. Association for Computational Linguistics: 2022, P. 1524–1535. https://doi.org/10.18653/v1/2022.findings-acl.120

6. *Тимофеева М.К., Ильина Д.В., Кононенко И.С.* Аргументативная разметка корпуса текстов научной интернет-коммуникации: жанровый анализ и исследование типовых моделей рассуждения с помощью платформы ArgNetBank Studio // Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. 2024. Т. 22, №1. С. 27–49. https://doi.org/10.25205/1818-7935-2024-22-1-27-49

7. *Shestakov V.K., Kononenko I.S., Sidorova E.A., Zagorulko Yu.A.* Assessing Inter-Annotator Agreement on Argumentative Markup // 2024 IEEE International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON). IEEE: 2024, P. 309–313. https://doi.org/10.1109/SIBIRCON63777.2024.10758535

8. *Akhmadeeva I., Sidorova E., Ilina D.* Argument mining in scientific communication: Comparative study // Internet and modern society. Human-computer communication. Cham: Springer Nature Switzerland, 2026. P. 152–166.
https://doi.org/10.1007/978-3-031-96177-9_13

9. *Beltagy I., Peters M. E., Cohan A.* Longformer: The long-document transformer //arXiv preprint arXiv:2004.05150. 2020.

AUTOMATIC EXTRACTION OF ARGUMENTATIVE RELATIONS FROM SCIENTIFIC COMMUNICATION TEXTS

Yu. A. Zagorulko[1] [0000-0002-7111-6524], E. A. Sidorova[2] [0000-0001-8731-3058],
I. R. Akhmadeeva[3] [0000-0002-7371-1087]

[1–3]*A.P. Ershov Institute of Informatics Systems of Siberian Branch of RAS, Novosibirsk, Russia*

[1]zagor@iis.nsk.su, [2]lsidorova @iis.nsk.su , [3]i.r.akhmadeeva@iis.nsk.su

Abstract

The complexity of the problem of extracting argumentative structures is associated with such problems as selecting argumentative segments, predicting long-range connections between non-contact segments, and training on data labeled with a low degree of inter-annotator consistency. In this paper, we consider an approach to extracting argumentative relations from fairly large texts related to scientific communication. A comparative analysis was performed of fine-tuning methods using a pretrained Longformer-type language model that takes into account long contexts and two methods that take into account annotator discrepancies in argument labeling by using the so-called soft labels obtained by uniformly smoothing labels and averaging expert assessments. The experiments were conducted on four datasets containing positive and negative examples of statement pairs (premise, conclusion) and differing in segmentation methods and average text size. The best results were obtained using the model with averaging expert assessments. At the same time, it is noted that the model using smoothed labels also increases the accuracy of classifiers, but worsens the recall.

Keywords*: argument mining, argumentative relation extraction, scientific communication, segmentation problem, soft label, label smoothing, language model.*

REFERENCES

1. *Meissner J.M., Thumwanit N., Sugawara S., Aizawa A.* Embracing Ambiguity: Shifting the Training Target of NLI Models // Proceedings of the 59th Annual Meeting

of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Online, August 2021. Association for Computational Linguistics: Vol. 2: Short Papers, P. 862–869.
https://doi.org/10.18653/v1/2021.acl-short.109

2. *Lukasik M., Bhojanapalli S., Menon A., Kumar S.* Does label smoothing mitigate label noise? // Proceedings of the 37th International Conference on Machine Learning, Virtual, 13–18 July 2020, Vol. 119, P. 6448–6458.
URL: https://proceedings.mlr.press/v119/lukasik20a.html

3. *Haque S., Bansal A., McMillan C.* Label smoothing improves neural source code summarization // 2023 IEEE/ACM 31st International Conference on Program Comprehension (ICPC), Melbourne, Australia, 15–16 May 2023. Institute of Electrical and Electronics Engineers: 2023. P. 101–112.
https://doi.org/10.1109/ICPC58990.2023.00025

4. *Szegedy C., Vanhoucke V., Ioffe S., Shlens J., Wojna Z.* Rethinking the Inception Architecture for Computer Vision // *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 27–30 June, 2016. Institute of Electrical and Electronics Engineers: P. 2818–2826. https://doi.org/10.1109/CVPR.2016.308

5. *Wang Y., Wang M., Chen Y., Tao S., Guo J., Su C., Zhang M., Yang H.* Capture Human Disagreement Distributions by Calibrated Networks for Natural Language Inference // Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland. May 2022. Association for Computational Linguistics: 2022, P. 1524–1535. https://doi.org/10.18653/v1/2022.findings-acl.120

6. *Timofeeva M.K., Ilina D.V., Kononenko I.S.* Argumentative Annotation of the Scientific Internet-Communication Corpus: Genre Analysis and Study of Typical Reasoning Models based on the ArgNetBank Studio Platform // NSU Vestnik. Series: Linguistics and Intercultural Communication. 2024. Vol. 22, No. 1. P. 27–49. (In Russ.) https://doi.org/10.25205/1818-7935-2024-22-1-27-49

7. *Shestakov V.K., Kononenko I.S., Sidorova E.A., Zagorulko Yu.A.* Assessing Inter-Annotator Agreement on Argumentative Markup // 2024 IEEE International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON). IEEE: 2024, P. 309–313. https://doi.org/10.1109/SIBIRCON63777.2024.10758535

8. *Akhmadeeva I., Sidorova E., Ilina D.* Argument mining in scientific communication: Comparative study // Internet and modern society. Human-computer communication. Cham: Springer Nature Switzerland, 2026. P. 152–166. https://doi.org/10.1007/978-3-031-96177-9_13

9. *Beltagy I., Peters M. E., Cohan A.* Longformer: The long-document transformer //arXiv preprint arXiv:2004.05150. 2020.

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 290 публикаций в области искусственного интеллекта, разработки интеллектуальных систем, инженерии знаний, онтологического моделирования и компьютерной лингвистики.

Yury Alekseevich ZAGORULKO – Cand. Sc. (Technology), Head of Laboratory at the A. P. Ershov Institute of Informatics Systems of Siberian Branch of RAS, an associate professor at Novosibirsk State University. List of scientific works includes more than 290 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

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, an associate professor at Novosibirsk State University. List of scientific works includes more than 160 peer-reviewed publications in the fields of Computational Linguistics, Intelligent System Development, Knowledge and Ontology Engineering.

email: lsidorova@iis.nsk.su

ORCID: 0000-0001-8731-3058

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, an assistant lecturer at the Department of Programming 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

Материал поступил в редакцию 15 октября 2025 года

УДК 004.89

НЕЙРОСИМВОЛИЧЕСКИЙ ПОДХОД К ДОПОЛНЕННОЙ ГЕНЕРАЦИИ ТЕКСТА НА ОСНОВЕ АВТОМАТИЗИРОВАННОЙ ИНДУКЦИИ МОРФОТАКТИЧЕСКИХ ПРАВИЛ

М. В. Исангулов[1] [0009-0006-3244-0328], А. М. Елизаров[2] [0000-0003-2546-6897], А. Р. Кунафин[3] [0009-0006-0495-265X], А. Р. Гатиатуллин[4] [0000-0003-3063-8147], Н. А. Прокопьев[5] [0000-0003-0066-7465]

[1, 2]Казанский (Приволжский) федерального университет, г. Казань, Россия
[3]Независимый исследователь
[4, 5]Академия наук Республики Татарстан, *г. Казань, Россия*

[1]marathon.our@gmail.com, [2]amelizarov@gmail.com, [3]aigizk@gmail.com, [4]ayrat.gatiatullin@gmail.com, [5]nikolai.prokopyev@gmail.com

Аннотация

Представлен гибридный нейросимволический метод, который объединяет большую языковую модель (LLM) и конечный автомат (FST) для обеспечения морфологической корректности при генерации текста на агглютинативных языках.

Система автоматически извлекает правила из корпусных данных: для локальных примеров словоформ LLM формирует цепочки морфологического разбора, которые затем агрегируются и упорядочиваются в компактные описания правил морфотактики (LEXC) и выбора алломорфов (regex). На этапе генерации LLM и FST работают совместно: если токен не распознается автоматом, LLM извлекает из контекста пару «лемма + теги», а FST реализует корректную поверхностную форму. В качестве набора данных использован корпус художественной литературы (~1600 предложений). Для списка из 50 существительных извлечено 250 словоформ. По предложенному алгоритму LLM сгенерировала 110 контекстных regex-правил вместе с LEXC-морфотактикой, на основе чего был скомпилирован FST, распознавший 170/250 форм (~70%). В прикладном тесте машинного перевода на подкорпусе из 300 предложений интеграция данного FST в цикл LLM повысила качество с BLEU 16.14 / ChrF 45.13 до BLEU 25.71 / ChrF 50.87 без дообучения переводчика. Подход применим к иным частям речи и другим

© М. В. Исангулов, А. М. Елизаров, А. Р. Кунафин, А. Р. Гатиатуллин, Н. А. Прокопьев, 2025.
Данная статья распространяется на условиях международной лицензии Creative Commons License Attribution 4.0 International (CC BY 4.0).

агглютинативным и малоресурсным языкам, где он может быть использован для наполнения словарных и грамматических ресурсов.

***Ключевые слова**: нейросимволический подход, большая языковая модель, конечные автоматы, двухуровневая морфология, LEXC морфотактика, машинный перевод, агглютинативные языки, башкирский язык.*

ВВЕДЕНИЕ

Морфологически сложные агглютинативные языки до сих пор остаются сложной областью для современных больших языковых моделей (LLM). Один корень может реализовывать десятки поверхностных форм за счет суффиксальной агглютинации, гармонии гласных, чередований и контекстных модификаций морфем. В условиях дефицита данных и неточной подсловной токенизации (BPE, unigram LM) модели часто не выделяют морфологически осмысленные сегменты; редкие суффиксальные цепочки нередко встречаются в обучении единично. Дополнительно агглютинативные языки слабо представлены в мультиязычных корпусах, что затрудняет обобщение поверхностных форм. В результате LLM порождают несуществующие словоформы, смешивают алломорфы и нарушают порядок морфем, что является проявлением «морфологического разрыва», когда смысл сохраняется, а форма деградирует [1]. На богатых ресурсами языках нейронные модели словоизменения (seq2seq/ трансформеры) справляются значительно лучше [2], однако их устойчивость на малоресурсных и агглютинативных системах остается ограниченной [3], что мотивирует гибридные решения.

Формальная линия работ по конечным автоматам восходит к двухуровневой морфологии Коскенниеми и инструментам Xerox [4, 5]. Конечные автоматы (FST) позволяют строго задать морфотактику (например, LEXC) и детерминированно реализовывать контекстные преобразования поверхностных форм. Такие грамматики теоретически порождают все допустимые формы и исключают недопустимые.

Современные открытые стеки (HFST LexC/TwolC; foma) поддерживают тот же парадигматический подход, включая компиляцию двухуровневых правил и разрешение конфликтов с весами, что делает FST практичными для морфологии

богатых языков. Тем не менее ручная разработка LEXC/правил алломорфии трудоемка и плохо масштабируется на новые части речи и языки. Популярные словарно-ориентированные корректоры орфографии, такие как Hunspell, используют форматы .dic (список слов с флагами) и .aff (аффиксальные директивы) и удобны для правки или проверки, но слабо генерализуются за пределы заданного словаря и не решают контекстный выбор граммем, поскольку опираются на перечисление слов и аффикс-паттернов, а не на полноценную морфологическую грамматику. В экосистеме Apertium доступны морфологические анализаторы и генераторы и shallow-transfer MT; для башкирского языка существует проект apertium-bak (bakmorph), однако типичный рабочий процесс остается лексикографически нагруженным и требует ручной поддержки XML словарей/парадигм.

Имеющиеся «нейро-символьные» стыковки, как правило, решают смежные, но иные задачи. В SGNMT конечные автоматы/решетки подключаются как «predictors» в декодере, то есть ограничивают поиск и добавляют внешние оценки, не гарантируя морфологически корректную поверхность и не выполняя выбор алломорфа в контексте [6]. В коррекции и спелл-чеке FST обычно выступает внешним источником разрешенных форм (аффиксальные словари, грамматики), не будучи интегрированным в сам генеративный цикл [7]. Существуют и двухшаговые NMT-схемы с выходом в формате lemma+tag и последующей детерминистической генерацией поверхности (например, с использованием SMOR для немецкого): такие подходы уменьшают разреженность, но разметка делается вручную, а FST не «учится» из корпуса и не встраивается как онлайн-валидатор/генератор [8]. Наконец, «обратное» направление — бутстрап нейронных анализаторов по готовым FST — демонстрирует рост покрытия и точности по сравнению с исходными автоматами, но не решает задачу индукции самих правил/лексиконов из корпусных свидетельств и их совместной работы с LLM в момент генерации [9]. В результате в текущем дискурсе отсутствуют работы, где а) LLM генерирует переносимую LEXC/regex-грамматику из корпусных примеров, б) этот автомат включен в сам процесс генерации для коррекции словоформ и в) показано заметное улучшение качества вывода на целевых корпусах.

В настоящей работе представлено решение описанных проблем в форме гибридного нейросимволического подхода — комбинации использования LLM и FST.

1. Из небольшой корпусной выборки извлекаются морфотактические описания (LEXC) и компактные контекстные правила (regex) выбора алломорфов: локальные правила, сгенерированные LLM по представленным формам, агрегируются, унифицируются и упорядочиваются от специфичных к обобщенным. Важное отличие от словарных систем (Hunspell и др.) является то, что конкретный список лемм служит лишь источником индукции, а полученные правила обобщаются на леммы вне словаря (например, шаблоны +ACC:+НЫ с контекстными правилами фонологической реализации применимы к любому существительному из класса), что уменьшает ручные затраты и повышает переносимость.

2. На этапе генерации текста компоненты работают в одном контуре: если токен не распознан автоматом, LLM извлекает из контекста пару «лемма + теги», а FST детерминированно реализует корректную поверхностную форму, тем самым сочетаются контекстная уместность и морфологическая валидность.

Эмпирически показаны значимый уровень корректного распознания словоформ, учитываемых в правилах, и прирост качества генерации текста на примере машинного перевода без дообучения модели (BLEU+9.57, ChrF+5.74 на подкорпусе художественного текста), при том, что грамматика остается компактной и объяснимой. Подход может быть применен к другим частям речи и агглютинативным, малоресурсным языкам, он предлагает практический путь к ускоренному пополнению словарных и грамматических ресурсов.

ГЕНЕРАЦИЯ ПРАВИЛ И АВТОМАТИЗИРОВАННЫЙ ПАЙПЛАЙН

Цель этого раздела — последовательно показать, как из небольшого параллельного корпуса автоматически извлекаются и анализируются словоформы, как на их основе синтезируются правила LEXC и regex и компонуется двунаправленный FST, который затем можно и анализировать (surface → analysis), и генерировать (lemma + tags → surface). Описаны также процесс индукции правил морфотактики LEXC и правил выбора алломофов regex, сборка

трансдьюсера и итеративная процедура расширения покрытия (схема пайплайна показана на рис. 1).

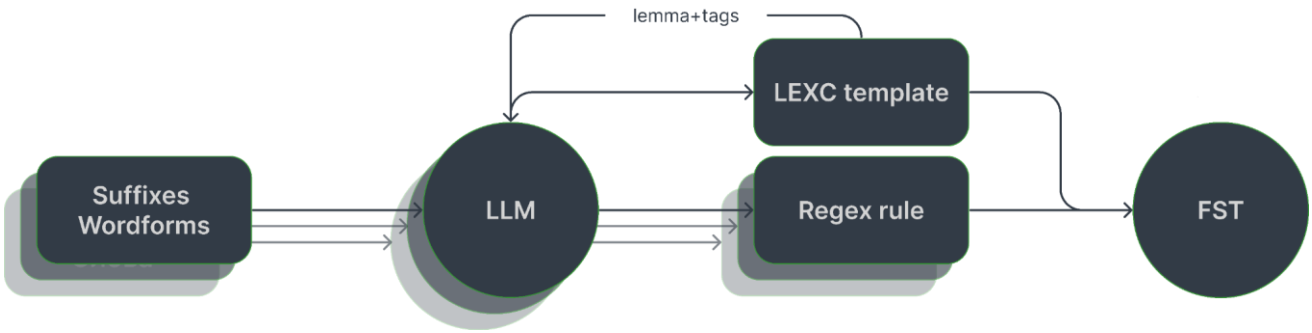


Рис. 1. Автоматизированная индукция: от словоформ к LEXC/regex и сборке FST.

Исходные данные — параллельное издание «Маленького принца» на башкирском и русском (~1600 предложений). Для иллюстрации приведен фрагмент:

Был китапты өлкән кешегә бағышлаған өсөн балаларҙан мине ғәфү итеүҙәрен һорайым. Прошу детей простить меня за то, что я посвятил эту книжку взрослому.

Для индукции был заранее отобран список целевых именных лемм:

бала кеше китап …

В настоящей работе сознательно фокус сделан только на существительных и описана их парадигма как последовательность состояний:

NStem → NNumber → NPoss → NCase → #

Корпус автоматически просматривается на предмет всех извлеченных словоформ для выбранных лемм в виде lemma, surface, позиция, локальный контекст. Пример структурированного вывода:

{"noun": "бала", "matches": [{"line_num": 4, "matched_form": "балаларҙан"},
{"line_num": 6, "matched_form": "балалар"}]}
{"noun": "йылға", "matches": [{"line_num": 865, "matched_form": "йылғалар"},
{"line_num": 882, "matched_form": "йылғаларға"}]}

Чтобы понять, какие алломорфы потребуются, были сгруппированы наблюдения по схожим суффиксальным хвостам и оценены их частоты. Это помогает выделить «опорные» аффиксы, для которых и будут записываться правила:

{"suffix": "лар", "count": 16, "occurrences": [{"form": "йылғалар", "noun": "йылға",
"line_num": 865}, {"form": "балалар", "noun": "бала", "line_num": 6}]}

На уровне морфотактики необходимы правила в формате LEXC, которые определяют допустимый порядок морфем. Такой шаблон сгенерирован с помощью LLM: верхний уровень задает путь NStem → NNumber → NPoss → NCase → #, а на surface-стороне вместо конкретных букв использованы шаблоны — абстрактные маркеры, которые позже «развернут» regex-правила. Например, +ЛАр означает множественное число, +ҺЫ — притяжательность с гармонией, +ГА/+ДА — классы онсета и гласной для дательного/местного, +НЫ — винительный с вариациями Н/ҙ/д/т и Ы/е по контексту. Эти шаблоны нужны, чтобы разделить ответственность: LEXC строго фиксирует порядок морфем, а выбор конкретных алломорфов определяется контекстным правилам (HFST/XFST regex), которые заменяют «заглавные» компоненты (Л, А, Ы, и т. п.) на нужные буквы в зависимости от окружения.

Минимальный фрагмент LEXC выглядит так:

```
LEXICON Root
бала   +N:0       NNumber ;
йылға  +N:0       NNumber ;
```

```
…
LEXICON NNumber
+SG:0      NPoss ;
+PL:+ЛАр    NPoss ;

LEXICON NPoss
+POSS0:0      NCase ;
+POSS1SG:+һЫ   NCase ;
+POSS2SG:+һЫ   NCase ;
+POSS3SG:+һЫ   NCase ;
+POSS1PL:+һЫ   NCase ;
+POSS2PL:+һЫ   NCase ;
+POSS3PL:+һЫ   NCase ;

LEXICON NCase
+NOM:0    # ;
+ACC:+НЫ  # ;
+GEN:+НЫ  # ;
+DAT:+ГА  # ;
+LOC:+ДА  # ;
+ABL:+ДА  # ;
```

Чтобы связать поверхности с анализами, каждой найденной форме нужна разметка «лемма + теги», которую размечает LLM, явно фиксируя формат и порядок тегов. Базовый системный промпт выглядит так:

```
"You label Bashkir NOUN analyses. Use this exact order:\n"
"Lexeme+N+{SG|PL}+{POSS1SG|POSS2SG|POSS3SG|POSS1PL|POSS2PL|POSS3PL|POSS0}?+{NOM|ACC|DAT|LOC|ABL|GEN}\n…"
```

В запрос также добавляется контекст из корпуса (оба языка):

```
prompt.append(f"- form: {it['form']} | noun: {it['noun']}  ba: {it['ba']}  ru: {it['ru']}")
```

На выходе получается:

```
балалар = "бала+N+PL+NOM"
баланы  = "бала+N+SG+ACC"
йылғаларға = "йылға+N+PL+DAT"
```

Далее по каждому «семейству» аффикса компактный пакет словоформ отправляется в LLM, чтобы сгенерировать правила, реализующие поверхностную алломорфию для соответствующего шаблона:

```
"You are designing HFST regex rules to realize noun surface allomorphy from
abstract LEXC tags …"
```

LLM выдает набор узконаправленных правил, которые затем объединяются и упорядочиваются «специфичное → основное». Примеры агрегированных правил для множественного числа и гармонии (схематично):

```
Л -> л || [ а | ә | ы | е | о | ө | я | э ] "+" _ [ A r ]
А -> а || [ а | о | у | ы ] ?* "+" [ л | т | д | ҙ ] _ [ r ]
```

Сборка трансдьюсера идет по цепочке: сначала компилируется LEXC, затем правила, после чего они последовательно композируются. В коде это отражается следующим образом:

```
# 1) компиляция морфотактики
lexc = hfst.compile_lexc_file(str(_built))
# 2) компиляция и композиция правил
# rules = ... compose_sequence(... hfst.regex(pattern) ...)
lexc.compose(rules)  # итоговый T = LEXC ∘ R1 ∘ R2 ∘ ... ∘ Rk
```

Получившийся FST на стороне анализа распознает данные словоформы и возвращает соответствующие разборы. После первой сборки автоматически вычленяются формы из корпуса, которые еще не распознаются, и повторяется цикл: LLM проставляет переходы, генерирует новые локальные правила, унифицирует, пересобирает. Итерации идут до тех пор, пока добавление правил

дает стабильный прирост покрытия, и размер автомата остается в заданных пределах.

На наборе из 50 лемм этот процесс привел к автомату, распознающему 170 из 250 уникальных словоформ (≈70%). В корпусе для этих лемм встретилось около 200 уникальных суффиксальных хвостов; итоговый стек правил включает порядка 110 regex-переписываний, покрывающих множественное (+ЛАр), притяжательность (1SG, 2SG, 3SG и 1PL, 2PL, 3PL) и падежи (ACC, GEN, DAT, LOC, ABL). Это обеспечивает переносимую LEXC/regex-грамматику, которая обобщается на леммы вне исходного списка и существенно автоматизирует построение конечных автоматов для именной морфологии.

ГИБРИДНАЯ АРХИТЕКТУРА ПЕРЕВОДА

Цель этого раздела — показать, как собранный в п. 2 двунаправленный FST практически используется вместе с LLM для автоматической правки морфологических ошибок при переводе. Двунаправленность важна: один и тот же автомат умеет анализировать поверхность (surface → lemma + tags) и генерировать корректную словоформу по анализу (lemma+tags → surface). Это позволяет не только проверять вывод модели, но и восстанавливать правильную форму там, где LLM дала «несуществующее» слово.

Формируется испытательный набор предложений из того же параллельного издания «Маленького принца» (~1600 пар предложений на русском и башкирском языках). Сначала извлекаем все предложения, содержащие извлеченные словоформы из целевой именной области, таких предложений набралось 300. Каждое из них переводится с русского на башкирский с помощью gpt-4o-mini, сохраняя исход и последующие правки для метрик. Базовая подсказка к переводу минимальна и нейтральна:

You are a professional translator. Translate the given Russian sentence into Bashkir.

Полученные гипотезы служат базовой линией качества; они же становятся входом для нашего гибридного цикла LLM ↔ FST. Основная идея такова: проверяется каждое сгенерированное с помощью FST слово; если автомата «нет» на эту поверхность, LLM возвращает «лемма + теги» для позиции, затем

генерируется корректная форма через FST и подставляется в перевод. Благодаря двунаправленности та же грамматика валидирует и исправляет (цикл показан на рис. 2).

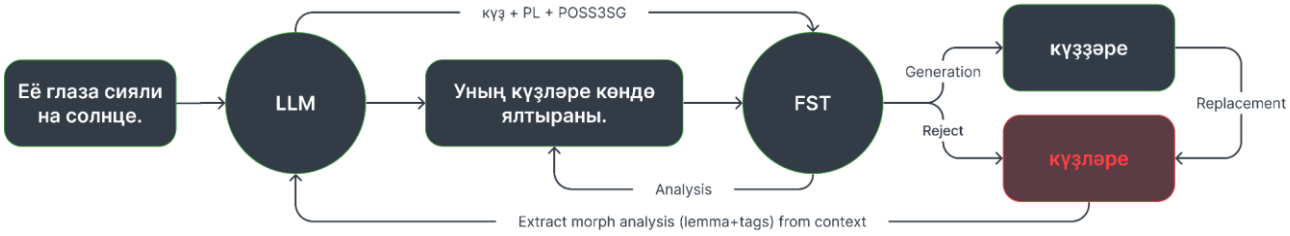


Рис. 2. Контур LLM ↔ FST при переводе: выявление нераспознанной формы, извлечение lemma+tags и генерация корректной поверхности.

Далее приведен пример работы цикла на одной фразе:

Запрос к переводчику:
Переведи с русского на башкирский: Её глаза сияли на солнце.

Исходный ответ gpt-4o-mini:
Уның күҙләре көндө ялтыраны.

Проблема:
күҙләре отсутствует в FST — форма несовместима с башкирской морфологией.

Уточнение к LLM (локальный контекст позиции):
Какая лемма и какие грамматические признаки у слова күҙләре в этом контексте?

Ответ:
күҙ + PL + POSS3SG

Генерация через FST:
күҙ + PL + POSS3SG → күҙҙәре

Итог:
Уның күҙҙәре көндө ялтыраны.

Смысл сохраняется, морфологическая ошибка снимается автоматически. Приведем еще один показательный случай с множественным числом:

До: Малайлар уҡытыусыга килде.
После: Малайҙар уҡытыусыға килде.

Тот же протокол формализуем простым псевдокодом:

```
for s in test_set:                    # 300 предложений
 hyp = LLM_translate(s.ru)            # gpt-4o-mini, базовый промпт
 for each token t at position p in hyp:
  if not FST.recognizes(t):
   (lemma, tags) = LLM_infer_analysis(s.ru, hyp, p)   # локальный контекст
   t_corr = FST.generate(lemma, tags)                 # lemma+tags → surface
   if t_corr exists:
    replace t with t_corr in hyp
 save {baseline: original LLM hyp, corrected: hyp}
```

Чтобы сравнить качество, для каждого сегмента сохраним базовую гипотезу и исправленную версию вместе с эталоном. Формат записи следующий:

{"line_num": 378, "translation": "Мин эште ташланым.", "fst_corrected": "Мин эшемде ташланым.", "reference": "Мин эшемде ташланым."}

На этом подкорпусе из 300 предложений (средняя длина — около 8 токенов или 60 символов) оценивается BLEU и ChrF. Базовый перевод gpt-4o-mini даёт BLEU 16.14 и ChrF 45.13; после правок через FST получили BLEU 25.71 и ChrF 50.87. Прирост составил +9.57 BLEU и +5.74 ChrF. Более резкий рост BLEU ожидаем: исправление одной именной группы зачастую «чинит» сразу несколько n-грамм подряд; ChrF реагирует умереннее, поскольку фиксирует локальные символные изменения (сводные результаты представлены на рис. 3).

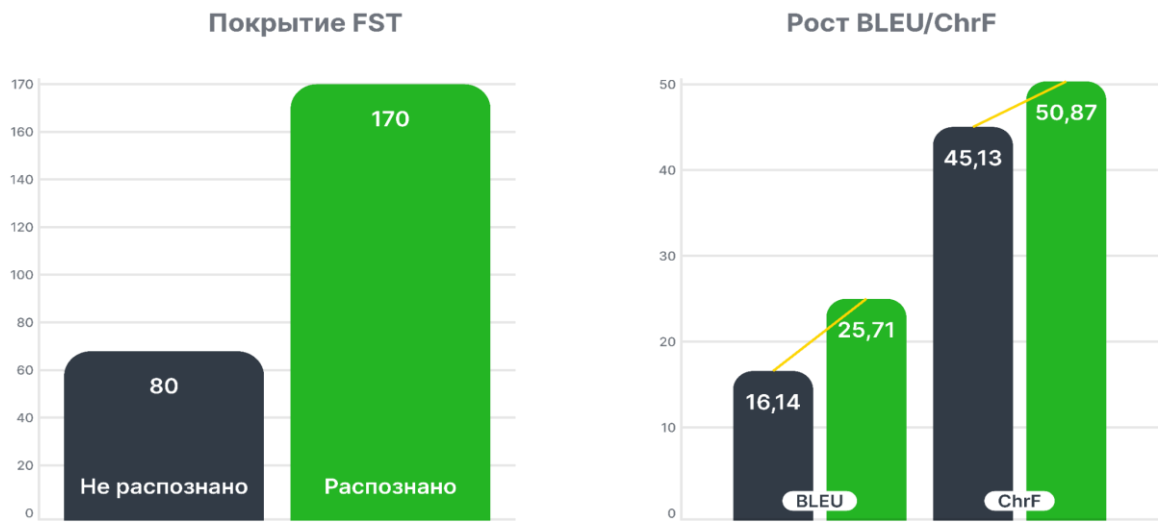


Рис. 3. Слева — покрытие FST по 250 словоформам; справа — прирост BLEU/ChrF после валидации через FST на 300 предложениях.

Важно, что gpt-4o-mini сам по себе относительно неплохо справляется с башкирским; на более слабых моделях выигрыш от данного подхода, вероятно, был бы еще заметнее. Поиск оптимальной базовой LLM находится за рамками настоящей работы, но сама архитектура к этому приспособлена: в гибридной схеме ключевая задача LLM состоит в том, чтобы извлекать лемму и теги, а не генерировать точную поверхность. Это означает, что компонент LLM можно заменять на более легкий/дешевый, обученный хуже на башкирском, FST всё равно возьмет на себя корректную реализацию поверхностной формы.

ЗАКЛЮЧЕНИЕ

Представлен гибридный нейросимволический подход к морфологически корректной генерации на агглютинативных языках: LLM извлекает из контекста пару «лемма + теги», а FST реализует и валидирует поверхностную форму. В автоматизированном режиме из корпусных данных извлекаются переносимые LEXC/regex-описания выбора алломорфов и фонологических чередований, что позволяет FST обобщаться на леммы вне исходного списка. Интеграция FST прямо в контур генерации дает заметный прирост качества (BLEU/ChrF) без дообучения переводчика, что подтверждает практичность подхода при ограниченных ресурсах.

Метод применим к задачам перевода, проверки орфографии и автоматизации процессов создания цифровых лингвистических ресурсов для малоресурсных языков, что особенно актуально для проектов, таких как «Тюркская морфема», где требуется предварительное заполнение грамматических шаблонов.

В будущем планируется расширять покрытие на другие морфологические категории (включая глаголы и другие части речи) и углублять автоматизацию индукции и верификации правил.

СПИСОК ЛИТЕРАТУРЫ

1. *Sproat R., Østling R.* The morphological gap between translation quality and surface accuracy // Proceedings of the WMT 2020 Conference. Online, 2020. P. 1015–1024.

2. *Kann K., Cotterell R., Schütze H.* Neural models of inflectional morphology // Proceedings of the 15th Conference of the European Chapter of the ACL (EACL 2017). Valencia, 2017. P. 322–334.

3. *Mielke S., Eisenstein J., Cotterell R.* Dialect-to-dialect translation and cross-dialect morphological robustness of language models // Transactions of the ACL. 2021. Vol. 9. P. 288–302.

4. *Koskenniemi K.* Two-level morphology: a general computational model for word-form recognition and production. Helsinki: University of Helsinki, Department of General Linguistics, 1983. 38 p.

5. *Beesley K.R., Karttunen L.* Finite-State Morphology. Stanford (CA): CSLI Publications, 2003. 550 p.

6. *Stahlberg F., Hasler E., Waite A.* SGNMT: A flexible NMT decoding toolkit for quick prototyping of new models // Proceedings of ACL System Demonstrations. Vancouver, 2017. P. 67–72.

7. *Hulden M.* FST-based grammar correction for richly inflected languages // Proceedings of ACL Workshop on Finite-State Methods. Montréal, 2012. P. 32–39.

8. *Tamchyna A., Bojar O.* Target-side context for morphological reinflection // Proceedings of the First Conference on Machine Translation (WMT 2016). Berlin, 2016. P. 586–594.

9. *Schwartz L., Liu S., Surrain S.* Bootstrapping a neural morphological analyzer from an existing FST // Proceedings of the ACL Workshop on Morphological Resources 2022. Seattle, 2022. P. 12–20.

NEURO-SYMBOLIC APPROACH TO AUGMENTED TEXT GENERATION VIA AUTOMATED INDUCTION OF MORPHOTACTIC RULES

M. V. Isangulov[1] [0009-0006-3244-0328], A. M. Elizarov[2] [0000-0003-2546-6897], A. R. Kunafin[3] [0009-0006-0495-265X], A. R. Gatiatullin[4] [0000-0003-3063-8147], N. A. Prokopyev[5] [0000-0003-0066-7465]

[1, 2]Kazan Federal University, Kazan, Russia

[3]Independent researcher

[4, 5]Academy of Sciences of the Republic of Tatarstan, Kazan, Russia

[1]marathon.our@gmail.com, [2]amelizarov@gmail.com, [3]aigizk@gmail.com

[4]ayrat.gatiatullin@gmail.com, [5]nikolai.prokopyev@gmail.com

Abstract

The work presents a hybrid neuro-symbolic method that combines a large language model (LLM) and a finite-state transducer (FST) to ensure morphological correctness in text generation for agglutinative languages. The system automatically extracts rules from corpus data: for local examples of word forms, the LLM produces sequences of morphological analyses, which are then aggregated and organized into compact descriptions of morphotactic rules (LEXC) and allomorph selection (regex). During generation, the LLM and FST operate jointly: if a token is not recognized by the automaton, the LLM derives a “lemma+tags” pair from the context, and the FST produces the correct surface form. A literary corpus (~1600 sentences) was used as the dataset. For a list of 50 nouns, 250 word forms were extracted. Using the proposed algorithm, the LLM generated 110 context-sensitive regex rules along with LEXC morphotactics, from which an FST was compiled that recognized 170/250 forms (~70%). In an applied machine translation test on a subcorpus of 300 sentences, integrating this FST into the LLM cycle improved quality from BLEU 16.14 / ChrF 45.13

to BLEU 25.71 / ChrF 50.87 without retraining the translator. The approach scales to other parts of speech (verbs, adjectives, etc.) as well as to other agglutinative and low-resource languages, where it can accelerate the development of lexical and grammatical resources.

***Keywords**: neuro-symbolic approach, large language model, finite-state transducers, two-level morphology, LEXC morphotactics, machine translation, agglutinative languages, Bashkir language.*

REFERENCES

1. *Sproat R., Østling R.* The morphological gap between translation quality and surface accuracy // Proceedings of the WMT 2020 Conference. Online, 2020. P. 1015–1024.

2. *Kann K., Cotterell R., Schütze H.* Neural models of inflectional morphology // Proceedings of the 15th Conference of the European Chapter of the ACL (EACL 2017). Valencia, 2017. P. 322–334.

3. *Mielke S., Eisenstein J., Cotterell R.* Dialect-to-dialect translation and cross-dialect morphological robustness of language models // Transactions of the ACL. 2021. Vol. 9. P. 288–302.

4. *Koskenniemi K.* Two-level morphology: a general computational model for word-form recognition and production. Helsinki: University of Helsinki, Department of General Linguistics, 1983. 38 p.

5. *Beesley K.R., Karttunen L.* Finite-State Morphology. Stanford (CA): CSLI Publications, 2003. 550 p.

6. *Stahlberg F., Hasler E., Waite A.* SGNMT: A flexible NMT decoding toolkit for quick prototyping of new models // Proceedings of ACL System Demonstrations. Vancouver, 2017. P. 67–72.

7. *Hulden M.* FST-based grammar correction for richly inflected languages // Proceedings of ACL Workshop on Finite-State Methods. Montréal, 2012. P. 32–39.

8. *Tamchyna A., Bojar O.* Target-side context for morphological reinflection // Proceedings of the First Conference on Machine Translation (WMT 2016). Berlin, 2016. P. 586–594.

9. *Schwartz L., Liu S., Surrain S.* Bootstrapping a neural morphological analyzer from an existing FST // Proceedings of the ACL Workshop on Morphological Resources 2022. Seattle, 2022. P. 12–20.

СВЕДЕНИЯ ОБ АВТОРАХ

ИСАНГУЛОВ Марат Вильданович окончил бакалавриат Института информационных технологий и интеллектуальных систем (ИТИС) Казанского (Приволжского) федерального университета в 2021 году, магистратуру ИТИС в 2023 г. В настоящее время – аспирант ИТИС.

Marat Vildanovich ISANGULOV graduated with a Bachelor's degree from Institute of Information Technology and Intelligent Systems (ITIS) of Kazan Federal University in 2021 and a Master's degree from ITIS in 2023. He is currently a PhD student at ITIS.

email: marathon.our@gmail.com

ORCID: 0009-0006-3244-0328

ЕЛИЗАРОВ Александр Михайлович – доктор физико-математических наук, профессор, заслуженный деятель науки Российской федерации, заслуженный деятель науки Республики Татарстан, профессор кафедры цифровой аналитики и технологий искусственного интеллекта Института информационных технологий и интеллектуальных систем Казанского (Приволжского) федерального университета.

ELIZAROV Alexander Mikhailovich – Doctor of Physics and Mathematics, Professor, Honoured Worker of Science of the Russia, Honoured Worker of Science of the Republic of Tatarstan, Kazan Federal University. Current scientific interests: data mining, recommender systems, cloud computing, knowledge extraction technologies.

email: amelizarov@gmail.com

ORCID: 0000-0003-2546-6897

КУНАФИН Айгиз Ражапович. Окончил Уфимский авиационный университет. Создатель башкироязычной умной колонки «Һомай» и одного из крупнейших открытых ресурсов для цифровизации башкирского языка. Победитель конкурса AI for Good Innovation Factory Malta (2025); представлял проект «Homai» на финале в Женеве (предфинал, 2025). Соавтор публикации WMT-2021 по машинному переводу тюркских языков, контрибьютор Apertium (Bashkir), эксперт программы ЮНЕСКО «Информация для всех» с 2022 года.

Aygiz Razhapovich KUNAFIN graduated from Ufa Aviation University. Creator of the Bashkir-language smart speaker “Һомай” and a major open-source platform for digitizing the Bashkir language. He is the winner of the AI for Good Innovation Factory Malta pitch competition in March 2025; represented the project “Homai” at the Geneva Grand Finale of the AI for Good Global Summit in July 2025 (pre-final round). He is a co-author of a WMT-2021 publication on machine translation for Turkic languages, contributor to Apertium (Bashkir); invited expert in UNESCO’s “Information for All” Programme since 2022.

email: aigizk@gmail.com
ORCID: 0009-0006-0495-265X

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 60 научных трудов.

Ayrat Rafizovich GATIATULLIN *g*raduated from Kazan State University in 1994, candidate in technical sciences (2002). He is a leading researcher at the Institute of Applied Semiotics of the Academy of Sciences of the Republic of Tatarstan. List of scientific works includes more than 60 publications.

email: ayrat.gatiatullin@gmail.com,
ORCID: 0000-0003-3063-8147

ПРОКОПЬЕВ Николай Аркадиевич. Кандидат технических наук. Окончил Институт вычислительной математики и информационных технологий Казанского федерального университета в 2015 году. Научный сотрудник Института прикладной семиотики Академии наук РТ. В списке научных трудов более 50 работ.

Nikolai Arkadievich PROKOPYEV candidate of Technical sciences. Graduated from the Institute of Computational Mathematics and Information Technologies of the Kazan Federal University in 2015. He is a researcher at the Institute of Applied Semiotics of the Academy of Sciences of the Republic of Tatarstan. List of scientific works includes more than 50 publications.

email: nikolai.prokopyev@gmail.com,

ORCID: 0000-0003-0066-7465

Материал поступил в редакцию 13 октября 2025 года

УДК 004.8

ОЦЕНКА НЕОПРЕДЕЛЕННОСТИ В ТРАНСФОРМЕРНЫХ ЦЕПЯХ НА ОСНОВЕ ПРИНЦИПА СОГЛАСОВАННОСТИ ЭФФЕКТИВНОЙ ИНФОРМАЦИИ

А. А. Красновский[0000-0001-6842-7340]

Университет Иннополис, г. Иннополис, Россия

a.a.krasnovsky@gmail.com

Аннотация

Механистическая интерпретируемость позволяет выявлять функциональные подграфы в больших языковых моделях (LLM), известные как трансформерные цепи (Transformer Circuits, TC), которые реализуют конкретные алгоритмы. Однако отсутствует формальный способ, позволяющий за один проход количественно оценить, когда активная цепь ведет себя согласованно и, следовательно, ее состояние может быть признано корректным. Опираясь на ранее предложенную автором пучково-теоретическую формализацию причинной эмерджентности (Krasnovsky, 2025), мы специализируем ее для трансформерных цепей и вводим безразмерную однопроходную оценку согласованности эффективной информации (Effective Information Consistency Score, EICS). EICS сочетает нормализованную несогласованность пучка, вычисляемую из локальных якобианов и активаций, с гауссовским прокси EI для причинной эмерджентности на уровне цепи, полученным из того же состояния прямого прохода. Такая конструкция является прозрачной (white-box), однопроходной и делает единицы измерения явными, так что оценка безразмерна. Представлены практические рекомендации по интерпретации оценки, учету вычислительных затрат (с быстрыми и точными режимами) и анализ простейшего примера для проверки на адекватность.

Ключевые слова*: механистическая интерпретируемость, трансформерные цепи, теория пучков, причинная эмерджентность, количественная оценка неопределенности, большие языковые модели (LLM).*

© А. А. Красновский, 2025.
Данная статья распространяется на условиях международной лицензии Creative Commons License Attribution 4.0 International (CC BY 4.0).

ВВЕДЕНИЕ

Основополагающей целью механистической интерпретируемости является восстановление (или реконструкция) алгоритмических компонентов LLM на мезоуровне («трансформерных цепей») [1, 2]. Эти подграфы («головы внимания», MLP и их пути) связаны с такими задачами, как копирование или индукция [1] и извлечение фактов [3]. После идентификации цепи возникает естественный вопрос: *функционирует ли она согласованно при данном входном сигнале?* Одна и та же модель может ответить на фактический запрос правильно или «галлюцинировать». В рамках настоящей работы предположим, что *степень причинной согласованности* активной цепи различается между этими режимами.

Кроме того, для построения концептуального базиса будем рассматривать идеи теории пучков и причинной эмерджентности, адаптируя их для трансформерных цепей [4—9]. В качестве решения мы предлагаем метрику EICS, которая вычисляется за один прямой проход и дает количественную оценку. Концептуально неопределенность трактуется как потеря причинно-следственной связности: высокая активность при низкой несогласованности означает надежность системы, а обратная картина — ее рискованность. В отличие от подходов к оценке неопределенности (Uncertainty Quantification, UQ) типа «черный ящик» [10—13], EICS идентифицирует конкретные механизмы, ответственные за возникновение неопределенности.

ОБЗОР СУЩЕСТВУЮЩИХ МЕТОДОВ

Механистическая интерпретируемость. Головы индукции и поведение обучения «в контексте» были описаны Олссоном и соавт. [1]. Недавние инструменты «графов атрибуций» (также известные как трассировка цепей) предоставляют процедуры обнаружения подграфов и их валидации [2]. Активно изучаются цепи, связанные со знаниями [3], взаимосвязь между локализацией и редактированием также является предметом всестороннего анализа [14].

Количественная оценка неопределенности (UQ). Методы UQ «черного ящика» включают распределенные свободные конформистские предсказания [10], пост-процессинг калибровки [11], глубокие ансамбли [12] и байесовские или

приближенно-байесовские методы тонкой настройки LLM, такие как Laplace-LoRA [13]. Наша задача состоит в обеспечении прозрачности за счет сигнала, генерируемого внутренними цепями.

Клеточные пучки и причинная эмерджентность. Клеточные пучки предлагают формализм для объединения локальных линейных отображений в глобально согласованные состояния, где степень несогласованности измеряется при помощи когомологий (кобоундариев) и операторов Лапласа — Ходжа [5, 6]. В свою очередь, причинная эмерджентность дает количественную оценку того, в каких случаях макромасштабные описания системы содержат больше эффективной информации, чем описания на уровне ее составных частей [7—9]. Мы адаптируем этот теоретический аппарат к анализу цепей трансформеров, вводя для этого явные и вычислимые показатели (прокси).

ОСНОВНЫЕ ТЕОРЕТИЧЕСКИЕ КОНЦЕПЦИИ

1. Трансформерные цепи

Рассмотрим нейросетевую архитектуру «трансформер» как ориентированный ациклический граф (DAG) $G = (V, E)$, где вершины соответствуют головам внимания или MLP, а ребра отражают поток информации. ТС — это подграф $G_M \subseteq G$, который предположительно реализует задачу. Для входа x каждая вершина $v \in V_M$ имеет активацию $a_v \in \mathbb{R}^{d_v}$. Мы будем обозначать как $e = (u \to v) \in E$ ориентированное ребро и использовать локальные линеаризации в наблюдаемых активациях.

2. Клеточные пучки на графах и вычислимая несогласованность

Определим клеточный пучок $\mathcal{F}$ на *подлежащей неориентированной* версии G_M (так что 1-коцепи живут на ребрах независимо от направления в DAG) со стеблями $\mathcal{F}(v) = \mathbb{R}^{d_v}$ и отображениями ограничений на ориентированных ребрах, задаваемыми якобианами, вычисленными в текущем состоянии:

$$\rho_e : \mathcal{F}(u) \to \mathcal{F}(v), \qquad \rho_{u\to v} := J_{u\to v} \equiv \left(\frac{\partial f_v}{\partial a_u}\right)_a.$$

Для 0-коцепи (назначения узлов) $s = \{s_v\}$ кобоундарий пучка $\delta^0 : C^0(G; f \to C^1(G; F)$ действует как

$$(\delta^0 s)_{(u\to v)} \;=\; \rho_{u\to v}\, s_u \;-\; s_v. \tag{1}$$

На графе (без 2-ячеек) $\delta^1 = 0$, следовательно, $H^1 \cong C^1/\operatorname{im}\delta^0$. Вместо того чтобы брать неканоническую «норму фактор-пространства», мы используем *нормализованную энергию несогласованности* из наблюдаемых активаций $a = \{a_v\}$:

$$C_{\mathrm{sh}}(G_M, a) \;=\; \frac{\left(\sum_{(u\to v)\in E_M} \|\rho_{u\to v} a_u - a_v\|_2^2\right)^{1/2}}{\varepsilon + \left(\sum_{(u\to v)\in E_M} \|a_u\|_2^2 + \|a_v\|_2^2\right)^{1/2}} \tag{2}$$

с малым $\varepsilon > 0$ для числовой устойчивости. Это безразмерная величина, которая равна 0 тогда и только тогда, когда a является (шумно) согласованным глобальным сечением. Можно опционально заменить a на оценку наименьших квадратов $\hat{s} = \operatorname{argmin}_s \sum_e \| \rho_e s_u - s_v \|^2$; обе операции — это однопроходные вычисления (на основе произведения вектора на якобиан и якобиана на вектор: VJP/JVP).

Вычисление JVP, инициированное узлами (эффективность). Чтобы эффективно вычислить (2), мы используем схему JVP, инициированную узлами: для каждого исходного узла $u \in V_M$ выполняется один JVP с семенем a_u для вычисления всех исходящих остаточных членов $\rho_{u\to v} a_u$ за один проход. Это уменьшает сложность с JVP для каждого ребра до $O(|V_M|)$ JVP (обычно 10—20 для средних цепей), сохраняя точное определение в (2).

3. Однопроходный гауссовский прокси для оценки эффективной информации

Истинная эффективная информация (EI) определяется интервенциями. Чтобы получить однопроходный прокси, предполагаем малыми изотропные локальные интервенции в текущем состоянии и аппроксимируем каждое отображение его якобианом. Для линейного отображения $y = Jx + \xi$ с изотропным x единичной дисперсии и малым аддитивным шумом ξ взаимная информация (в натах) пропорциональна $\frac{1}{2}\log\det(I + \alpha J^{\top} J)$ с масштабом $\alpha > 0$. Поэтому определим

$$\mathrm{EI}_G(J) := \tfrac{1}{2}\log\det(I + \alpha J^{\top} J),\;\; \Delta\mathrm{EI}_G(G_M) := \mathrm{EI}_G(J_M) \;-\; \textstyle\sum_{v\in V_M} \mathrm{EI}_G\,(J_v), \tag{3}$$

где J_M — макро-якобиан от входов цепи к ее выходам (полученный путем линеаризации составного подграфа). Далее используем положительную часть $\Delta\mathrm{EI}_G^+ = \max(0, \Delta\mathrm{EI}_G)$ и необязательную нормализацию $\widetilde{\Delta\mathrm{EI}}_G := \Delta\mathrm{EI}_G^+ / \big(\varepsilon + \mathrm{EI}_G(J_M)\big)$, чтобы удерживать оценки в $[0,1)$.

4. Гауссовская прокси-оценка эффективной информации – вывод и замечания по реализации

Определение модели. Рассмотрим локальное, линейное описание цепи вокруг наблюдаемого состояния прямого прохода. Пусть $x \in \mathbb{R}^n$ обозначает малое стохастическое вмешательство на входах цепи, а $y \in \mathbb{R}^m$ — выходы цепи. Мы аппроксимируем

$$y \;=\; J\,x + \xi, \qquad x \sim \mathcal{N}(0, \sigma_x^2 I_n),\;\; \xi \sim \mathcal{N}\big(0, \sigma_\xi^2 I_m\big). \tag{4}$$

Взаимная информация. Для линейного гауссовского канала с независимыми гауссовскими входом и шумом

$$I(x;y) \;=\; \tfrac{1}{2}\log\det\big(I_m + \sigma_x^2/\sigma_\xi^2\, JJ^\top\big) \;=\; \tfrac{1}{2}\log\det\big(I_n + \sigma_x^2/\sigma_\xi^2\, J^\top J\big) \tag{5}$$

Определение прокси. Мы используем гауссовский прокси EI из (3); для цепи G_M возникновение и его нормализованная положительная часть задаются, как в п. 3.3,

$$\widetilde{\Delta\mathrm{EI}}_G \;=\; \frac{\max(0, \Delta\mathrm{EI}_G)}{\varepsilon + \mathrm{EI}_G(J_M)}. \tag{6}$$

Инвариантность, чувствительность и приближение при малых α. $\mathrm{EI}_G(J)$ зависит только от сингулярных значений J. Его чувствительность к α равна

$$\frac{\partial}{\partial\alpha}\left(\tfrac{1}{2}\log\det(I + \alpha J^\top J)\right) \;=\; \tfrac{1}{2}\,\mathrm{tr}[(I + \alpha J^\top J)^{-1} J^\top J],$$

и для $\alpha\,\sigma_{\max}^2 \ll 1$

$$\frac{1}{2}\log\det(I + \alpha J^\top J) \;\approx\; \frac{\alpha}{2}\,\|\,J\,\|_F^2.$$

См. также (5) для связи со взаимной информацией линейного гауссовского канала и (4) для постановки линейной модели.

Вычисление. Используем разложение Холецкого или собственное разложение для малых n; для больших n используем оцениватели лог-детерминанта Hutch++ или Ланцоша только с JVP/VJP-произведениями. Остаточные связи обработаем путем построения линейного блочного оператора или вычисления лог-детерминанта с помощью методов Крылова.

МЕТОД: ОЦЕНКА СОГЛАСОВАННОСТИ ЭФФЕКТИВНОЙ ИНФОРМАЦИИ

1. Определение

Даны G_M, активации a и якобианы ребер $\{\rho_{u\to v}\}$ из одного прямого прохода. Определим

$$\mathrm{EICS}(G_M; a) \;=\; \frac{\widetilde{\Delta\mathrm{EI}}_G(G_M)}{1+C_{\mathrm{sh}}(G_M,a)}. \tag{7}$$

Высокие значения EICS означают (i) сильную макроуровневую интеграцию информации относительно частей и (ii) низкое внутреннее несогласие на ребрах.

Почему это устраняет предыдущие проблемы. 1) Мы никогда не берем норму фактор-пространства H^1; а измеряем *энергию несогласия* (2) напрямую и безразмерно. 2) Термин EI – это четко сформулированный гауссовский прокси лог-определителя; единицы измерения – наты, которые становятся безразмерными через нормализацию. 3) DAG не содержат направленных циклов, но несогласованность пучка остается осмысленной на неориентированном 1-скелете.

2. Практическое использование и интерпретация

По конструкции $C_{\mathrm{sh}} \geq 0$ и $\widetilde{\Delta\mathrm{EI}}_G \in [0,1)$, следовательно, $\mathrm{EICS} \in [0,1)$. Используем тренировочное подмножество, чтобы выбрать порог τ (AUROC/F1), а также сообщаем компоненты $1/(1+C_{\mathrm{sh}})$ и $\widetilde{\Delta\mathrm{EI}}_G$, чтобы диагностировать факторы. Для α либо устанавливаем $\alpha = 1$ (априорное отношение сигнал — шум), либо выбираем α так, чтобы значение $\frac{1}{2}\log\det(I + \alpha J_M^{\top} J_M)$ оставалось в целевом межквартильном диапазоне и избегало насыщения.

ТЕОРЕТИЧЕСКИЕ СВОЙСТВА

Допущение 1 (Локальная линейность и ограниченность). *Вдоль G_M отображения локально линейны, якобианы $\{\rho_e\}$ липшицевы в окрестности a и нормы*

операторов ограничены. Лапласиан Ходжа пучка $L = \delta^{0\dagger}\delta^{0}$ *(с внутренним произведением, индуцированным весами ребер) имеет спектральный зазор* $\lambda_2(L) > 0$ [5].

Утверждение 1 (Вычислимость за один проход). *При выполнении Допущения 1 как* $C_{sh}(G_M, a)$*, так и* $\widetilde{\Delta EI}_G(G_M)$ *являются детерминированными функциями одного прямого прохода и его произведений якобиана на вектор. Следовательно, для вычисления* $EICS$ *требуется константное (*$O(1)$*) число прямых проходов.*

Базис доказательства. δ^0 строится из $\{\rho_e\}$, вычисленных в a. Как остаточный член (2), так и термины лог-детерминанта (3) являются функциями этих объектов. Для вычисления не требуется стохастическое моделирование по входным данным.

Утверждение 2 (Устойчивость к малым возмущениям вне цепи (оценка)). *Пусть* y *обозначает выходы цепи. Рассмотрим аддитивное внешнее возмущение* η*, которое входит в* G_M *с усилением не более* γ *в норме оператора. При утверждении 1*

$$\| \hat{s} - s^{\star} \| \leq \frac{\gamma}{\lambda_2(L)} \| \eta \|, \qquad \| \Delta y \| \leq \kappa \| \hat{s} - s^{\star} \|$$

для некоторой локальной константы Липшица κ*. В частности, малые значения* C_{sh} *(означающие высокие значения* $\lambda_2(L)$*) приводят к сужению интервалов оценок.*

О роли $\lambda_2(L)$. Оценка масштабируется как $1/\lambda_2(L)$. На практике $\lambda_2(L)$ зависит от i) связности и ii) весов ребер, индуцируемых локальными якобианами. Мы рекомендуем: a) *сообщать* $\lambda_2(L)$ для каждой цепи; b) *нормировать веса ребер* по операторным нормам $\rho_{u\to v}$ и c) при необходимости *регуляризовать* $L \leftarrow L + \beta I$, когда эмпирическое $\lambda_2(L)$ близко к нулю, — это ужесточает практическую оценку, не изменяя C_{sh} или EICS.

АЛГОРИТМ

1. Однопроходный EICS для трансформерной цепи

Алгоритм 1. Однопроходный EICS для трансформерной цепи

1: **Вход:** модель $\mathcal{M}$, вход x, цепь $G_M = (V_M, E_M)$, масштаб $\alpha > 0$.
2: **Выход:** $\mathrm{EICS}(G_M; a)$.
3: **Прямой проход и активации:** выполняем $\mathcal{M}(x)$ и записываем $\{a_v\}_{v \in V_M}$.
4: **Якобианы ребер:** для каждого $(u \to v) \in E_M$ вычисляем $\rho_{u \to v} = (\partial f_v / \partial a_u)_a$ с помощью VJP/JVP.
5: **Несогласованность пучка:** вычисляем $C_{\mathrm{sh}}(G_M, a)$ по формуле (2). *Реализация:* используем *JVP, инициированные узлами* (одна JVP на исходный узел u), чтобы вычислить все $\rho_{u \to v} a_u$ для исходящих ребер.
6: **Гауссовский прокси EI:** строим макро-якобиан J_M и якобианы узлов $\{J_v\}$. Вычисляем $\widetilde{\Delta \mathrm{EI}}_G$ по формулам (3), (6).
Быстрый режим (ранжирование): приближение для малых α; используем методы Hutch++ или Ланцоша с 4–8 зондами для каждого J_v и 8–12 для J_M.
Точный режим (малые блоки): вычисляем $\mathrm{logdet}(I + \alpha J^\top J)$ через разложение Холецкого или SVD.
7: **Оценка:** возвращаем $\mathrm{EICS} = \widetilde{\Delta \mathrm{EI}}_G / (1 + C_{\mathrm{sh}})$.

2. Вычислительные затраты и масштабирование

Порядок величины. С использованием JVP, инициированных узлами, вычисление C_{sh} требует $O(|V_M|)$ JVP (примерно 10–20 для средних цепей). В *быстром* режиме EI (приближение Фробениуса для малых α + зондирование Hutchinson) общее количество работ автоград обычно составляет ~ 50–200 JVP/VJP-произведений на ограниченном подграфе, или около ~ 2–6 эквивалентов прямого прохода. В *точном* режиме EI (большие α или явные факторизации) ожидайте ~ 5–15 эквивалентов прямого прохода. Расчет выполняется без применения методов Монте-Карло по входным данным: все величины детерминированно вычисляются из одного прямого прохода.

Параметры масштабирования. a) Пакетно инициируйте JVP для всех узлов; b) ограничьте EI топ-k сингулярными направлениями через Ланцоша; c) кешируйте промежуточные линейные отображения вдоль TC; d) предпочитайте малые α для задач ранжирования.

ПРЕДЛАГАЕМАЯ ВАЛИДАЦИЯ

1. Протокол оценки

Мы описываем протокол оценки для задачи фактических вопроса–ответа, используя рабочий процесс графа атрибуции [2] для идентификации цепи извлечения фактов G_{fact}. Предлагаем два набора: (A) вопросы с проверяемыми ответами; (B) адверсариальные или провоцирующие галлюцинации подсказки [15]. Ожидается, что у множества A будет более высокий EICS (низкий C_{sh}, положительное $\widetilde{\Delta\mathrm{EI}}_G$), а у множества B — пониженные оценки. Сравниваемые подходы включают лог-вероятность, энтропию, дисперсию глубоких ансамблей [12] и размеры конформных наборов [10].

2. Базовые методы и абляционный анализ

1) Корреляция активаций на ребрах (EAC). Средняя корреляция Пирсона между a_u и a_v по $(u \to v) \in E_M$ (по измерениям).

2) Остаток выравнивания на ребрах (EAR).

$$\frac{1}{|E_M|} \sum_{(u \to v)} \| \hat{\rho}_{u \to v} a_u - a_v \|_2$$

с оценкой наименьших квадратов по каждому ребру $\hat{\rho}_{u \to v}$ (без связи пучка).

Абляции. (A1) $1/(1 + C_{\mathrm{sh}})$ только; (A2) $\widetilde{\Delta\mathrm{EI}}_G$ только.

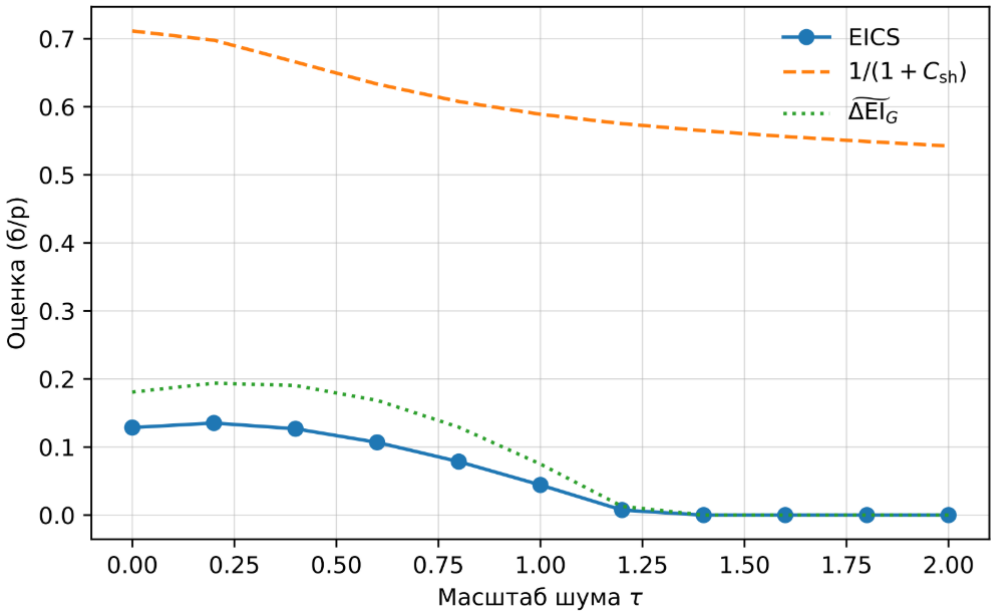


Рис. 1. Простейшая проверка адекватности результатов для цепи из 6 узлов с двумя параллельными ветвями. По мере увеличения шума узлов τ несогласованность пучка C_{sh} возрастает (а $1/(1+C_{\mathrm{sh}})$ падает). Мы также уменьшаем выравнивание между ветвями при увеличении τ (декогерентность ребер), что вызывает уменьшение прокси возникновения $\widetilde{\Delta \mathrm{EI}}_G$ и общей EICS. Кривые показывают средние значения по начальным генераторам. Определения следуют (2), (3) и (7).

3. Простейшая проверка на адекватность (аналитический и симуляционный протоколы)

Настройка. 6-узловая прямоугольная ТС с линейными блоками и аддитивным гауссовским шумом на ребрах; изменяйте шум на подмножестве ребер. Вычисляйте C_{sh}, $\widetilde{\Delta \mathrm{EI}}_G$ EICS и базовые методы для разных начальных генераторов.

Аналитическая проверка (малые α). При $\alpha = \sigma_x^2/\sigma_\xi^2$ увеличение аддитивного шума уменьшает термины EI; шум на ребрах увеличивает остатки в (2). Следовательно, EICS уменьшается с шумом, что соответствует интуиции (см. рис. 1 для простейшей проверки на адекватность, иллюстрирующей эти тенденции).

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ И ОГРАНИЧЕНИЙ ИССЛЕДОВАНИЯ

Зависимость от цепи. EICS имеет смысл лишь настолько, насколько корректно задан G_M.

Линеаризация. Основанный на якобианах пучок и прокси EI предполагают локальную линейность.

Стоимость. JVP, инициированные узлами, быстрые лог-детерминанты и топ-k направления смягчают затраты.

Место EICS среди методов UQ типа «черный ящик». EICS предоставляет *механистические* доказательства, дополняя калибровочные и конформные инструменты.

ЗАКЛЮЧЕНИЕ

Представлена реализация концепции применения пучковой и причинно-эмерджентной перспективы к трансформерным цепям как практическая, однопроходная оценка (EICS). Заменив плохо определенные нормы когомологий на нормализованную энергию несогласованности и определив гауссовский лог-детерминантный прокси EI, получим, что EICS является как вычислимой, так и безразмерной оценкой. Кроме того, детально описаны практические рекомендации, анализ вычислительных затрат для быстрых и точных режимов и простейшая проверка на адекватность.

ПРИЛОЖЕНИЯ

Код простейшей проверки на адекватность для рис. 1

В приведенном ниже скрипте воспроизводится рис. 1. Он реализует примитивную двухветвевую модельную цепь, вычисляет $\boldsymbol{C}_{\mathbf{sh}}$ (2), нормализованную прокси-оценку эмерджентности $\boldsymbol{\Delta\widetilde{\mathbf{EI}}_G}$ (3) и $\mathbf{EICS}$ (7) как функции масштаба шума $\boldsymbol{\tau}$.

```
import numpy as np, matplotlib.pyplot as plt
EPS, D, N_SEEDS = 1e-8, 32, 100
TAUS = np.linspace(0.0, 2.0, 11)
alpha, align = 1.0, 0.9  # фиксированное отношение сигнал-шум; высокая начальная согласованность ветвей

def rand_matrix(d, scale=0.8, rng=None):
    rng = np.random.default_rng() if rng is None else rng
    return scale * rng.normal(size=(d, d)) / np.sqrt(d)

def ei_proxy(J, alpha):  # 0.5 * сумма log(1 + alpha * sigma^2)
    s = np.linalg.svd(J, compute_uv=False)
    return 0.5 * np.sum(np.log1p(alpha * (s**2)))
```

```
def build_branch_mats(D=32, align=0.9, rng=None):
    rng = np.random.default_rng(123) if rng is None else rng
    U = rand_matrix(D, 0.8, rng); A = rand_matrix(D, 0.9, rng); W =
rand_matrix(D, 0.9, rng)
    W13 = U;  W23 = (1-align)*rand_matrix(D,0.8,rng) + align*U
    W34 = A;  W35 = (1-align)*rand_matrix(D,0.9,rng) + align*A
    W46 = W;  W56 = (1-align)*rand_matrix(D,0.9,rng) + align*W
    return W13, W23, W34, W35, W46, W56

def metrics_at_tau(tau, rng):
    W13,W23,W34,W35,W46,W56 = build_branch_mats(D, align, rng)
    # Декогерентность ребер: уменьшаем согласованность между ветвями с ростом
tau
    h = min(1.0, tau/2.0)
    nrg = np.random.default_rng(rng.integers(10**9))
    W56 = (1-h)*W56 + h*rand_matrix(D, 0.9, nrg)
    W35 = (1-h)*W35 + h*rand_matrix(D, 0.9, nrg)

    # Две параллельные подцепи (части), макро-оператор — их сумма
    Jb1 = W46 @ W34 @ W13
    Jb2 = W56 @ W35 @ W23
    JM  = Jb1 + Jb2

    EI_macro = ei_proxy(JM, alpha)
    EI_parts = ei_proxy(Jb1, alpha) + ei_proxy(Jb2, alpha)
    dEI_g = max(0.0, EI_macro - EI_parts) / (EPS + EI_macro)

    a1 = rng.normal(size=D); a2 = rng.normal(size=D)
    a3 = W13@a1 + W23@a2; a4 = W34@a3; a5 = W35@a3; a6 = W46@a4 + W56@a5
    a1o = a1 + tau*rng.normal(size=D); a2o = a2 + tau*rng.normal(size=D)
    a3o = a3 + tau*rng.normal(size=D); a4o = a4 + tau*rng.normal(size=D)
    a5o = a5 + tau*rng.normal(size=D); a6o = a6 + tau*rng.normal(size=D)

    edges = [(a1o,a3o,W13),(a2o,a3o,W23),(a3o,a4o,W34),
             (a3o,a5o,W35),(a4o,a6o,W46),(a5o,a6o,W56)]
    num = den = 0.0
    for au,av,W in edges:
        r = W@au - av; num += r@r; den += au@au + av@av
    Csh = np.sqrt(num) / (EPS + np.sqrt(den))
    EICS = dEI_g / (1.0 + Csh)
    return Csh, dEI_g, EICS

C_m, d_m, S_m = [], [], []
C_e, d_e, S_e = [], [], []
for tau in TAUS:
    Cs, ds, Ss = [], [], []
    for k in range(N_SEEDS):
```

```
        rng = np.random.default_rng(1000 + k)
        Csh, dEIg, EICS = metrics_at_tau(tau, rng)
        Cs.append(Csh); ds.append(dEIg); Ss.append(EICS)
    Cs, ds, Ss = map(np.array, (Cs,ds,Ss))
    C_m.append(Cs.mean()); d_m.append(ds.mean()); S_m.append(Ss.mean())
    C_e.append(Cs.std(ddof=1)/np.sqrt(N_SEEDS))
    d_e.append(ds.std(ddof=1)/np.sqrt(N_SEEDS))
    S_e.append(Ss.std(ddof=1)/np.sqrt(N_SEEDS))

TAUS = np.array(TAUS); C_m=np.array(C_m); d_m=np.array(d_m); S_m=np.array(S_m)
C_e=np.array(C_e); d_e=np.array(d_e); S_e=np.array(S_e)
plt.figure(figsize=(6.2,4.2))
plt.plot(TAUS, S_m, '-o', label='EICS')
invC = 1.0/(1.0 + C_m)
plt.plot(TAUS, invC, '--', label=r'$1/(1+C_{\mathrm{sh}})$')
plt.plot(TAUS, d_m, ':', label=r'$\widetilde{\Delta \mathrm{EI}}_G$')
plt.xlabel(r'Масштаб шума $\tau$'); plt.ylabel('Оценка (безразмерно)')
plt.title('Простейшая проверка на адекватность: влияние шума на EICS и компо-
ненты')
plt.grid(True, linewidth=0.5, alpha=0.5); plt.legend(frameon=False)
plt.tight_layout(); plt.savefig('fig_toy_noise_curve.pdf',
bbox_inches='tight')
```

СПИСОК ЛИТЕРАТУРЫ

1. *Olsson C., Elhage N., Nanda N., et al.* In-context Learning and Induction Heads. 2022. arXiv : 2209.11895.

2. *Anthropic*. Circuit Tracing / Attribution Graphs: Methods & Applications: Transformer Circuits Team. 2025. Access mode: https://transformer-circuits.pub/2025/attribution-graphs/ (дата обращения: 2025-08-20).

3. *Yao Y., Zhang N., Xi Z., Wang M., Xu Z., Deng S., and Chen H.* Knowledge Circuits in Pretrained Transformers // Advances in Neural Information Processing Systems (NeurIPS). 2024. Vol. 37. P. 118571–118602.

4. *Krasnovsky A.A.* Sheaf-Theoretic Causal Emergence for Resilience Analysis in Distributed Systems. 2025. arXiv: 2503.14104.

5. *Hansen J., Ghrist R.* Toward a Spectral Theory of Cellular Sheaves // Journal of Applied and Computational Topology. 2019. Vol. 3, No. 4. P. 315–358.

6. *Robinson M.* Topological Signal Processing. Springer, 2014.

7. *Rosas F.E., Mediano P.A.M., Jensen H.J., Seth A.K., Barrett A.B., Carhart-Harris R.L., and Bor D.* Reconciling Emergences: An Information-Theoretic Approach to Identify Causal Emergence in Multivariate Data // PLOS Computational Biology. 2020. Vol. 16, No. 12. P. e1008289.

8. *Tononi G., Sporns O.* Measuring Information Integration // BMC Neuroscience. 2003. Vol. 4. P. 31.

9. *Oizumi M., Albantakis L., Tononi G.* From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0 // PLOS Computational Biology. 2014. Vol. 10, No. 5. P. e1003588.

10. *Angelopoulos A.N., Bates S.* A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification. 2021. arXiv: 2107.07511.

11. *Guo C., Pleiss G., Sun Y., and Weinberger K.Q.* On Calibration of Modern Neural Networks // Proceedings of the 34th International Conference on Machine Learning (ICML). PMLR. 2017. P. 1321–1330.

12. *Lakshminarayanan B., Pritzel A., Blundell C.* Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles // Advances in Neural Information Processing Systems (NeurIPS). 2017. Vol. 30.

13. Bayesian Low-rank Adaptation for Large Language Models (Laplace-LoRA). 2023. ICLR 2024 version. arXiv: 2308.13111.

14. Does Localization Inform Editing? Surprising Differences in Causality-Based Localization vs. Knowledge Editing in Language Models / Hase P., Bansal M., Kim B., and Ghandeharioun A. // Advances in Neural Information Processing Systems (NeurIPS). 2023. Vol. 36. P. 17643–17668.

15. *Huang L., Yu W., Ma W., Zhong W., Feng Z., Wang H., Chen Q., Peng W., Feng X., Qin B., et al.* A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions // ACM Transactions on Information Systems. 2025. Vol. 43, No. 2. P. 1–55.

MEASURING UNCERTAINTY IN TRANSFORMER CIRCUITS WITH EFFECTIVE INFORMATION CONSISTENCY

A. A. Krasnovsky[0000-0001-6842-7340]

Innopolis University, Innopolis, Russia

a.a.krasnovsky@gmail.com

Abstract

Mechanistic interpretability has identified functional subgraphs within large language models (LLMs), known as Transformer Circuits (TCs), that appear to implement specific algorithms. Yet we lack a formal, single-pass way to quantify when an active circuit is behaving coherently and thus likely trustworthy. Building on the author's prior sheaf-theoretic formulation of causal emergence (Krasnovsky, 2025), we specialize it to transformer circuits and introduce the single-pass, dimensionless Effective-Information Consistency Score (EICS). EICS combines (i) a *normalized sheaf inconsistency* computed from local Jacobians and activations, with (ii) a *Gaussian EI proxy* for circuit-level causal emergence derived from the same forward state. The construction is white-box, single-pass, and makes units explicit so that the score is dimensionless. We further provide practical guidance on score interpretation, computational overhead (with fast and exact modes), and a toy sanity-check analysis.

Keywords*:* mechanistic interpretability, ransformer circuits, sheaf theory, causal emergence, uncertainty quantification, large language models (LLMs).

REFERENCES

1. *Olsson C., Elhage N., Nanda N., et al.* In-context Learning and Induction Heads. 2022. arXiv: 2209.11895.

2. *Anthropic*. Circuit Tracing / Attribution Graphs: Methods & Applications: Transformer Circuits Team. 2025. Access mode: https://transformer-circuits.pub/2025/attribution-graphs/ (accessed: 2025-08-20).

3. *Yao Y., Zhang N., Xi Z., Wang M., Xu Z., Deng S., and Chen H.* Knowledge Circuits in Pretrained Transformers // Advances in Neural Information Processing Systems (NeurIPS). 2024. Vol. 37. P. 118571–118602.

4. *Krasnovsky A.A.* Sheaf-Theoretic Causal Emergence for Resilience Analysis in Distributed Systems. 2025. arXiv : 2503.14104.

5. *Hansen J., Ghrist R.* Toward a Spectral Theory of Cellular Sheaves // Journal of Applied and Computational Topology. 2019. Vol. 3, No. 4. P. 315–358.

6. *Robinson M.* Topological Signal Processing. Springer, 2014.

7. *Rosas F.E., Mediano P.A.M., Jensen H.J., Seth A.K., Barrett A.B., Carhart-Harris R.L., and Bor D.* Reconciling Emergences: An Information-Theoretic Approach to Identify Causal Emergence in Multivariate Data // PLOS Computational Biology. 2020. Vol. 16, No. 12. P. e1008289.

8. *Tononi G., Sporns O.* Measuring Information Integration // BMC Neuroscience. 2003. Vol. 4. P. 31.

9. *Oizumi M., Albantakis L., Tononi G.* From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0 // PLOS Computational Biology. 2014. Vol. 10, No. 5. P. e1003588.

10. *Angelopoulos A.N., Bates S.* A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification. 2021. arXiv : 2107.07511.

11. *Guo C., Pleiss G., Sun Y., and Weinberger K.Q.* On Calibration of Modern Neural Networks // Proceedings of the 34th International Conference on Machine Learning (ICML). PMLR. 2017. P. 1321–1330.

12. *Lakshminarayanan B., Pritzel A., Blundell C.* Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles // Advances in Neural Information Processing Systems (NeurIPS). 2017. Vol. 30.

13. Bayesian Low-rank Adaptation for Large Language Models (Laplace-LoRA). 2023. ICLR 2024 version. arXiv : 2308.13111.

14. Does Localization Inform Editing? Surprising Differences in Causality-Based Localization vs. Knowledge Editing in Language Models / Hase P., Bansal M., Kim B., and Ghandeharioun A. // Advances in Neural Information Processing Systems (NeurIPS). 2023. Vol. 36. P. 17643–17668.

15. *Huang L., Yu W., Ma W., Zhong W., Feng Z., Wang H., Chen Q., Peng W., Feng X., Qin B., et al.* A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions // ACM Transactions on Information Systems. 2025. Vol. 43, No. 2. P. 1–55.

СВЕДЕНИЯ ОБ АВТОРЕ

КРАСНОВСКИЙ Анатолий Анатольевич – аспирант Университета Иннополис. Исследовательская работа сфокусирована на фундаментальных вопросах интерпретируемости сложных систем, а также на разработке и применении методов математического моделирования для их анализа. Академические изыскания подкреплены более чем десятилетним опытом работы в IT-индустрии, где занимался проектированием и разработкой высоконагруженных распределенных систем. Имеет степень магистра с отличием в области прикладной математики и компьютерных наук.

Anatoly Anatolievich KRASNOVSKY is a Ph.D. student at Innopolis University. His research focuses on fundamental questions in the interpretability of complex systems, as well as the development and application of mathematical modeling methods for their analysis. His academic pursuits are grounded in over a decade of experience in the IT industry, where he designed and developed high-load distributed systems. He holds an M.S. in Applied Mathematics and Computer Science with honors.

email: a.a.krasnovsky@gmail.com

ORCID: 0000-0001-6842-7340

Материал поступил в редакцию 13 октября 2025 года

УДК 004.89

АБСТРАКТИВНАЯ СУММАРИЗАЦИЯ НОВОСТЕЙ ВНЕШНЕЙ ТОРГОВЛИ НА ОСНОВЕ НОВОГО СПЕЦИАЛИЗИРОВАННОГО КОРПУСА ДАННЫХ

Д. А. Лютова[1] [0009-0008-7049-5957], В. А. Малых[2] [0000-0002-4508-2527]

[1]Всероссийская академия внешней торговли, г. Москва, Россия

[1, 2]Университет ИТМО, г. Санкт-Петербург, Россия

[2]Международный университет информационных технологий, г. Алматы, Казахстан

[1]lyutovad@gmail.com, [2]valentin.malykh@phystech.edu

Аннотация

Представлен TradeNewsSum — корпус для абстрактивной генерации аннотаций к новостям внешней торговли, охватывающий русско- и англоязычные публикации из профильных источников. Все рефераты подготовлены вручную по унифицированным правилам. Проведены эксперименты с дообучением трансформерных и seq2seq-моделей и автоматическую оценку по схеме LLM-as-a-judge. Наилучшие результаты показала LLaMA 3.1 в режиме инструкционного промптинга, продемонстрировав высокие значения по метрикам, включая фактологическую полноту.

Ключевые слова: *абстрактивное реферирование, многоязычный корпус, новости внешней торговли, санкции, торговые режимы, TradeNewsSum, трансформеры, большие языковые модели, LLM-as-a-judge, NER-оценка сущностей.*

ВВЕДЕНИЕ

Автоматическое реферирование становится важным инструментом для обработки новостного потока, особенно в сфере внешней торговли, где требуются краткие и точные изложения сообщений о санкциях, соглашениях и торговых режимах. При этом существующие корпуса в основном англоязычные,

© Д. А. Лютова, В. А. Малых, 2025.
Данная статья распространяется на условиях международной лицензии Creative Commons License Attribution 4.0 International (CC BY 4.0).

тематически общие и редко содержат качественные абстрактивные аннотации на русском языке, что ограничивает обучение и валидацию многоязычных моделей. TradeNewsSum восполняет этот разрыв: корпус включает тексты на русском и английском языках, собранные с релевантных площадок, и вручную аннотированные краткие рефераты, пригодные для обучения, сравнения и прикладной оценки генерации.

Наш вклад состоит в создании, аннотировании и подробном описании корпуса TradeNewsSum, демонстрации его практической ценности на серии экспериментов с моделями seq2seq, трансформерными и большими языковыми моделями, в том числе в режиме инструкционного промптинга, и сравнительной оценке моделей как классическими метриками ROUGE, BERTScore, NER-F1, так и в парадигме LLM-as-a-judge.

Результаты подтверждают эффективность корпуса для обучения и оценки моделей в задачах автоматического реферирования в домене внешнеэкономических новостей.

ОБЗОР СУЩЕСТВУЮЩИХ РАБОТ

Для обучения и оценки абстрактивного реферирования широко применяются англоязычные корпуса CNN/DailyMail [20], XSum [14], Newsroom [8], NYT [18] и MultiNews [3], но они слабо адаптированы к внешнеэкономической тематике. В CNN/DailyMail аннотации сгенерированы автоматически и ограничены по качеству [21]; XSum дает однофразовые рефераты с частыми искажениями смысла; Newsroom (≈1.3 млн пар) объединяет разнородные источники и страдает позиционным смещением [11]; MultiNews (56 тыс. примеров) охватывает лишь английский язык [6].

Среди многоязычных наборов XL-Sum [10] (44 языка) фактически использует первое предложение статьи, а MassiveSumm [22] (92 языка) построен автоматически, содержит аннотации низкого качества. Русскоязычные Gazeta [9], RIA [6], Lenta.ru [25] в основном содержат заголовки или метатеги, без полноценных абстрактивных рефератов. MLSum [19] дает ручные аннотации на шести языках (включая русский), но без англоязычных параллелей.

Таким образом, датасеты ограничены тематически, лингвистически и структурно; почти нет полноценных многоязычных корпусов с аннотациями

по внешней экономике, рефераты часто фрагментарны или автоматически извлечены, кросс-языковая оценка затруднена. Русский язык в датасетах представлен слабо. Это снижает применимость существующих моделей к анализу внешнеторговой повестки и подчеркивает необходимость моделей, адаптированных к специализированным многоязычным корпусам.

С учетом этих ограничений особенно важны разработка и адаптация моделей под тематически специализированные и многоязычные корпуса. Ранние решения строились на Seq2Seq с вниманием [1, 17], но качество было достаточно невысоким. Улучшения дали Pointer-Generator [21] и обучение с подкреплением [15], уменьшив повторы и оптимизировав метрики напрямую. Переход к моделям transformer [23] и предобученным энкодер-декодерам (BART [12], T5 [16], PEGASUS [26]) обеспечил значительные результаты на CNN/DailyMail и XSum. Современный этап — это большие языковые модели (БЯМ), например GPT-3 [6], GPT-4, Claude, DeepSeek и др., способные решать задачу по инструкции [7]; при наличии пар предпочтений для донастройки используют DPO, что повышает качество суммаризации [24].

Качество обычно оценивают метриками лексико-семантического сходства — ROUGE [22], METEOR [2], BERTScore [27] — и через извлечение сущностей (NER) для фактологической полноты [4]. Набирают популярность методы с участием БЯМ (GPTScore, G-Eval и аналоги) с более высокой корреляцией с оценками людей [5, 13]; также применяют BLEURT, SummaC, FactCC, MAUVE. Большинство подходов разработано для английского языка и требуют адаптации к задачам на других, включая русский.

КОРПУС TradeNewsSum

В рамках проведенного исследования создан специализированный корпус TradeNewsSum[1], ориентированный на абстрактивное реферирование внешнеэкономических новостей. Корпус включает 59395 записей за 2020–2025 г. В него отбирались тексты с содержательной информацией о трансграничных экономических взаимодействиях (экспорт-импорт, инвестиции, санкции, логистические инициативы, гуманитарная помощь и др.).

[1] https://huggingface.co/datasets/lyutovad/TradeNewsSum

Каждая публикация снабжена кратким вручную аннотированным абстрактивным резюме и метаданными: языком оригинала, датой, ссылкой на источник и списком упомянутых стран.

Структура корпуса включает следующие поля:

text — исходный текст публикации,

summary_orig_lang — реферат на языке оригинала,

summary_translated — его перевод на второй язык,

orig_lang — язык оригинала (ru или en),

locations — список стран,

url — источник,

dates — дата публикации.

Корпус является двуязычным: 67% записей представлены на русском языке, 33% — на английском. Для обучения моделей и оценки производительности использовалось стандартное стратифицированное разбиение: обучающая, валидационная и тестовая выборки (80/10/10).

Особенность корпуса — высокая доля абстрактивных аннотаций, сформулированных вручную, что отличает его от большинства русскоязычных ресурсов, основанных на автоматических заголовках.

Сбор новостей проводился из 257 специализированных источников: государственных, агентских, отраслевых и деловых — по темам внешней торговли, санкций, инвестиций и макроэкономики. Основу корпуса составляют публикации на русском и английском языках с официальных сайтов, международных агентств (например, Reuters, Xinhua), деловых СМИ (РБК, «Коммерсантъ»), отраслевых платформ и агрегаторов (UN Comtrade, Eurasianet); актуальность и доступность источников регулярно проверялись вручную.

Тексты варьируются по сложности: от простых однотипных событий в одной стране до многокомпонентных материалов с несколькими странами, товарами и числовыми показателями. Соответственно, аннотации колеблются от кратких до развернутых (до 500–600 знаков), что позволяет обучать модели на разной степени контекстной насыщенности.

Каждая публикация сопровождается составленным специалистом рефератом по следующим формализованным правилам: информационно-деловой

стиль, акцент на сущностях (страны, товары, числовые значения), исключение вводных слов, цитат, оценочных суждений и ссылок; относительные формулировки времени заменяются точными датами по дате публикации. Аннотирование проходило в три этапа: первичная разметка, кросс-проверка и финальное утверждение с участием эксперта; при расхождениях применяется согласование. Все аннотациисоставлены экспертами без автогенерации и извлечения метаданных.

Для корпуса TradeNewsSum рассчитаны количественные характеристики по языкам (русский, английский) и сплитам (обучающая, валидационная, тестовая). В табл. 1 приведены длины текстов и рефератов (в словах и предложениях), показатели словарного разнообразия и лексическое перекрытие между текстами и рефератами.

Средняя длина англоязычных текстов существенно превышает русскоязычные: ≈360 слов против ≈175, при этом средняя длина рефератов остается стабильной — ≈53 слова для английского языка и ≈39 для русского. Количество предложений согласуется с длинами: англоязычные публикации содержат около 17 предложений, русскоязычные — около 11, тогда как рефераты на обоих языках состоят в среднем из 2.6–2.7 предложений. Коэффициент сжатия подтверждает различия: для английского он ниже (≈0.15) по сравнению с русским (≈0.22), что указывает на более агрессивное сжатие англоязычных изложений. В терминах словарного разнообразия английская часть корпуса богаче: абсолютные и средние показатели уникальных слов и лемм выше. Одновременно доля совпадающих лемм между текстом и рефератом больше в русскоязычной выборке, что свидетельствует о более экстрактивном характере русских аннотаций; англоязычные рефераты чаще используют перефразирование и обобщение, что важно учитывать при выборе и настройке моделей.

Для дополнительной оценки качества эталонных аннотаций была расчитана метрика сохранения сущностей (NER-precision, recall, F1). Результаты показывают высокую точность (≈0.85), при относительно низкой полноте (≈0.3), что отражает характер текстов абстрактов: в рефератах сохраняются ключевые акторы, но опускаются второстепенные детали. F1 этих показателей составляет 0.39–0.45.

Табл. 1. Статистика по текстам и рефератам: слова, предложения, леммы и пересечения

Метрика	Тренировочная часть		Валидационная часть		Тестовая часть	
	рус.	англ.	рус.	англ.	рус.	англ.
Число пар	32041	15475	4005	1934	4005	1935
Мин./Макс. слов (текст)	8/3335	7/4021	20/3101	11/2764	10/3134	9/2706
Мин./Макс. слов (реферат)	5/376	7/389	6/292	8/271	6/211	8/260
Ср. слов (текст)	176.0	361.5	173.5	362.9	178.4	358.4
Ср. слов (реферат)	39.0	52.8	38.4	53.9	39.6	52.9
Ср. предлож. (текст)	1.2	16.9	11.1	16.9	11.3	16.6
Ср. предлож. (реферат)	2.6	2.6	2.6	2.7	2.7	2.6
Коэф. сжатия[2]	0.222	0.146	0.221	0.149	0.222	0.148
УЛ[3] (реферат)	25760	21621	9792	8291	10005	8153
Совпадающие УЛ	25405	20898	9696	8107	9897	8042
Доля новых лемм[4]	0.014	0.033	0.01	0.022	0.011	0.014

[2] Коэффициент сжатия = Ср. слов (реферат) / Ср. слов (текст)

[3] УЛ – уникальные леммы

[4] Доля новых лемм = 1 - Совпадающие УЛ / УЛ (реферат)

Таким образом, TradeNewsSum представляет собой лингвистически разнородный корпус с элементами как экстрактивного, так и абстрактивного реферирования, что делает его подходящим для обучения seq2seq-моделей и оценки генерации в многоязычном контексте. Благодаря структуре, языковому охвату и качеству аннотаций он может использоваться в задачах реферирования, NER, графового анализа, мониторинга внешней торговли, а также в научных и образовательных целях.

ЭКСПЕРИМЕНТЫ И РЕЗУЛЬТАТЫ

Для оценки качества была использована комбинированная метрика, объединяющая показатели лексико-семантического сходства (ROUGE, METEOR, BERTScore) и фактологической полноты (NER):

$$\text{weighted} = \text{base}_{\text{score}} \cdot \text{ner}_{\text{coef}},$$

где

$$\text{base}_{\text{score}} = 0.20 \cdot \text{BERTScore}_{\text{F1}} + 0.15 \cdot \text{METEOR} + 0.05 \cdot \text{ROUGE}_1 + \\ + 0.10 \cdot \text{ROUGE}_2 + 0.15 \cdot \text{ROUGE}_{\text{Lsum}} + 0.30 \cdot \text{NER}_{\text{F1}},$$

$$\text{ner}_{\text{coef}} = \min\big(1.0, \max(0.6, 0.6 + 0.4 \cdot \text{NER}_{\text{R}})\big).$$

Предложенная метрика weighted учитывает и поверхностное, и семантическое совпадения с эталоном, а также полноту передачи ключевых сущностей, критически важную для новостных задач. Сглаженный штраф ner_{coef} повышает устойчивость к частичным пропускам и делает оценку более гибкой. Для оценки без дообучения была использована тестовая выборка TradeNewsSum (5940 текстов, RU+EN). Наилучшие результаты дали mbart-large-cc25 (EN) и mbart_ru_sum_gazeta (RU); pegasus-cnn_dailymail также показала хорошие результаты на английских данных, особенно по NER. ruGPT3 продемонстрировала минимальную точность. Полные значения приведены в табл. 2.

Табл. 2. Результаты моделей без дообучения на тестовой части корпуса

Модель	язык	weighted	$ROUGE_{Lsum}$	METEOR	$BERTScore_{F1}$	NER_{F1}
mT5_multilingual_XLSum	ru	0.2112	0.142	0.1849	0.8732	0.237
	en	0.2643	0.2336	0.1801	0.8913	0.338
mbart-large-cc25	ru	0.3867	0.2949	0.4318	0.9079	0.47
	en	0.4252	0.4086	0.4296	0.9084	0.494
mbart-large-50-many-to-many-mmt	ru	0.3098	0.2537	0.3756	0.8932	0.329
	en	0.1325	0.0685	0.0482	0.8361	0.067
mbart_ru_sum_gazeta	ru	0.4464	0.4007	0.5332	0.9259	0.528
pegasus-cnn_dailymail	en	0.3904	0.378	0.3524	0.9015	0.493
rugpt3large_based_on_gpt2	ru	0.17	0.0706	0.139	0.8408	0.118
	en	0.132	0.0741	0.0902	0.8217	0.045

Для повышения качества генерации были дообучены модели pointer_generator, mBART, NLLB, mT5 и LLaMA на 47516 парах новостей и аннотаций из корпуса TradeNewsSum. Для LLaMA3:8B-Instruct использовался режим инструкционного инференса без дообучения. Все модели демонстрируют значительное улучшение качества по сравнению с результатами «из коробки», особенно по метрикам ROUGE и NER_{F1}. Наивысших значений достигла LLaMA, подтвердив потенциал инструкционного подхода. Подробные результаты представлены в табл. 3.

Табл. 3. Результаты моделей после дообучения на корпусе TradeNewsSum

Модель	язык	weighted	$ROUGE_{Lsum}$	METEOR	$BERTScore_{F1}$	NER_{F1}
pointer_ generator	ru	0.2916	0.0505	0.2659	0.8261	0.391
	en	0.2559	0.0784	0.023	0.8211	0.396
mbart-large-50-many-to-many-mmt	ru	0.5991	0.5093	**0.7335**	0.9533	0.707
	en	0.5502	0.5427	0.5712	0.9344	0.643
nllb-200-distilled-600M	ru	0.5948	0.49	0.7265	**0.9528**	0.704
	en	0.5225	0.5178	0.5307	0.93	0.618
mT5_ multilingual_ XLSum	ru	0.4776	0.4143	0.5152	0.9451	0.65
	en	0.4539	0.4911	0.4162	0.9292	0.574
llama3.1:8b-instruct	ru	**0.6269**	**0.5406**	0.5718	0.9448	**0.728**
	en	**0.6099**	**0.583**	**0.6231**	**0.94**	**0.741**

Для оценки использовался подход LLM-as-a-judge, при котором другая БЯМ анализирует итоговую аннотацию на основе оригинального текста, правил генерации и заданной инструкции. Модель возвращала оценку по критериям точности (faithfulness), соблюдения структуры (structure_adherence), качества языка (style_and_grammar), наличия критических ошибок (critical_violations) и текстового комментария (comment). Такая схема позволяет формализованно оценить качество генерации и выявить потенциальные ошибки без участия человека.

Для снижения вычислительных затрат мы оценивали «медианные» и «сложные» случаи (длинные тексты, высокая числовая насыщенность, множество локаций). Для автоматической оценки итоговых аннотаций использовались две модели: GigaChat Lite и DeepSeek. Модель GigaChat обработала 89.5% примеров (10.5% отказов, главным образом из-за политически чувствительных сюжетов); средние оценки по точности, структуре и стилю превышали 4 балла, критические нарушения не выявлены. Модель DeepSeek оценила все

237 аннотаций, зафиксировав критические ошибки в 36% случаев — в основном из-за неполноты содержания и нарушений структуры (особенно в материалах о санкциях и конфликтах), при этом качество языка стабильно высокое (см. табл. 4).

Табл. 4. Средние оценки качества рефератов по результатам оценки GigaChat и DeepSeek

Критерий	GigaChat Lite	DeepSeek
Валидные ответы	212 (89.5%)	237 (100%)
Отказы от оценки	10.5%	0%
Точность передачи содержания	4.11	3.69
Структура и следование правилам	4.11	3.85
Язык и стиль	4.12	4.82
Критические нарушения	0%	36% (85 из 237)

Русскоязычные аннотации получают более высокие оценки по точности и структуре, а доля критических нарушений у них значительно ниже: 14.4% против 50.7% для англоязычных текстов. Это указывает на языковую асимметрию качества и подтверждает необходимость дополнительной постобработки англоязычных саммари (табл. 5).

Табл. 5. Сравнение качества рефератов на русском и английском языках

Показатель	Русский язык	Английский язык
Точность передачи содержания	4.23	3.32
Структура и следование правилам	4.11	3.66
Язык и стиль	4.89	4.82
Критические нарушения	14.4%	50.7%

ОГРАНИЧЕНИЯ И НАПРАВЛЕНИЯ РАЗВИТИЯ

Несмотря на высокую проработку, корпус TradeNewsSum имеет ряд ограничений по языковому охвату, тематике и структуре аннотаций. В текущей версии преобладают русско- и англоязычные материалы; доля французских и португальских текстов составляет менее 0.5% и не анализируется в основной части, что сужает возможности многоязычных исследований. Не все сущности (например, компании и товарные группы) отмечены явно, а аннотации не всегда охватывают все фактологические элементы. Отсутствует система версионирования правок, а стратификация по языку усложняет событийный анализ.

В дальнейшем планируется увеличить долю французских и португальских материалов и добавить испанские и китайские новости, ввести явную разметку ключевых сущностей, улучшить методологию аннотирования и внедрить контроль версий, что позволит сформировать полноформатный мультиязычный ресурс.

ЗАКЛЮЧЕНИЕ

Представлен специализированный корпус TradeNewsSum, ориентированный на генерацию аннотаций к новостям внешней торговли. Корпус охватывает русско- и англоязычные публикации из профильных источников и снабжен подготовленными экспертами рефератами по унифицированным правилам. Проведены эксперименты с дообучением трансформерных и seq2seq-моделей и автоматическая оценка качества по схеме LLM-as-a-judge. Наилучшие результаты продемонстрировала модель LLaMA 3.1 в режиме инструкционного промптинга, показав высокие значения по всем метрикам, включая фактологическую полноту.

Полученные результаты подтверждают применимость предложенного подхода к генерации кратких содержаний в профессиональной новостной повестке. Корпус может использоваться для задач построения дайджестов, анализа торговых потоков и графового моделирования международных отношений. Таким образом, TradeNewsSum является практико-ориентированным ресурсом для исследовательских и аналитических задач, а выявленные ограничения формируют направления его дальнейшего развития.

СПИСОК ЛИТЕРАТУРЫ

1. *Bahdanau D. et al*. End-to-end attention-based large vocabulary speech recognition // 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2016. P. 4945–4949.

2. *Banerjee S., Lavie A.* METEOR: An automatic metric for MT evaluation with improved correlation with human judgments // Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. 2005. P. 65–72.

3. *Fabbri A. R. et al.* Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model // arXiv preprint arXiv:1906.01749. 2019.

4. *Fischer T., Remus S., Biemann C.* Measuring faithfulness of abstractive summaries // Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022). 2022. P. 63–73.

5. *Fu J. et al.* Gptscore: Evaluate as you desire // arXiv preprint arXiv:2302.04166. 2023.

6. *Gavrilov D., Kalaidin P., Malykh V.* Self-attentive model for headline generation // Advances in Information Retrieval: 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14–18, 2019, Proceedings, Part II 41. Springer International Publishing, 2019. P. 87–93.

7. *Goyal T., Li J. J., Durrett G.* News summarization and evaluation in the era of gpt-3 // arXiv preprint arXiv:2209.12356. 2022.

8. *Grusky M., Naaman M., Artzi Y.* Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies // arXiv preprint arXiv:1804.11283. 2018.

9. *Gusev I.* Dataset for automatic summarization of Russian news // Artificial Intelligence and Natural Language: 9th Conference, AINL 2020, Helsinki, Finland, October 7–9, 2020, Proceedings 9. Springer International Publishing, 2020. P. 122–134.

10. *Hasan T. et al.* XL-sum: Large-scale multilingual abstractive summarization for 44 languages // arXiv preprint arXiv:2106.13822. 2021.

11. *Kryściński W. et al.* Neural text summarization: A critical evaluation // arXiv preprint arXiv:1908.08960. 2019.

12. *Lewis M. et al.* Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension // arXiv preprint arXiv:1910.13461. 2019.

13. *Liu Y. et al.* G-eval: NLG evaluation using gpt-4 with better human alignment // arXiv preprint arXiv:2303.16634. 2023.

14. *Narayan S., Cohen S. B., Lapata M.* Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization // arXiv preprint arXiv:1808.08745. 2018.

15. *Paulus R., Xiong C., Socher R.* A deep reinforced model for abstractive summarization // arXiv preprint arXiv:1705.04304. 2017.

16. *Raffel C. et al.* Exploring the limits of transfer learning with a unified text-to-text transformer // Journal of machine learning research. 2020. Vol. 21, No. 140. P. 1–67.

17. *Rush A.M., Chopra S., Weston J.* A neural attention model for abstractive sentence summarization // arXiv preprint arXiv:1509.00685. 2015.

18. *Sandhaus E.* The New York Times Annotated Corpus Overview [Electronic resource]. Philadelphia: Linguistic Data Consortium, 2008. (LDC Catalog No. LDC2008T19). https://gwern.net/doc/ai/dataset/2008-sandhaus.pdf (accessed: 21.05.2025).

19. *Scialom T. et al.* MLSUM: The multilingual summarization corpus // arXiv preprint arXiv:2004.14900. 2020.

20. *See A., Liu P. J., Manning C.D.* A Neural Attention Model for Abstractive Sentence Summarization [Electronic resource]. 2016. https://github.com/abisee/cnn-dailymail (accessed 07.04.2025).

21. *See A., Liu P.J., Manning C.D.* Get to the point: Summarization with pointer-generator networks // arXiv preprint arXiv:1704.04368. 2017.

22. *Varab D., Schluter N.* MassiveSumm: a very large-scale, very multilingual, news summarisation dataset // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021. P. 10150–10161.

23. *Vaswani A. et al.* Attention is all you need // Advances in neural information processing systems. 2017. Vol. 30.

24. *Xin L., Liutova D., Malykh V.* Cross-Language Summarization in Russian and Chinese Using the Reinforcement Learning // International Conference on Analysis of Images, Social Networks and Texts. Cham: Springer Nature Switzerland, 2024. P. 179–192.

25. *Yutkin M.* Lenta.Ru News Dataset [Electronic resource]. 2018. Available at: https://github.com/yutkin/Lenta.Ru-News-Dataset (accessed 04.05.2025).

26. *Zhang J. et al.* Pegasus: Pre-training with extracted gap-sentences for abstractive summarization // International conference on machine learning. PMLR, 2020. P. 11328–11339.

27. *Zhang T. et al.* Bertscore: Evaluating text generation with bert // arXiv preprint arXiv:1904.09675. 2019.

ABSTRACTIVE SUMMARIZATION FOR TRADE NEWS ANALYSIS BASED ON A NEW DOMAIN-SPECIFIC DATASET

D. A. Liutova[1] [0009-0008-7049-5957], V. A. Malykh[2] [0000-0002-4508-2527]

[1]Russian Foreign Trade Academy, Moscow, Russia
[1, 2]ITMO University, Saint Petersburg, Russia
[2]International IT University, Almaty, Kazakhstan
[1]lyutovad@gmail.com, [2]valentin.malykh@phystech.edu

Abstract

We present TradeNewsSum—a corpus for abstractive summarization of international trade news—covering Russian- and English-language publications from domain-specific sources. All summaries are manually prepared following unified guidelines. We conducted experiments with fine-tuning transformer and seq2seq models and performed automatic evaluation using the LLM-as-a-judge scheme. LLaMA 3.1 in instruction-prompting mode achieved the best results, showing high scores across metrics, including factual completeness.

Keywords: *abstractive summarization, multilingual corpus, international trade news, sanctions, trade regimes, TradeNewsSum, transformers, large language models, LLM-as-a-judge, NER-based entity evaluation.*

REFERENCES

1. *Bahdanau D. et al*. End-to-end attention-based large vocabulary speech recognition // 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2016. P. 4945–4949.

2. *Banerjee S., Lavie A.* METEOR: An automatic metric for MT evaluation with improved correlation with human judgments // Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. 2005. P. 65–72.

3. *Fabbri A. R. et al.* Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model // arXiv preprint arXiv:1906.01749. 2019.

4. *Fischer T., Remus S., Biemann C.* Measuring faithfulness of abstractive summaries // Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022). 2022. P. 63–73.

5. *Fu J. et al.* Gptscore: Evaluate as you desire // arXiv preprint arXiv:2302.04166. 2023.

6. *Gavrilov D., Kalaidin P., Malykh V.* Self-attentive model for headline generation // Advances in Information Retrieval: 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14–18, 2019, Proceedings, Part II 41. Springer International Publishing, 2019. P. 87–93.

7. *Goyal T., Li J. J., Durrett G.* News summarization and evaluation in the era of gpt-3 // arXiv preprint arXiv:2209.12356. 2022.

8. *Grusky M., Naaman M., Artzi Y.* Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies // arXiv preprint arXiv:1804.11283. 2018.

9. *Gusev I.* Dataset for automatic summarization of Russian news // Artificial Intelligence and Natural Language: 9th Conference, AINL 2020, Helsinki, Finland, October 7–9, 2020, Proceedings 9. Springer International Publishing, 2020. P. 122–134.

10. *Hasan T. et al.* XL-sum: Large-scale multilingual abstractive summarization for 44 languages // arXiv preprint arXiv:2106.13822. 2021.

11. *Kryściński W. et al.* Neural text summarization: A critical evaluation // arXiv preprint arXiv:1908.08960. 2019.

12. *Lewis M. et al.* Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension // arXiv preprint arXiv:1910.13461. 2019.

13. *Liu Y. et al.* G-eval: NLG evaluation using gpt-4 with better human alignment // arXiv preprint arXiv:2303.16634. 2023.

14. *Narayan S., Cohen S. B., Lapata M.* Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization // arXiv preprint arXiv:1808.08745. 2018.

15. *Paulus R., Xiong C., Socher R.* A deep reinforced model for abstractive summarization // arXiv preprint arXiv:1705.04304. 2017.

16. *Raffel C. et al.* Exploring the limits of transfer learning with a unified text-to-text transformer // Journal of machine learning research. 2020. Vol. 21, No. 140. P. 1–67.

17. *Rush A.M., Chopra S., Weston J.* A neural attention model for abstractive sentence summarization // arXiv preprint arXiv:1509.00685. 2015.

18. *Sandhaus E.* The New York Times Annotated Corpus Overview [Electronic resource]. Philadelphia: Linguistic Data Consortium, 2008. (LDC Catalog No. LDC2008T19). https://gwern.net/doc/ai/dataset/2008-sandhaus.pdf (accessed: 21.05.2025).

19. *Scialom T. et al.* MLSUM: The multilingual summarization corpus // arXiv preprint arXiv:2004.14900. 2020.

20. *See A., Liu P. J., Manning C.D.* A Neural Attention Model for Abstractive Sentence Summarization [Electronic resource]. 2016. https://github.com/abisee/cnn-dailymail (accessed 07.04.2025).

21. *See A., Liu P.J., Manning C.D.* Get to the point: Summarization with pointer-generator networks // arXiv preprint arXiv:1704.04368. 2017.

22. *Varab D., Schluter N.* MassiveSumm: a very large-scale, very multilingual, news summarisation dataset // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021. P. 10150–10161.

23. *Vaswani A. et al.* Attention is all you need // Advances in neural information processing systems. 2017. Vol. 30.

24. *Xin L., Liutova D., Malykh V.* Cross-Language Summarization in Russian and Chinese Using the Reinforcement Learning // International Conference on Analysis of Images, Social Networks and Texts. Cham: Springer Nature Switzerland, 2024. P. 179–192.

25. *Yutkin M.* Lenta.Ru News Dataset [Electronic resource]. 2018. Available at: https://github.com/yutkin/Lenta.Ru-News-Dataset (accessed 04.05.2025).

26. *Zhang J. et al.* Pegasus: Pre-training with extracted gap-sentences for abstractive summarization // International conference on machine learning. PMLR, 2020. P. 11328–11339.

27. *Zhang T. et al.* Bertscore: Evaluating text generation with bert // arXiv preprint arXiv:1904.09675. 2019.

СВЕДЕНИЯ ОБ АВТОРАХ

ЛЮТОВА Дарья Андреевна — выпускница магистратуры Университета ИТМО 2025 года по направлению «Искусственный интеллект», аспирантка ИТМО. Исследователь в области обработки естественного языка и аналитики новостных потоков. Область научных интересов: большие языковые модели и методы обработки текста.

Daria LYUTOVA — M.Sc. graduate (2025) in Artificial Intelligence from ITMO University and a Ph.D. student at ITMO. She is a researcher in natural language processing and news analytics. Research interests include large language models and natural language processing.

email: lyutovad@gmail.com
ORCID: 0009-0008-7049-5957

Малых Валентин Андреевич, закончил МФТИ в 2009 году. В 2019 году защитил кандидатскую диссертацию по специальности 05.13.11. В настоящее время является доцентом ВШЦК ИТМО, а также профессором-исследователем в МУИТ. Область научных интересов: обработка текстов, большие языковые модели.

Valentin Malykh graduated from MIPT in 2009. In 2019, he defended his PhD in technical sciences. Valentin is currently an assistant professor at the Digital Culture department, ITMO University and research professor at IITU University. Research interests: natural language processing, large language models.

email: valentin.malykh@phystech.edu
ORCID: 0000-0002-4508-2527

Материал поступил в редакцию 11 октября 2025 года

УДК 004.81

ИССЛЕДОВАНИЕ КВАНТОВАНИЯ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ: ОЦЕНКА ЭФФЕКТИВНОСТИ С АКЦЕНТОМ НА РУССКОЯЗЫЧНЫЕ ЗАДАЧИ

Д. Р. Пойманов[1] [0009-0001-5390-915X], М. С. Шутов[2] [0009-0009-0530-5034]

[1]*Московский государственный университет им. М. В. Ломоносова, г. Москва, Россия*

[2]*Московский физико-технический институт (национальный исследовательский университет), г. Москва, Россия*

[1]poimanovdr@my.msu.ru, [2]mihailshutov105@gmail.com

Аннотация

Квантование стало ключевой техникой сжатия и ускорения больших языковых моделей (LLM). Несмотря на то, что исследования низкобитного квантования активно развиваются применительно к англоязычным LLM, его влияние на морфологически богатые и разнородные по ресурсам языки, включая русский, остается изученным значительно хуже. Поэтому требуются дополнительные исследования этого вопроса в связи с развитием высокоэффективных русскоязычных и многоязычных LLM.

Мы провели систематическое исследование квантования предобученных моделей в эффективные 2.0—4.25 бита на параметр для современных русскоязычных LLM различного масштаба от 4 до 32 млрд параметров (4 B и 32 B). Экспериментальная часть охватывает как стандартное равномерное квантование, так и специализированные низкобитные форматы. Полученные результаты выявили несколько ключевых тенденций: i) устойчивость русскоязычных LLM к квантованию варьируется в зависимости от архитектуры и размера модели; ii) 4-битное квантование демонстрирует высокую надежность, особенно при использовании продвинутых форматов; iii) 3-битное и 2-битное квантования оказались наиболее чувствительными к указанным калибровки. Полученные эмпирические данные демонстрируют необходимость учета домена модели при использовании различных методов квантования.

© Д. Р. Пойманов, М. С. Шутов, 2025.
Данная статья распространяется на условиях международной лицензии Creative Commons License Attribution 4.0 International (CC BY 4.0).

***Ключевые слова**: квантование нейросетей, сжатие и оптимизация больших языковых моделей.*

ВВЕДЕНИЕ

Большие языковые модели (LLM) сегодня выступают ключевым инструментом в обработке естественного языка, обеспечивая передовые результаты в таких задачах, как ответы на вопросы [1], ведение диалога [2], генерация кода [3] и рассуждения [4–6]. Однако стремительное увеличение размера моделей — до десятков и сотен миллиардов параметров — сопровождается резким ростом потребностей в вычислительных ресурсах и памяти, что делает проблему эффективного развертывания одной из центральных для современной исследовательской повестки. Одним из наиболее эффективных подходов к решению проблемы является квантование, при котором веса моделей преобразуются в более компактные представления. Такой подход позволяет существенно ускорить процесс инференса и снизить затраты памяти [7–10]. Недавние разработки методов квантования, включая GPTQ, AWQ, SmoothQuant и QTIP, показали, что при использовании специализированных алгоритмов квантования возможно сохранить высокую производительность даже в условиях агрессивного сжатия [9–11].

Несмотря на значительные достижения в области квантования нейросетей, большинство исследований сосредоточено на англоязычных или многоязычных LLM, что оставляет существенный пробел в изучении моделей для русского языка. В последние годы российское NLP-сообщество представило ряд крупномасштабных открытых моделей, включая T-Pro 2.0 [12], YaGPT [13] и RuAdaptQwen [14]. Эти разработки расширяют и адаптируют возможности многоязычных базовых архитектур, таких как Qwen [15], к задачам, ориентированным на русский язык. Хотя названные модели демонстрируют конкурентоспособные результаты на ряде бенчмарков, они, как правило, распространяются в форматах полной точности (FP16 или BF16), а доступные версии с квантованием, если они вообще присутствуют, нередко сопровождаются заметной деградацией качества. Это ограничивает практическое применение русскоязычных LLM в условиях дефицита ресурсов — например, на мобильных устройствах или в промышленных системах с требованиями к низкой задержке по памяти.

Особый интерес и актуальность исследования подкрепляются тем, что ведущие исследовательские группы все чаще выпускают модели в квантованных форматах по умолчанию. Так, компания OpenAI представила оптимизированные 4-битовые версии своих моделей [16], а в DeepSeek показано [17], что и обучение, и инференс модели непосредственно с использованием 8-битной арифметики позволяют достигать сопоставимой производительности при существенном росте эффективности. Эти тенденции свидетельствуют о том, что в будущем практическое внедрение LLM может опираться не столько на модели с высокой числовой точностью, сколько на тщательно спроектированные низкобитные представления. В то же время для русскоязычных LLM данное направление остается недостаточно изученным: отсутствует систематическая оценка того, как различные методы квантования и разрядность влияют на качество моделей при решении ключевых лингвистических и логических задач.

В настоящей работе мы стремимся восполнить этот пробел, проводя всестороннее исследование влияния квантования на LLM, адаптированные к русскому языку. Особое внимание уделено моделям, охватывающим различные масштабы и архитектуры, а именно на Qwen3-4B, RuAdaptQwen3-4B, Qwen3-32B, RuAdaptQwen3-32B и T-Pro 2.0-32B. Для каждой из названных моделей мы провели серию экспериментов по квантованию: i) скалярное равномерное квантование после обучения (Post Training Quantization, PTQ) в 4, 3 и 2 бита на параметр; ii) векторное квантование в эффективные 2 бита на вес методом QTIP и iii) квантование весов модели в специализированные форматы MXFP4 и MXINT4. Оценка деградации модели проведена на бенчмарках, включающих общеязыковые (PIQA, MMLU) и русскоязычные (PIQA-RU, MMLU-RU) наборы данных, что позволило выявить взаимосвязь между стратегией квантования, битностью и языковой адаптацией.

Статья состоит из двух основных частей.

1. В первой части представлено первое систематическое исследование квантования для LLM, адаптированных для русского языка, в котором освещены актуальные проблемы, а также проведено сравнение с моделями, ориентированными на английский язык.

2. Вторая часть посвящена детальному сравнительному анализу компромиссов между различными подходами к квантованию, включая скалярное квантование с калибровкой, агрессивное векторное квантование и наивное квантование в специализированные форматы. При этом было учтено влияние двух типов калибровочных данных — англоязычного корпуса RedPajama и русскоязычного T-Wix.

Мы считаем, что настоящее исследование формирует методологическую основу для последующей разработки эффективных методов адаптации и внедрения специализированных больших языковых моделей с учетом различных лингвистических особенностей.

ОБЗОР СУЩЕСТВУЮЩИХ МЕТОДОВ

Современные методы квантования можно разделить на две группы в зависимости от стратегии сжатия (PTQ и QAT) и методологии сжатия (скалярное и векторное квантование). PTQ (Post-Training Quantization, квантование после обучения) относится к подходам, при которых модель сжимается после того, как она уже прошла обучение. Методы PTQ являются высокоэффективными с точки зрения вычислительных затрат, поскольку дорогостоящий этап обучения уже завершен. QAT (Quantization-Aware Training, Обучение с учетом квантования) — это обучение, как правило, всех параметров модели с имитацией эффектов квантования. Хотя подходы QAT обычно обеспечивают более высокую эффективность сжатия, они требуют значительно больше вычислительных ресурсов даже в сравнении с обучением без квантования.

В настоящем исследовании мы делаем акцент на PTQ-методах. Рассмотрим сначала ключевые принципы PTQ, включая адаптивное округление (adaptive rounding), поблочную оптимизацию (block-wise fine-tuning), эффективное дообучение модели (parameter-efficient end-to-end fine-tuning or PEFT), а также алгоритмы скалярного и векторного квантования.

Адаптивное округление (*Adaptive Rounding*) представляет собой метод посттрейнинговой квантизации, при котором вместо стандартного округления к ближайшему значению используется итеративный алгоритм, поэтапно кванти-

зирующий подмножество весов слоя LLM. На каждом шаге алгоритм минимизирует ошибки, возникающие в выходных активациях слоя вследствие квантизации на предыдущих итерациях. Наиболее известные подходы к квантизации LLM с использованием адаптивного округления — GPTQ [8] и LDLQ [18] — рассматривают выход каждого линейного слоя в качестве локальной целевой функции при минимизации ошибки квантизации, что позволяет проводить процедуру квантизации модели параллельно и без большой калибрационной выборки.

Блочный PTQ [19, 20] представляет собой продвинутую технику дистилляции знаний, при которой блоки квантованной модели (студента) обучаются на основе соответствующих блоков оригинальной модели (учителя) с использованием функции потерь, наложенной на активации. Такой подход значительно сокращает вычислительный граф в процессе оптимизации и, как следствие, снижает затраты ресурсов.

Эффективное дообучение. После этапа блочного PTQ используют метод PEFT [21, 22] для восстановления качества квантованной модели и оптимизации ее производительности. В отличие от традиционного обучения, которое корректирует все параметры сети, PEFT фокусируется только на небольшом поднаборе важных параметров, которые в наибольшей степени влияют на точность вывода модели.

В контексте квантования LLM наибольший интерес представляет квантование весов линейных слоев, поскольку эти слои содержат большую часть параметров модели (обычно >95% весов модели). Квантование линейных слоев позволяет существенно снизить как вычислительные затраты на умножение матриц, так и объем занимаемой памяти. В настоящее время наиболее широко используются методы скалярного квантования, однако в условиях низкобитного сжатия (2 бита на вес или меньше) методы векторного квантования позволяют получать меньшую просадку в точности.

Скалярное квантование. При скалярном квантовании (Scalar quantization) каждое вещественное значение параметра модели заменяется на значение из дискретного множества уровней квантования. Таким образом, непрерывное пространство параметров аппроксимируется конечным набором возможных значений, что позволяет значительно сократить объем хранимых данных. В контексте

сжатия нейронных сетей задача квантования заключается в минимизации искажения, возникающего при замене исходных весов их квантованными аналогами. В простейшем случае минимизируется ошибка, определяемая как разность между исходным и квантованными параметрами, однако при низкобитном сжатии становится необходим учет изменения отклика модели на калибровочном наборе данных, чтобы сохранить качество предсказаний.

В процессе скалярного квантования каждому параметру сопоставляется индекс соответствующего уровня квантования, который сохраняется в низкобитном формате. Например, при 4-битном квантовании два параметра могут быть закодированы в одном байте памяти. Это обеспечивает компактное хранение параметров и возможность восстановления приближенного вектора весов модели при необходимости.

Векторное квантование. В отличие от скалярного квантования, векторное квантование (*Vector Quantization, VQ*) предполагает квантование целых векторов весов нейронной сети, а не отдельных скалярных параметров. Каждая группа весов квантуемого слоя (вектор весов) заменяется одним из векторов из заранее определенного набора — кодовой книги (*codebook*). В результате квантования каждому вектору весов сопоставляется индекс выбранного вектора кодовой книги, что позволяет эффективно хранить параметры в сжатом виде. В результате достигается значительное уменьшение объема модели при умеренной потере точности. В настоящее время ведущими методами низкобитного квантования LLM (эффективные 2 бита на вес и ниже) с точки зрения компромисса между сжатием и качеством модели являются именно методы векторного квантования — AQLM [23], Quip# [24] и QTIP [25].

В нашем экспериментальном исследовании мы использовали методы как скалярного, так и векторного квантования. Мы реализовали методологии, представленные в оригинальных работах, однако скорректировали некоторые гиперпараметры и наборы калибрационных датасетов для соответствия целям исследования.

ЭКСПЕРИМЕНТЫ

Нами были использованы открытые языковые модели и наборы данных на русском и английском языках. Для анализа влияния различных методов квантования мы рассмотрели как подходы, основанные на калибровке (EfficientQAT, QTIP), так и методы без калибровки (квантование весов в форматы Microscaling). В качестве оценки деградации моделей после квантования были использованы открытые англоязычные и русскоязычные бенчмарки.

Модели

Оценим разнообразный набор современных больших языковых моделей, включая как мультиязычные, так и варианты, адаптированные для русского языка. В частности, рассмотрим модель Qwen3-4B [26] и ее адаптированную для русского языка версию RuAdaptQwen3-4B [27], представляющие собой компактные модели на основе архитектуры трансформер для решения задач обработки естественного языка. В качестве больших моделей исследуем Qwen3-32B [26] и T-Pro-2.0-32 [21], причем последняя является одной из самых мощных открытых LLM, ориентированных на русский язык и выпущенных на данный момент. Такой выбор позволяет нам сравнить эффекты квантования для моделей различных размеров (4 B и 32 B) и для разных языковых доменов (многоязычные и оптимизированные для русского языка модели).

Методы квантования нейросетей

В экспериментах по квантованию больших языковых моделей мы применили несколько методов, которые продемонстрировали свою эффективность как в академических исследованиях, так и в промышленных приложениях, в частности, скалярное квантование на основе Learning Step Quantization (LSQ) в рамках EfficientQAT [19], преобразование весов модели в числовой формат Microscaling [28] и QTIP [25], современный метод векторного квантования, который использует trellis сжатие для максимально эффективного распределения битов при векторном квантовании. Краткие описания каждого метода даны ниже (см. рис. 1–3).

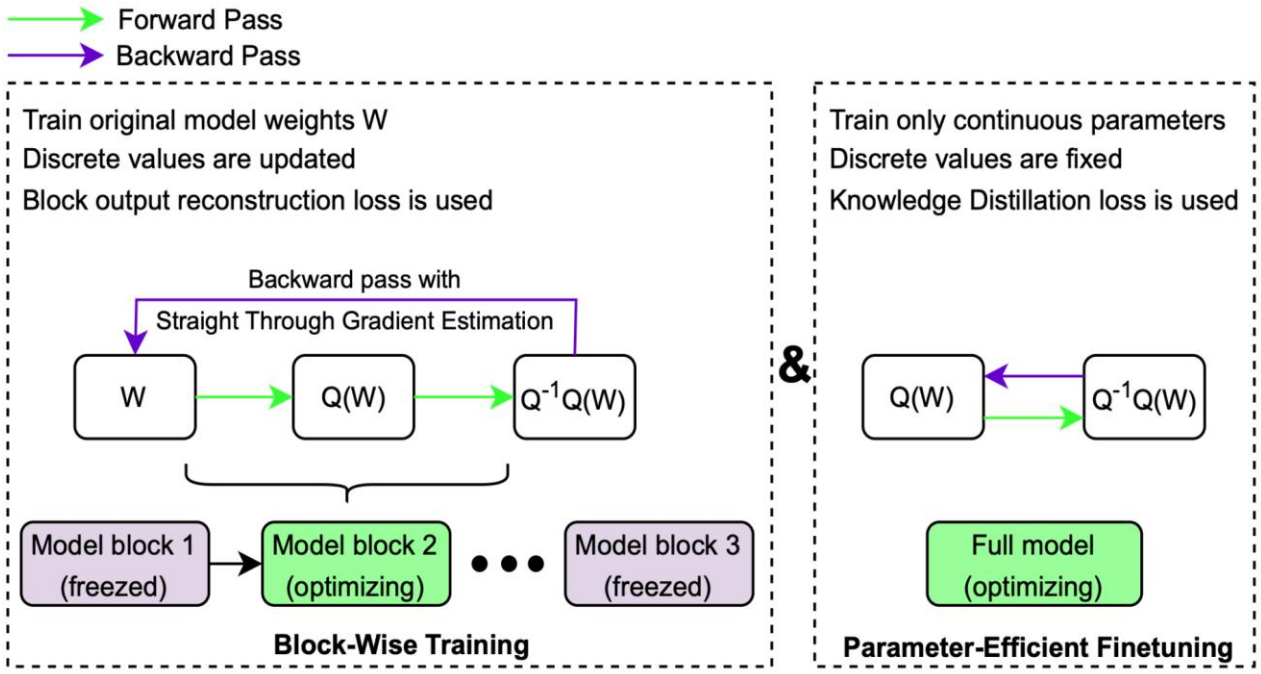


Рис. 1. Пайплайн квантования EfficientQAT [19]. Посттрейнинговое квантование предполагает 2 этапа: 1) поблочная дистилляция знаний и 2) эффективное по затрачиваемым ресурсам дообучение модели на выходы оригинальной модели.

EfficientQAT — это метод обучения, изначально разработанный для скалярного квантования и основанный на двухэтапной схеме: i) блочная дистилляция знаний из оригинальной модели (учителя) в квантованную (ученик) и ii) дообучение ограниченного числа параметров (PEFT) квантованной модели. Такой подход представляет собой ресурсоэффективный вариант постобучающего квантования, позволяющий сохранять качество практически без потерь даже при агрессивном квантовании, требуемом при жестких вычислительных ограничениях.

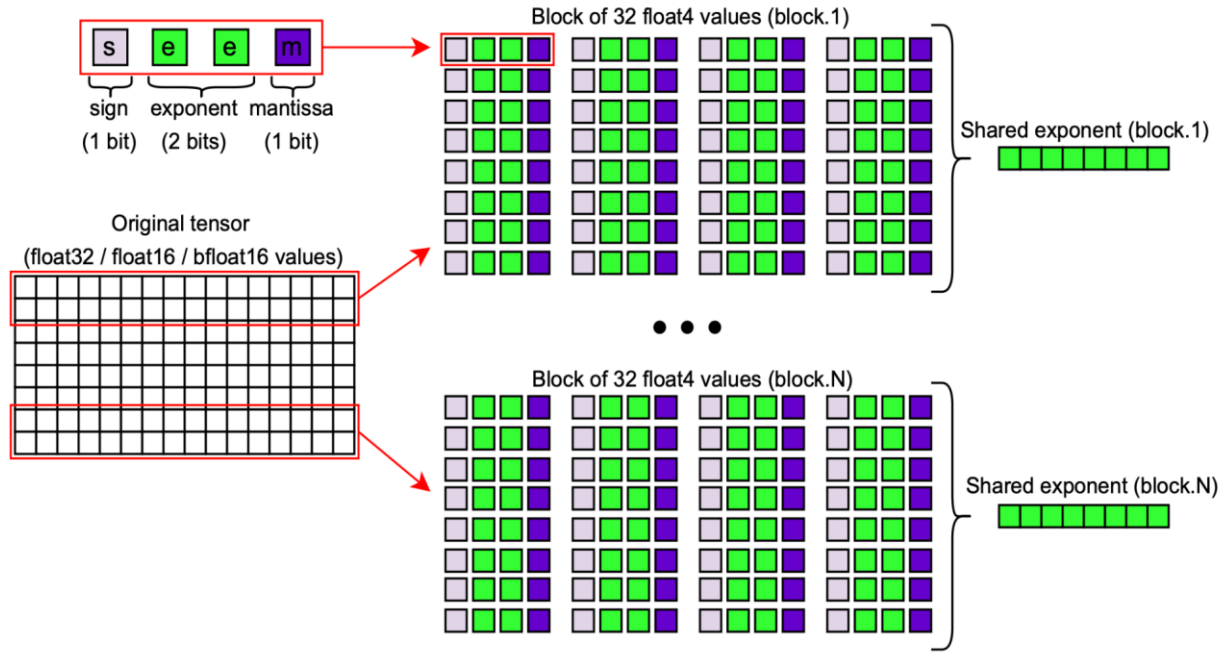


Рис. 2. Визуализация квантования тензора bfloat16 значений в MXFP4 формат [28]. Каждый блок из 32 оригинальных весов представляется в виде 32 float4 чисел и одной общей экспоненты.

Microscaling формат данных. Формат MX (Microscaling) представляет собой специализированное числовое представление, разработанное для снижения объемов памяти и ускорения инференса. Матрица в таком формате разделена на группы элементов, каждый из которых хранится в низкоразрядном формате (например, FP4 или INT4), при этом каждой группе присваивается коэффициент масштабирования, являющийся общей экспонентой для всех элементов группы. Формат MX нативно поддерживается в последних поколениях графических процессоров NVIDIA (например, в архитектуре Blackwell), что упрощает как процесс сжатия, так и аппаратное выполнение вычислений.

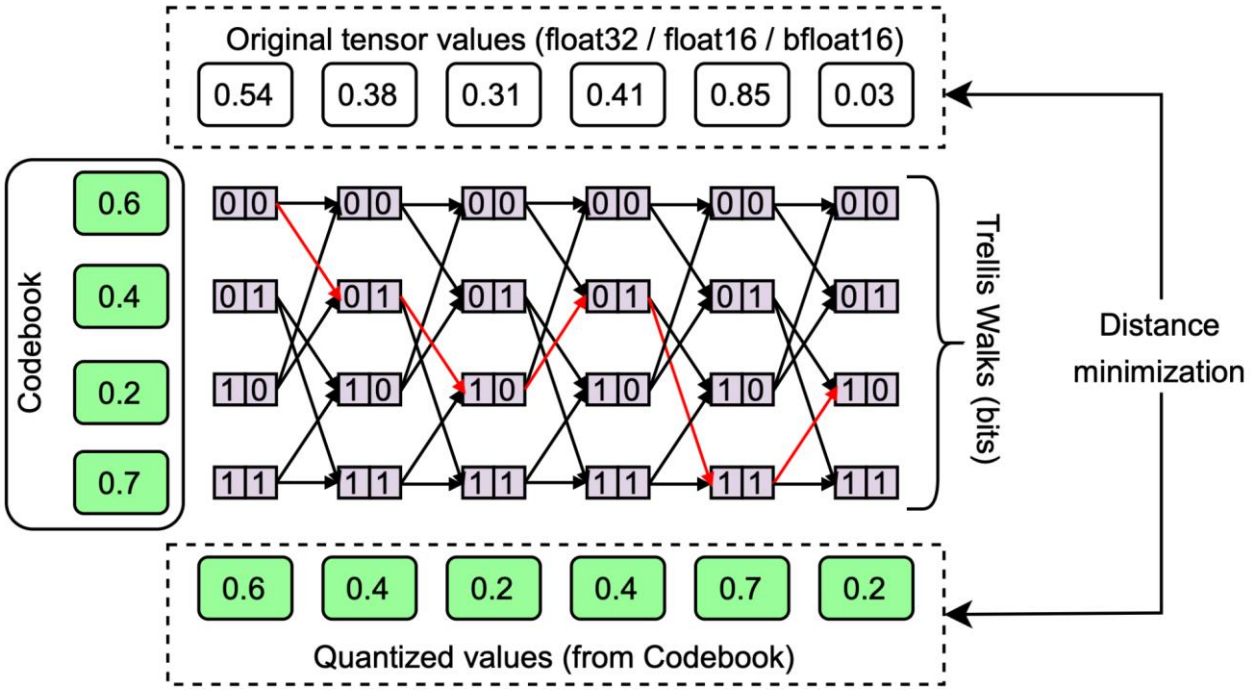


Рис. 3. Визуализация получения квантованных весов методом векторного квантования QTIP [25]. Индекс вектора формируется так, что биты соседних скаляров в векторе переиспользуются. Для получения оптимальной конфигурации используется алгоритм Витерби (Viterbi).

QTIP представляет собой современный метод векторного квантования, в котором используются умножения весов и активаций LLM на рандомизированные матрицы Адамара для снижения числа выбросов, а затем применяется высокоэффективное векторное квантование преобразованных весов с использованием техники адаптивного округления BlockLDLQ. Такой подход позволяет сжимать LLM до эффективных 2 бит на вес при сохранении качества генерации текстов сравнимым с оригинальной моделью. На данный момент QTIP можно рассматривать как одну из наиболее передовых технологий векторного PTQ для крупномасштабных LLM.

Во всех наших экспериментах квантованию подвергались только веса линейных слоев трансформер-блоков в соответствии с наиболее распространенными практиками квантования больших языковых моделей. Мы изучили как скалярное квантование (равномерное квантование в 2, 3 и 4 бита, Microscaling-фор-

маты), так и низкобитное векторное квантование (QTIP). В случае скалярного квантования в 3 и 4 бита мы применили групповое масштабирование и смещение весов с размером группы 128 параметров, что дало эффективные 3.25 и 4.25 бит на вес соответственно. При квантовании в скалярные 2 бита мы использовали размер группы в 64 элемента (эффективные 2.5 бит на вес), а также стандартные Microscaling форматы (MXFP4 и MXINT4) с группами по 32 элемента с компактными масштабами (эффективные 4.25 бит на вес). Векторное квантование QTIP реализует наиболее агрессивное сжатие — до эффективных 2.0 бит на вес.

Данные для обучения

Были использованы два набора данных для калибровки. RedPajama [29] — это открытый корпус для предварительного обучения, созданный с целью воспроизведения и расширения исходного набора веб-источников и систематизации знаний (CommonCrawl, C4, GitHub, Wikipedia, Books, arXiv, StackExchange). Первоначально он был выпущен как «чистый» набор данных объемом 1,2 ТБ (V1), а позднее расширен до версии RedPajama-V2, включающей 30 ТБ отфильтрованных токенов (более 100 ТБ в необработанном виде) на пяти языках и дополненной метаданными качества для отбора. В наших экспериментах в качестве калибровочного корпуса мы использовали 8 млн токенов RedPajama (~40 Mb текста), преимущественно англоязычных, отражающих широкий спектр веб-ресурсов и источников знаний.

Другой набор, T-Wix [30], представляет собой российский набор данных для дообучения нейросетей, ориентированный на выполнение инструкций и задач по рассуждению на определенные темы. Корпус включает два раздела: «Общий» (468 тыс. примеров, охватывающих математику, естественные науки, программирование, общие знания, ролевые сценарии и др.) и «Рассуждения» (31 тыс. примеров с подробными пошаговыми трассировками решений). В наших экспериментах T-Wix был использован для калибровки моделей в соответствии с русскоязычным распределением данных и форматами инструкций, что позволило сопоставимо оценить влияние калибровки при различных режимах квантования.

Бенчмарки

Для оценки мы применили комбинацию тестов, а именно: вычислили перплексию квантованных языковых моделей и их точность ответов на вопросы различной сложности. Для оценки качества языкового моделирования были задействованы корпус WikiText [31] и его русский аналог WikiText-Ru [32], которые обеспечивают стандартизированную оценку перплексии модели на естественном тексте. Для оценки эффективности рассуждений и ответов на вопросы были использованы MMLU [33] и PIQA [34], два широко используемых английских бенчмарка, охватывающих многозадачные знания и рассуждение на общие темы. Кроме того, мы включили их русские адаптации — MMLU-Ru [35] и PIQA-Ru [36] — для специальной проверки устойчивости моделей на русском языке. Для запуска тестов была применена стандартизированная библиотека lm-evaluation-harness [37]. В совокупности этот набор тестов позволил нам измерить как общую перплексию, так и зависимость точности выполнения конкретных задач в зависимости от языка.

РЕЗУЛЬТАТЫ

Представим результаты для четырех моделей Qwen3-4B, RuAdaptQwen3-4B, Qwen3-32B и T-Pro 2.0-32B в четырех режимах квантования: скалярное квантование методом LSQ от 2 до 4 бит и векторное квантование QTIP в режиме 2 бита на вес. Для каждого режима калибровка выполнялась с использованием двух текстовых корпусов (RedPajama и T-Wix) по 8 M токенов. Показатели качества включают в себя метрику качества языковых моделей – перплексию – на датасете WikiText (с предложениями на английском языке) и WikiText-RU (с предложениями, переведенными на русский язык), а также долю правильных ответов на выбранных QA (Question Answering) бенчмарках MMLU, MMLU-Ru, PIQA и PIQA-Ru для тестирования качества. Сначала обсудим перплексию, потом точность ответов на вопросы, а затем проанализируем зависимость качества моделей от калибровочных данных, отдельно уделив внимание влиянию векторного и скалярного квантования при 2 битах. Полные результаты представлены в табл. 1–3.

Во всех четырех моделях скалярное квантование до 4 бит на вес в значительной степени сохраняло качество работы моделей (метрика близка к модели

в оригинальном BF16-формате) как на корпусе WikiText, так и на WikiText-RU. В частности, модель T-Pro-IT-2.0-32B демонстрирует наименьший рост перплексии, оставаясь близкой к полной точности на обоих корпусах, в то время как модели объемом 4 B оказались несколько более чувствительными, особенно на корпусе WikiText-RU. Модель Qwen3-32B проявила такую же устойчивость, характерную для крупных моделей: ее перплексия в 4-битном режиме аномально снижается на русскоязычных данных. В целом квантование до 4 битов является надежным вариантом для развертывания всех исследуемых моделей, при этом относительное отклонение в качестве для вариантов объемом 4 B больше на корпусе WikiText-RU, чем на WikiText.

Снижение битности до 3 бит приводит к значительному росту перплексии на обоих корпусах у всех моделей. Снижение качества более выражено на корпусе WikiText-RU, что указывает на повышенную чувствительность к квантованию для русского языка. Большие модели (Qwen3-32B, T-Pro-IT-2.0-32B) сохраняют бóльшую стабильность по сравнению с Qwen3-4B, однако разрыв в качестве между английским и русским языками увеличивается по сравнению с 4-битным режимом. Это позволяет предположить, что морфологическая сложность русского языка и различия в токенизации усиливают шум, вызванный агрессивным квантованием.

Табл. 1. Результаты оценки квантованных моделей с числом параметров 4 B

Точность	Данные	W2	W2-Ru	MMLU	MMLU-Ru	PIQA	PIQA-Ru
Qwen3-4B							
FP16	–	11.7	6.94	70.0	62.1	76.3	64.3
scalar 4 bit	RedPaj.	12.3	7.34	68.5	61.0	75.7	63.4
	T-Wix	12.5	7.27	68.8	61.3	75.8	63.7
scalar 3 bit	RedPaj.	15.9	9.10	63.2	54.2	74.5	60.7
	T-Wix	15.9	8.78	61.6	54.2	74.4	62.7

scalar 2 bit	RedPaj.	14.8	12.0	48.6	38.6	70.8	56.7
	T-Wix	18.5	8.56	47.2	41.6	70.2	60.4
QTIP 2 bit	RedPaj.	13.5	11.5	52.7	37.7	73.1	57.5
	T-Wix	19.9	8.21	49.6	42.2	72.9	61.8
RuadaptQwen3-4B							
FP16	–	9.09	11.0	68.9	62.6	77.5	67.7
scalar 4 bit	RedPaj.	9.44	11.6	67.4	60.9	77.4	66.6
	T-Wix	9.49	11.6	68.0	60.7	77.3	67.3
scalar 3 bit	RedPaj.	11.2	14.5	61.6	54.2	75.4	64.3
	T-Wix	11.4	14.0	59.9	52.2	75.9	64.5
scalar 2 bit	RedPaj.	13.5	24.8	48.3	37.3	71.9	59.3
	T-Wix	16.4	18.9	45.2	39.6	70.3	61.9
QTIP 2 bit	RedPaj.	12.4	19.8	49.0	38.3	67.9	53.1
	T-Wix	14.6	17.9	45.7	41.4	69.6	55.4

Табл. 2. Результаты оценки квантованных моделей с числом параметров 32 B

Точность	Данные	W2	W2-Ru	MMLU	MMLU-Ru	PIQA	PIQA-Ru
Qwen3-32B							
FP16	–	6.67	4.44	81.9	76.6	81.9	70.5
scalar 4 bit	RedPaj.	6.80	4.38	81.2	76.0	81.7	70.0
	T-Wix	6.82	4.37	81.0	76.1	81.9	70.4
scalar 3 bit	RedPaj.	7.59	4.88	80.3	74.4	80.9	69.2
	T-Wix	7.65	4.78	80.0	74.5	81.7	70.4
scalar 2 bit	RedPaj.	9.88	6.04	72.1	64.2	77.3	65.1
	T-Wix	9.24	5.24	72.0	67.2	78.1	67.1
QTIP 2 bit	RedPaj.	9.32	12.4	71.2	65.1	77.2	66.7
	T-Wix	13.0	7.99	70.4	66.5	77.9	67.1
T-pro-it-2.0							
FP16	–	5.53	6.96	83.6	78.7	81.9	71.2
scalar 4 bit	RedPaj.	5.67	7.18	83.1	77.1	81.8	71.0
	T-Wix	5.66	7.21	82.2	77.9	81.9	71.0
scalar 3 bit	RedPaj.	6.20	8.19	81.3	75.1	81.6	69.9
	T-Wix	6.30	8.03	80.3	75.2	81.8	70.7
scalar 2 bit	RedPaj.	7.53	16.6	73.7	63.9	78.3	65.8
	T-Wix	8.56	11.4	73.8	68.5	78.0	68.1

QTIP 2 bit	RedPaj.	7.77	34.3	72.6	65.0	78.2	66.9
	T-Wix	10.4	13.9	71.5	66.2	78.9	67.4

При использовании скалярного 2-битного квантования наблюдается существенный рост перплексии, особенно на корпусе WikiText-RU. Наиболее заметное увеличение демонстрируют модели с 4 B параметров, в то время как 32 B-модели остаются относительно более стабильными, хотя их качество также заметно снижается по сравнению с 3-битным режимом. На практике 2-битное скалярное квантование оказывается чрезмерно агрессивным для сохранения качества генерации текстов на русском языке и требует применения методов дообучения.

На QA бенчмарках 4-битное скалярное квантование сохраняет точность, незначительно уступая несжатым моделям. Для моделей Qwen3-32B и T-Pro-IT-2.0-32B высокие показатели на MMLU сохраняются, а на MMLU-RU наблюдается лишь минимальный регресс. Аналогичная картина наблюдается для бенчмарков PIQA и PIQA-RU. Модели с 4 B параметрами демонстрируют несколько большее снижение точности, причем их русскоязычные варианты (MMLU-RU, PIQA-RU) страдают сильнее, чем англоязычные. Тем не менее это снижение остается в пределах, приемлемых для большинства практических применений.

При 3-битной точности все четыре показателя точности снижаются во всех моделях. Снижение незначительно для моделей 32B и более заметно для Qwen3-4B. Показатели на русскоязычных тестах (MMLU-RU, PIQA-RU) падают больше, чем на англоязычных, что соответствует наблюдениям, сделанным ранее по перплексии, и подчеркивает, что ведение рассуждений и анализ текстов моделями в русском языке более чувствительны к шуму квантования.

Табл. 3. Результаты оценки квантованных моделей с числом параметров 32 B (scalar vs microscaling)

Точность	W2	W2-Ru	W2	W2-Ru
	Qwen3-32B		T-pro-it-2.0	
FP16	6.67	4.44	5.53	6.96
Scalar 4bit	6.80	4.38	5.67	7.18
	6.82	4.37	5.66	7.21
MXFP4	11.5	10.4	11.7	8.36
MXINT4	12.7	8.55	5.83	7.51

При скалярном 2-битном квантовании снижение точности на QA становится существенным, особенно на бенчмарках MMLU-RU и PIQA-RU. Для моделей объемом 4 B данный режим квантования, как правило, не соответствует приемлемым порогам качества для промышленного развертывания без дополнительной адаптации. Хотя 32 B-модели по-прежнему превосходят 4 B-модели при той же разрядности, наблюдаемое снижение качества по сравнению с 3-битным режимом подтверждает, что прямое 2-битное квантование является рискованным для QA задач.

Сравнение скалярного и векторного квантования при 2 битах. Во всех четырех моделях скалярное 2-битное квантование демонстрирует незначительное превосходство над векторным как по перплексии, так и по точности на QA бенчмарках. Однако векторное квантование обеспечивает более низкую среднюю эффективную разрядность, что приводит к дополнительной экономии памяти (эффективные 2.5 бит в случае скалярного квантования с группами по 64 элемента против эффективных 2 бит QTIP). Компромисс между методами особенно заметен на русскоязычных бенчмарках (WikiText-RU, MMLU-RU, PIQA-RU), где скалярное 2-битное квантование сохраняет бóльшую стабильность, в особенности

для моделей с 4 B параметрами. В случае более крупных 32 B-моделей разрыв между методами сокращается, и векторное квантование, откалиброванное с помощью русскоязычного корпуса T-Wix, в некоторых случаях приближается по качеству к скалярному. Таким образом, на практике скалярное 2-битное квантование остается предпочтительным, когда приоритетной задачей является сохранение качества модели, тогда как векторное квантование становится привлекательной альтернативой в случаях, требующих максимальной эффективности использования памяти.

Корреляции и характер ошибок. Мы наблюдаем положительную корреляцию между ростом перплексии и снижением точности на QA бенчмарках, причем на русскоязычных задачах эта связь выражена сильнее. Анализ ошибок на бенчмарках MMLU-RU и PIQA-RU показал, что низкобитное скалярное квантование увеличивает количество ошибок, связанных с согласованием (например, падежным, родовым), пониманием устойчивых выражений, а также с обработкой слов с длинными формами. Эта картина является закономерной и соответствует лингвистическим особенностям русского языка.

ЗАКЛЮЧЕНИЕ

Проведенное исследование дает первую систематическую оценку квантования больших языковых моделей, адаптированных под русский язык. Наши результаты показали, что хотя 4-битное квантование остается высоконадежным даже для морфологически сложного русского языка, переход на 3 бита вызывает значительное снижение устойчивости в качестве при генерации текстов, особенно при сжатии небольших моделей и тестирования их на специализированных русскоязычных бенчмарках. При 2-битном квантовании наивные подходы оказываются практически неприменимыми для русскоязычных LLM, однако калибровки на русскоязычных корпусах данных частично восстанавливают производительность. В целом оптимизированные для русского языка модели, такие как T-Pro-2.0-32B, демонстрируют более высокую устойчивость к квантованию по сравнению с мультиязычными аналогами, что подчеркивает важность как масштаба модели, так и языковой адаптации. Эти результаты свидетельствуют о том,

что успешное развертывание сжатых русскоязычных LLM требует не только технических инноваций в области квантования, но и тщательного учета языково-специфичной калибровки и оценки.

СПИСОК ЛИТЕРАТУРЫ

1. *Shavrina T. et al.* RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020. P. 4717–4726. https://doi.org/10.18653/v1/2020.emnlp-main.381

2. *Mendonça J., Lavie A., Trancoso I.* On the Benchmarking of LLMs for Open-Domain Dialogue Evaluation // Proceedings of the 6th Workshop on NLP for Conversational AI (NLP4ConvAI 2024). 2024. P. 1–12. https://doi.org/10.48550/arXiv.2407.03841

3. *Liu J. et al.* Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation //Advances in Neural Information Processing Systems. 2023. Vol. 36. P. 21558–21572. https://doi.org/10.48550/arXiv.2305.01210

4. *Hendrycks D. et al.* Measuring massive multitask language understanding, 2021 // International Conference on Learning Representations. 2021. https://doi.org/10.48550/arXiv.2009.03300

5. *Clark P. et al.* Think you have solved question answering? try arc, the ai2 reasoning challenge // arXiv preprint arXiv:1803.05457. 2018. https://doi.org/10.48550/arXiv.1803.05457

6. *Zellers R. et al.* HellaSwag: Can a Machine Really Finish Your Sentence? // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019. P. 4791–4800. https://doi.org/10.48550/arXiv.1905.07830

7. *Dettmers T. et al.* Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale // Advances in neural information processing systems. 2022. Vol. 35, P. 30318–30332. https://doi.org/10.48550/arXiv.2208.07339

8. *Frantar E. et al.* OPTQ: Accurate post-training quantization for generative pre-trained transformers // 11th International Conference on Learning Representations. 2023. https://doi.org/10.48550/arXiv.2210.17323

9. *Lin J. et al.* Awq: Activation-aware weight quantization for on-device llm compression and acceleration // Proceedings of machine learning and systems. 2024. Vol. 6. P. 87–100. https://doi.org/10.1145/3714983.3714987

10. *Xiao G. et al.* Smoothquant: Accurate and efficient post-training quantization for large language models // International conference on machine learning. PMLR, 2023. P. 38087 –38099. https://doi.org/10.48550/arXiv.2211.10438

11. *Tseng A. et al.* Qtip: Quantization with trellises and incoherence processing // Advances in Neural Information Processing Systems. 2024. Vol. 37. P. 59597–59620. https://doi.org/10.48550/arXiv.2406.11235

12. *T-Tech.* T-pro-2.0. – Hybrid reasoning model based on Qwen3-32B // HuggingFace.co: The collaboration platform. 2025. URL: https://huggingface.co/t-tech/T-pro-it-2.0

13. *Yandex company.* YandexGPT // HuggingFace.co: The collaboration platform. 2025. URL: https://huggingface.co/yandex/YandexGPT-5-Lite-8B-instruct

14. *Tikhomirov M., Chernyshev D.* Facilitating large language model russian adaptation with learned embedding propagation // Journal of Language and Education. 2024. Vol. 10. No. 4 (40). P. 130–145. https://doi.org/10.48550/arXiv.2412.21140

15. *Team Q. et al.* Qwen2 technical report // arXiv preprint arXiv:2407.10671. 2024. Vol. 2. P. 3. https://doi.org/10.48550/arXiv.2407.10671

16. *Agarwal S. et al.* gpt-oss-120b & gpt-oss-20b Model Card // arXiv e-prints. 2025. P. arXiv: 2508.10925. https://doi.org/10.48550/arXiv.2508.10925

17. *Liu A. et al.* DeepSeek-V3 Technical Report // arXiv e-prints. 2024. P. arXiv: 2412.19437. https://doi.org/10.48550/arXiv.2412.19437

18. *Chee J. et al.* Quip: 2-bit quantization of large language models with guarantees // Advances in Neural Information Processing Systems. 2023. Vol. 36, P. 4396 – 4429. https://doi.org/10.48550/arXiv.2307.13304

19. *Chen M. et al.* Efficientqat: Efficient quantization-aware training for large language models // Annual Meeting of the Association for Computational Linguistics. 2025. Vol. 1. P. 10081–10100. https://doi.org/10.48550/arXiv.2407.11062

20. *Shao W. et al.* OmniQuant: Omnidirectionally Calibrated Quantization for Large Language Models // The Twelfth International Conference on Learning Representations. 2024. https://doi.org/10.48550/arXiv.2308.13137

21. *Hu E. J. et al.* Lora: Low-rank adaptation of large language models // International Conference on Machine Learning. 2022. Vol. 1, No. 2. P. 3. https://doi.org/10.48550/arXiv.2106.09685

22. *Han Z. et al.* Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey // arXiv e-prints. 2024. P. arXiv: 2403.14608. https://doi.org/10.48550/arXiv.2403.14608

23. *Egiazarian V. et al.* Extreme compression of large language models via additive quantization // Proceedings of the 41st International Conference on Machine Learning. 2024. P. 12284–12303. https://doi.org/10.48550/arXiv.2401.06118

24. *Tseng A. et al.* QuIP#: Even Better LLM Quantization with Hadamard Incoherence and Lattice Codebooks // International Conference on Machine Learning. PMLR, 2024. P. 48630–48656. https://doi.org/10.48550/arXiv.2402.04396

25. *Tseng A. et al.* Qtip: Quantization with trellises and incoherence processing // Advances in Neural Information Processing Systems. 2024. Vol. 37. P. 59597–59620. https://doi.org/10.48550/arXiv.2406.11235

26. *Yang A. et al.* Qwen3 technical report // arXiv e-prints. 2025. P. arXiv: 2505.09388. https://doi.org/10.48550/arXiv.2505.09388

27. *Achiam J. et al.* GPT-4 Technical Report // arXiv e-prints. 2023. arXiv: 2303.08774. https://doi.org/10.48550/arXiv.2303.08774

28. *Darvish Rouhani B. et al.* Microscaling data formats for deep learning // arXiv e-prints. 2023. P. arXiv: 2310.10537. https://doi.org/10.48550/arXiv.2310.10537

29. *Weber M. et al.* Redpajama: an open dataset for training large language models // Advances in neural information processing systems. 2024. Vol. 37. P. 116462–116492. https://doi.org/10.52202/079017-3697

30. *Potapov A.* T-Wix – Russian supervised fine-tuning (SFT) dataset // HuggingFace.co: The collaboration platform. 2025. URL: https://huggingface.co/datasets/t-tech/T-Wix

31. *Merity S. et al.* Pointer Sentinel Mixture Models // International Conference on Learning Representations. 2017. https://doi.org/10.48550/arXiv.1609.07843

32. *Korablinov V., Braslavski P.* RuBQ: A Russian dataset for question answering over Wikidata // International Semantic Web Conference. Cham: Springer International Publishing. 2020. P. 97–110. https://doi.org/10.1007/978-3-030-62466-8_7

33. *Li H. et al.* CMMLU: Measuring massive multitask language understanding in Chinese // Findings of the Association for Computational Linguistics. 2024. P. 11260–11285. https://doi.org/10.48550/arXiv.2306.09212

34. *Bisk Y. et al.* Piqa: Reasoning about physical commonsense in natural language // Proceedings of the AAAI conference on artificial intelligence. 2020. Vol. 34. №. 05. P. 7432–7439. https://doi.org/10.1609/aaai.v34i05.6239

35. *Fenogenova A. et al.* MERA: A Comprehensive LLM Evaluation in Russian //Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024. P. 9920–9948. https://doi.org/10.18653/v1/2024.acl-long.534

36. *Chirkin A. et al.* RusConText Benchmark: A Russian Language Evaluation Benchmark for Understanding Context // ACL 2025 Student Research Workshop. 2025. https://aclanthology.org/2025.acl-srw.91/

37. *EleutherAI.* Language Model Evaluation Harness // Zenodo. 2024. v0.4.3. https://zenodo.org/records/10256836

EXPLORING POST-TRAINING QUANTIZATION OF LARGE LANGUAGE MODELS WITH A FOCUS ON RUSSIAN EVALUATION

D. Poimanov[1] [0009-0001-5390-915X], M. Shutov[2] [0009-0009-0530-5034]

[1]*Lomonosov Moscow State University, Moscow, Russia*

[2]*Moscow Institute of Science and Technology, Moscow, Russia*

[1]poimanovdr@my.msu.ru, [2]mihailshutov105@gmail.com

Abstract

The rapid adoption of large language models (LLMs) has made quantization a central technique for enabling efficient deployment under real-world hardware and memory constraints. While English-centric evaluations of low-bit quantization are increasingly available, much less is known about its effects on morphologically rich and resource-diverse languages such as Russian. This gap is particularly important given the recent emergence of high-performing Russian and multilingual LLMs. In this work, we conduct a systematic study of 2-, 3-, and 4-bit post-training quantization (PTQ) for

state-of-the-art Russian LLMs across different model scales (4B and 32B). Our experimental setup covers both standard uniform quantization and specialized low-bit formats, as well as lightweight finetuning for recovery in the most extreme 2-bit setting. Our findings highlight several important trends: (i) the tolerance of Russian LLMs to quantization differs across model families and scales; (ii) 4-bit quantization is generally robust, especially when advanced formats are used; (iii) 3-bit models expose sensitivity to calibration data and scaling strategies; and (iv) 2-bit models, while severely degraded under naive PTQ, can be partially restored through short finetuning. Empirical results show that the model's domain must be considered when using different quantization techniques.

Keywords*: neural networks quantization, compression and optimization of large language models.*

REFERENCES

1. *Shavrina T. et al.* RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020. P. 4717–4726. https://doi.org/10.18653/v1/2020.emnlp-main.381

2. *Mendonça J., Lavie A., Trancoso I.* On the Benchmarking of LLMs for Open-Domain Dialogue Evaluation // Proceedings of the 6th Workshop on NLP for Conversational AI (NLP4ConvAI 2024). 2024. P. 1–12. https://doi.org/10.48550/arXiv.2407.03841

3. *Liu J. et al.* Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation //Advances in Neural Information Processing Systems. 2023. Vol. 36. P. 21558–21572. https://doi.org/10.48550/arXiv.2305.01210

4. *Hendrycks D. et al.* Measuring massive multitask language understanding, 2021 // International Conference on Learning Representations. 2021. https://doi.org/10.48550/arXiv.2009.03300

5. *Clark P. et al.* Think you have solved question answering? try arc, the ai2 reasoning challenge // arXiv preprint arXiv:1803.05457. 2018. https://doi.org/10.48550/arXiv.1803.05457

6. *Zellers R. et al.* HellaSwag: Can a Machine Really Finish Your Sentence? // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019. P. 4791–4800. https://doi.org/10.48550/arXiv.1905.07830

7. *Dettmers T. et al.* Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale // Advances in neural information processing systems. 2022. Vol. 35, P. 30318–30332. https://doi.org/10.48550/arXiv.2208.07339

8. *Frantar E. et al.* OPTQ: Accurate post-training quantization for generative pre-trained transformers // 11th International Conference on Learning Representations. 2023. https://doi.org/10.48550/arXiv.2210.17323

9. *Lin J. et al.* Awq: Activation-aware weight quantization for on-device llm compression and acceleration // Proceedings of machine learning and systems. 2024. Vol. 6. P. 87–100. https://doi.org/10.1145/3714983.3714987

10. *Xiao G. et al.* Smoothquant: Accurate and efficient post-training quantization for large language models // International conference on machine learning. PMLR, 2023. P. 38087 –38099. https://doi.org/10.48550/arXiv.2211.10438

11. *Tseng A. et al.* Qtip: Quantization with trellises and incoherence processing // Advances in Neural Information Processing Systems. 2024. Vol. 37. P. 59597–59620. https://doi.org/10.48550/arXiv.2406.11235

12. *T-Tech.* T-pro-2.0. – Hybrid reasoning model based on Qwen3-32B // HuggingFace.co: The collaboration platform. 2025. URL: https://huggingface.co/t-tech/T-pro-it-2.0

13. *Yandex company.* YandexGPT // HuggingFace.co: The collaboration platform. 2025. URL: https://huggingface.co/yandex/YandexGPT-5-Lite-8B-instruct

14. *Tikhomirov M., Chernyshev D.* Facilitating large language model russian adaptation with learned embedding propagation // Journal of Language and Education. 2024. Vol. 10. No. 4 (40). P. 130–145. https://doi.org/10.48550/arXiv.2412.21140

15. *Team Q. et al.* Qwen2 technical report // arXiv preprint arXiv:2407.10671. 2024. Vol. 2. P. 3. https://doi.org/10.48550/arXiv.2407.10671

16. *Agarwal S. et al.* gpt-oss-120b & gpt-oss-20b Model Card // arXiv e-prints. 2025. P. arXiv: 2508.10925. https://doi.org/10.48550/arXiv.2508.10925

17. *Liu A. et al.* DeepSeek-V3 Technical Report // arXiv e-prints. 2024. P. arXiv: 2412.19437. https://doi.org/10.48550/arXiv.2412.19437

18. *Chee J. et al.* Quip: 2-bit quantization of large language models with guarantees // Advances in Neural Information Processing Systems. 2023. Vol. 36, P. 4396 – 4429. https://doi.org/10.48550/arXiv.2307.13304

19. *Chen M. et al.* Efficientqat: Efficient quantization-aware training for large language models // Annual Meeting of the Association for Computational Linguistics. 2025. Vol. 1. P. 10081–10100. https://doi.org/10.48550/arXiv.2407.11062

20. *Shao W. et al.* OmniQuant: Omnidirectionally Calibrated Quantization for Large Language Models // The Twelfth International Conference on Learning Representations. 2024. https://doi.org/10.48550/arXiv.2308.13137

21. *Hu E. J. et al.* Lora: Low-rank adaptation of large language models // International Conference on Machine Learning. 2022. Vol. 1, No. 2. P. 3. https://doi.org/10.48550/arXiv.2106.09685

22. *Han Z. et al.* Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey // arXiv e-prints. 2024. P. arXiv: 2403.14608. https://doi.org/10.48550/arXiv.2403.14608

23. *Egiazarian V. et al.* Extreme compression of large language models via additive quantization // Proceedings of the 41st International Conference on Machine Learning. 2024. P. 12284–12303. https://doi.org/10.48550/arXiv.2401.06118

24. *Tseng A. et al.* QuIP#: Even Better LLM Quantization with Hadamard Incoherence and Lattice Codebooks // International Conference on Machine Learning. PMLR, 2024. P. 48630–48656. https://doi.org/10.48550/arXiv.2402.04396

25. *Tseng A. et al.* Qtip: Quantization with trellises and incoherence processing // Advances in Neural Information Processing Systems. 2024. Vol. 37. P. 59597–59620. https://doi.org/10.48550/arXiv.2406.11235

26. *Yang A. et al.* Qwen3 technical report // arXiv e-prints. 2025. P. arXiv: 2505.09388. https://doi.org/10.48550/arXiv.2505.09388

27. *Achiam J. et al.* GPT-4 Technical Report // arXiv e-prints. 2023. arXiv: 2303.08774. https://doi.org/10.48550/arXiv.2303.08774

28. *Darvish Rouhani B. et al.* Microscaling data formats for deep learning // arXiv e-prints. 2023. P. arXiv: 2310.10537. https://doi.org/10.48550/arXiv.2310.10537

29. *Weber M. et al.* Redpajama: an open dataset for training large language models // Advances in neural information processing systems. 2024. Vol. 37. P. 116462–116492. https://doi.org/10.52202/079017-3697

30. *Potapov A.* T-Wix – Russian supervised fine-tuning (SFT) dataset // HuggingFace.co: The collaboration platform. 2025. URL: https://huggingface.co/datasets/t-tech/T-Wix

31. *Merity S. et al.* Pointer Sentinel Mixture Models // International Conference on Learning Representations. 2017. https://doi.org/10.48550/arXiv.1609.07843

32. *Korablinov V., Braslavski P.* RuBQ: A Russian dataset for question answering over Wikidata // International Semantic Web Conference. Cham: Springer International Publishing. 2020. P. 97–110. https://doi.org/10.1007/978-3-030-62466-8_7

33. *Li H. et al.* CMMLU: Measuring massive multitask language understanding in Chinese // Findings of the Association for Computational Linguistics. 2024. P. 11260–11285. https://doi.org/10.48550/arXiv.2306.09212

34. *Bisk Y. et al.* Piqa: Reasoning about physical commonsense in natural language // Proceedings of the AAAI conference on artificial intelligence. 2020. Vol. 34. №. 05. P. 7432–7439. https://doi.org/10.1609/aaai.v34i05.6239

35. *Fenogenova A. et al.* MERA: A Comprehensive LLM Evaluation in Russian //Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024. P. 9920–9948. https://doi.org/10.18653/v1/2024.acl-long.534

36. *Chirkin A. et al.* RusConText Benchmark: A Russian Language Evaluation Benchmark for Understanding Context // ACL 2025 Student Research Workshop. 2025. https://aclanthology.org/2025.acl-srw.91/

37. *EleutherAI.* Language Model Evaluation Harness // Zenodo. 2024. v0.4.3. https://zenodo.org/records/10256836

СВЕДЕНИЯ ОБ АВТОРАХ

ПОЙМАНОВ Дмитрий Романович. Окончил магистратуру кафедры математических методов прогнозирования (ММП) факультета вычислительной математики и кибернетики (ВМК) Московского государственного университета имени М.В. Ломоносова в 2024 году. В настоящее время – аспирант ММП ВМК МГУ.

Dmitrii Romanovich POIMANOV. Graduated from the Master's program of the Department of Mathematical Methods of Forecasting (MMP) at the Faculty of Computational Mathematics and Cybernetics (CMC) of Lomonosov Moscow State University in 2024. He is currently a PhD student at MSU.

email: poimanovdr@my.msu.ru

ORCID: 0009-0001-5390-915X

ШУТОВ Михаил Сергеевич. Окончил магистратуру ФПМИ МФТИ в 2024 году. В настоящее время – аспирант ФПМИ МФТИ.

Mikhail Sergeevich SHUTOV. Graduated from the Master's program at the Moscow Institute of Physics and Technology (MIPT), Faculty of Applied Mathematics and Computer Science in 2024. He is currently a Ph.D. student at the same faculty.

email: mihailshutov105@gmail.com

ORCID: 0009-0009-0530-5034

Материал поступил в редакцию 11 октября 2025 года

УДК 004.056

СОКРЫТИЕ В СМЫСЛЕ: СЕМАНТИЧЕСКОЕ КОДИРОВАНИЕ ДЛЯ ГЕНЕРАТИВНО-ТЕКСТОВОЙ СТЕГАНОГРАФИИ

О. Ю. Рогов[1] [0000-0001-9672-2427], Д. Е. Инденбом[2] [0009-0001-9444-6075], Д. С. Корж[3] [0009-0000-6614-120X], Д. В. Пугачёва[4] [0000-0002-4285-1001], В. А. Воронов[5] [0000-0003-3835-6144], Е. В. Тутубалина[6] [0000-0001-7936-0284]

[1, 3, 4, 6]*Институт искусственного интеллекта, г. Москва, Россия*

[1, 2, 5]*Московский физико-технический институт, г. Долгопрудный, Россия*

[1, 3]*Московский технический университет связи и информатики, г. Москва, Россия*

[6]*Высшая школа экономики, г. Москва, Россия*

[6]*Казанский (Приволжский) федеральный университет, г. Казань, Россия*

[1]rogov@airi.net, [2]indenbom.de@phystech.edu, [3]korzh@airi.net, [4]daria.pugacheva@skoltech.ru, [5]v-vor@yandex.ru, [6]tutubalina@airi.net

Аннотация

В статье предложена новая система для генерации стеганографического текста, скрывающая двоичные сообщения в семантически связном естественном языке с помощью скрытого пространства, обусловливающего большие языковые модели (LLM). Секретные сообщения сначала кодируются в непрерывные векторы с помощью обученного отображения двоичного кода в скрытое пространство, которое используется для управления генерацией текста посредством донастройки префикса. В отличие от предыдущих методов стеганографии на уровне токенов или синтаксиса, наш метод позволяет избежать явной манипуляции словами и вместо этого работает полностью в скрытом семантическом пространстве, что обеспечивает более плавные и менее заметные результаты. На стороне получателя скрытое представление восстанавливается из сгенерированного текста и декодируется обратно в исходное сообщение.
В качестве ключевого теоретического вклада мы предоставляем гарантию надежности: если восстановленный скрытый вектор находится в пределах ограниченного расстояния от изначального, обеспечивается точное восстановление сооб-

© О. Ю. Рогов, Д. Е. Инденбом, Д. С. Корж, Д. В. Пугачёва, В. А. Воронов, Е. В. Тутубалина, 2025.
Данная статья распространяется на условиях международной лицензии Creative Commons License Attribution 4.0 International (CC BY 4.0).

щения, причем граница определяется константой Липшица декодера и минимальным отступом логитов. Этот формальный результат предлагает принципиальный подход к компромиссу между надежностью и емкостью в скрытых стеганографических системах. Эмпирическая оценка как на синтетических данных, так и в практических предметных областях, таких как отзывы на Amazon, показывает, что наш метод достигает высокой точности восстановления сообщений (выше 91%), высокую плавность текста и конкурентоспособную емкость до 6 бит на элемент предложения, сохраняя при этом устойчивость к нейронному стегоанализу. Эти результаты демонстрируют, что генерация со скрытым условием предлагает безопасный и практичный путь для встраивания информации в современные LLM.

***Ключевые слова**: стеганография, семантическое кодирование, языковые модели, донастройка префиксов, граф знаний, генерация естественного языка, скрытое обусловливание, нейронный стегоанализ.*

ВВЕДЕНИЕ

Способность скрытно встраивать информацию в естественный язык играет ключевую роль в безопасной коммуникации и цифровых водяных знаках. Традиционные методы стеганографии работают на уровне символов, слов или синтаксиса, часто вводя статистические артефакты или заметные возмущения в сгенерированный текст. С широким распространением мощных больших языковых моделей (LLM) появилась новая возможность встраивать информацию в сам смысл языка.

Мы представляем новую структуру для семантической стеганографии, которая скрывает сообщения не в выборе токенов на поверхностном уровне, а в скрытой семантической структуре.

Наш метод кодирует двоичные сообщения в виде плотных путей, каждый из которых представляет собой связную концептуальную структуру (например, «астронавт», «исследует», «планета», «удивлен»). Эти пути отображаются в непрерывные скрытые векторы с помощью кодировщика, а затем проецируются в пространство входных векторов LLM. Посредством донастройки префикса (prefix

tuning) полученный вектор обусловливает языковую модель генерировать плавные, похожие на человеческие предложения, которые сохраняют заданную семантику.

На этапе декодирования сгенерированный текст анализируется с помощью распознавания именованных сущностей (Named Entity Recognition — NER) и семантической маркировки ролей для восстановления исходного концептуального пути. Затем этот восстановленный путь отображается обратно в скрытое пространство и декодируется в исходное сообщение с помощью обратного отображения графа. Поскольку наш метод использует внутреннее семантическое выравнивание LLM, он позволяет избежать прямой манипуляции токенами и остается устойчивым в условиях «черного ящика», когда внутренние логиты или распределения выборки недоступны.

Мы проверили наш подход на синтетическом и открытом наборе структурированных текстов, продемонстрировав конкурентоспособную битовую емкость (5–6 бит на семантическую единицу), высокую лингвистическую естественность и устойчивость к современным системам стегоанализа. Насколько нам известно, это первая стеганографическая структура, которая согласовывает кодирование графа знаний с префиксным обусловливанием LLM для плавной и безопасной генерации текста.

Помимо эмпирических результатов, мы предоставляем формальные гарантии надежности [1] для восстановления сообщений в скрытом пространстве. В частности, мы анализируем условия, при которых двоичные сообщения, встроенные в непрерывные скрытые векторы, могут быть надежно декодированы после возмущений, например, возникающих во время генерации или извлечения. Наш анализ дает жесткие ограничения на допустимую величину возмущения с точки зрения константы Липшица декодера и отступа логитов от порога принятия решения. Этот результат устанавливает принципиальный компромисс между стабильностью скрытого кодирования и точностью декодирования и может послужить основой для будущих разработок доказательно безопасных или сертифицировано устойчивых стеганографических систем.

1.ЛИТЕРАТУРНЫЙ ОБЗОР

1.1. Генеративно-текстовая стеганография

Генеративно-текстовая стеганография использует предварительно обученные языковые модели для создания естественно выглядящих текстов, которые скрывают секретную информацию с помощью механизма стеганографического отображения. В таких системах качество как языковой модели, так и стратегии отображения играет критическую роль в обеспечении уровня скрытности и возможности восстановления данных.

В 2012 г. Эрнан Моральдо [2] предпринял одну из первых попыток создания генератора текста с встроенным зашифрованным сообщением. Для этого было предложено использовать языковую модель на основе цепи Маркова. В описанном методе в рамках марковской цепи анализируется корпус текстов для определения вероятностей переходов между токенами (словами) в пределах одного предложения. На основе этих вероятностей выбор каждого следующего генерируемого токена автоматически соотносится с кодируемой группой токенов. И каждый последующий переход сужает выбор кодируемых единиц информации вплоть до одной конкретной. Однако данный метод упирается в ограниченную емкость знаний марковской цепи, что приводит к неестественности генерируемого текста.

Позднее в работе Фанга и др. [3] был представлен новый основополагающий подход, в котором словарь *V* разделяется на 2^b непересекающихся групп $[V_1, V_2, \dots, V_{2^b}]$ для кодирования *b*-битных сегментов секретного сообщения. Во время генерации выбирается токен с наибольшей вероятностью в соответствующей группе. Последующие исследования Янга и др. [4, 5] продемонстрировали, что более совершенные языковые модели, такие как LSTM и BERT, значительно повышают как естественность, так и безопасность генерации стеганографического текста.

Зиглер и др. [6] и Дай и др. [7] применили GPT-2 в качестве базовой языковой модели и ввели отдельные стеганографические отображения, адаптированные к ее распределению токенов. Их результаты показали, что выбор функции

отображения оказывает значительное влияние на заметность и емкость стеганографических систем.

С точки зрения криптографии, Чжан и др. [8] предложили адаптивную динамическую группировку (Adaptive Dynamic Grouping — ADG), доказательно безопасный подход, который рекурсивно встраивает секретные биты посредством адаптивной группировки токенов в словаре. Из последних исследований отметим работу Динга и др. [9], где авторы представили метод Discop, который сохраняет исходное распределение токенов, копируя его в процессе встраивания, что обеспечивает высокие показатели незаметности и безопасности.

1.2. Большие языковые модели

Большие языковые модели (Large Language Model — LLM) продемонстрировали высокую эффективность в широком спектре генеративных задач. В недавних работах было исследовано их использование для генерации различных типов структурированных или размеченных данных, включая табличные записи [10], тройки для графов знаний [11] и пары предложений [12].

Большинство ранних подходов полагалось на простые классовые префиксы и zero-shot промпты для генерации данных в определенной области. Для улучшения устойчивости SuperGen [13] и ZeroGen [14] использовали LLM для генерации обучающих данных для задач классификации текста и включили устойчивые к шуму методы обучения [15] для устранения несоответствий в сгенерированных метках. SunGen [16] еще больше улучшил генерацию за счет определения весов, соответствующих качеству данных, для взвешенного использования синтетических примеров во время обучения, чтобы повысить общую эффективность.

Параллельно с этим недавние достижения в области промпт-инжиниринга сделали генерацию на основе LLM более контролируемой. Чэнь и др. [17] предложили подход мягкой донастройки промптов, применимый к LLM типа «белый ящик» с доступом к ключу случайной генерации. Ю и др. [18] расширили эту идею на окружения типа «черный ящик» и API к LLM (например, ChatGPT), продемонстрировав, что высококачественная генерация данных может быть достигнута без размеченных примеров или доступа к внутренним компонентам модели.

2. МЕТОД

2.1. Вводная информация

На рис. 1 представлена общая схема работы предлагаемого метода стеганографии. Ниже последовательно описывается каждый элемент системы.

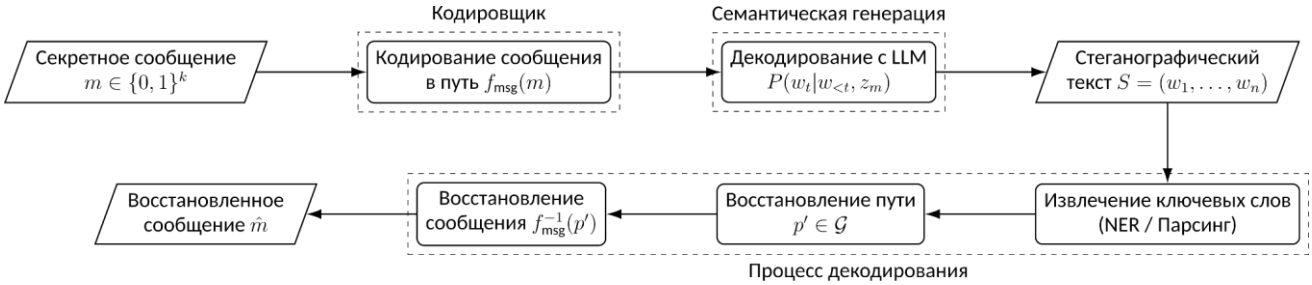


Рис. 1. Схема предлагаемого метода

Пусть $m \in \{0,1\}^k$ — секретное двоичное сообщение фиксированной длины k. Сообщение отображается в скрытый вектор $z_m \in R^d$ с помощью детерминированного или обучаемого кодировщика $f_{\text{enc}}: \{0,1\}^k \to R^d$. Это кодирование выполняет роль условного сигнала для генерации и сохраняется согласованным между отправителем и получателем. Для эффективной работы с языковой моделью (Language Model — LM) вектор z_m может быть спроецирован в пространство эмбеддингов декодера с помощью небольшого многослойного перцептрона (Multi-Layer Perceptron — MLP).

Скрытое обусловливание. Теперь рассмотрим LM, которая генерирует последовательность $S = (w_1, \ldots, w_n)$ из словаря D с помощью моделирования условных вероятностей:

$$P(S \mid z_m) = \prod_{t=1}^{n} P(w_t \mid w_{<t}, z_m),$$

где z_m — скрытый вектор, полученный из двоичного сообщения m. Данное обусловливание реализуется посредством донастройки префикса или прямого добавления эмбеддинга в скрытое состояние модели.

Генерация стеганографического текста. Во время генерации скрытый вектор z_m используется для направления модели на создание плавного текста, в ко-

торый неявно встроено сообщение. Модель обучается таким образом, чтобы изменения в z_m влияли на генерацию восстанавливаемым и декодируемым образом, без явного добавления битов сообщения в токены. Результатом является естественно выглядящее предложение *S*, которое кодирует *m* в своей скрытой семантической структуре.

Декодирование. Получив сгенерированное стеганографическое предложение S, получатель использует ту же языковую модель, чтобы получить соответствующий скрытый вектор $\hat{z}$. Затем применяется функция декодера $f_{dec}: R^d \to \{0,1\}^k$ для восстановления сообщения:

$$\widehat{m} = f_{\mathrm{dec}}(\hat{z}),$$

где $\widehat{m}$ — восстановленная битовая строка. Для правильного декодирования восстановленный скрытый вектор должен оставаться в пределах ϵ-шара исходного закодированного вектора $\|z_m - \hat{z}\|_2^2 \leq \epsilon$.

Система оптимизирована целиком (end-to-end) для достижения максимального качества текста при сохранении точного восстановления сообщения. В процессе обучения используется комбинированная функция потерь, включающая стандартное слагаемое, отвечающее за моделирование языка, и компонент, проверяющий восстановление данных:

$$L = L_{\mathrm{NLL}} + \lambda \cdot \|f_{\mathrm{dec}}(\hat{z}) - m\|_1, \qquad (1)$$

где L_{NLL} — отрицательный логарифм правдоподобия сгенерированного текста, а λ — коэффициент регуляризации, балансирующий плавность и восстанавливаемость.

2.2. Гарантии надежности

Лемма 1 (Скрытая надежность). *Пусть* $m \in \{0,1\}^k$ *— двоичное сообщение, закодированное в скрытый вектор* $z_m \in R^d$. *Предположим, что получатель принимает некоторое приближение* $\hat{z}$, *удовлетворяющее* $\| \hat{z} - z_m \|_2 \leq \delta$.

Пусть $f_{\mathrm{dec}}: R^d \to R^k$ *— декодер, являющийся* L*-липшицевым отображением*. Следовательно,

$$\|f_{\mathrm{dec}}(z_1) - f_{\mathrm{dec}}(z_2)\|_\infty \leq L\|z_1 - z_2\|_2 \text{ для любого } z_1,\ z_2 \in R^d.$$

Предположим, что f_{dec} генерирует логиты, такие что для истинного скрытого z_m:

$$\mathrm{round}\big(f_{\mathrm{dec}}(z_m)\big) = m, \quad \min_{1\le i\le k} |[f_{\mathrm{dec}}(z_m)]_i - 0.5| \ge \eta > 0,$$

где η — минимальный битовый отступ. Если $\delta < \eta/L$, то декодированное сообщение в точности соответствует m:

$$\mathrm{round}\big(f_{\mathrm{dec}}(\hat{z})\big) = m.$$

Доказательство. По условию Липшица и условию восстановления,

$$\max_{1\le i\le k} |[f_{\mathrm{dec}}(\hat{z})]_i - [f_{\mathrm{dec}}(z_m)]_i| = \| f_{\mathrm{dec}}(\hat{z}) - f_{\mathrm{dec}}(z_m) \|_\infty \le L\delta < \eta.$$

Это означает, что для каждого бита $i \in \{1, \dots, k\}$

$$|[f_{\mathrm{dec}}(\hat{z})]_i - [f_{\mathrm{dec}}(z_m)]_i| < \eta.$$

По предположению о битовом отступе в z_m:

- Если $m_i = 0$, то $[f_{\mathrm{dec}}(z_m)]_i \le 0.5 - \eta$, и, следовательно:

$$[f_{\mathrm{dec}}(\hat{z})]_i < [f_{\mathrm{dec}}(z_m)]_i + \eta \le (0.5 - \eta) + \eta = 0.5.$$

- Если $m_i = 1$, то $[f_{\mathrm{dec}}(z_m)]_i \ge 0.5 + \eta$, и, следовательно:

$$[f_{\mathrm{dec}}(\hat{z})]_i > [f_{\mathrm{dec}}(z_m)]_i - \eta \ge (0.5 + \eta) - \eta = 0.5.$$

В обоих случаях $[f_{\mathrm{dec}}(\hat{z})]_i$ находится строго на правильной стороне от 0.5. Следовательно,

$$\mathrm{round}([f_{\mathrm{dec}}(\hat{z})]_i) = m_i \text{ для любого } i$$

и $\mathrm{round}\big(f_{\mathrm{dec}}(\hat{z})\big) = m$.

3. ЭКСПЕРИМЕНТЫ И ИХ ОБСУЖДЕНИЕ

Мы оцениваем предложенную стеганографическую систему с помощью доступной легкой языковой модели, обусловленной на вектора двоичных сообщений. Наша цель — оценить способность системы генерировать плавный, незаметный стеганографический текст, при этом обеспечивая точное восстановление сообщения.

Каждое секретное сообщение m представляет собой строку из 64 бит, выбранную случайным образом из равномерного распределения, т. е.

$m \in \{0,1\}^{64}$. Двоичное сообщение отображается в скрытый вектор $z_m \in R^{128}$ с помощью обучаемой нейросети кодировщика f_{enc}, состоящей из двух полносвязанных слоев со скрытой размерностью 256 и функцией активации ReLU. Формально,

$$z_m = \text{ReLU}(W_2 \cdot \text{ReLU}(W_1 m + b_1) + b_2).$$

Полученный скрытый вектор преобразуют и проецируют в последовательность эмбеддингов для префикса в форме (prefix_length, hidden_dim), где prefix_length = 10 и hidden_dim = 768.

В качестве базовой языковой модели мы используем GPT-Neo-125M от EleutherAI. Условный префикс реализуется с помощью адаптеров LoRA через библиотеку peft. Эмбеддинги префикса добавляются в начало потока эмбеддингов на вход модели, чтобы сформировать обусловленный контекст для генерации.

Обучение проводится с помощью оптимизатора AdamW со скоростью обучения (learning rate) 10^{-4}, размером батча, равным 16, и максимальной длиной последовательности, равной 64. Мы дообучаем модель в течение 5 эпох на синтетическом наборе данных из $5 \cdot 10^4$ пар сообщений и текстов. Модель обучается целиком (end-to-end), чтобы минимизировать функцию потерь из уравнения (1). Мы эмпирически подобрали $\lambda = 1.0$ для всех экспериментов.

Мы измеряем качество текста с помощью перплексии на модели GPT-2 и метрики BERTScore (F1) по отношению к эталонным текстам из тренировочного домена. Точность передачи (recovery) оценивается как процент двоичных сообщений, восстановленных со 100% битовой точностью с помощью f_{dec} из сгенерированного текста. Кроме того, битрейт рассчитывается как общее количество успешно декодированных битов, разделенных на количество сгенерированных токенов (bits per token).

Табл. 1 эмпирически подтверждает лемму 1, демонстрируя, что наш метод обеспечивает наджное восстановление сообщений в условиях практических ограничений. Практически идеальные показатели восстановления (91.2% для синтетических данных и 89.5% для данных Amazon) подтверждают, что декодер сохраняет достаточный отступ η в пространстве логитов, демонстрируя при этом выпол-

нение условия Липшица для константы L — в точности теоретического требования для соблюдения $\delta < \eta/L$. Более высокая восстанавливаемость синтетических данных соответствует их более низкой перплексии (17.3 *против* 26.7), что указывает на более предсказуемую генерацию текста, которая, по сути, сохраняет более крупные отступы η от порога принятия решений. Важно, что конкурентоспособные битрейты (3.9–3.4 битов на токен) доказывают, что эти гарантии надежности не мешают практической емкости. Оптимизируя геометрию скрытого пространства для максимизации η при минимизации L, мы достигаем условия стабильности леммы без потери плотности информации. Незначительное снижение точности восстановления в отзывах на Amazon отражает дополнительную сложность для обусловливания с помощью промпта, которая ужесточает соотношение η/L, но все же сохраняет запас прочности значительно выше порога ошибки.

Таблица 1. Количественное сравнение с базовыми решениями на синтетическом наборе данных и наборе отзывов на Amazon для фиксированного отображения словаря (Fixed Vocabulary Mapping – FVM), генератора на основе марковской цепи (Markov Chain Generator – MCG) и предлагаемого метода.

Method	**Perplexity ↓**	**F1 ↑**	**Recovery ↑**	**Bits per token ↑**
FVM–Synth	42.1	0.741	72.3%	2.0
FVM–Amazon	48.9	0.702	68.0%	1.8
MCG–Synth	35.8	0.780	76.1%	1.7
MCG–Amazon	43.2	0.735	71.4%	1.5
Ours–Synth	**17.3**	**0.890**	**91.2%**	**3.9**
Ours–Amazon	**26.7**	**0.870**	**89.5%**	**3.4**

Наш метод значительно превосходит оба базовых подхода: фиксированное отображение словаря (Fixed Vocabulary Mapping — FVM) и генератор на основе марковской цепи (Markov Chain Generator — MCG) — по всем оцениваемым показателям. По сравнению с FVM, который опирается на статичную группировку токенов без контекстной адаптации, наш подход показывает более чем на 20 баллов выше по метрике BERTScore F1 и улучшает коэффициент восстановления сообщений почти на 20% как на синтетических, так и на практических наборах данных. Аналогично, хотя MCG производит более естественные результаты, чем FVM,

благодаря своему вероятностному моделированию, ему не хватает семантической согласованности, и он предлагает только ограниченную битовую емкость (от 1.5 до 1.7 битов на токен), что намного ниже 3.4–3.9 битов на токен, достигаемых нашей моделью, направляемой в скрытом пространстве. Эти улучшения обеспечиваются генерацией, обусловленной в скрытном непрерывном пространстве, что позволяет достигнуть более богатой выразительности и более надежного встраивания сообщений без использования жестких лексических ограничений.

Лемма 1 устанавливает формальную гарантию надежности восстановления дискретных сообщений, которая согласуется с несколькими теоретическими подходами и расширяет их. Наше требование отступа η отражает принцип ограничивающих рамок в нейронном кодировании, где низкоразмерные представления достигают устойчивости, ограничивая динамику безопасными областями. Здесь η определяет именно такую область в пространстве логитов, гарантируя, что возмущения остаются в границах принятия решений.

Условие Липшица для константы L формализует стабильность при вмешательствах, что является краеугольным камнем современных унифицированных теорий надежности [8]. Рассматривая восстановление как относительную стабильность по отношению к l_2-ограниченным возмущениям, мы определяем надежность как целевую устойчивость при вмешательствах. Зависящий от данных параметр η позволяет избежать искусственного масштабирования 2^k, что отвечает критике чрезмерно консервативных теоретических ограничений. Эмпирические показатели восстановления из табл. 1 демонстрируют эксплуатационную жизнеспособность, показывая, что оптимизированные скрытые пространства обеспечивают компромисс между эффективностью и надежностью, предсказанный геометрической теорией обучения (geometric learning theory).

3.1. Ограничения

Хотя наша система демонстрирует высокую емкость, плавность и устойчивость при генерации стеганографического текста, она имеет ряд ограничений, которые открывают возможности для будущих исследований.

Во-первых, наш текущий декодер предполагает доступ к одной и той же языковой модели или модели с сопоставимой скрытой структурой. Это может

ограничить восстановление сообщений между вариантами моделей или системами на основе API, где внутренние представления различаются. Во-вторых, кодировщик и декодер между бинарным и скрытым пространствами обучаются на синтетических данных или на данных из конкретной области, которые могут плохо обобщаться для генерации в открытом домене без доменной адаптации. Наконец, наша теоретическая гарантия надежности предполагает фиксированный декодер с известной константой Липшица и разделяющим отступом. На практике надежная оценка этих параметров может быть нетривиальной, особенно для высокоразмерных скрытых пространств или в условиях интенсивных враждебных возмущений, что является предметом текущей работы.

Кроме того, хотя наш метод позволяет избежать возмущений на уровне токенов, он все же может создавать небольшие сдвиги в распределении, обнаруживаемые с помощью продвинутых моделей стегоанализа, обученных на скрытых или стилистических признаках. Наконец, наш подход требует умеренных вычислительных ресурсов для донастройки префикса и обусловливания на вектор, а расширение этого метода на очень большие модели (например, GPT5) или мультимодальные окружения может создать проблемы с масштабированием. Мы считаем, что эти ограничения могут быть устранены с помощью таких техник, как независимое от модели кодирование, динамическая калибровка декодера и состязательное дообучение для стеганографической инвариантности.

ЗАКЛЮЧЕНИЕ

Несмотря на добавленные ограничения на естественность и используемый словарный запас, модель сохраняет высокий показатель восстановления сообщений (89.5%) и приемлемый уровень перплексии. Битрейт несколько снижается из-за более длинных и связных предложений, необходимых для соответствия предметной области, но остается конкурентоспособным по сравнению с предыдущими разработками. Результаты подчеркивают масштабируемость и надежность нашего подхода к встраиванию на основе векторов в практических сценариях генерации текста.

Благодарности

Работа выполнена при частичной поддержке Программы стратегического академического лидерства Казанского (Приволжского) федерального университета («ПРИОРИТЕТ-2030»).

СПИСОК ЛИТЕРАТУРЫ

1. *Karimov E., Varlamov A., Ivanov D., Korzh D., and Rogov O.Y. Novel.* LossEnhanced Universal Adversarial Patches for Sustainable Speaker Privacy. — 2025. — 2505.19951.

2. *Moraldo H.H.* An Approach for Text Steganography Based on Markov Chains // ArXiv. 2014. Vol. abs/1409.0915.

3. *Fang T.*, *Jaggi M., Argyraki K.* Generating steganographic text with LSTMs // arXiv preprint arXiv:1705.10742. 2017.

4. *Yang Z.-L., Guo X.-Q., Chen Z.-M., Huang Y.-F., Zhang Y.-J*. RNN-stega: Linguistic steganography based on recurrent neural networks // IEEE Transactions on Information Forensics and Security. 2018. Vol. 14, No. 5. P. 1280–1295.

5. *Yang Z.-L., Zhang S.-Y., Hu Y.-T., Hu Z.-W., Huang Y.-F.* VAE-Stega: linguistic steganography based on variational auto-encoder // IEEE Transactions on Information Forensics and Security. 2020. Vol. 16. P. 880–895.

6. *Ziegler Z., Deng Y., Rush A. M*. Neural Linguistic Steganography // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019. P. 1210–1215.

7. *Dai F.Z., Cai Z.* Towards near-imperceptible steganographic text // arXiv preprint arXiv:1907.06679. 2019.

8. *Zhang S., Yang Z., Yang J., Huang Y.* Provably Secure Generative Linguistic Steganography// Findings of the AssociationforComputational Linguistics: ACLIJCNLP 2021. 2021. P. 3046–3055.

9. Ding J., Chen K., Wang Y., Zhao N., Zhang W., Yu N. Discop: Provably Secure Steganography in Practice Based on “Distribution Copies” // 2023 IEEE Symposium on Security and Privacy (SP) / IEEE Computer Society. 2023. P. 2238– 2255.

10. *Borisov V., Seßler K., Leemann T., Pawelczyk M., Kasneci G.* Languagemodels are realistic tabular data generators // arXiv preprint arXiv:2210.06280. 2022.

11. *Chia Y.K., Bing L., Poria S., Si L.* RelationPrompt: Leveraging prompts to generate synthetic data for zero-shot relation triplet extraction // arXiv preprint arXiv:2203.09101. 2022.

12. *Schick T.*, *Schütze H.* Generating datasets with pretrained language models // arXiv preprint arXiv:2104.07540. 2021.

13. *Meng Y., Huang J., Zhang Y., Han J.* Generating training data with language models: Towards zero-shot language understanding // Advances in Neural Information Processing Systems. 2022. Vol. 35. P. 462–477.

14. *Ye J.*, *Gao J., Li Q., Xu H., Feng J., Wu Z., Yu T., Kong L.* Zerogen: Efficient zero-shot learning via dataset generation // arXiv preprint arXiv:2202.07922. 2022.

15. *Wang Y., Ma X., Chen Z., Luo Y., Yi J., Bailey J.* Symmetric cross entropy for robust learning with noisy labels // Proceedings of the IEEE/CVF international conference on computer vision. 2019. P. 322–330.

16. *Gao J., Pi R., Yong L., Xu H., Ye J., Wu Z., Zhang W., Liang X., Li Z., Kong L.* Self-guided noise-free data generation for efficient zero-shot learning // International Conference on Learning Representations (ICLR 2023). 2023.

17. *Chen D.*, *Lee C., Lu Y., Rosati D., Yu Z.* Mixture of Soft Prompts for Controllable Data Generation // arXiv preprint arXiv:2303.01580. 2023.

18. *Yu Y., Zhuang Y., Zhang J., Meng Y., Ratner A., Krishna R., Shen J., Zhang C.* Large language model as attributed training data generator: A tale of diversity and bias // arXiv preprint arXiv:2306.15895. 2023.

HIDING IN MEANING: SEMANTIC ENCODING FOR GENERATIVE TEXT STEGANOGRAPHY

O. Y. Rogov[1] [0000-0001-9672-2427], D. E. Indenbom[2] [0009-0001-9444-6075],
D. S. Korzh[3] [0009-0000-6614-120X], D. V. Pugacheva[4] [0000-0002-4285-1001],
V.A. Voronov[5] [0000-0003-3835-6144], E.V. Tutubalina[6] [0000-0001-7936-0284]

[1, 3, 4, 6]Artificial Intelligence Research Institute*, Moscow, Russia*
[1, 2 ,5]Moscow Institute of Physics and Technology, Dolgoprudny, Russia
[1, 3]Moscow Technical University of Communications and Informatics, Moscow, Russia
[6]HSE University, Moscow, Russia
[6]Kazan Federal University, Kazan, Russia

[1]rogov@airi.net, [2]indenbom.de@phystech.edu, [3]korzh@airi.net, [4]daria.pugacheva@skoltech.ru, [5]v-vor@yandex.ru, [6]tutubalina@airi.net

Abstract

We propose a novel framework for steganographic text generation that hides binary messages within semantically coherent natural language using latent-space conditioning of large language models (LLMs). Secret messages are first encoded into continuous vectors via a learned binary-to-latent mapping, which is used to guide text generation through prefix tuning. Unlike prior token-level or syntactic steganography, our method avoids explicit word manipulation and instead operates entirely within the latent semantic space, enabling more fluent and less detectable outputs. On the receiver side, the latent representation is recovered from the generated text and decoded back into the original message. As a key theoretical contribution, we provide a robustness guarantee: if the recovered latent vector lies within a bounded distance of the original, exact message reconstruction is ensured, with the bound determined by the decoder's Lipschitz continuity and the minimum logit margin. This formal result offers a principled view of the reliability–capacity trade-off in latent steganographic systems. Empirical evaluation on both synthetic data and real-world domains such as Amazon reviews shows that our method achieves high message recovery accuracy (above 91%), strong text fluency and competitive capacity up to 6 bits per sentence element while maintaining resilience against neural steganalysis. These findings demonstrate that latent

conditioned generation offers a secure and practical pathway for embedding information in modern LLMs.

Keywords*: steganography, semantic encoding, language models, prefix tuning, knowledge graphs, natural language generation, latent conditioning, neural steganalysis.*

REFERENCES

1. *Karimov E., Varlamov A., Ivanov D., Korzh D., and Rogov O.Y. Novel.* LossEnhanced Universal Adversarial Patches for Sustainable Speaker Privacy. — 2025. — 2505.19951.

2. *Moraldo H.H.* An Approach for Text Steganography Based on Markov Chains // ArXiv. 2014. Vol. abs/1409.0915.

3. *Fang T.*, *Jaggi M., Argyraki K.* Generating steganographic text with LSTMs // arXiv preprint arXiv:1705.10742. 2017.

4. *Yang Z.-L., Guo X.-Q., Chen Z.-M., Huang Y.-F., Zhang Y.-J*. RNN-stega: Linguistic steganography based on recurrent neural networks // IEEE Transactions on Information Forensics and Security. 2018. Vol. 14, No. 5. P. 1280–1295.

5. *Yang Z.-L., Zhang S.-Y., Hu Y.-T., Hu Z.-W., Huang Y.-F.* VAE-Stega: linguistic steganography based on variational auto-encoder // IEEE Transactions on Information Forensics and Security. 2020. Vol. 16. P. 880–895.

6. *Ziegler Z., Deng Y., Rush A. M*. Neural Linguistic Steganography // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019. P. 1210–1215.

7. *Dai F.Z., Cai Z.* Towards near-imperceptible steganographic text // arXiv preprint arXiv:1907.06679. 2019.

8. *Zhang S., Yang Z., Yang J., Huang Y.* Provably Secure Generative Linguistic Steganography// Findings of the AssociationforComputational Linguistics: ACLIJCNLP 2021. 2021. P. 3046–3055.

9. Ding J., Chen K., Wang Y., Zhao N., Zhang W., Yu N. Discop: Provably Secure Steganography in Practice Based on “Distribution Copies” // 2023 IEEE Symposium on Security and Privacy (SP) / IEEE Computer Society. 2023. P. 2238– 2255.

10. *Borisov V., Seßler K., Leemann T., Pawelczyk M., Kasneci G.* Languagemodels are realistic tabular data generators // arXiv preprint arXiv:2210.06280. 2022.

11. *Chia Y.K., Bing L., Poria S., Si L.* RelationPrompt: Leveraging prompts to generate synthetic data for zero-shot relation triplet extraction // arXiv preprint arXiv:2203.09101. 2022.

12. *Schick T.*, *Schütze H.* Generating datasets with pretrained language models // arXiv preprint arXiv:2104.07540. 2021.

13. *Meng Y., Huang J., Zhang Y., Han J.* Generating training data with language models: Towards zero-shot language understanding // Advances in Neural Information Processing Systems. 2022. Vol. 35. P. 462–477.

14. *Ye J.*, *Gao J., Li Q., Xu H., Feng J., Wu Z., Yu T., Kong L.* Zerogen: Efficient zero-shot learning via dataset generation // arXiv preprint arXiv:2202.07922. 2022.

15. *Wang Y., Ma X., Chen Z., Luo Y., Yi J., Bailey J.* Symmetric cross entropy for robust learning with noisy labels // Proceedings of the IEEE/CVF international conference on computer vision. 2019. P. 322–330.

16. *Gao J., Pi R., Yong L., Xu H., Ye J., Wu Z., Zhang W., Liang X., Li Z., Kong L.* Self-guided noise-free data generation for efficient zero-shot learning // International Conference on Learning Representations (ICLR 2023). 2023.

17. *Chen D.*, *Lee C., Lu Y., Rosati D., Yu Z.* Mixture of Soft Prompts for Controllable Data Generation // arXiv preprint arXiv:2303.01580. 2023.

18. *Yu Y., Zhuang Y., Zhang J., Meng Y., Ratner A., Krishna R., Shen J., Zhang C.* Large language model as attributed training data generator: A tale of diversity and bias // arXiv preprint arXiv:2306.15895. 2023.

СВЕДЕНИЯ ОБ АВТОРАХ

РОГОВ Олег Юрьевич — получил степень магистра в МГУ и степень кандидата наук в области математического моделирования и физики в Центре фотоники Российской академии наук. В настоящее время является соруководителем проекта машинного зрения в Университете Шарджи и руководителем исследовательской группы в AIRI. Его научные интересы включают эволюционные алгоритмы и глубокое обучение, слияние информации и принятие решений, а также основы проектирования и оценки производительности систем визуализации, обработку сигналов и обнаружение аномалий.

Oleg Yurievich ROGOV — received a Master's degree from MSU and a Ph.D. degree in mathematical modeling and physics at the Photonics Centre of the Russian Academy of Sciences. He is currently the co-PI of the medical vision project at the University of Sharjah and head of a research group at AIRI. His research interests include evolutionary algorithms and deep learning, information fusion and decision making, fundamentals of imaging system design and performance evaluation, signal processing, anomaly detection and estimation.

email: rogov@airi.net

ORCID: 0000-0001-9672-2427

ИНДЕНБОМ Дмитрий Евгеньевич — окончил бакалавриат и получил степень магистра по направлению «Прикладная математика и физика» в МФТИ. В настоящее время является аспирантом МФТИ и работает над диссертацией по теме «Методы защиты информации при использовании больших языковых моделей». Его научные интересы включают цифровую маркировку синтетических данных, скрытое шифрование сообщений в генерируемом тексте и интерпретацию работы языковых моделей.

Dmitrii Evgenievich INDENBOM — received a Bachelor's and a Master's degrees in applied mathematics and physics from MIPT. He is currently working on his dissertation on "Information security methods in the context of large language models" at the graduate school of MIPT. His research interests include digital watermarking of synthetic data, secret messages embedding in generated text, and interpretation of language models.

email: indenbom.de@phystech.edu

ORCID: 0009-0001-9444-6075

Корж Дмитрий Сергеевич — окончил бакалавриат в МФТИ и получил степень магистра в области наук о данных в Сколковском институте наук и технологий; продолжает научную работу в аспирантуре Сколтеха. В настоящее время является младшим научным сотрудником в группе «Доверенные и безопасные интеллектуальные системы» в AIRI и в «Лаборатории безопасного искусственного интеллекта» МТУСИ. Его научные интересы включают методы устойчивости нейронных сетей, безопасность алгоритмов глубокого обучения, прикладные задачи в звуковом домене.

Dmitrii Sergeevich Korzh — received a Bachelor's degree from MIPT and a Master's degree in data science from the Skolkovo Institute of Science and Technology, where he continues his scientific work in graduate school. He is currently a junior researcher in the Reliable and Secure Intelligent Systems group at AIRI and at the Laboratory of Secure Artificial Intelligence at MTUCI. His research interests include methods of robustness of neural networks, the security of deep learning algorithms, and applied problems in the audio domain.

email: korzh@airi.net
ORCID: 0009-0000-6614-120X

ПУГАЧЁВА Дарья Валерьевна — окончила бакалавриат в МФТИ, получила степени магистра по направлениям «Прикладные математика и физика» и «Математика и компьютерные науки» в МФТИ и в Сколковском институте наук и технологий, получила степень кандидата физико-математических наук, защитив диссертацию на тему «Лазерно-плазменное ускорение поляризованных заряженных частиц» в МФТИ. В настоящее время работает научным сотрудником в исследовательской группе прикладного NLP в AIRI. Ее научные интересы включают вопросы генерализации и устойчивости моделей для управления воплощенными агентами, комбинаторную оптимизацию, графовые нейронные сети.

Darya Valeryaevna PUGACHEVA — received a Bachelor's degree from MIPT and completed a Master's degrees in applied mathematics and physics and in mathematics and computer science at MIPT and at the Skolkovo Institute of Science and Technology. She obtained a PhD in physics and mathematics after defending her dissertation on "Laser–plasma

acceleration of polarized charged particles" at MIPT. She is currently a researcher in the Domain-specific NLP research group at AIRI. Her research interests include improving the generalization and robustness of models for the control of embodied agents, combinatorial optimization, and graph neural networks.

email: Daria.Pugacheva@skoltech.ru

ORCID: 0000-0002-4285-1001

ВОРОНОВ Всеволод Александрович — окончил Иркутский государственный университет, получил степень кандидата технических наук Институте проблем управления РАН. В настоящее время является заведующим лабораторией комбинаторной геометрии Кавказского математического центра Адыгейского государственного университета, работает старшим научным сотрудником лаборатории комбинаторных и геометрических структур Московского физико-технического института. Его научные интересы включают теорию графов, дискретную геометрию, динамические системы и машинное обучение.

Vsevolod Alexandrovich VORONOV — graduated from Irkutsk State University and received a PhD in technical sciences at the Institute of Control Sciences of RAS. He is the head of the Laboratory of Combinatorial Geometry at the Caucasus Mathematical Center of the Adyghe State University and a senior researcher in the Laboratory of Combinatorial and Geometric Structures at MIPT. His research interests include graph theory, discrete geometry, dynamical systems, and machine learning.

email: v-vor@yandex.ru

ORCID: 0000-0003-3835-6144

ТУТУБАЛИНА Елена Викторовна — получила степень кандидата физико-математических наук в Институте системного программирования им. В.П. Иванникова РАН и степень доктора компьютерных наук, защитив диссертацию на тему “Модели и методы автоматической обработки неструктурированных данных в биомедицинской области” в ВШЭ. В настоящее время является руководителем научной группы Domain-specific NLP в Институте AIRI, старшим научным сотрудником ИСП РАН и Казанского федерального университета. Ее научные интересы включают машинное обучение, обработку естественного языка, и исследование генерализации и устойчивости языковых моделей.

Elena Viktorovna TUTUBALINA — received a PhD in physics and mathematics at the Ivannikov Institute for System Programming of the Russian Academy of Sciences and the degree of Doctor of Computer Science after defending a dissertation on “Models and methods for automatic processing of unstructured data in the biomedical domain” at HSE. She currently leads the Domain-specific NLP research group at AIRI and is a Senior Researcher at the Ivannikov Institute for System Programming of the Russian Academy of Science and at the Kazan Federal University. Her research interests include machine learning, natural language processing, and studies of generalization and robustness of language models.

email: tutubalina@airi.net

ORCID: 0000-0001-7936-0284

Материал поступил в редакцию 14 октября 2025 года

УДК 004.81

УСЛОВНАЯ ГЕНЕРАЦИЯ ЭЛЕКТРОКАРДИОГРАММ С ПОМОЩЬЮ ИЕРАРХИЧЕСКИХ ВАРИАЦИОННЫХ АВТОКОДИРОВЩИКОВ

И. А. Свиридов[1] [0009-0009-5912-1118], К. С. Егоров[2] [0009-0006-6991-4136]

[1, 2]*Sber AI Lab, г. Москва, Россия*

[1]ianatosviridov@sberbank.ru, [2]Egorov.K.Ser@sberbank.ru

Аннотация

Сердечно-сосудистые заболевания являются одной из основных причин смертности. Автоматический анализ электрокардиограмм (ЭКГ) может существенно облегчить работу врачей, но его эффективность ограничена нехваткой и несбалансированностью данных. Создание синтетических ЭКГ помогает частично решить эти проблемы. Хотя чаще всего для этого применяются генеративно-состязательные сети (GAN), но последние исследования показали, что вариационные автокодировщики (VAE) могут обеспечивать сопоставимое качество.

В работе представлена модель cNVAE-ECG — модификация Nouveau VAE (NVAE), способная генерировать 12 отведений 10-секундных ЭКГ с различными патологиями. Используя компактную схему работы с каналами и встроенные представления классов для условной генерации, cNVAE-ECG улучшает результаты в задачах бинарной и multi-label классификации, обеспечивая прирост метрики AUROC до 2% по сравнению с моделями на основе GAN. Модель представлена в открытом доступе: https://github.com/univanxx/cNVAE_ECG.

***Ключевые слова**: ЭКГ, вариационный автокодировщик, условная генерация, GAN.*

ВВЕДЕНИЕ

Сердечно-сосудистые заболевания (ССЗ) по-прежнему занимают первое место среди причин смертности во всем мире, и анализ ЭКГ играет ключевую роль в их диагностике и профилактике [1]. Современные методы глубокого обучения показывают высокий потенциал в автоматической интерпретации ЭКГ [2], однако их применение ограничено несколькими факторами: i) малым объемом доступ-

© И. А. Свиридов, К. С. Егоров, 2025.
Данная статья распространяется на условиях международной лицензии Creative Commons License Attribution 4.0 International (CC BY 4.0).

ных данных из-за требований к конфиденциальности и юридическим ограничениям [3], ii) сильным дисбалансом классов — некоторые патологии встречаются значительно реже [4], iii) наличием шумов и артефактов в записях [5].

Создание синтетических данных представляет собой перспективное направление, которое позволяет увеличить обучающие выборки, выровнять распределение классов и снизить стоимость сбора новых данных — при условии, что сгенерированные ЭКГ реалистично воспроизводят все 12 отведений и могут быть сгенерированы для конкретных патологий [6].

Хотя большинство существующих исследований основано на генеративно-состязательных сетях (GAN), вариационные автокодировщики (VAE) [7], включая архитектуру NVAE [8], ориентированную на изображения, остаются малоизученными в контексте генерации ЭКГ. Мы предлагаем **cNVAE-ECG** — условное расширение NVAE, адаптированное для генерации 12-отведенных ЭКГ. Модель доступна в открытом доступе и демонстрирует превосходство в задачах классификации патологий и обучения с переносом (transfer learning), достигая прироста AUROC до 2% по сравнению с базовыми моделями GAN.

ОБЗОР ЛИТЕРАТУРЫ

Ранние методы синтеза ЭКГ в основном опирались на GAN. Архитектуры вроде PGAN [9] и ProEGAN-MS [10] генерировали отдельные удары или одноотведенные сигналы, улучшая качество за счет продвинутых функций потерь, поэтапного обучения и физиологических ограничений [11]. Эти подходы создавали реалистичные фрагменты ЭКГ, но не могли моделировать зависимость между отведениями и плохо масштабировались на длинные записи. Позднее были предложены условные GAN, где генерация выполнялась с учетом класса заболевания [12]. Однако и здесь оставались трудности при создании длительных многоканальных сигналов. Модели WaveGAN* и Pulse2Pulse [13] стали первыми, которые сгенерировали полноценные 12-отведенные 10-секундные ЭКГ, а MLCGAN [14] показала, что условная генерация может повысить качество последующей классификации.

Альтернативой GAN стали диффузионные модели [15, 16], обладающие высокой гибкостью, но требующие значительных вычислительных ресурсов [17]. Вариационные автокодировщики (VAE), напротив, обучаются стабильнее и быстрее [18, 19]. Недавние работы [20–22] подтвердили их потенциал для генерации многоканальных ЭКГ, однако большинство таких моделей не учитывает классы патологий. В ответ на это мы разработали модель cNVAE-ECG, которая сочетает иерархическую структуру NVAE с условной генерацией по классам заболеваний.

КРАТКАЯ СПРАВКА ПО ВАРИАЦИОННЫМ АВТОКОДИРОВЩИКАМ

Для условной генерации стандартных ЭКГ потребовалось разработать архитектуру на базе NVAE, способную работать с одномерными 12-канальными сигналами и генерировать их в соответствии с заданным классом.

Вариационные автокодировщики (VAE) — это стохастические архитектуры, использующие вариационный вывод для аппроксимации распределения данных $p_\theta(x)$, параметризуемого через θ, и порождения выборок из латентного представления $p_\theta(z)$. Они состоят из двух блоков: энкодера, отображающего входы в сжатое представление $p_\theta(z|x)$, и декодера, восстанавливающего $p_\theta(x|z)$ после сэмплирования из $p_\theta(z)$.

Недостаток VAE в виде более низкого качества генерации по сравнению с GAN можно смягчить продвинутыми приемами вариационного вывода, например иерархической генерацией. В этом подходе латентные переменные разделяются на группы $z = \{z_1, z_2, \dots, z_L\}$ с априорным распределением $p(z) = \prod_l p(z_l|z_{l-1})$ и аппроксимирующим апостериорным распределением $q(z|x) = \prod_l (z_l|z_{l-1}, x)$, где каждый условный фактор моделируется факторизованным нормальным распределением.

Однако иерархические VAE часто страдают от неустойчивости: исчезающие слои, «взрывающиеся» градиенты (exploding gradients). NVAE решает эти вопросы с помощью «снизу вверх» энкодера, «сверху вниз» декодера и техник стабилизации, включая дискретизированные смеси логистических распределений [23] и учет межканальных зависимостей, достигая передовых результатов на MNIST, CIFAR-10 и CelebA [24–27]. Опираясь на эти усовершенствования, мы адаптируем NVAE к одномерным многоканальным сигналам ЭКГ и вводим cNVAE-ECG для условной генерации.

ПРЕДЛОЖЕННЫЙ ПОДХОД cNVAE-ECG

Ниже описана архитектура cNVAE-ECG, представленная на рис. 1.

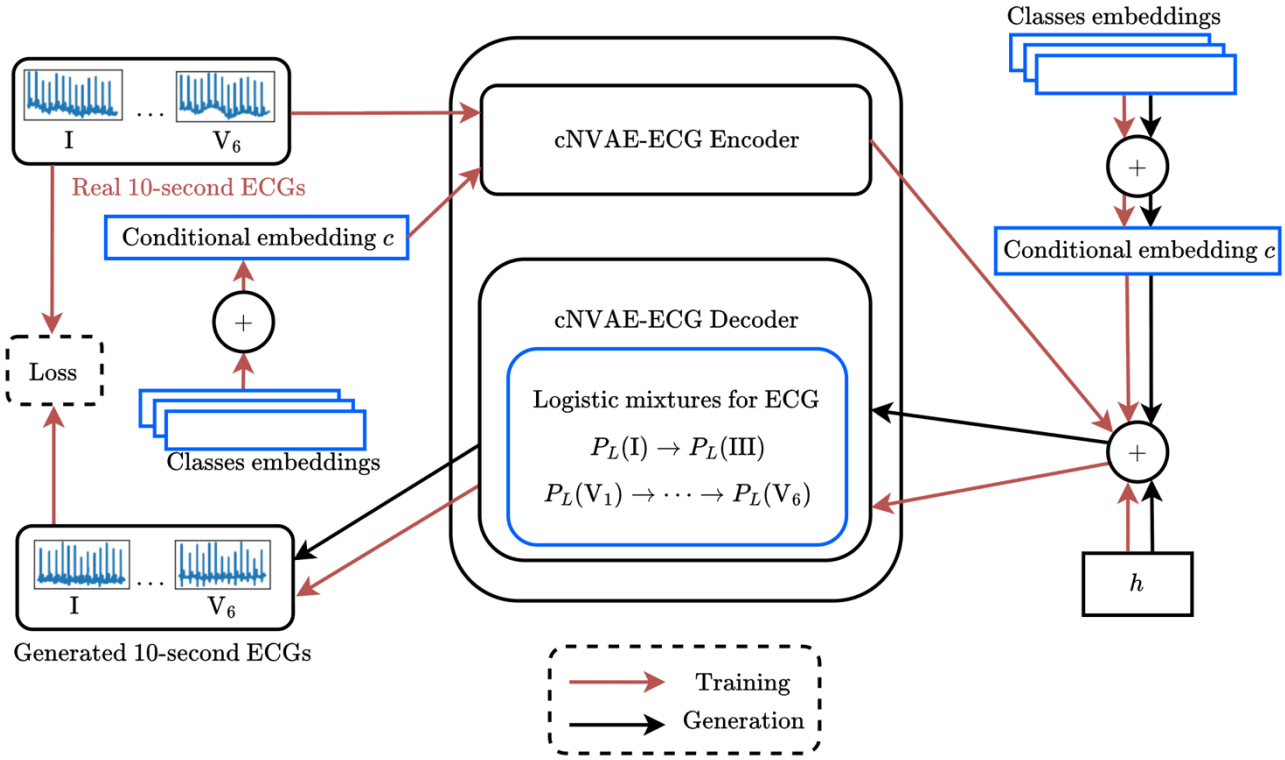


Рис. 1. Архитектура предложенного подхода cNVAE-ECG

1. Трансформация каналов (отведений)

Одной из ключевых особенностей стандартного представления через 12 основных отведений – 6 отведений конечностей $(I, II, III, aVR, aVL, aVF)$ и 6 отведений грудной клетки $(V_1, \ldots, V_6)$ – является взаимосвязь между отведениями конечностей через закон Эйнтховена [28]

$$I + III = II$$

и уравнения Гольдбергера [28]

$$aVL = \frac{I - III}{2}, \qquad -aVR = \frac{I + II}{2}, \quad aVF = \frac{II + III}{2}$$

Таким образом, используя эти уравнения, генерацию шести отведений конечностей I, II, III, aVR, aVL можно заменить генерацией только двух — I и III. В результате в процессе работы модели формируется восемь отведений, которые

затем преобразуются в полный набор из двенадцати. Для генерации этих отведений была определена процедура работы со смесями распределений и формирования каналов, описанная в предыдущем разделе.

Изначально каждое логистическое распределение $P_L(m_1),..,P_L(m_K)$ в смеси возвращает выходные значения методом обратного сэмплирования:

$$P_L(m_i) = \mu_{m_i} + s_{m_i} \log\left(\frac{u}{1-u}\right),$$

где μ_{m_i} и s_{m_i} являются параметрами m_i^{th} логистического распределения, $u \sim \text{Uniform}(0,1)$.

Для этого мы предлагаем алгоритм построения смеси распределений для работы с восемью одномерными ЭКГ-сигналами, точнее — переопределения способа выбора параметров распределений. Связь между каналами (отведениями) определяется следующим образом.

1) Первое отведение I генерируется как $P_L(C_{I_i}) = P_L\left(\mu_{I_i}(C_{I_i}), s_{I_i}(C_{I_i})\right)$, где C_{I_i} — контекстный тензор, получаемый из смеси логистических распределений для отведения I.
2) После этого отведение III генерируется с параметрами $\mu_{III_i}(C_{III_i}, I_i) = \mu_{III_i}(C_{III_i}) + \beta\mu_{I_i}(C_{I_i})$ и $s_{III_i}(C_{III_i}, I_i) = s_{III_i}(C_{III_i}) + \beta s_{I_i}(C_{I_i})$.
3) Грудное отведение V_1 затем генерируется независимо от отведений конечностей с параметрами $\mu_{V_{1,i}}(C_{V_{1i}})$ и $\mu_{V_{1,i}}(C_{V_{1i}})$, так как оно отражает информацию о вертикальных плоскостях, тогда как конечностные — о горизонтальных.
4) Далее, отведение V_2 генерируется с параметрами $\mu_{V_{2,i}}(C_{V_{2,i}}, V_{1,i}) = \mu_{V_{2,i}}(C_{V_{2,i}}) + \alpha(V_{2,i})\mu_{V_{1,i}}(C_{V_{1,i}})$, $s_{V_{2,i}}(C_{V_{2,i}}, V_{1,i}) = s_{V_{2,i}}(C_{V_{2,i}}) + \alpha(V_{2,i}) \cdot s_{V_{1,i}}(C_{V_{1,i}})$.
5) Оставшиеся отведения генерируются согласно параметрам $\mu_{V_{k,i}}(C_{V_{k,i}}, V_{k,i}, V_{k-1,i}, \cdots, V_{1,i}) = \mu_{V_{k,i}}(C_{V_{k,i}}) + \sum_{j=1}^{k-1} \alpha(V_{j_k,i})\mu_{V_{j,i}}(C_{V_{j,i}})$, $s_{V_{k,i}}(C_{V_{k,i}}, V_{k,i}, V_{k-1,i}, \cdots, V_{1,i}) = s_{V_{k,i}}(C_{V_{k,i}}) + \sum_{j=1}^{k-1} \alpha(V_{j_k,i})\, s_{V_{j,i}}\left(C_{V_{j,i}}\right)$.

2. Использование одномерных сверток

Так как каждая ЭКГ представляет собой одномерный сигнал длиной 5000 отсчетов (10 с при частоте 500 Гц), мы адаптировали NVAE для работы с одномерными свертками. Дополнительно был протестирован вариант с двухмерным представлением сигнала в виде спектрограммы STFT [29], однако он показал худшие результаты в последующих задачах. При этом были скорректированы гиперпараметры и операции выравнивания, чтобы корректно обрабатывать сигнал полной длины.

3. Условная генерация

В задаче условной генерации одна модель обучается порождать выборки, соответствующие конкретным классам, вместо того чтобы строить отдельные архитектуры для каждого из них. Такой подход позволяет модели изучать общие закономерности, сохраняя при этом особенности, характерные для отдельных классов [12, 14, 15, 21]. Обычно это реализуется добавлением эмбеддинга класса. В NVAE на верхнем уровне иерархии используется глобальный обучаемый вектор h. В нашей модели cNVAE-ECG этот вектор обогащен эмбеддингом класса c, который в случае многометочной (multi-label) классификации представляет собой сумму эмбеддингов всех присутствующих меток. Вектор c подается как в энкодер, так и в декодер, что позволяет модели формировать глобальное представление пространства ЭКГ, одновременно учитывая признаки конкретных патологий.

ПОСТАНОВКА ЭКСПЕРИМЕНТОВ

1. Задачи и конкуренты

Мы оценивали cNVAE-ECG по решениям двух последующих задач: 1) бинарная классификация сигналов ЭКГ для выявления патологий и 2) многометочная (multi-label) классификация в режиме обучения с переносом (transfer learning) [30, 31]. Решение первой задачи проверяет, улучшает ли добавление сгенерированных сигналов метрики на тестовой выборке, а решение второй — повышает ли предобучение на синтетических ЭКГ способность модели к обобщению. В качестве классификатора использовалась XResNet1d101 [30] — модификация ResNet,

доказавшая эффективность при анализе ЭКГ и сохраняющая информативные признаки при умеренных вычислительных затратах.

В обеих задачах обучающие выборки дополнялись сгенерированными сигналами в разных пропорциях, и измерялось их влияние на качество моделей. В задаче бинарной классификации тестировалось добавление как всех, так и только патологических сигналов, чтобы оценить влияние на баланс классов. В задаче обучения с переносом модели предварительно обучались на сбалансированных синтетических данных, затем они дообучались на целевых наборах и оценивались по метрике AUROC.

Для сравнения cNVAE-ECG сопоставлялась с условными версиями двух моделей на основе GAN — WaveGAN* и Pulse2Pulse [34]. Эти базовые модели были модифицированы для поддержки эмбеддингов классов. Все модели обучались в течение 134 ч на GPU A100, чтобы обеспечить достоверность и сопоставимость экспериментов, а также учитывать экологическую нагрузку, связанную с обучением глубоких нейросетей [32].

2. Датасеты

Мы использовали данные ЭКГ из набора PhysioNet/CinC Challenge 2021 — PTB-XL [33] (21837 записей), Georgia [34] (10344) и Ningbo [35] (34905). Все записи были стандартизированы до длины 10 с при частоте 500 Гц.

Набор PTB-XL применялся для обучения модели cNVAE-ECG, а также генерации синтетических сигналов для предобучения и выполнения задачи бинарной классификации. Для решения этой задачи мы выбрали инфаркт миокарда (MI), поскольку это один из крупнейших классов в наборе данных PTB-XL, отличающийся от синусового ритма (SR). Датасеты Georgia и Ningbo использовались для экспериментов по обучению с переносом в многометочных задачах.

Во всех наборах данных мы выбрали одни и те же семь патологий средней и редкой встречаемостей: отклонение электрической оси влево (LAD), гипертрофия левого желудочка (LVH), фибрилляция предсердий (AF), синусовая тахикардия (STach), атриовентрикулярная блокада I степени (IAVB), синусовая брадикардия (SB) и аномалии зубца T (TAb). Все диагнозы были классифицированы по терминологии SNOMED CT [36].

Для удаления выбросов применялась перцентильная фильтрация (2.5–97.5 %). Табл. 1 суммирует обучающую выборку PTB-XL и показывает для каждого используемого класса количество реальных ЭКГ и дополнительно сгенерированных синтетических примеров, созданных cNVAE-ECG для предобучения. Это позволило оценить поведение методов в различных условиях.

Табл. 1. Распределение классов в датасете PTB-XL

Название класса	SR	MI	LAD	TAb	LVH	AF	STach	SB	IAVB
Реальные сигналы	12.243	3.300	3.276	1.556	908	849	502	456	445
Сгенерированные сигналы	0	8.943	8.967	10.687	11.335	11.394	11.741	11.787	11.798

РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ

1. Количественные результаты

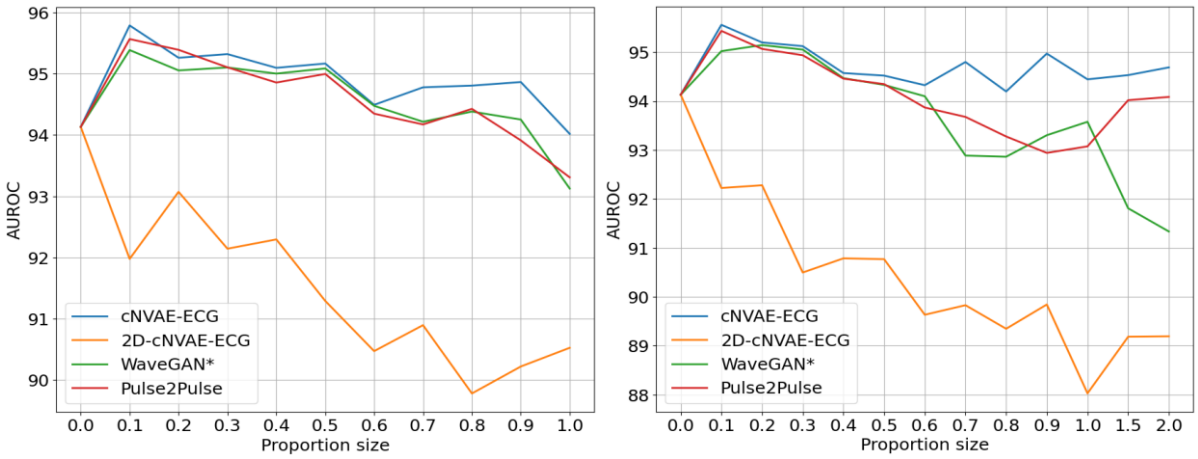


Рис. 2. Результаты экспериментов по задаче обогащения обучающей выборки PTB-XL для бинарной классификации: класса инфаркта миокарда (слева) и обоих классов, включая синусовый ритм (справа)

Левая часть рис. 2 показывает, что при добавлении только сигналов инфаркта миокарда (MI) для снижения дисбаланса классов cNVAE-ECG показывает наилучшие результаты при малых пропорциях, обеспечивая прирост AUROC

до 2%. Напротив, версия 2D-cNVAE-ECG ухудшает результаты. Кроме того, с увеличением доли синтетических данных качество падает, вероятно, из-за ограниченного разнообразия исходных примеров MI, что делает сгенерированные записи менее обобщаемыми при масштабировании.

Как показано на правой части рис. 2, добавление сигналов, сгенерированных моделью cNVAE-ECG для обоих классов, улучшает значение AUROC во всех пропорциях в задаче бинарной классификации. Модель cNVAE-ECG снова превосходит WaveGAN* и Pulse2Pulse при большинстве соотношений, достигая максимального прироста AUROC до 1,5% при долях 0.1–0.3 и 0.9. Эти улучшения показывают, что небольшое добавление синтетических сигналов действует как регуляризатор, а большее количество повышает разнообразие паттернов.

Табл. 2. Значения AUROC (%) на датасете Georgia в зависимости от стратегии предобучения (усреднено по всем пропорциям)

Имя класса	Без предобучения	Реальные данные	Предложенный метод (cNVAE-ECG)	WaveGAN*	Pulse2Pulse
LAD	93.80	94.67	**95.23**	95.19	93.98
Tab	89.77	92.07	**92.40**	91.23	88.40
LVH	92.51	97.33	**97.99**	96.30	94.45
AF	91.26	93.15	**93.67**	91.44	90.71
STach	98.46	**99.43**	99.39	98.99	98.37
SB	86.70	**88.19**	87.99	86.42	83.97
IAVB	91.18	**93.81**	93.17	92.26	89.73

Из табл. 2 видно, что использование предобучения для датасета Georgia в целом улучшает качество модели по всем классам. Особенно предобучение, обогащенное cNVAE-ECG, демонстрирует наилучшие результаты среди методов, основанных на GAN, превосходя базовое предобучение для четырех самых распространенных (согласно табл. 1) классов в наборе, использованном при обучении cNVAE-ECG. Это наблюдение указывает, что cNVAE-ECG уловила зависимости для

редких классов слабее, чем для более распространенных. Кроме того, в некоторых случаях модели Pulse2Pulse и WaveGAN* показывают худшие результаты, чем без предобучения вовсе.

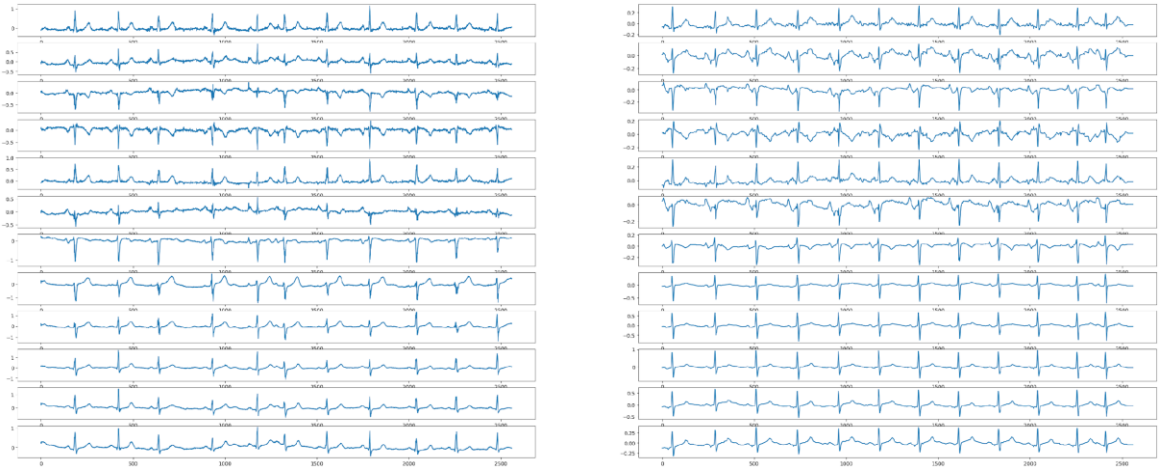
Табл. 3. Значения AUROC (%) на датасете Ningbo в зависимости от стратегии предобучения (усреднено по всем пропорциям)

Имя класса	Без предобучения	Реальные данные	Предложенный метод (cNVAE-ECG)	Wave-GAN*	Pulse2Pulse
LAD	97.79	97.84	**98.02**	97.73	97.62
TAb	88.86	89.55	**89.65**	89.59	88.55
LVH	91.58	91.63	**92.08**	90.73	89.67
STach	98.49	99.02	**99.50**	99.07	99.05
SB	99.71	**99.80**	99.79	99.74	99.71
IAVB	96.74	**97.56**	97.38	97.24	96.38

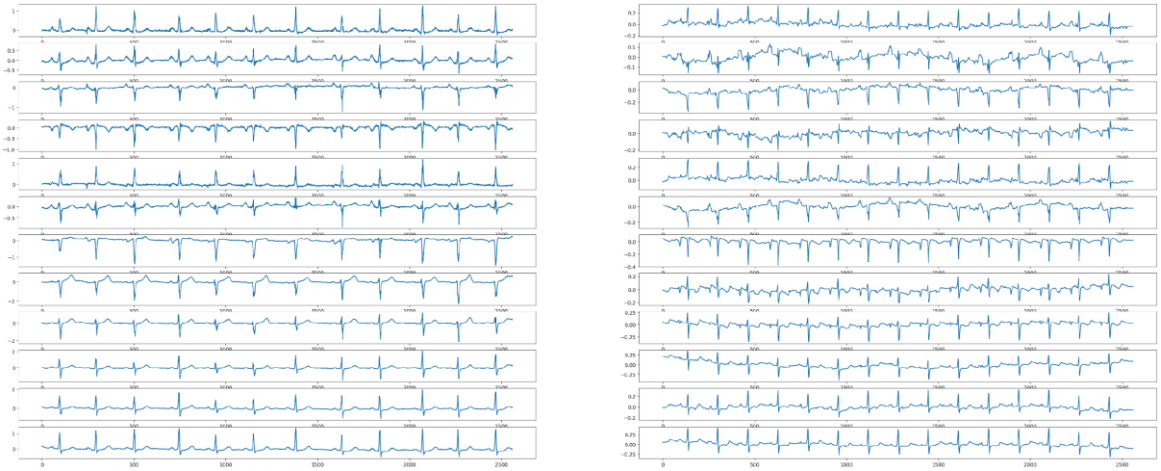
Табл. 3 показывает, что для большинства классов добавление сигналов, сгенерированных с помощью cNVAE-ECG и моделей, подобных GAN, дает приемлемый прирост качества на тестовом наборе Ningbo по сравнению с результатами без предобучения. Модель cNVAE-ECG снова демонстрирует лучшие результаты по всем классам по сравнению с другими генеративными методами. Однако для редких классов, таких как SB и IAVB, качество оказалось ниже, чем при использовании исходного предобучения. С учетом предыдущих результатов можно заключить, что для этих классов у cNVAE-ECG недостаточно данных, чтобы успешно генерировать соответствующие типы ЭКГ-сигналов.

2. Качественные результаты

На рис. 3 представлены реальные и сгенерированные сигналы ЭКГ для синусового ритма (SR) и инфаркта миокарда (MI). Несмотря на наличие отдельных артефактов, сгенерированные сигналы воспроизводят основные волновые характеристики, наблюдаемые в обучающих данных.



а)



б)

Рис. 3. Результаты генерации ЭКГ-сигналов (в правой части) по сравнению с оригинальными сигналами (в левой части) для классов: а) синусовый ритм, б) инфаркта миокарда

Мы также провели более детальное сравнение одного сердечного цикла для отведения I синусового ритма (см. на рис. 4).

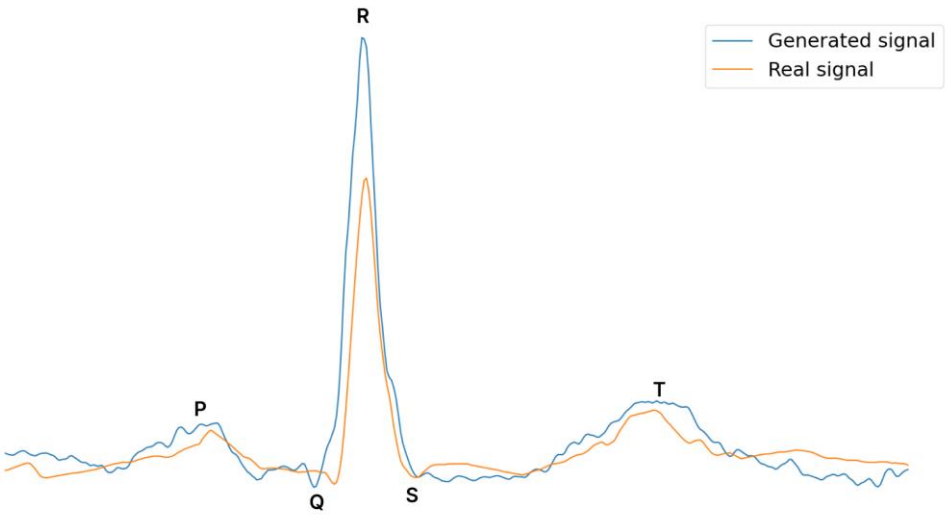


Рис. 4. Сравнение реального и сгенерированного одного сердечного цикла для отведения I в классе синусового ритма

Это сравнение показывает, что модель cNVAE-ECG успешно воспроизводит основную структуру и характерные особенности реального ЭКГ-сигнала, такие как зубцы P, Q, R, S и T [37].

ЗАКЛЮЧЕНИЕ

Мы представили cNVAE-ECG — модификацию NVAE для условной генерации 12-отведенных 10-секундных ЭКГ. Модель продемонстрировала стабильное улучшение показателя AUROC по сравнению с базовыми моделями на основе GAN (WaveGAN*, Pulse2Pulse) как в задаче бинарной классификации, так и в задаче многометочного обучения с переносом. Предобучение с использованием cNVAE-ECG улучшило перенос на датасеты Georgia и Ningbo, особенно для классов средней частоты, хотя редкие патологии остаются сложными для генерации. Сгенерированные ЭКГ в целом сохраняют ключевые волновые характеристики при наличии лишь незначительных артефактов. В дальнейшем планируется провести клиническую валидацию и реализовать модель в федеративной среде для повышения ее обобщающей способности и устойчивости [38].

СПИСОК ЛИТЕРАТУРЫ

1. *Tsao C.W., Aday A.W., Almarzooq Z.I., et al.* Heart Disease and Stroke Statistics–2023 Update: A Report from the American Heart Association // Circulation. 2023. Vol. 147, No. 8. P. e93–e621.

2. *Liu X., Wang H., Li Z., Qin L.* Deep learning in ECG diagnosis: A review // Knowledge-Based Systems. 2021. Vol. 227. P. 107187.

3. *Gerke S., Minssen T., Cohen G.* Ethical and legal challenges of artificial intelligence-driven healthcare // Artificial Intelligence in Healthcare / Eds. A. Bohr, K. Memarzadeh. Academic Press, 2020. P. 295–336.

4. *Reyna M.A., Sadr N., Alday E.A.P., et al.* Will two do? Varying dimensions in electrocardiography: The PhysioNet/Computing in Cardiology Challenge 2021 // Computing in Cardiology. 2021. Vol. 48. P. 1–4.

5. *Friesen G., Jannett T., Jadallah M., Yates S., Quint S., Nagle H.* A comparison of the noise sensitivity of nine QRS detection algorithms // IEEE Transactions on Biomedical Engineering. 1990. Vol. 37, No. 1. P. 85–98.

6. *Maron B.J., Friedman R.A., Kligfield P., et al.* Assessment of the 12-lead ECG as a screening test for detection of cardiovascular disease in healthy general populations of young people (12–25 years of age) // Circulation. 2014. Vol. 130, No. 15. P. 1303–1334.

7. *Kingma D.P., Welling M.* Auto-encoding variational Bayes. 2022.

8. *Vahdat A., Kautz J.* NVAE: A deep hierarchical variational autoencoder. 2020.

9. *Golany T., Radinsky K.* PGANs: Personalized generative adversarial networks for ECG synthesis to improve patient-specific deep ECG classification // Proceedings of the AAAI Conference on Artificial Intelligence. 2019. Vol. 33, No. 1. P. 557–564.

10. *Yang H., Liu J., Zhang L., Li Y., Zhang H.* ProEGAN-MS: A progressive growing generative adversarial networks for electrocardiogram generation // IEEE Access. 2021. Vol. 9. P. 52089–52100.

11. *Golany T., Radinsky K., Freedman D.* SimGANs: Simulator-based generative adversarial networks for ECG synthesis to improve deep ECG classification // Proceedings of the 37th International Conference on Machine Learning (PMLR). 2020. Vol. 119. P. 3597–3606.

12. *Nankani D., Baruah R.D.* Investigating deep convolution conditional GANs for electrocardiogram generation // 2020 International Joint Conference on Neural Networks (IJCNN). 2020. P. 1–8.

13. *Thambawita V., Isaksen J.L., Hicks S.A., et al.* DeepFake electrocardiograms using generative adversarial networks are the beginning of the end for privacy issues in medicine // Scientific Reports. 2021. Vol. 11. P. 21896.

14. *Wu J., Wang L., Pan H., Wang B.* MLCGAN: Multi-lead ECG synthesis with multi label conditional generative adversarial network // ICASSP 2023 – IEEE International Conference on Acoustics, Speech and Signal Processing. 2023. P. 1–5.

15. *Alcaraz J.M.L., Strodthoff N.* Diffusion-based conditional ECG generation with structured state space models. 2023.

16. *Ho J., Jain A., Abbeel P.* Denoising diffusion probabilistic models // CoRR. 2020. abs/2006.11239.

17. *Dhariwal P., Nichol A.* Diffusion models beat GANs on image synthesis. 2021.

18. *Xia Y., Wang W., Wang K.* ECG signal generation based on conditional generative models // Biomedical Signal Processing and Control. 2023. Vol. 82. P. 104587.

19. *El-Kaddoury M., Mahmoudi A., Himmi M.M.* Deep generative models for image generation: A practical comparison between variational autoencoders and generative adversarial networks // Mobile, Secure, and Programmable Networking. Cham: Springer, 2019. P. 1–8.

20. *Kuznetsov V., Moskalenko V., Gribanov D., Zolotykh N.* Interpretable feature generation in ECG using a variational autoencoder // Frontiers in Genetics. 2021. Vol. 12. P. 638191.

21. *Sang Y., Beetz M., Grau V.* Generation of 12-lead electrocardiogram with subject-specific, image-derived characteristics using a conditional variational autoencoder // 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI). 2022. P. 1–5.

22. *Beetz M., Banerjee A., Sang Y., Grau V.* Combined generation of electrocardiogram and cardiac anatomy models using multi-modal variational autoencoders // 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI). 2022. P. 1–4.

23. *Salimans T., Karpathy A., Chen X., Kingma D.P.* PixelCNN++: Improving the PixelCNN with discretized logistic mixture likelihood and other modifications. 2017.

24. *Deng L.* The MNIST database of handwritten digit images for machine learning research [Best of the Web] // IEEE Signal Processing Magazine. 2012. Vol. 29, No. 6. P. 141–142.

25. *Krizhevsky A.* Learning multiple layers of features from tiny images. 2009.

26. *Larsen A.B.L., Sønderby S.K., Larochelle H., Winther O.* Autoencoding beyond pixels using a learned similarity metric. 2016.

27. *Karras T., Aila T., Laine S., Lehtinen J.* Progressive growing of GANs for improved quality, stability, and variation. 2018.

28. *Malmivuo J., Plonsey R.* Bioelectromagnetism. 15. 12-Lead ECG System. 1975. P. 277–289.

29. *Griffin D., Lim J.* Signal estimation from modified short-time Fourier transform // IEEE Transactions on Acoustics, Speech, and Signal Processing. 1984. Vol. 32, No. 2. P. 236–243.

30. *Strodthoff N., Wagner P., Schaeffter T., Samek W.* Deep learning for ECG analysis: Benchmarks and insights from PTB-XL // IEEE Journal of Biomedical and Health Informatics. 2021. Vol. 25, No. 5. P. 1519–1528.

31. *Donahue C., McAuley J., Puckette M.* Adversarial audio synthesis. 2019.

32. *Wu C.-J., Raghavendra R., Gupta U., et al.* Sustainable AI: Environmental implications, challenges and opportunities // ArXiv. 2021. abs/2111.00364.

33. *Wagner P., Strodthoff N., Bousseljot R.-D., et al.* PTB-XL, a large publicly available electrocardiography dataset // Scientific Data. 2020. Vol. 7. P. 154.

34. *Goldberger A., Amaral L., Glass L., et al.* PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals // Circulation. 2000. Vol. 101. P. e215–e220.

35. *Zheng J., Chu H., Struppa D., et al.* Optimal multi-stage arrhythmia classification approach // Scientific Reports. 2020. Vol. 10.

36. *El-Sappagh S., Franda F., Ali F., Kwak K.-S.* SNOMED CT standard ontology based on the ontology for general medical science // BMC Medical Informatics and Decision Making. 2018. Vol. 18, No. 1. P. 76.

37. *Berkaya S.K., Uysal A.K., Gunal E.S., et al.* A survey on ECG analysis // Biomedical Signal Processing and Control. 2018. Vol. 43. P. 216–235.

38. *Zhang M., Wang Y., Luo T.* Federated learning for arrhythmia detection of non-IID ECG // 2020 IEEE 6th International Conference on Computer and Communications (ICCC). 2020. P. 1176–1180.

CONDITIONAL ELECTROCARDIOGRAM GENERATION USING HIERARCHICAL VARIATIONAL AUTOENCODERS

I. A. Sviridov[1] [0009-0009-5912-1118], K. S. Egorov[2] [0009-0006-6991-4136]

[1, 2]*Sber AI Lab, Moscow, Russia*

[1]ianatosviridov@sberbank.ru, [2]Egorov.K.Ser@sberbank.ru

Abstract

Cardiovascular diseases remain the leading cause of mortality, and automated electrocardiogram (ECG) analysis can ease clinical workloads but is limited by scarce and imbalanced data. Synthetic ECG can mitigate these issues, and while most methods use Generative Adversarial Networks (GANs), recent work show variational autoencoders (VAEs) perform comparably. We introduce **cNVAE-ECG**, a conditional Nouveau VAE (NVAE) that generates high-resolution, 12-lead, 10-second ECGs with multiple pathologies. Leveraging a compact channel-generation scheme and class embeddings for multi-label conditioning, cNVAE-ECG improves downstream binary and multi-label classification, achieving up to a 2% AUROC gain in transfer learning over GAN-based models.

Keywords*: ECG, variational autoencoder, conditional generation, GAN.*

REFERENCES

1. *Tsao C.W., Aday A.W., Almarzooq Z.I., et al.* Heart Disease and Stroke Statistics–2023 Update: A Report from the American Heart Association // Circulation. 2023. Vol. 147, No. 8. P. e93–e621.

2. *Liu X., Wang H., Li Z., Qin L.* Deep learning in ECG diagnosis: A review // Knowledge-Based Systems. 2021. Vol. 227. P. 107187.

3. *Gerke S., Minssen T., Cohen G.* Ethical and legal challenges of artificial intelligence-driven healthcare // Artificial Intelligence in Healthcare / Eds. A. Bohr, K. Memarzadeh. Academic Press, 2020. P. 295–336.

4. *Reyna M.A., Sadr N., Alday E.A.P., et al.* Will two do? Varying dimensions in electrocardiography: The PhysioNet/Computing in Cardiology Challenge 2021 // Computing in Cardiology. 2021. Vol. 48. P. 1–4.

5. *Friesen G., Jannett T., Jadallah M., Yates S., Quint S., Nagle H.* A comparison of the noise sensitivity of nine QRS detection algorithms // IEEE Transactions on Biomedical Engineering. 1990. Vol. 37, No. 1. P. 85–98.

6. *Maron B.J., Friedman R.A., Kligfield P., et al.* Assessment of the 12-lead ECG as a screening test for detection of cardiovascular disease in healthy general populations of young people (12–25 years of age) // Circulation. 2014. Vol. 130, No. 15. P. 1303–1334.

7. *Kingma D.P., Welling M.* Auto-encoding variational Bayes. 2022.

8. *Vahdat A., Kautz J.* NVAE: A deep hierarchical variational autoencoder. 2020.

9. *Golany T., Radinsky K.* PGANs: Personalized generative adversarial networks for ECG synthesis to improve patient-specific deep ECG classification // Proceedings of the AAAI Conference on Artificial Intelligence. 2019. Vol. 33, No. 1. P. 557–564.

10. *Yang H., Liu J., Zhang L., Li Y., Zhang H.* ProEGAN-MS: A progressive growing generative adversarial networks for electrocardiogram generation // IEEE Access. 2021. Vol. 9. P. 52089–52100.

11. *Golany T., Radinsky K., Freedman D.* SimGANs: Simulator-based generative adversarial networks for ECG synthesis to improve deep ECG classification // Proceedings of the 37th International Conference on Machine Learning (PMLR). 2020. Vol. 119. P. 3597–3606.

12. *Nankani D., Baruah R.D.* Investigating deep convolution conditional GANs for electrocardiogram generation // 2020 International Joint Conference on Neural Networks (IJCNN). 2020. P. 1–8.

13. *Thambawita V., Isaksen J.L., Hicks S.A., et al.* DeepFake electrocardiograms using generative adversarial networks are the beginning of the end for privacy issues in medicine // Scientific Reports. 2021. Vol. 11. P. 21896.

14. *Wu J., Wang L., Pan H., Wang B.* MLCGAN: Multi-lead ECG synthesis with multi label conditional generative adversarial network // ICASSP 2023 – IEEE International Conference on Acoustics, Speech and Signal Processing. 2023. P. 1–5.

15. *Alcaraz J.M.L., Strodthoff N.* Diffusion-based conditional ECG generation with structured state space models. 2023.

16. *Ho J., Jain A., Abbeel P.* Denoising diffusion probabilistic models // CoRR. 2020. abs/2006.11239.

17. *Dhariwal P., Nichol A.* Diffusion models beat GANs on image synthesis. 2021.

18. *Xia Y., Wang W., Wang K.* ECG signal generation based on conditional generative models // Biomedical Signal Processing and Control. 2023. Vol. 82. P. 104587.

19. *El-Kaddoury M., Mahmoudi A., Himmi M.M.* Deep generative models for image generation: A practical comparison between variational autoencoders and generative adversarial networks // Mobile, Secure, and Programmable Networking. Cham: Springer, 2019. P. 1–8.

20. *Kuznetsov V., Moskalenko V., Gribanov D., Zolotykh N.* Interpretable feature generation in ECG using a variational autoencoder // Frontiers in Genetics. 2021. Vol. 12. P. 638191.

21. *Sang Y., Beetz M., Grau V.* Generation of 12-lead electrocardiogram with subject-specific, image-derived characteristics using a conditional variational autoencoder // 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI). 2022. P. 1–5.

22. *Beetz M., Banerjee A., Sang Y., Grau V.* Combined generation of electrocardiogram and cardiac anatomy models using multi-modal variational autoencoders // 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI). 2022. P. 1–4.

23. *Salimans T., Karpathy A., Chen X., Kingma D.P.* PixelCNN++: Improving the PixelCNN with discretized logistic mixture likelihood and other modifications. 2017.

24. *Deng L.* The MNIST database of handwritten digit images for machine learning research [Best of the Web] // IEEE Signal Processing Magazine. 2012. Vol. 29, No. 6. P. 141–142.

25. *Krizhevsky A.* Learning multiple layers of features from tiny images. 2009.

26. *Larsen A.B.L., Sønderby S.K., Larochelle H., Winther O.* Autoencoding beyond pixels using a learned similarity metric. 2016.

27. *Karras T., Aila T., Laine S., Lehtinen J.* Progressive growing of GANs for improved quality, stability, and variation. 2018.

28. *Malmivuo J., Plonsey R.* Bioelectromagnetism. 15. 12-Lead ECG System. 1975. P. 277–289.

29. *Griffin D., Lim J.* Signal estimation from modified short-time Fourier transform // IEEE Transactions on Acoustics, Speech, and Signal Processing. 1984. Vol. 32, No. 2. P. 236–243.

30. *Strodthoff N., Wagner P., Schaeffter T., Samek W.* Deep learning for ECG analysis: Benchmarks and insights from PTB-XL // IEEE Journal of Biomedical and Health Informatics. 2021. Vol. 25, No. 5. P. 1519–1528.

31. *Donahue C., McAuley J., Puckette M.* Adversarial audio synthesis. 2019.

32. *Wu C.-J., Raghavendra R., Gupta U., et al.* Sustainable AI: Environmental implications, challenges and opportunities // ArXiv. 2021. abs/2111.00364.

33. *Wagner P., Strodthoff N., Bousseljot R.-D., et al.* PTB-XL, a large publicly available electrocardiography dataset // Scientific Data. 2020. Vol. 7. P. 154.

34. *Goldberger A., Amaral L., Glass L., et al.* PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals // Circulation. 2000. Vol. 101. P. e215–e220.

35. *Zheng J., Chu H., Struppa D., et al.* Optimal multi-stage arrhythmia classification approach // Scientific Reports. 2020. Vol. 10.

36. *El-Sappagh S., Franda F., Ali F., Kwak K.-S.* SNOMED CT standard ontology based on the ontology for general medical science // BMC Medical Informatics and Decision Making. 2018. Vol. 18, No. 1. P. 76.

37. *Berkaya S.K., Uysal A.K., Gunal E.S., et al.* A survey on ECG analysis // Biomedical Signal Processing and Control. 2018. Vol. 43. P. 216–235.

38. *Zhang M., Wang Y., Luo T.* Federated learning for arrhythmia detection of non-IID ECG // 2020 IEEE 6th International Conference on Computer and Communications (ICCC). 2020. P. 1176–1180.

СВЕДЕНИЯ ОБ АВТОРАХ

СВИРИДОВ Иван Анатольевич – получил степень бакалавра по прикладной математике и информатике в МГТУ им. Н.Э. Баумана (Москва, Россия) в 2021 году и степень магистра по компьютерным наукам в Высшей школе экономики (Москва, Россия) в 2023 году. С 2023 года по настоящее время работает исследователем в Центре практического искусственного интеллекта Sber AI Lab (Москва, Россия), где занимается исследованиями в области искусственного интеллекта в медицине. Области его научных интересов: большие языковые модели, многоагентные системы, генеративное моделирование, биосигналы.

Ivan Anatolevich SVIRIDOV – received a B.S. degree in applied mathematics and informatics from Bauman Moscow State Technical University, Moscow, Russia, in 2021 and an M.S. degree in computer science from the Higher School of Economics, Moscow, Russia, in 2023. From 2023 to the present, he has been a researcher at Sber AI Lab in Moscow, Russia, mainly focused on AI in medicine. His research interests are: large language models, multi-agent systems, generative modeling, and biosignals.

email: ianatosviridov@sberbank.ru

ORCID 0009-0009-5912-1118

ЕГОРОВ Константин Сергеевич – получил степень магистра в Московском государственном техническом университете им. Н.Э. Баумана (Москва, Россия). Свою карьеру начал инженером-электронщиком, специализируясь на слаботочных системах, с 2010 по 2018 год. С 2018 года работает исследователем в Центре практического искусственного интеллекта Sber AI Lab, где основное внимание уделяет медицинским сигналам, таким как ЭКГ, ЭЭГ и PPG.

Konstantin Sergeevich EGOROV – received a master's degree from Bauman Moscow State Technical University, Moscow, Russia. He began his career as an electronics engineer, specializing in low-current systems from 2010 to 2018. Since 2018, he has been a researcher at Sber AI Lab, where his primary focus is on medical signals such as ECG, EEG, and PPG.

email: Egorov.K.Ser@sberbank.ru

ORCID 0009-0006-6991-4136

Материал поступил в редакцию 12 октября 2025 года

УДК 004.8

ГДЕ НАХОДЯТСЯ ЛУЧШИЕ ПРИЗНАКИ? ПОСЛОЙНЫЙ АНАЛИЗ СЛОЕВ ТРАНСФОРМЕРА ДЛЯ ЭФФЕКТИВНОЙ КЛАССИФИКАЦИИ ЭНДОСКОПИЧЕСКИХ ИЗОБРАЖЕНИЙ

А. Таха[1] [0009-0006-6346-4162], Р. А. Лукманов[2] [0000-0001-9257-7410]

[1, 2]Университет Иннополис, г. Иннополис, Россия

[1]Центр искусственного интеллекта университета Иннополис, г. Иннополис, Россия

[1]a.taha@innopolis.university, [2]r.lukmanov@innopolis.university

Аннотация

В поисках путей развития медицинского искусственного интеллекта показано, что предварительно обученный Vision Transformer с линейным классификатором может достигать высокой и конкурентоспособной производительности в классификации эндоскопических изображений. Представлен систематический послойный анализ, который выявляет источник наиболее важных признаков, оспаривая общепринятую эвристику использования только последнего слоя. Установлен отчетливый феномен «пика перед концом», когда поздне-промежуточный слой предлагает более обобщаемое представление для последующей медицинской задачи. На стандартных наборах данных Kvasir и HyperKvasir предложенный подход с малым количеством параметров не только получить достаточно высокую точность, но и значительно сокращает вычислительные затраты. Полученные работы могут быть рекомендованы в качестве практического руководства по эффективному использованию признаков общих базовых моделей в клинических условиях.

Ключевые слова*: классификация эндоскопических изображений, замороженный кодировщик, извлечение признаков, послойный анализ, визуальный трансформер (ViT), перенос обучения, самоконтролируемое обучение (SSL), медицинский искусственный интеллект.*

© А. Таха, Р. А. Лукманов, 2025.
Данная статья распространяется на условиях международной лицензии Creative Commons License Attribution 4.0 International (CC BY 4.0).

ВВЕДЕНИЕ

Эндоскопия желудочно-кишечного тракта (ЖКТ) является краеугольным камнем в диагностике и лечении широкого спектра заболеваний, от воспалительных заболеваний кишечника (ВЗК) до предотвращения колоректального рака путем удаления предраковых полипов [1–3]. Однако эффективность эндоскопии ограничена человеческой интерпретацией, при этом частота пропущенных аденом во время колоноскопии достигает 26% [4, 5]. Для снижения таких диагностических ошибок в качестве перспективного решения появились системы автоматизированной диагностики (САД), основанные на искусственном интеллекте (ИИ) [6].

Современное состояние в эндоскопической САД представлено моделями глубокого обучения (ГО), в частности сверточными нейронными сетями (СНС) [7], такими как ResNet [8], а в последнее время и Vision Transformers (ViT) [9]. Типичная методология применения этих моделей — это полное дообучение (full fine-tuning), при котором модель, предварительно обученная на крупномасштабном наборе данных (например, ImageNet [10]), адаптируется к эндоскопической задаче путем переобучения всех ее параметров или обучения с нуля на специфических эндоскопических наборах данных. Хотя это и дает хорошие результаты, практическое внедрение является серьезным препятствием. Полное дообучение и обучение с нуля требуют значительных ресурсов ГП и длительного времени обучения, что создает барьер для исследовательских и клинических учреждений [11]. Кроме того, эти модели требуют большого количества данных, а стоимость приобретения больших наборов медицинских данных, аннотированных экспертами, является еще одной проблемой в области анализа медицинских изображений [12].

Для преодоления этих проблем в данной работе исследована более эффективная парадигма: использование предварительно обученной модели в качестве фиксированного экстрактора признаков. В этом подходе глубокий кодировщик остается замороженным, а обучается только простой, легковесный неглубокий декодер на высокоуровневых признаках, извлеченных из кодировщика. Этот метод значительно сокращает количество обучаемых параметров и уменьшает время обучения с часов или дней до минут, а также решает проблему нехватки

данных. Успех этого подхода основан на предположении, что крупномасштабные, предварительно обученные кодировщики с богатыми, обобщаемыми признаками достаточно натренированы для решения последующих медицинских задач без дальнейшей модификации.

Эта парадигма ставит два фундаментальных вопроса:

1) Может ли вычислительно простая модель, состоящая из фиксированного кодировщика и неглубокого декодера, достичь или даже превзойти производительность сложных, полностью дообученных систем в классификации эндоскопических изображений?

2) Если да, то где в этом кодировщике находятся лучшие признаки для этой задачи?

Хотя выбор признаков является известной техникой, обычно оно выполняется из последнего слоя, и систематический анализ качества признаков по всей глубине сети отсутствовал. Наша основная гипотеза заключается в том, что оптимальное представление признаков находится не в последнем слое, а в промежуточном слое i^*. Мы можем формально определить это как задачу оптимизации, где мы стремимся найти индекс слоя i^*, который минимизирует потери на валидации $\mathcal{L}_{\text{val}}$ для декодера $g_{\theta_i^*}$, обученного на признаках из этой слоя:

$$i^* = \arg\min_{i \in \{1,\dots,N\}} \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{val}}}\left[\mathcal{L}_{\text{CE}}\left(g_{\theta_i^*}\left(\mathbf{h}_i(x)\right), y\right)\right],$$

где $\mathbf{h}_i(x)$ – вектор признаков из слоя i, θ_i^* – оптимальные веса декодера для этого слоя.

Целью настоящей работы являются:

- **достижение высокой производительности с эффективной моделью:** мы демонстрируем, что простой классификатор, обученный на признаках из оптимального слоя замороженного кодировщика, достигает отличных результатов на наборах данных HyperKvasir и других бенчмарках;
- **новый послойный анализ признаков:** мы представим, насколько нам известно, первый систематический послойный анализ качества признаков из замороженного кодировщика для эндоскопической классификации на стандартных отраслевых бенчмарках;

- **количественная оценка эффективности и визуальное подтверждение:** мы количественно оценим экономию вычислительных ресурсов в нашем подходе и предоставим качественные доказательства с помощью визуализации t-SNE с целью подтверждения разделимости классов извлеченных вложений, отражает важность трансферного обучения и позволяет создавать интерпретируемые визуализации признаков..

2. СВЯЗАННЫЕ РАБОТЫ

Рассмотрим ландшафт глубокого обучения в эндоскопии ЖКТ, от доминирующей парадигмы полного дообучения до более эффективных стратегий трансферного обучения.

2.1. Полное дообучение в эндоскопии

Доминирующей парадигмой в анализе эндоскопических изображений является полное дообучение, при котором все параметры модели, предварительно обученной на общем наборе данных ImageNet обновляются на целевых медицинских данных для достижения самых современных (SOTA) результатов [13]. На бенчмарк-наборах данных Kvasir [14] и HyperKvasir [15] полностью дообученные CHC, такие как DenseNet-201 [16] и ResNet-101, продемонстрировали точность классификации, превышающую 95–97%. В последнее время Vision Transformers (ViT) в задачах эндоскопической классификации достигли даже лучших результатов, чем CHC [17]. Однако это сопряжено с большими вычислительными затратами, что является значительным препятствием для быстрого экспериментирования и клинического внедрения [18].

2.2. Эффективное трансферное обучение и извлечение признаков

Чтобы смягчить вычислительную нагрузку полного дообучения, были разработаны более эффективные стратегии. Методы параметроэффективного дообучения (PEFT), такие как LoRA [19] или техники разреживания [20], обновляют лишь небольшую долю параметров модели, что позволяет снизить затраты на обучение. Другим подходом является извлечение признаков, при котором весь предварительно обученный кодировщик замораживается и обучается только простой классификационный декодер на признаках, которые он производит. Этот метод,

также известный как линейное зондирование (linear probing), когда используется линейный слой [21], сокращает время обучения с часов до минут.

Однако извлечение признаков в медицинской визуализации обычно основывалось на эвристике использования только последнего слоя кодировщика. Этот подход упускает богатую, специфичную для задачи информацию, доступную в промежуточных слоях [22, 23]. Насколько нам известно, систематический послойный анализ для определения оптимального источника признаков для эндоскопической классификации не проводился, он и рассматривается в настоящей работе.

3. МЕТОДОЛОГИЯ

Наша методология использует простой конвейер для выделения качества характеристик предварительно обученных признаков в качестве основной переменной. Наша экспериментальная структура включает замороженную основу Vision Transformer (ViT) для генерации признаков, процесс послойного извлечения и неглубокую обучаемую классификационную надстройку.

3.1. Архитектурный конвейер

Предлагаемая архитектура показана на рис. 1. Входное эндоскопическое изображение сначала проходит через предварительно обученный и полностью замороженный кодировщик ViT. Мы можем формально определить кодировщик Φ как композицию из N блоков трансформера, $\Phi = L_N \circ \ldots \circ L_1$. Затем мы перехватываем выходную карту признаков из определенного промежуточного слоя L_i, где $i \in \{1,2, \ldots ,24\}$. Эта высокоразмерная карта признаков $\mathbf{z}_i(x) = (L_i \circ \ldots \circ L_1)(x)$ агрегируется в единый вектор признаков $\mathbf{h}_i(x)$, который служит входом для неглубокого обучаемого декодера, ответственного за конечное предсказание класса. Конечное распределение вероятностей $\hat{y}$ задается формулой

$$\hat{y} = \text{softmax}(\mathbf{W}\mathbf{h}_i(x) + \mathbf{b}),$$

где $\{\mathbf{W}, \mathbf{b}\}$ – единственные обучаемые параметры модели. Весь этот процесс повторяется независимо для каждого из 24 слоев кодировщика.

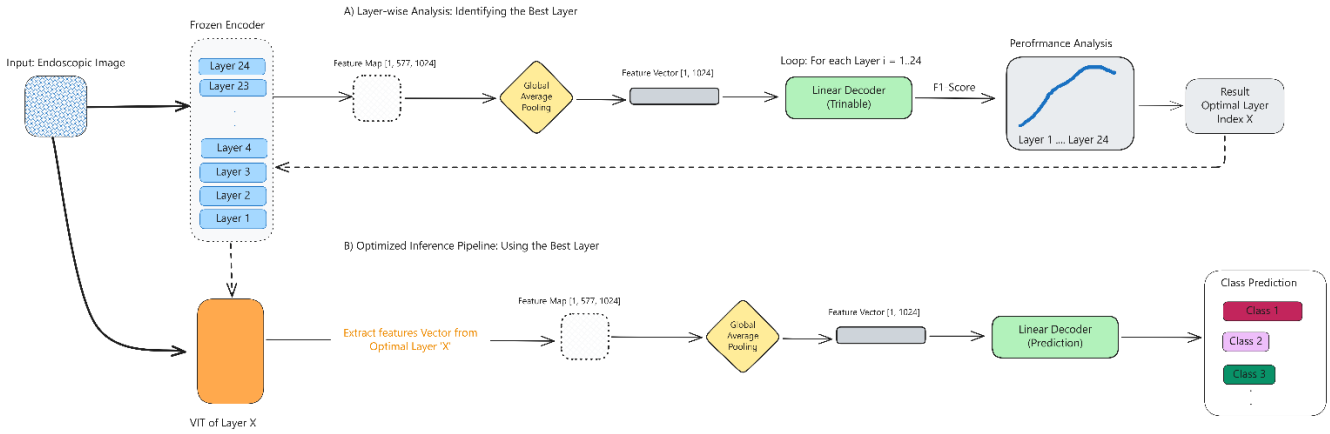


Рис. 1. Обзор предлагаемого конвейера классификации.

Примечание. Входное изображение обрабатывается замороженным кодировщиком Vision Transformer (ViT). Мы извлекаем вложения токенов патчей из определенного промежуточного слоя *X*. Эти вложения агрегируются с помощью глобального среднего пулинга для формирования единого вектора признаков, который затем подается в минималистичный линейный декодер для окончательной классификации. Декодер является единственным обучаемым компонентом в архитектуре.

3.2. Наборы данных и предварительная обработка

Для обеспечения надежности и обобщаемости полученных результатов мы проверили верификацию предложенного метода на трех широко известных публичных наборах данных: Kvasir-V1 и Kvasir-V2 [14], которые являются сбалансированными наборами данных, содержащими 4000 и 8000 изображений соответственно, по 8 классам находок в ЖКТ, и HyperKvasir [15], крупномасштабный набор данных, из которого мы формируем задачу классификации на 8 классах из 8531 изображения, чтобы соответствовать нашим другим экспериментам. Для всех экспериментов мы разделяем данные на обучающий (80%) и валидационный (20%) наборы, используя стратифицированную выборку для сохранения распределения классов в обеих частях. Все изображения изменяются до нативного разрешения кодировщика 336× 336 пикселей и нормализуются с использованием стандартного среднего значения и стандартного отклонения ImageNet [10].

3.3. Замороженный кодировщик и послойное извлечение признаков

$$\mathcal{L}_{\text{pretrain}} = \frac{1}{2B}\sum_{i=1}^{B}\left(\mathcal{L}_i^{(v\to u)} + \mathcal{L}_i^{(u\to v)}\right),$$

где

$$\mathcal{L}_i^{(v\to u)} = -\log\frac{\exp\left(\frac{\text{sim}(\mathbf{v}_i,\mathbf{u}_i)}{\tau}\right)}{\sum_{j=1}^{B}\exp\left(\frac{\text{sim}(\mathbf{v}_i,\mathbf{u}_j)}{\tau}\right)},$$

$$\mathcal{L}_i^{(u\to v)} = -\log\frac{\exp\left(\frac{\text{sim}(\mathbf{u}_i,\mathbf{v}_i)}{\tau}\right)}{\sum_{j=1}^{B}\exp\left(\frac{\text{sim}(\mathbf{u}_i,\mathbf{v}_j)}{\tau}\right)}.$$

Здесь $\mathcal{L}_i^{(v\to u)}$ – потери от изображения к тексту, которые приближают вложение изображения $\mathbf{v}_i$ к соответствующему текстовому вложению $\mathbf{u}_i$; $\mathcal{L}_i^{(u\to v)}$ – потери от текста к изображению, которые обеспечивают близость текстового вложения $\mathbf{u}_i$ к парному изображению $\mathbf{v}_i$. Это было сделано для создания надежного общего пространства вложений.

Для проведения послойного анализа мы регистрируем прямой хук (forward hook) на каждом из 24 остаточных блоков ViT. Для такого входного изображения выход блока на слое i представляет собой тензор формы [1, 577, 1024], соответственно представляющий размер батча, длину последовательности и размерность признаков. Длина последовательности 577 состоит из одного токена [CLS] и 576 токенов патчей (24×24).

Для принципиального сравнения по всем слоям мы отбрасываем специализированный токен [CLS], репрезентативная система которого используется для обработки последнего слоя и не является однородной по всей глубине сети. Вместо этого мы сосредотачиваемся на сетке из 576 токенов патчей, которая представляет пространственную карту признаков изображения на любом данном слое i. Мы агрегируем эти вложения патчей с помощью глобального среднего пулинга

[24] для получения единого, концептуально последовательного вектора признаков, что позволяет провести справедливую оценку основных визуальных признаков на каждой глубине.

3.4. Легковесный классификационный декодер

Мы используем простой декодер: один полносвязный линейный слой без скрытых слоев или нелинейных активаций. Он напрямую отображает 1024-мерный вектор признаков из кодировщика в N_{classes} выходных логитов для классификации.

Этот выбор намеренно минимизирует количество обучаемых параметров, что изолирует вклад в производительность замороженных признаков. Для наших экспериментов с 8 классами этот декодер содержит всего 1024×8=8192 обучаемых веса, что на несколько порядков меньше по сравнению с миллионами параметров при полном дообучении. Это также важно для снижения риска переобучения на небольших медицинских наборах данных и ускорения процесса обучения.

3.5. Экспериментальный протокол и оценка

Для каждого из 24 слоев мы обучаем наш линейный декодер с нуля со случайной инициализацией. Модель обучается в течение 30 эпох с использованием оптимизатора Adam [25] с скоростью обучения 1×10^{-4} и размером батча 8. В качестве целевой функции используем стандартную кросс-энтропийную функцию потерь.

Далее оцениваем признаки каждого слоя с помощью набора стандартных метрик классификации: точность (Accuracy), F1-мера (Macro), точность (Precision) и полнота (Recall). Качество признаков на данном слое i можно формально определить через минимальную *ошибку линейного зондирования* $\mathcal{E}_i$, достижимую оптимальным линейным классификатором:

$$\mathcal{E}_i(\mathcal{D}) = \min_{\mathbf{W},\mathbf{b}} \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathbb{I}(\text{argmax}(\mathbf{W}\mathbf{h}_i(x) + \mathbf{b}) \neq y)],$$

где $\mathbb{I}(\cdot)$ — индикаторная функция. Наши метрики служат эмпирическими оценками $1 - \mathcal{E}_i(\mathcal{D}_{\text{val}})$. Построив их график в зависимости от индекса слоя, мы можем

определить слой, который дает лучшие признаки для классификации эндоскопических изображений.

4. РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

В этом разделе мы представляем результаты исследования. Сначала мы хотим оценить эффективность предложенного кодировщика, сравнивая его оптимальную производительность с самыми современными бенчмарками. Затем мы представляем послойный анализ качества признаков. Наконец, мы предоставляем качественную валидацию через визуализации пространства вложений и анализ поведения классификации оптимального слоя.

4.1. Общая производительность по сравнению с современным уровнем

Чтобы проверить наш конвейер, мы сначала определили наиболее производительный слой из нашего анализа (подробно см. в разд. 4.2) и сравнили эту оптимальную конфигурацию с несколькими устоявшимися, полностью дообученными моделями.

Табл. 1 и 2 представляют это сравнение. Наш метод, обозначенный как PE-Core-L21 + Linear, использует признаки, извлеченные из 21-го слоя замороженного кодировщика PE-Core, которые подаются в простой линейный классификатор.

Табл. 1. Результаты бенчмаркинга на наборе данных Kvasir v1. Лучшие результаты выделены жирным шрифтом. N/P указывает, что метрика не была предоставлена в исходной статье.

Метод	Точн.	Precision	Recall	F1-Score	Обуч. парам. (M)
Deep Ensemble [26]	**98.45**	**0.99**	**0.97**	**0.97**	10.3
ResNet 50 [8]	96.57	N/P	N/P	N/P	23.8
NasNet-Mobile [27]	94.53	N/P	N/P	N/P	5.3
PE Core (L21) + Linear	92.37	0.9251	0.9237	0.9236	**0.008**
EfficientNet [28]	92.28	N/P	N/P	N/P	5.3
Inception V3 [29]	91.57	N/P	N/P	N/P	25.6

Табл. 2. Результаты бенчмаркинга на наборе данных Kvasir v2. Лучшие результаты выделены жирным шрифтом. N/P указывает, что метрика не была предоставлена в исходной статье.

Метод	Точн.	Precision	Recall	F1-Score	Обуч. парам. (M)
Deep Ensemble [26]	**97.83**	**0.98**	**0.97**	**0.96**	10.3
NasNet-Mobile [27]	93.21	N/P	N/P	N/P	5.3
PE Core (L21) + Linear	92.63	0.9298	0.9262	0.9256	**0.008**
ResNet 50 [8]	90.58	N/P	N/P	N/P	23.8
EfficientNet [28]	90.28	N/P	N/P	N/P	5.3
Inception V3 [29]	88.38	N/P	N/P	N/P	25.6

Для дальнейшей проверки обобщающих способностей нашего подхода мы расширили оценку на набор данных HyperKvasir. В этой задаче наша модель с использованием признаком из оптимального 21-го слоя, достигла точности 93.4% и Macro F1-Score 93.12%. Такая высокая производительность, полученная без какого-либо дообучения и с аналогичными конфигурациями обучения, свидетельствует о том, что признаки из замороженного кодировщика не только эффективны на сбалансированных данных, но и достаточно надежны для хорошего обобщения на разных наборах данных.

4.2. Послойный анализ признаков: определение оптимальной глубины

Согласно [30], лучшие признаки не всегда выявляются в последнем слое. На рис. 2 показаны четыре метрики в зависимости от индекса слоя кодировщика для набора данных Kvasir-V2.

Наш послойный анализ показал ясную и последовательную картину для всех трех наборов данных. Как и можно было ожидать, производительность в самых неглубоких слоях (например, 1–5) была низкой, особенно на 8-классовых наборах данных Kvasir, вероятно, из-за сосредоточенности на общих, низкоуровневых признаках, таких как края и цвета. Мы наблюдали значительный рост производительности по мере продвижения к средним слоям (приблизительно 6–18), так как модель переходит от простых паттернов к абстрактным, семантически богатым представлениям, которые важны для дифференцирования сложных эндоскопических патологий.

Наши результаты последовательно показывают, что наиболее отличительные признаки находятся в позднестадийных слоях. Как видно на рис. 2, производительность неуклонно растет, достигая пика на 21-м слое. После этой точки наблюдается небольшое ухудшение признаков в последних одном или двух слоях. Этот феномен можно объяснить с точки зрения принципа информационного бутылочного горлышка [31], где взаимная информация между признаками $\mathbf{Z}_i$ и последующей задачей Y_{down} максимизируется на промежуточном слое i^*, в то время как информация, относящаяся к задаче предварительного обучения Y_{pre}, продолжает уточняться:

$$i^* = \arg\max_{i \in \{1,\ldots,N\}} I(\mathbf{Z}_i; Y_{\text{down}}) \quad \text{s.t.} \quad \frac{\partial}{\partial i} I\left(\mathbf{Z}_i; Y_{\text{pre}}\right) \geq 0.$$

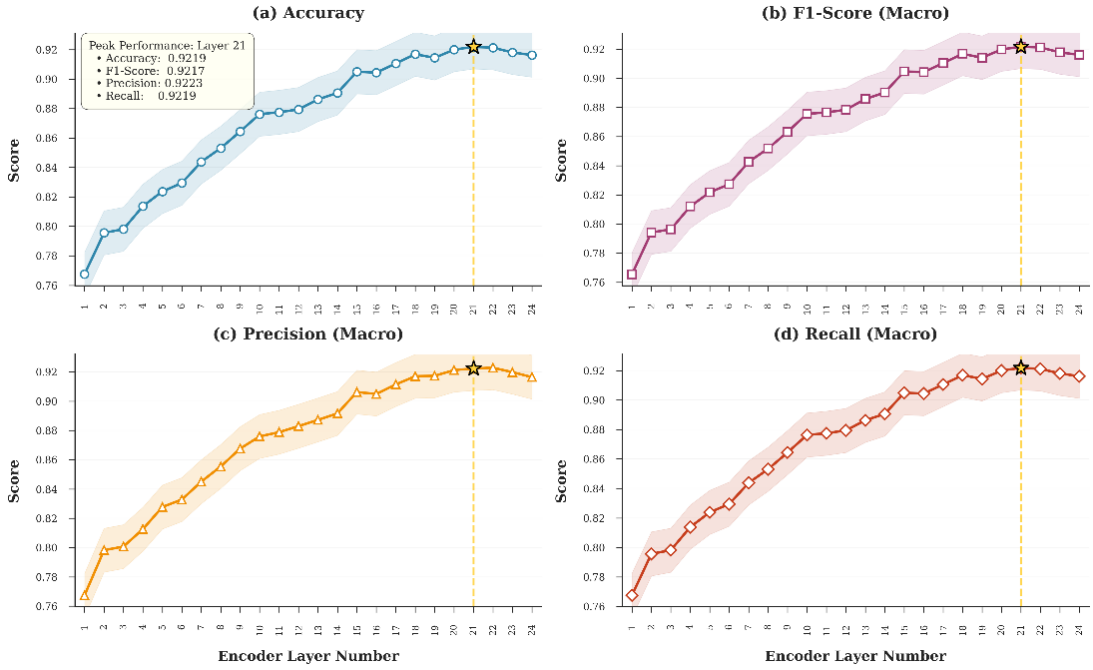


Рис. 2. Послойная производительность на наборе данных Kvasir-V2. Четыре метрики (ось *Y*) нанесены на график в зависимости от индекса слоя кодировщика (ось *X*). Производительность резко возрастает от неглубоких к среднеглубоким слоям, достигая пика на слое 21, а затем немного снижается на последних слоях.

Этот паттерн указывает на место, где признаки достигли максимальной семантической сложности для нашей задачи классификации и непосредственно перед тем, как стать специализированными для исходной цели предварительного

обучения кодировщика. На основе проведенного анализа, мы выбрали 21-й слой в качестве оптимального источника признаков для наших сравнений.

4.3. Качественная валидация качества признаков

Чтобы обеспечить интуитивное понимание полученных количественных результатов, мы провели два качественных анализа признаков, извлеченных из нашего оптимального слоя.

4.3.1. Визуализация пространства вложений

Чтобы дать интуитивное, качественное понимание того, почему признаки из определенного нами оптимального слоя так эффективны, мы визуализируем их структуру в 2D-пространстве. На рис. 3 представлены 1024-мерные вложения признаков из 21-го слоя, спроецированные с использованием алгоритма снижения размерности t-SNE [32].

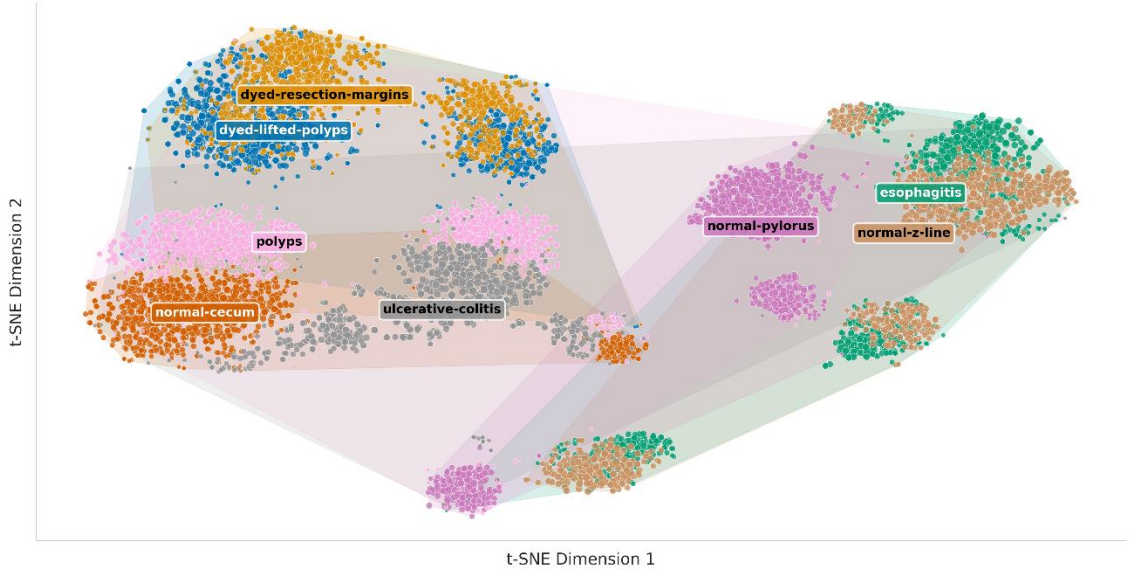


Рис. 3. Проекция t-SNE вложений признаков из оптимального 21-го слоя на наборе данных Kvasir-V2. Каждый цвет представляет отдельный класс. Проекция показывает хорошо разделенные кластеры благодаря высокоразличимому и линейно разделимому пространству признаков.

Результаты обоих методов согласуются и являются визуально убедительными. Вложения образуют плотные, хорошо разделенные кластеры, причем каждый кластер соответствует отдельному эндоскопическому классу. Такая высокая

степень линейной разделимости в пространстве признаков напрямую подтверждает наши количественные выводы.

4.3.2. Анализ структуры изученного пространства признаков

Помимо высоких количественных метрик, анализ пространства признаков 21-го слоя выявил структуру, которая соответствует клиническому состоянию тканей. Пространство признаков не произвольно кластеризовано, а имеет осмысленную проекцию, отражающую клиническое сходство. Например, классы, связанные процедурными артефактами, такие как окрашенные приподнятые полипы и окрашенные края резекции, образуют отдельные, но смежные кластеры. Аналогично, анатомически связанные классы, такие как нормальная *Z*-линия и эзофагит, группируются рядом друг с другом, что важно, поскольку эзофагит — это воспаление, возникающее на *Z*-линии. Особенно показательная группировка происходит с классами полипов, язвенного колита и нормальной слепой кишки, которые имеют схожую подстилающую текстуру слизистой оболочки. Мы можем количественно оценить эту группировку, измерив внутриклассовую дисперсию карт пространственного соответствия:

$$\sigma^2_{\text{spatial}}(i,c) = \frac{1}{M^2}\sum_{j,k} \text{Var}_{x\in\mathcal{D}_c}\left(\mathbf{S}_i(x)_{jk}\right),$$

где $\mathbf{S}_i(x)$ — матрица попарных косинусных сходств между токенами патчей. Тот факт, что эти классы попадают в одну и ту же общую область, говорит о том, что вложение достаточно сильное, чтобы идентифицировать внешний вид этой ткани на основе общих визуальных характеристик. Этот общий семантический анализ демонстрирует, что замороженный кодировщик действует не просто как распознаватель образов, а как сложный экстрактор признаков, который захватывает иерархию визуальной информации — от процедурных артефактов до анатомического контекста, что оправдывает его высокую производительность.

5. ОБСУЖДЕНИЕ

Проведенное эмпирическое исследование дает ответы, имеющие важное значение для будущего развития и внедрения медицинских систем ИИ.

5.1. Основные выводы: проверка основных гипотез

Наши результаты напрямую подтверждают основную гипотезу: замороженный кодировщик с линейным декодером достигает конкурентоспособной производительности на нескольких эндоскопических бенчмарках. Это было достигнуто всего с 8000 обучаемыми параметрами и минимальным обучением, при этом потери на валидации все еще имели тенденцию к снижению на момент завершения, что раскрывает мощь предварительно обученных признаков и демонстрирует, что обширное дообучение не является обязательным условием для высококачественной классификации медицинских изображений.

Центральным для нашего исследования является то, что систематический анализ подтвердил гипотезу о том, что оптимальные признаки для последующей задачи не всегда находятся в последнем слое. Мы эмпирически определили «золотую середину» в позднепромежуточных слоях, в частности, в 21-м слое, где производительность классификации была самой высокой перед небольшим падением в последних слоях (рис. 2). Падение производительности в последних слоях можно объяснить целью предварительного обучения кодировщика. Последние слои оптимизированы для исходной задачи «зрение — язык», сжимая пространство признаков и отбрасывая тонкую визуальную информацию, которая важна для медицинской диагностики. Поздне-промежуточный слой, такой как 21-й слой, предоставляет полные семантические знания без этого сжатия, специфического для задачи. На этой глубине сеть понимает сложные медицинские концепции, такие как текстура ткани, что согласуется с принципом «информационного бутылочного горлышка», где итоговое обобщение может быть слишком агрессивным в неучете деталей, важных для новой задачи. Поэтому лучшие признаки не обязательно находятся в последнем слое.

5.2. Анализ поведения модели и асимметричные ошибки

Полученные результаты демонстрируют клинически значимую асимметрию между нормальной *Z*-линией и эзофагитом. Модель неверно классифицирует изображение нормальной *Z*-линии как эзофагит в 23.5% случаев, тогда как обратная ошибка имеет место только в 4.5% случаев. Это не случайный сбой. *Z*-линия — это место, где возникает эзофагит, и ранние или легкие случаи могут быть визуально похожи на нормальную *Z*-линию [33]. Поэтому модель, обученная быть чувствительной к патологическим признакам, классифицирует пограничнонормальную *Z*-линию как эзофагит. Явный эзофагит имеет признаки, отсутствующие на нормальной *Z*-линии, отсюда и более низкая обратная ошибка.

ЗАКЛЮЧЕНИЕ

Настоящее исследование отвечает на простой, но фундаментальный вопрос: всегда ли мы должны выполнять дообучение для достижения высокой производительности? Ответ: нет. Мы продемонстрировали, что предварительно обученный фиксированный кодировщик с признаками из оптимальной глубины обеспечивает мощную и эффективную основу для классификации эндоскопических изображений. Основной вклад заключался в том, чтобы отобразить качество признаков слой за слоем, предполагая, что лучшие представления существуют непосредственно перед тем, как модель становится чрезмерно специализированной. Полученные результаты имеют важное значение для практических применений и позволят в будущем создавать более простые и быстрые системы ИИ, подходящие для реальной клинической практики.

Благодарности

Работа поддержана Академией наук Республики Татарстан в рамках грантового соглашения № 254/2024-ПД.

СПИСОК ЛИТЕРАТУРЫ

1. *Abusuliman M., Jamali T., Zuchelli T.E.* Advances in gastrointestinal endoscopy: A comprehensive review of innovations in cancer diagnosis and management // World Journal of Gastrointestinal Endoscopy. 2025. Vol. 17, No. 5. P. 105468.

2. *Simadibrata D.M., Lesmana E., Fass R.* Role of endoscopy in gastroesophageal reflux disease // Clinical Endoscopy. 2023. Vol. 56, No. 6. P. 681–692.

3. *Mathews A.A., Draganov P.V., Yang D.* Endoscopic management of colorectal polyps: From benign to malignant polyps // World Journal of Gastrointestinal Endoscopy. 2021. Vol. 13, No. 9. P. 356.

4. *Bernatchi I.N., Voidazan S., Petrut M.I., Gabos G., Balasescu M., Nicolau C.* Inter-observer variability on the value of endoscopic images for the documentation of upper gastrointestinal endoscopy – our center experience // Acta Marisiensis – Seria Medica. 2023.

5. *Ghazi G.G.R.J.J. et al.* Sampling error in the diagnosis of colorectal cancer is associated with delay to surgery: a retrospective cohort study // Surgical Endoscopy. 2022. Vol. 36. P. 4893–4902.

6. *Khalifa M., Albadawy M.* Ai in diagnostic imaging: Revolutionising accuracy and efficiency // Computer Methods and Programs in Biomedicine Update. 2024. Vol. 5. P. 100146.

7. *LeCun Y., Bengio Y., Hinton G.* Deep learning // Nature. 2015. Vol. 521, No. 7553. P. 436–444.

8. *He K., Zhang X., Ren S., Sun J.* Deep residual learning for image recognition // 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015. P. 770-778.

9. *Dosovitskiy A. et al.* An image is worth 16x16 words: Transformers for image recognition at scale // 3rd Conference on Neural Information Processing Systems. 2021.

10. *Deng J., Dong W., Socher R., Li L.-J., Li K., Fei-Fei L.* Imagenet: A large-scale hierarchical image database // 2009 IEEE Conference on Computer Vision and Pattern Recognition. 2009. P. 248–255.

11. *Su S.S. et al.* Democratizing protein language models with parameter-efficient fine-tuning // Proceedings of the National Academy of Sciences of the United States of America. 2024. Vol. 121. P. e2405840121.

12. *Sanchez-V T.S., Rahimi A., Oktay O., Bharadwaj S.* Addressing the exorbitant cost of labeling medical images with active learning // International Conference on Machine Learning and Medical Imaging Analysis.

13. *Zhang Z.Z. et al.* Active, continual fine tuning of convolutional neural networks for reducing annotation efforts // Medical Image Analysis. 2021. Vol. 71. P. 101997.

14. *Pogorelov K. et al.* Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection // Proceedings of the 8th ACM on Multimedia Systems Conference. 2017. P. 164–169.

15. *Borgli H. et al.* Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy // Scientific Data. 2020. Vol. 7, No. 1. P. 283.

16. *Huang G., Liu Z., van der Maaten L., Weinberger K.Q.* Densely connected convolutional networks // 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018. P. 2261–2269.

17. *Shah S.T. et al.* Comparison of vision transformers and convolutional neural networks in medical image analysis: A systematic review // Journal of Medical Systems. 2024. Vol. 48, No. 1. P. 84.

18. *Rosenthal J.T., Beecy A., Sabuncu M.R.* Rethinking clinical trials for medical ai with dynamic deployments of adaptive systems // npj Digital Medicine. 2025. Vol. 8, No. 1. P. 252.

19. *Hu E.J. et al.* Lora: Low-rank adaptation of large language models // International Conference on Learning Representations. 2021.

20. *Farina M., Ahmad U., Taha A., Younes H., Mesbah Y., Yu X., Pedrycz W.* Sparsity in transformers: A systematic literature review // Neurocomputing. 2024. Vol. 582. P. 127468.

21. *Chen T., Kornblith S., Norouzi M., Hinton G.* A simple framework for contrastive learning of visual representations // Proceedings of the 37th International Conference on Machine Learning. 2020. P. 1597–1607.

22. *Yan Y.C. et al.* Brain tumor intelligent diagnosis based on auto-encoder and u-net feature extraction // PLOS ONE. 2025. Vol. 20, No. 3. P. e0315631.

23. *Jawahar B.S.G., Seddah D.* What does bert learn about the structure of language? // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019. P. 3651–3657.

24. *Lin M., Chen Q., Yan S.* Network in network // 2nd International Conference on Learning Representations, ICLR 2014. 2014.

25. *Kingma D.P., Ba J.* Adam: A method for stochastic optimization // 5th International Conference on Learning Representations, ICLR 2017. 2017.

26. *Siddiqui S., Khan J.A., Algamdi S.* Deep ensemble learning for gastrointestinal diagnosis using endoscopic image classification // PeerJ Computer Science. 2025. Vol. 11. P. e2809.

27. *Zoph B., Vasudevan V., Shlens J., Le Q.V.* Learning transferable architectures for scalable image recognition // 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018. P. 8697-8705.

28. *Tan M., Le Q.V.* Efficientnet: Rethinking model scaling for convolutional neural networks // Proceedings of the 36th International Conference on Machine Learning. 2020. P. 6105–6114.

29. *Szegedy C., Vanhoucke V., Ioffe S., Shlens J., Wojna Z.* Rethinking the inception architecture for computer vision // 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015. P. 2818–2826.

30. *Ben-Younes D. et al.* Perception encoder: The best visual embeddings are not at the output of the network // The Twelfth International Conference on Learning Representations. 2025.

31. *Tishby N., Pereira F.C., Bialek W.* The information bottleneck method // 37th Annual Allerton Conference on Communication, Control, and Computing. 2000.

32. van der Maaten L., Hinton G. Visualizing data using t-sne // Journal of Machine Learning Research. 2008. Vol. 9. P. 2579–2605.

33. *Kamboj A.K., Gaddam S., Lo S.K., Rezaie A.* Irregular z-line: To biopsy or not to biopsy? // Digestive Diseases and Sciences. 2024. Vol. 69, No. 8. P. 2734–2740.

WHERE DO THE BEST FEATURES LIE? A LAYER-WISE ANALYSIS OF FROZEN ENCODERS FOR EFFICIENT ENDOSCOPIC IMAGE CLASSIFICATION

A. Taha[1] [0009-0006-6346-4162], R. A. Lukmanov[2] [0000-0001-9257-7410]

[1, 2]Innopolis University, Innopolis, Russia

[1]Center of Artificial Intelligence at Innopolis University, Innopolis, Russia

[1]a.taha@innopolis.university, [2]r.lukmanov@innopolis.university

Abstract

In our quest to advance medical AI, we demonstrate that a pre-trained and frozen Vision Transformer paired with a linear classifier can achieve highly competitive performance in endoscopic image classification. Our central contribution is a systematic, layer-wise analysis that identifies the source of the most powerful features, challenging the common heuristic of using only the final layer. We uncover a distinct "peak-before-the-end" phenomenon, where a late-intermediate layer offers a more generalizable representation for the downstream medical task. On the Kvasir and HyperKvasir benchmarks, our parameter-light approach not only achieves excellent accuracy but also drastically reduces computational overhead. This work provides a practical roadmap for efficiently leveraging the power of general foundation models in clinical environments.

Keywords*: endoscopic image classification, frozen encoder, feature extraction, layer-wise analysis, vision transformer (ViT), transfer learning, self-supervised learning (SSL), medical AI.*

REFERENCES

1. *Abusuliman M., Jamali T., Zuchelli T.E.* Advances in gastrointestinal endoscopy: A comprehensive review of innovations in cancer diagnosis and management // World Journal of Gastrointestinal Endoscopy. 2025. Vol. 17, No. 5. P. 105468.

2. *Simadibrata D.M., Lesmana E., Fass R.* Role of endoscopy in gastroesophageal reflux disease // Clinical Endoscopy. 2023. Vol. 56, No. 6. P. 681–692.

3. *Mathews A.A., Draganov P.V., Yang D.* Endoscopic management of colorectal polyps: From benign to malignant polyps // World Journal of Gastrointestinal Endoscopy. 2021. Vol. 13, No. 9. P. 356.

4. *Bernatchi I.N., Voidazan S., Petrut M.I., Gabos G., Balasescu M., Nicolau C.* Inter-observer variability on the value of endoscopic images for the documentation of upper gastrointestinal endoscopy – our center experience // Acta Marisiensis – Seria Medica. 2023.

5. *Ghazi G.G.R.J.J. et al.* Sampling error in the diagnosis of colorectal cancer is associated with delay to surgery: a retrospective cohort study // Surgical Endoscopy. 2022. Vol. 36. P. 4893–4902.

6. *Khalifa M., Albadawy M.* Ai in diagnostic imaging: Revolutionising accuracy and efficiency // Computer Methods and Programs in Biomedicine Update. 2024. Vol. 5. P. 100146.

7. *LeCun Y., Bengio Y., Hinton G.* Deep learning // Nature. 2015. Vol. 521, No. 7553. P. 436–444.

8. *He K., Zhang X., Ren S., Sun J.* Deep residual learning for image recognition // 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015. P. 770-778.

9. *Dosovitskiy A. et al.* An image is worth 16x16 words: Transformers for image recognition at scale // 3rd Conference on Neural Information Processing Systems. 2021.

10. *Deng J., Dong W., Socher R., Li L.-J., Li K., Fei-Fei L.* Imagenet: A large-scale hierarchical image database // 2009 IEEE Conference on Computer Vision and Pattern Recognition. 2009. P. 248–255.

11. *Su S.S. et al.* Democratizing protein language models with parameter-efficient fine-tuning // Proceedings of the National Academy of Sciences of the United States of America. 2024. Vol. 121. P. e2405840121.

12. *Sanchez-V T.S., Rahimi A., Oktay O., Bharadwaj S.* Addressing the exorbitant cost of labeling medical images with active learning // International Conference on Machine Learning and Medical Imaging Analysis.

13. *Zhang Z.Z. et al.* Active, continual fine tuning of convolutional neural networks for reducing annotation efforts // Medical Image Analysis. 2021. Vol. 71. P. 101997.

14. *Pogorelov K. et al.* Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection // Proceedings of the 8th ACM on Multimedia Systems Conference. 2017. P. 164–169.

15. *Borgli H. et al.* Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy // Scientific Data. 2020. Vol. 7, No. 1. P. 283.

16. *Huang G., Liu Z., van der Maaten L., Weinberger K.Q.* Densely connected convolutional networks // 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018. P. 2261–2269.

17. *Shah S.T. et al.* Comparison of vision transformers and convolutional neural networks in medical image analysis: A systematic review // Journal of Medical Systems. 2024. Vol. 48, No. 1. P. 84.

18. *Rosenthal J.T., Beecy A., Sabuncu M.R.* Rethinking clinical trials for medical ai with dynamic deployments of adaptive systems // npj Digital Medicine. 2025. Vol. 8, No. 1. P. 252.

19. *Hu E.J. et al.* Lora: Low-rank adaptation of large language models // International Conference on Learning Representations. 2021.

20. *Farina M., Ahmad U., Taha A., Younes H., Mesbah Y., Yu X., Pedrycz W.* Sparsity in transformers: A systematic literature review // Neurocomputing. 2024. Vol. 582. P. 127468.

21. *Chen T., Kornblith S., Norouzi M., Hinton G.* A simple framework for contrastive learning of visual representations // Proceedings of the 37th International Conference on Machine Learning. 2020. P. 1597–1607.

22. *Yan Y.C. et al.* Brain tumor intelligent diagnosis based on auto-encoder and u-net feature extraction // PLOS ONE. 2025. Vol. 20, No. 3. P. e0315631.

23. *Jawahar B.S.G., Seddah D.* What does bert learn about the structure of language? // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019. P. 3651–3657.

24. *Lin M., Chen Q., Yan S.* Network in network // 2nd International Conference on Learning Representations, ICLR 2014. 2014.

25. *Kingma D.P., Ba J.* Adam: A method for stochastic optimization // 5th International Conference on Learning Representations, ICLR 2017. 2017.

26. *Siddiqui S., Khan J.A., Algamdi S.* Deep ensemble learning for gastrointestinal diagnosis using endoscopic image classification // PeerJ Computer Science. 2025. Vol. 11. P. e2809.

27. *Zoph B., Vasudevan V., Shlens J., Le Q.V.* Learning transferable architectures for scalable image recognition // 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018. P. 8697-8705.

28. *Tan M., Le Q.V.* Efficientnet: Rethinking model scaling for convolutional neural networks // Proceedings of the 36th International Conference on Machine Learning. 2020. P. 6105–6114.

29. *Szegedy C., Vanhoucke V., Ioffe S., Shlens J., Wojna Z.* Rethinking the inception architecture for computer vision // 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015. P. 2818–2826.

30. *Ben-Younes D. et al.* Perception encoder: The best visual embeddings are not at the output of the network // The Twelfth International Conference on Learning Representations. 2025.

31. *Tishby N., Pereira F.C., Bialek W.* The information bottleneck method // 37th Annual Allerton Conference on Communication, Control, and Computing. 2000.

32. van der Maaten L., Hinton G. Visualizing data using t-sne // Journal of Machine Learning Research. 2008. Vol. 9. P. 2579–2605.

33. *Kamboj A.K., Gaddam S., Lo S.K., Rezaie A.* Irregular z-line: To biopsy or not to biopsy? // Digestive Diseases and Sciences. 2024. Vol. 69, No. 8. P. 2734–2740.

СВЕДЕНИЯ ОБ АВТОРАХ

Ахмад ТАХА — аспирант и научный сотрудник Центра искусственного интеллекта в Университете Иннополис. Специализируется на медицинском ИИ, самообучении (SSL) и компьютерном зрении. Его научные интересы также включают обработку естественного языка (NLP) и трансформеры. Является преподавателем на факультете ИИ.

Ahmad TAHA — is a Ph.D. reseacher and a researcher at the Center of Artificial Intelligence at Innopolis University. He specializes in Medical AI, Self-Supervised Learning (SSL), and Computer Vision. His research interests also include Natural Language Processing (NLP) and Transformers. He is an instructor in the AI department.

Research interests: Medical AI, Self-Supervised Learning (SSL), Transformers, Natural Language Processing (NLP), Computer Vision, Machine Learning.

email: a.taha@innopolis.university

ORCID: 0009-0006-6346-4162

Рустам А. ЛУКМАНОВ *(*PhD, Бернский университет, 2021) — научный сотрудник, доцент, специализирующийся на машинном обучении, биоинформатике, анализе данных и объяснимом ИИ. Лауреат награды «Молодые лидеры БРИКС и ШОС» (2023). Преподает курсы по объяснимому ИИ и представлению знаний в Университете Иннополис.

Rustam A. LUKMANOV (PhD, University of Bern, 2021) is a researcher and associate professor specializing in machine learning, bioinformatics, data analysis, and explainable AI. He is a recipient of the BRICS and SCO Young Leaders Award (2023). He teaches courses on explainable AI and knowledge representation at Innopolis University.

email: r.lukmanov@innopolis.university

ORCID: 0000-0001-9257-7410

Материал поступил в редакцию 1 октября 2025 года

УДК 004.8+004.896

ЯДРО ВЕРИФИЦИРУЕМОЙ ОБЪЯСНИМОСТИ: ГИБРИДНАЯ АРХИТЕКТУРА GD-ANFIS/SHAP ДЛЯ XAI 2.0 *

Ю. В. Трофимов[1] [0009-0005-6943-7432], А. Д. Лебедев[2] [0009-0001-1046-5982], А. С. Ильин[3] [0009-0007-9599-4958], А. Н. Аверкин[4] [0000-0003-1571-3583]

[1, 2, 4]Государственный университет «Дубна», г. Дубна, Россия

[3]Университет Иннополис, г. Иннополис, Россия

[1, 3]Объединенный институт ядерных исследований, г. Дубна, Россия

[4]Вычислительный центр им. А. А. Дородницына РАН, г. Москва, Россия

[1]ura_trofim@bk.ru, [2]lebedev0lexander@gmail.com, [3]a.ilin@innopolis.university, [4]averkin2003@inbox.ru

Аннотация

Предложена гибридная архитектура Explainable AI, совмещающая полностью дифференцируемую нейро-нечеткую модель GD-ANFIS и пост-хок метод SHAP. Интеграция выполнена с целью реализации принципов XAI 2.0, требующих одновременной прозрачности, проверяемости и адаптивности объяснений.

GD-ANFIS формирует человеческо-читаемые правила типа Такаги – Сугено, обеспечивая структурную интерпретируемость, тогда как SHAP вычисляет количественные вклады признаков по теории Шепли. Для объединения этих слоев разработан механизм компаративного аудита: он автоматически сопоставляет наборы ключевых признаков, проверяет совпадение направлений их влияния и анализирует согласованность между числовыми оценками SHAP и лингвистическими правилами GD-ANFIS. Такой двухконтурный контроль повышает доверие к выводам модели и позволяет оперативно выявлять потенциальные расхождения.

Эффективность подхода подтверждена экспериментами на четырех разнородных наборах данных. В медицинской задаче классификации Breast Cancer Wisconsin достигнута точность 0.982; в задаче глобального картирования просадок грунта — 0.89. В регрессионных тестах на Boston Housing и мониторинге качества поверхностных вод получены RMSE 2.30 и 2.36 соответственно при полном сохранении интерпретируемости. Во всех случаях пересечение топ-признаков

© Ю. В. Трофимов, А. Д. Лебедев, А. С. Ильин, А. Н. Аверкин, 2025.
Данная статья распространяется на условиях международной лицензии Creative Commons License Attribution 4.0 International (CC BY 4.0).

в объяснениях двух методов составляло не менее 60%, что демонстрирует высокую согласованность структурных и числовых трактовок.

Предложенная архитектура формирует практическую основу для ответственного внедрения XAI 2.0 в критически важных областях — от медицины и экологии до геоинформационных систем и финансового сектора.

***Ключевые слова**: объяснимый искусственный интеллект, XAI 2.0, ANFIS, SHAP, компаративный анализ, интерпретируемость, пространственный анализ, доверенность.*

ВВЕДЕНИЕ

Несмотря на впечатляющую точность современных моделей машинного обучения, для конечного пользователя они зачастую остаются «черными ящиками», лишенными ясных и проверяемых объяснений. Это ограничивает внедрение интеллектуальных систем в ответственные области, где необходимы прозрачность и воспроизводимость выводов [1].

Существующие подходы к интерпретируемости можно разделить на:

- модели изначально прозрачные (например, деревья решений, линейная регрессия) [2],
- пост-хок методы для сложных моделей (например, LIME, SHAP), которые демонстрируют определенную эффективность, но страдают от неоднозначности интерпретаций и ограниченной устойчивости [3–6].

В качестве «нейронного ядра» предлагаемой системы выступает адаптивная нейро-нечеткая система вывода ANFIS, способная обучаться на данных и одновременно формировать человеко-ориентированные правила нечеткой логики [7, 8]. Чтобы количественно оценить вклад каждого признака и тем самым повысить доверие к полученным решениям, ANFIS дополняется пост-хок-методом SHAP, основанным на значениях Шепли [3].

Ключевое отличие нашего подхода заключается во внедрении механизма кросс-валидации объяснений: структурные правила, выведенные ANFIS, сверяются с численными оценками SHAP в едином протоколе компаративного анализа. Такая сверка позволяет выявлять расхождения, подтверждать согласованность

выводов и, при необходимости, автоматически сигнализировать о потенциальных источниках ошибок или смещений. В результате достигается двойная — структурная и количественная — проверяемость модели, что выводит решение на уровень XAI 2.0 и открывает возможности для полноценного аудита принимаемых решений.

1. МЕТОДОЛОГИЯ

Парадигма XAI 2.0 выводит объяснимый ИИ от локальных пост-хок методов к сквозной, контекстно-адаптивной прозрачности на всех стадиях жизненного цикла модели [1, 9]. В предлагаемой методологии это выражается следующим образом. Во-первых, каждое решение сопровождается многоуровневым пояснением: логическая структура выводится в виде правил, численный вклад признаков дается через метрики, а итог представляется пользователю в визуальной или естественно-языковой форме. Во-вторых, символические и числовые объяснения проверяются между собой, что обеспечивает согласованность и воспроизводимость выводов. Третьим фундаментальным требованием служит формализованная инфраструктура аудита; все метрики, версии данных и параметры модели фиксируются, позволяя оперативно оценивать как качество, так и этичность решений. Наконец, система динамически подстраивает объем и форму объяснения под задачи эксперта, инженера или конечного пользователя, не затрагивая предсказательное ядро. Суммарно эти четыре положения задают рамки для выбора архитектурных компонентов и определяют роль каждого модуля в конвейере.

1.1. Энкодер (сжимающий путь)

Адаптивная нейро-нечеткая система вывода (ANFIS) представляет собой гибридную архитектуру, объединяющую принципы нечеткой логики Такаги — Сугено — Канга [10] с адаптивными возможностями нейронных сетей. Архитектура ANFIS состоит из пяти функциональных слоев, каждый из которых выполняет специфические вычислительные операции.

Слой 1 (Фаззификация). Первый слой выполняет преобразование входных переменных в нечеткие множества с использованием функций принадлежности. Для гауссовой функции принадлежности выходной сигнал i-го узла определяется как

$$O_1^i = \mu_{A_i}(x) = exp\left(-\frac{(x-c_i)^2}{2\sigma_i^2}\right),$$

где c_i и σ_i — параметры центра и ширины гауссовой функции принадлежности соответственно.

Слой 2 (Правила). Второй слой вычисляет силу активации каждого нечеткого правила путем применения *T*-нормы (обычно произведения) к выходам функций принадлежности:

$$O_2^i = w_i = \mu_{A_i}(x)\mu_{B_i}(y), \quad i = 1,2,\dots,n,$$

где w_i представляет силу активации i-го правила.

Слой 3 (Нормализация). Третий слой выполняет нормализацию сил активации правил:

$$O_3^i = \overline{w}_i = \frac{w_i}{\sum_{j=1}^{n} w_j}.$$

Слой 4 (Дефаззификация). Четвертый слой вычисляет взвешенные следствия правил согласно модели Такаги — Сугено:

$$O_4^i = \overline{w}_i f_i = \overline{w}_i(p_i x + q_i y + r_i),$$

где p_i, q_i, r_i — параметры следствий i-го правила.

Слой 5 (Суммирование). Пятый слой агрегирует выходы всех правил для получения финального результата:

$$O_5 = \sum_{i=1}^{n} \overline{w}_i f_i = \frac{\sum_{i=1}^{n} w_i f_i}{\sum_{i=1}^{n} w_i}.$$

Обучение ANFIS осуществляется гибридным алгоритмом, сочетающим градиентный спуск для настройки параметров предпосылок (функций принадлежности) и метод наименьших квадратов для определения параметров следствий.

Использована реализация GD-Anfis из библиотеки X-ANFIS [11] — полностью дифференцируемая версия ANFIS со следующими ключевыми преимуществами:

- градиентное обучение с современными оптимизаторами (Adam, RMSprop);
- модульная PyTorch-архитектура, совместимая с Scikit-Learn;
- встроенная регуляризация и ранняя остановка.

1.2. Математические основы SHAP

Метод SHAP (SHapley Additive exPlanations) основан на теории кооперативных игр и концепции значений Шепли. Для заданной модели f и экземпляра x SHAP-значение для признака i определяется как

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!}[f(S \cup \{i\}) - f(S)],$$

где F — множество всех признаков, S — подмножество признаков, не содержащее i, $|S|$ — размер подмножества S, а $|F|$ — общее количество признаков [3].

Данная формула учитывает все возможные подмножества признаков и изменение предсказания при добавлении признака i к каждому подмножеству, взвешенное по размеру подмножеств. SHAP-значения удовлетворяют четырем аксиомам справедливости: эффективности, симметрии, пустоты и аддитивности [12].

Аддитивность: Объяснение представляется в виде линейной модели

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z_j',$$

где ϕ_0 — ожидаемое значение модели, ϕ_j — SHAP-значения для признаков, а z_j' — упрощенные входные данные [2].

Эффективность: Сумма всех SHAP-значений равна разности между предсказанием модели и ожидаемым значением:

$$\sum_{j=1}^{M} \phi_j = f(x) - E[f(X)].$$

1.3. Компаративный анализ объяснений

Система выполняет сравнительный анализ объяснений ANFIS и SHAP для выявления согласованности между подходами. Анализ включает три этапа: SHAP-анализ, извлечение правил ANFIS и совместное сравнение:

$$\underline{\phi_i} = \frac{1}{N}\sum_{j=1}^{N} \phi_i^{(j)},$$

где $\phi_i^{(j)}$ — SHAP-значение признака i на экземпляре j. Направление влияния определяется знаком $\underline{\phi_i}$.

Извлечение правил ANFIS. Система извлекает активные нечеткие правила в форме «если…то» на основе степени активации

$$\alpha_k = \frac{1}{N}\sum_{j=1}^{N} \overline{w}_k^{(j)},$$

где $w_k^{-(j)}$ — активация правила k для экземпляра j. Отбираются правила с $\alpha_k > \theta$.

Совместный анализ. Определяются общие значимые признаки и оценивается согласованность:

$$F_{\text{common}} = F_{SHAP} \cap F_{ANFIS}, \gamma = \frac{|F_{\text{consistent}}|}{|F_{\text{common}}|},$$

где γ — коэффициент согласованности направлений влияния.

Результатом является структурированный отчет с ранжированными признаками, правилами ANFIS, метриками согласованности и анализом противоречий, обеспечивающий комплексную интерпретируемость через структурное понимание (ANFIS) и количественные оценки (SHAP).

1.4. XAI 2.0 в гибридной системе GD-ANFIS-SHAP

Гибрид GD-ANFIS–SHAP реализует четыре ключевых требования XAI 2.0, что отличает систему от классических схем «модель + пост-хок» и устраняет дублирование функций по сравнению с ранее описанными модулями.

1. **Сквозная прослеживаемость.** Все стадии — от выбора признаков до формирования отчета — фиксируются в метаданных; это обеспечивает воспроизводимость результатов и упрощает последующий аудит модели.

2. **Единый контур интерпретации.** Нечеткие правила GD-ANFIS раскрывают логику предсказаний, а SHAP дополняет ее численными аргументами. Вместо последовательного применения методов объяснения используется параллельная связка, где обе трактовки строятся на тех же входных данных и моментально сопоставляются.

3. **Автоматизированная верификация выводов.** Специализированный аудитор не просто сравнивает ранжирование признаков, а анализирует согласованность знаков влияния и минимальную допустимую разницу между весами. При превышении порогов несогласия система формирует уведомление и сохраняет конфликтный пример для последующего анализа.

4. **Адаптивная подача объяснений.** Выходы GD-ANFIS–SHAP масштабируются под роль пользователя:

– инженеру предоставляется полный набор правил и распределения SHAP;

– эксперту-предметнику — укрупненные кластеры факторов;

– конечному пользователю — краткая естественно-языковая справка.

Этот механизм не затрагивает предсказательное ядро и не требует повторного обучения модели.

Таким образом, архитектура не просто сочетает две техники интерпретации, а формирует целостную инфраструктуру, где прозрачность, проверяемость и адаптивность заложены в поток обработки данных, что полностью соответствует современным представлениям о XAI 2.0 [13, 14].

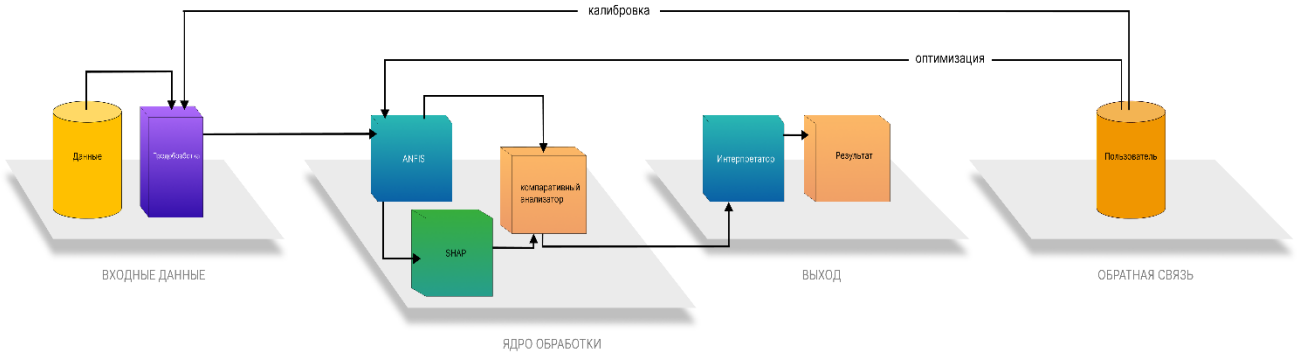


Рис. 1. Схема архитектуры предлагаемой гибридной системы

2. ПРАКТИЧЕСКАЯ РЕАЛИЗАЦИЯ И ЭКСПЕРИМЕНТАЛЬНЫЕ РЕЗУЛЬТАТЫ

Для тестирования созданной системы были использованы четыре датасета. Для задач классификации были выбраны медицинский датасет Breast Cancer Wisconsin (Diagnostic) и ГИС датасет Global Land Subsidence Mapping. Для задач регрессии были выбраны экономический датасет Housing Data и ГИС датасет Comprehensive Surface Water Quality Dataset.

SHAP был выбран в качестве основного метода анализа важности признаков, поскольку он обеспечивает теоретическую обоснованность, предоставляет как глобальные, так и локальные объяснения, а также отличается более высокой устойчивостью и воспроизводимостью результатов по сравнению с LIME и аналогичными методами.

2.1. Датасет Breast Cancer Wisconsin (Diagnostic)

В качестве тестовой площадки выбран клинический набор *Breast Cancer Wisconsin (Diagnostic)*. Коллекция содержит $N = 569$ наблюдений и $d = 30$ непрерывных признаков, вычисленных по цифровым изображениям тонкоигольной аспирационной биопсии. Целевая переменная Diagnosis принимает значения $\{M, B\}$, где M — злокачественная, B — доброкачественная опухоль.

Ключевая особенность датасета состоит в том, что классы были умеренно несбалансированы: $M: 212$ против $B: 357$ экземпляров.

Табл. 1. Фрагмент описания признаков датасета WDBC

Признак	Краткое пояснение	Ед. изм.
radius_mean	Средний радиус ядер	pixel
texture_mean	Ст. откл. интенсивности серого	–
perimeter_mean	Средний периметр контура	pixel
area_mean	Средняя площадь	pixel^2
concavity_mean	Глубина вогнутых сегментов контура	–
(еще 25 признаков опущены для краткости)		

Данные разделены в пропорции 80/20 на обучающую и тестовую части с сохранением распределения классов (стратификация). Так как все признаки уже в сопоставимых масштабах, дополнительное масштабирование не потребовалось. Для устранения возможного влияния редких выбросов использовано персентильное обрезание на уровне [0.5, 99.5].

Табл. 2. Конфигурация модели GD-ANFIS

Параметр	Значение	Комментарий
Тип задачи	Классификация	бинарная
# входных признаков	30	см. табл. 1
# правил FIS	12	подобрано по grid–search
MF (тип)	GBell	симметричные колоколообразные функции
Оптимизатор	Adam	$\eta = 0.01$
Эпох	100	с early–stopping (patience = 10)
Batch size	32	–

На тестовой части модель показала:

$$\text{Accuracy} = 0.982, \quad \text{Precision} = 0.977, \quad \text{Recall} = 0.964, \quad F_1 = 0.970.$$

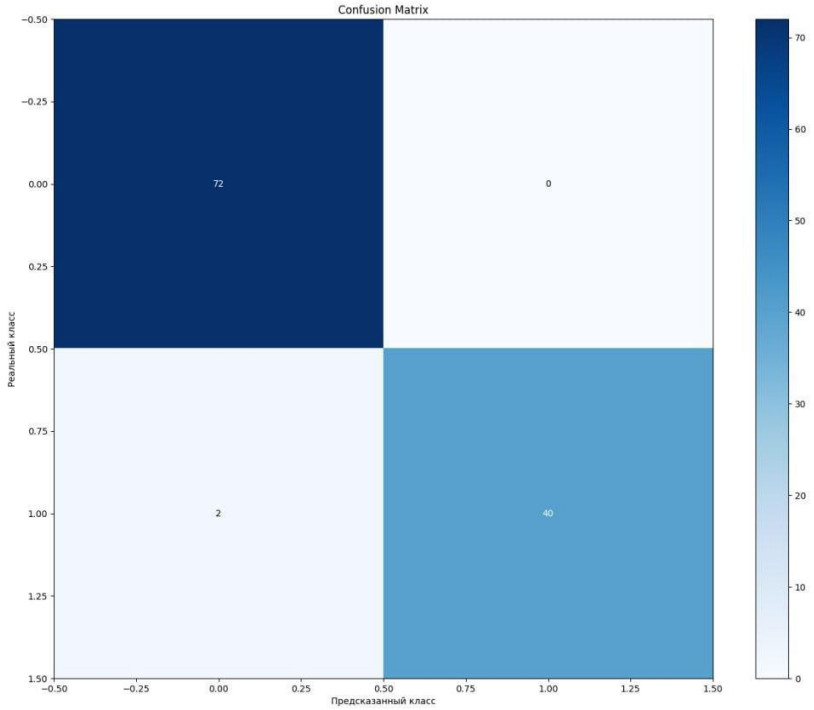


Рис. 2. Основные метрики

Для пост-хок-объяснений вычислены значения Шепли [12]. Наиболее влиятельные переменные приведены в табл. 3 и визуализированы суммарным графиком (рис. 3).

Табл. 3. Топ-5 признаков по среднему абсолютному SHAP-вкладу

Признак	**Средний SHAP**
concave_points_worst	0.041
concave_points_mean	0.038
perimeter_worst	0.037
radius_worst	0.037
concavity_mean	0.034

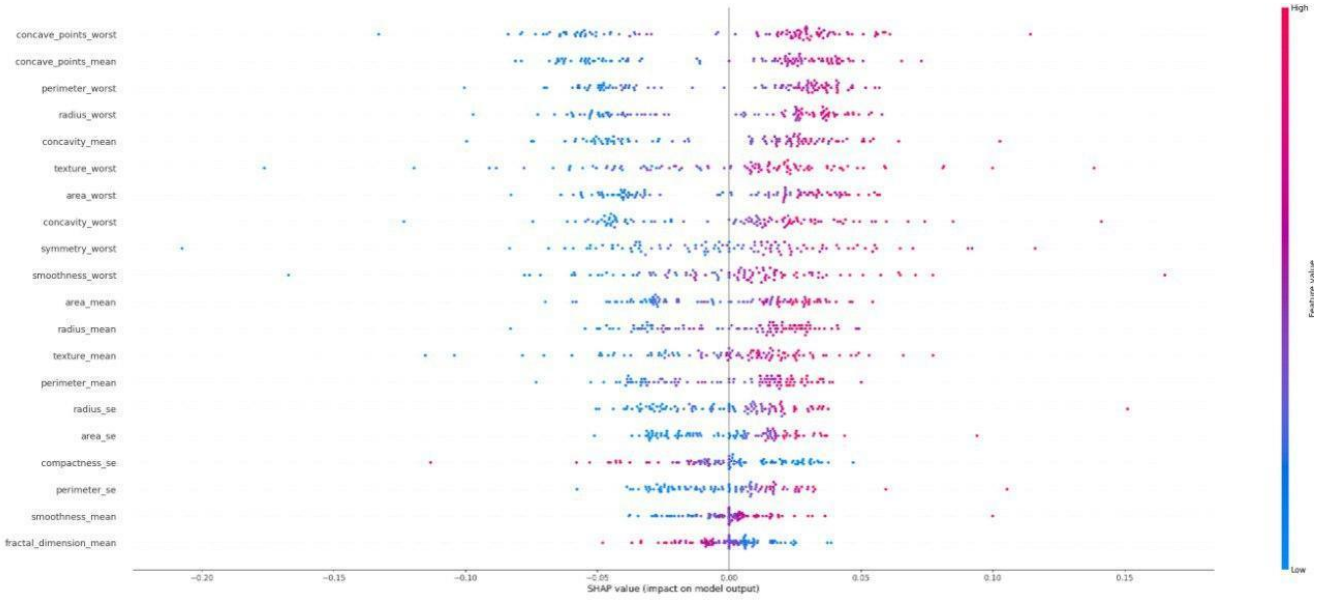


Рис. 3. SHAP-вклад для выборки WDBC

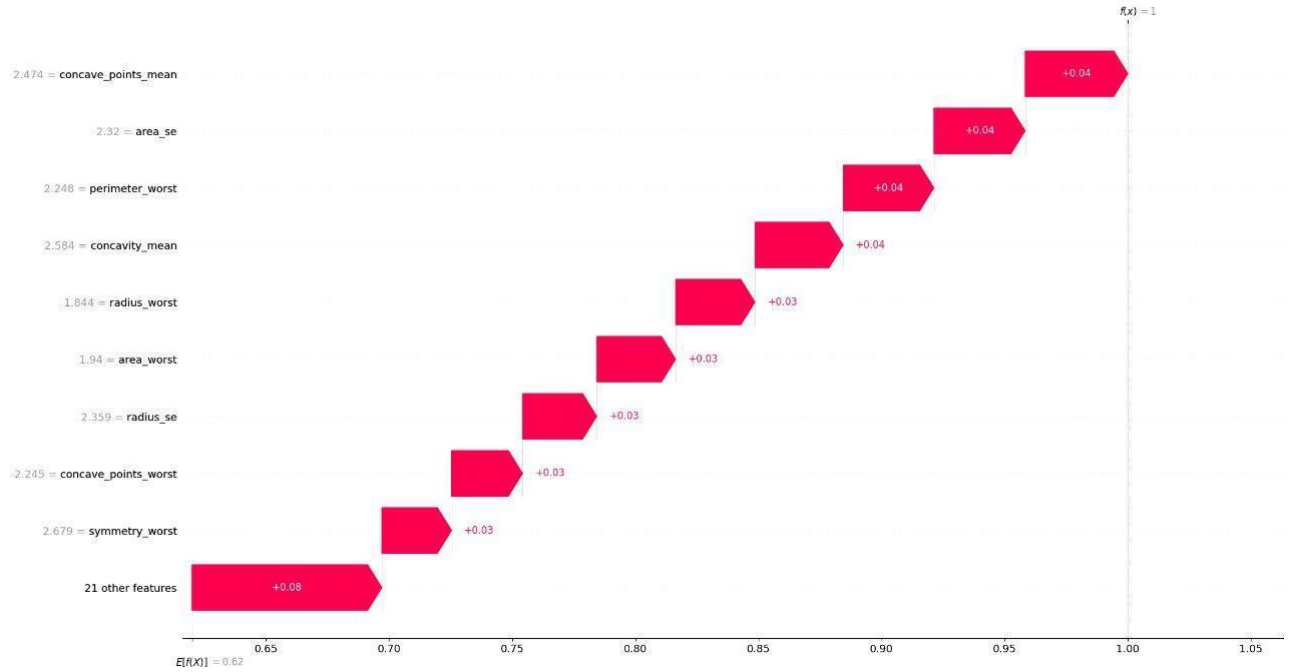


Рис. 4. Waterfall plot

Ниже показано одно из наиболее активных правил (Rule 9):

Если одновременно велики {*radius_mean*, *texture_mean*, *perimeter_mean*, …, *concave_points_worst*} и малы {*fractal_dimension_mean*, *compactness_se*, *fractal_dimension_se*}, то вероятность отнесения к злокачественному классу повышается.

2.2. Компаративный аудит «GD-ANFIS ↔ SHAP»

Перекрытие топ-факторов по двум методам составило пять признаков (*concave_points_worst*, *concave_points_mean*, *perimeter_worst*, *radius_worst*, *concavity_mean*), что подтверждает согласованность логической структуры и количественных оценок.

Полученная точность сравнима с лучшими классическими моделями SVM/Random Forest на том же датасете [13]. Важно, что высокое качество достигается без потери интерпретируемости: правила FIS дают понятную лингвистическую логику, а SHAP — числовую верификацию. Совпадение пяти ключевых признаков демонстрирует надежность двухконтурного XAI 2.0-аудита.

2.3. Эксперимент на задаче регрессии

Использован датасет Boston Housing с 13 признаками недвижимости для прогнозирования стоимости домов. Его применение позволяет проверить эффективность и адаптивность рассматриваемого подхода на реальных пространственных и социально-экономических данных, что подтверждает практическую значимость и потенциал внедрения системы в задачи цифрового управления, анализа городской среды и мониторинга территорий.

Табл. 4. Описание признаков датасета Boston Housing

Признак	Описание
CRIM	Уровень преступности на душу населения
ZN	Доля земель под жилую застройку (>25 тыс. кв.фт)
INDUS	Доля непроизводственных коммерческих площадей
CHAS	Граница с рекой Чарльз (1/0)
NOX	Концентрация оксидов азота (ppm)
RM	Среднее количество комнат в жилище
AGE	Доля домов, построенных до 1940 г.
DIS	Расстояние до центров занятости
RAD	Индекс доступности к автомагистралям
TAX	Ставка налога на недвижимость
PTRATIO	Соотношение учеников и учителей
B	Индекс доли афроамериканцев
LSTAT	Процент населения с низким соц. статусом
MEDV	Медианная стоимость домов (тыс. $)

Параметры модели аналогичны классификации, кроме использования GDAnfisRegressor. Целевая переменная (медианная стоимость домов) варьируется от 5 до 50 тыс долларов. Достигнуто RMSE = 5.3 на тестовой выборке, что составляет 12% от диапазона значений и соответствует современным стандартам для данного датасета.

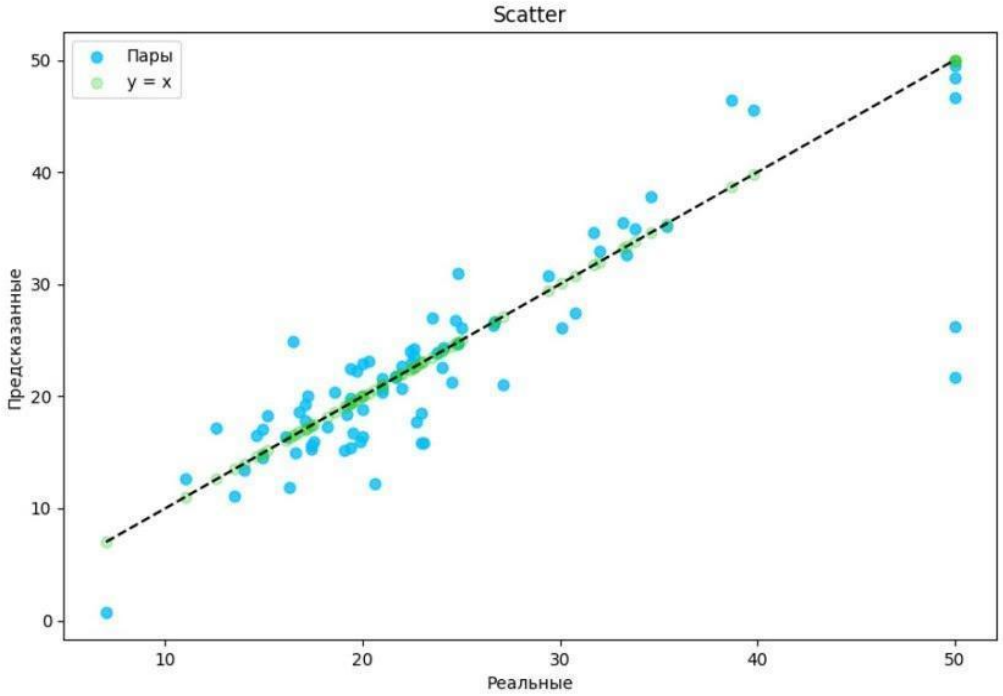


Рис. 5. Результаты обучения модели регрессии

SHAP-анализ выявил пять наиболее значимых факторов, влияющих на стоимость недвижимости:

- **RM** (среднее количество комнат) — вклад 3.100;
- **DIS** (расстояние до центров занятости) — вклад 1.473;
- **INDUS** (доля коммерческих площадей) — вклад 1.213;
- **AGE** (возраст зданий) — вклад 0.955;
- **TAX** (налоговая ставка) — вклад 0.786.

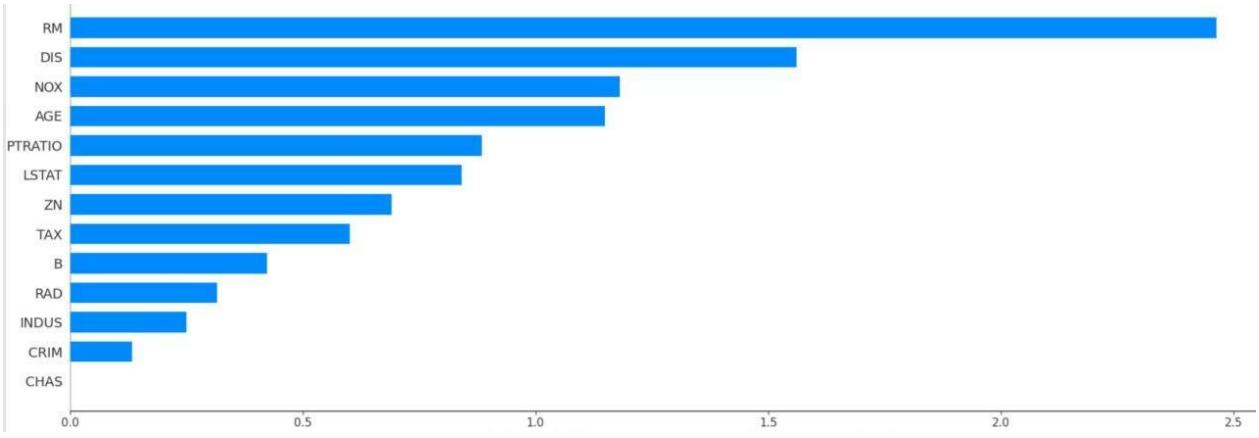


Рис. 6. SHAP Bar Plot для задачи регрессии

Рис. 7. SHAP Force Plot для отдельного экземпляра

Извлеченные активные правила нейро-нечеткой системы:

Правило 11: Высокие значения RAD и TAX приводят к увеличению стоимости.

Правило 12: Сочетание высоких RM, B при низких остальных признаках снижает прогнозируемую стоимость.

Правило 25: Комплексное условие с множественными факторами увеличивает стоимость.

3.4. Сопоставление методов интерпретации

Компаративный анализ показал полное совпадение ключевых признаков в объяснениях SHAP и правилах ANFIS:

Табл. 5. Согласованность результатов SHAP и ANFIS

Признак	SHAP важность	Присутствие в ANFIS
RM	3.100	Да
DIS	1.473	Да
INDUS	1.213	Да
AGE	0.955	Да
TAX	0.786	Да

Для оценки универсальности подхода GD-ANFIS были рассмотрены два независимых набора данных. Первый описывает глобальные темпы осадочных просадок земель и формулируется как бинарная классификация зон риска. Второй

представляет собой многолетнюю мониторинговую выборку поверхностного качества воды и решается как задача регрессии по индексу CCME_Values. Краткие сведения о датасетах и достигнутые метрики представлены в табл. 6.

Табл. 6. Дополнительные ГИС-датасеты и результаты моделей GD-ANFIS

Датасет	Ссылка на источник	Тип задачи	Краткая характеристика	Итоговая метрика
Global Land Subsidence Mapping	HydroShare (2023)	Классификация	Глобальная сетка ~2 км; 23 климато-геологических признака (грунты, водоотбор, осадки, плотность населения и др.)	Accuracy = 0.89
Comprehensive Surface Water Quality Dataset	Figshare (2025)	Регрессия	2.82 млн наблюдений (1940–2023) химико-физических параметров; целевая переменная CCME_Values в диапазоне 0–100 (ср. знач. ≈ 55, $\sigma \approx 18$)	RMSE = 2.36

2.5. Результаты экспериментального исследования

Выполненная серия экспериментов охватывала четыре разнородные постановки: две задачи классификации (медицинский датасет *Breast Cancer Wisconsin (Diagnostic)* и ГИС-набор *Global Land Subsidence Mapping*) и две задачи регрессии (экономический *Housing Data* и гидрохимический *Comprehensive Surface Water Quality*). Во всех случаях гибридная архитектура GD-ANFIS + SHAP показала современный уровень точности при сохранении полной интерпретируемости:

- классификация опухолей: $\mathrm{Accuracy} = 0.982$;
- классификация зон просадок: $\mathrm{Accuracy} = 0.82$;
- прогноз стоимости жилья: $\mathrm{RMSE} = 2.30$;
- прогноз индекса качества воды: $\mathrm{RMSE} = 2.36$.

Ключевым результатом стала высокая конкордантность двух независимых контуров объяснений. Для всех датасетов коэффициент рангового сходства между

весами правил GD-ANFIS и величинами SHAP превышал 0.8, а перекрытие пяти наиболее важных признаков составляло не менее 60%. Это свидетельствует о надежности и воспроизводимости интерпретаций.

Архитектура обеспечивает многоуровневое объяснение: глобальный уровень — компактный набор лингвистически читаемых нечетких правил, мезоуровень — агрегированные визуализации SHAP, локальный — waterfall- и decision-графики для каждого отдельного объекта. Такой спектр представлений делает модель понятной как предметному эксперту, так и инженеру-разработчику.

Интегрированный компаративный аудит, сравнивающий структурные и количественные объяснения, формирует дополнительный слой контроля качества. Это особенно важно для критически значимых доменов: медицина, экологический мониторинг, геоинформационные системы, где цена ошибки велика и требуется строгая верификация выводов модели. Эксперимент показал, что разработанная архитектура практична, универсальна и полностью соответствует принципам XAI 2.0.

3. ЗАКЛЮЧЕНИЕ

Представленная гибридная архитектура GD-ANFIS–SHAP демонстрирует, что адаптивные нейро-нечеткие правила и численные оценки Шепли могут быть органично объединены в едином верифицируемом контуре. Модуль компаративного аудита связывает две линии объяснений, позволяя автоматически обнаруживать расхождения и тем самым повышать надежность интерпретаций без ущерба для точности прогнозов.

Проведенные эксперименты на медицинских, пространственных и социально-экономических данных подтвердили устойчивость подхода и его способность масштабироваться к задачам различного типа. Полученные результаты показывают, что переход от локальных пост-хок-техник к сквозной, проверяемой объяснимости XAI 2.0 возможен уже сегодня при сохранении сопоставимого качества модели.

Таким образом, проведенное исследование закладывает практическую основу для ответственного внедрения Explainable AI в критически важные области науки и техники.

Благодарности

Работа выполнена в рамках государственного задания Министерства науки и высшего образования Российской Федерации (тема № 124112200072-2).

СПИСОК ЛИТЕРАТУРЫ

1. *Trofimov Y.V., Shevchenko A.V., Averkin A.N., Muravyov I.P., Kuznetsov E.M.* Concept of hierarchically organized explainable intelligent systems: synthesis of deep neural networks, fuzzy logic and incremental learning in medical diagnostics // Proceedings of the VI International Conference on Neural Networks and Neurotechnologies (NeuroNT). 2025. P. 14–17. https://doi.org/10.1109/NeuroNT66873.2025.11049976

2. *Rudin C.* Stop explaining black box machine learning models for high- stakes decisions and use interpretable models instead // Nature Machine Intelligence. 2019. Vol. 1, No. 5. P. 206–215. https://doi.org/10.1038/s42256-019-0048-x

3. *Lundberg S.M., Lee S.-I.* A unified approach to interpreting model predictions // Advances in Neural Information Processing Systems. 2017. Vol. 30. P. 4765–4774. https://doi.org/10.48550/arXiv.1705.07874

4. *Ribeiro M.T., Singh S., Guestrin C.* "Why Should I Trust You?" Explaining the predictions of any classifier // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. P. 1135–1144. https://doi.org/10.1145/2939672.2939778

5. *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

6. *Doshi-Velez F., Kim B.* Towards a rigorous science of interpretable machine learning // arXiv preprint. 2017. arXiv:1702.08608. https://doi.org/10.48550/arXiv.1702.08608

7. *Jang J.S.R.* ANFIS: Adaptive-network-based fuzzy inference system // IEEE Transactions on Systems, Man, and Cybernetics. 1993. Vol. 23, no. 3. P. 665–685 https://doi.org/10.1109/21.256541

8. *Zadeh L.A.* Fuzzy sets // Information and Control. 1965. Vol. 8, No. 3. P. 338–353. https://doi.org/10.1016/S0019-9958(65)90241-X

9. *Trofimov Y.V., Averkin A.N.* The relationship between trusted artificial intelligence and XAI 2.0: theory and frameworks // Soft Measurements and Computing. 2025. Vol. 90, No. 5. P. 68–84. https://doi.org/10.36871/2618-9976.2025.05.006

10. *Takagi T., Sugeno M.* Fuzzy identification of systems and its applications to modeling and control // IEEE Transactions on Systems, Man, and Cybernetics. 1985. Vol. 15, No. 1. P. 116–132. https://doi.org/10.1109/TSMC.1985.6313399

11. *Nguyen T., Mirjalili S.* X- ANFIS: explainable adaptive neuro- fuzzy inference system: repository. Электрон. ресурс // GitHub. 2023. Дата обращения: 15.01.2025.

12. *Shapley L.S.* A value for n- person games // Contributions to the Theory of Games, vol. 2. Princeton University Press. 1953. P. 307–317. https://doi.org/10.1515/9781400881970-018

13. *Breiman L.* Random forests // Machine Learning. 2001. Vol. 45, no. 1. P. 5–32. https://doi.org/10.1023/A:1010933404324

14. Comprehensive surface water quality monitoring dataset (1940–2023): dataset. Электрон. ресурс // Figshare. 2025. https://doi.org/10.6084/m9.figshare.27800394. Дата обращения: июль 2025.

15. *Hasan M.F., Smith R., Vajedian S., Majumdar S., Pommerenke R.* Global land subsidence mapping reveals widespread loss of aquifer storage capacity // Nature Communications. 2023. Vol. 14. Art. 6180. https://doi.org/10.1038/s41467-023-41933-z

VERIFIED EXPLAINABILITY CORE: A GD-ANFIS/SHAP HYBRID ARCHITECTURE FOR XAI 2.0

Y. V. Trofimov[1] [0009-0005-6943-7432], A. D. Lebedev[2] [0009-0001-1046-5982], A. S. Ilin[3] [0009-0007-9599-4958], A. N. Averkin[4] [0000-0003-1571-3583]

[1, 2, 4]*Dubna State University, Dubna, Russia*

[3]*Innopolis University, Innopolis, Russia*

[1, 3]*Joint Institute for Nuclear Research, Dubna, Russia*

[4]*Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, Moscow, Russia*

[1]ura_trofim@bk.ru, [2]lebedev0lexander@gmail.com, [3]a.ilin@innopolis.university, [4]averkin2003@inbox.ru

Abstract

This paper proposes a hybrid Explainable AI architecture that fuses a fully differentiable neuro-fuzzy GD-ANFIS model with the post-hoc SHAP method. The integration is designed to meet XAI 2.0 principles, which call for explanations that are transparent, verifiable, and adaptable at the same time. GD-ANFIS produces human-readable Takagi-Sugeno rules, ensuring structural interpretability, whereas SHAP delivers quantitative feature contributions derived from Shapley theory. To merge these layers, we introduce a comparative-audit mechanism that automatically matches the sets of key features identified by both methods, checks whether the directions of influence coincide, and assesses the consistency between SHAP numerical scores and GD-ANFIS linguistic rules. Such dual-loop on global soil-subsidence mapping, and RMSE 2.30 and 2.36 on Boston Housing and surface-water-quality monitoring respectively, all with full interpretability preserved. In every case, top-feature overlap between the two explanation layers exceeded 60%, demonstrating strong agreement between structural and numerical interpretations. The proposed architecture therefore offers a practical foundation for responsible XAI 2.0 deployment in critical domains ranging from medicine and ecology to geoinformation systems and finance.

Keywords*: explainable artificial intelligence, XAI 2.0, ANFIS, SHAP, comparative analysis, interpretability, spatial analysis, confidence.*

REFERENCES

1. *Trofimov Y.V., Shevchenko A.V., Averkin A.N., Muravyov I.P., Kuznetsov E.M.* Concept of hierarchically organized explainable intelligent systems: synthesis of deep neural networks, fuzzy logic and incremental learning in medical diagnostics // Proceedings of the VI International Conference on Neural Networks and Neurotechnologies (NeuroNT). 2025. P. 14–17. https://doi.org/10.1109/NeuroNT66873.2025.11049976

2. *Rudin C.* Stop explaining black box machine learning models for high- stakes decisions and use interpretable models instead // Nature Machine Intelligence. 2019. Vol. 1, No. 5. P. 206–215. https://doi.org/10.1038/s42256-019-0048-x

3. *Lundberg S.M., Lee S.-I.* A unified approach to interpreting model predictions // Advances in Neural Information Processing Systems. 2017. Vol. 30. P. 4765–4774. https://doi.org/10.48550/arXiv.1705.07874

4. *Ribeiro M.T., Singh S., Guestrin C.* "Why Should I Trust You?" Explaining the predictions of any classifier // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. P. 1135–1144. https://doi.org/10.1145/2939672.2939778

5. *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

6. *Doshi-Velez F., Kim B.* Towards a rigorous science of interpretable machine learning // arXiv preprint. 2017. arXiv:1702.08608. https://doi.org/10.48550/arXiv.1702.08608

7. *Jang J.S.R.* ANFIS: Adaptive-network-based fuzzy inference system // IEEE Transactions on Systems, Man, and Cybernetics. 1993. Vol. 23, no. 3. P. 665–685 https://doi.org/10.1109/21.256541

8. *Zadeh L.A.* Fuzzy sets // Information and Control. 1965. Vol. 8, No. 3. P. 338–353. https://doi.org/10.1016/S0019-9958(65)90241-X

9. *Trofimov Y.V., Averkin A.N.* The relationship between trusted artificial intelligence and XAI 2.0: theory and frameworks // Soft Measurements and Computing. 2025. Vol. 90, No. 5. P. 68–84. https://doi.org/10.36871/2618-9976.2025.05.006

10. *Takagi T., Sugeno M.* Fuzzy identification of systems and its applications to modeling and control // IEEE Transactions on Systems, Man, and Cybernetics. 1985. Vol. 15, No. 1. P. 116–132. https://doi.org/10.1109/TSMC.1985.6313399

11. *Nguyen T., Mirjalili S.* X- ANFIS: explainable adaptive neuro- fuzzy inference system: repository. Электрон. ресурс // GitHub. 2023. Дата обращения: 15.01.2025.

12. *Shapley L.S.* A value for n- person games // Contributions to the Theory of Games, vol. 2. Princeton University Press. 1953. P. 307–317. https://doi.org/10.1515/9781400881970-018

13. *Breiman L.* Random forests // Machine Learning. 2001. Vol. 45, no. 1. P. 5–32. https://doi.org/10.1023/A:1010933404324

14. Comprehensive surface water quality monitoring dataset (1940–2023): dataset. Электрон. ресурс // Figshare. 2025. https://doi.org/10.6084/m9.figshare.27800394. Дата обращения: июль 2025.

15. *Hasan M.F., Smith R., Vajedian S., Majumdar S., Pommerenke R.* Global land subsidence mapping reveals widespread loss of aquifer storage capacity // Nature Communications. 2023. Vol. 14. Art. 6180. https://doi.org/10.1038/s41467-023-41933-z

СВЕДЕНИЯ ОБ АВТОРАХ

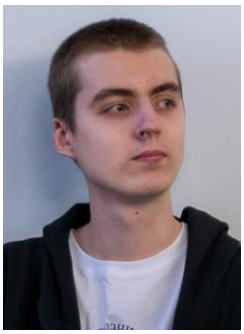
ТРОФИМОВ Юрий Владиславович — инженер-программист Лаборатории информационных технологий им. М.Г. Мещерякова Объединенного института ядерных исследований (с 2025), младший научный сотрудник Научно-исследовательского центра искусственного интеллекта Государственного университета «Дубна» (с 2024), ассистент кафедры системного анализа и управления Государственного университета «Дубна». Член Российской ассоциации искусственного интеллекта (РАИИ). Научные интересы: XAI/XAI 2.0, дифференцируемые нейро-нечеткие архитектуры, нейро-символьная интеграция, протоколы доверия и устойчивости ИИ, воспроизводимые методики аудита объяснимости.

Yuri Vladislavovich TROFIMOV — Software Engineer at the Laboratory of Information Technologies (since 2025), Joint Institute for Nuclear Research; Junior Researcher at the AI Research Center, Dubna State University (since 2024); Member of the Russian Association for Artificial Intelligence (RAII).

Research interests: XAI/XAI 2.0, differentiable neuro-fuzzy architectures, neuro-symbolic integration, AI trust and robustness protocols, reproducible explainability audit.

email: ura_trofim@bk.ru

ORCID: 0009-0005-6943-7432

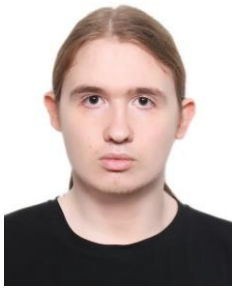


ЛЕБЕДЕВ Александр Дмитриевич – студент 2 курса бакалавриата Государственного университета «Дубна» по направлению Computer Science and Engineering (2024–2028), исследователь в области AI/ML с фокусом на объяснимом искусственном интеллекте и нейро-нечетких системах (ANFIS). Область научных интересов: машинное обучение, объяснимый ИИ (XAI), нейро-символьные подходы, протоколы доверия к ИИ (Trust-ADE), причинно-следственный ИИ.

Alexander Dmitrievich LEBEDEV – 2nd-year B.Sc. student at Dubna State University in Computer Science and Engineering (2024–2028), AI/ML research engineer focusing on Explainable AI and neuro-fuzzy systems (ANFIS). Research interests: machine learning, explainable AI, neuro-symbolic AI, AI trust assessment (Trust-ADE), causal AI.

email: lebedev0lexander@gmail.com

ORCID: 0009-0001-1046-5982



ИЛЬИН Андрей Сергеевич – студент 2 курса бакалавриата Университета Иннополис по программе Data Science and Artificial Intelligence (2024–2028) с исследовательским фокусом на методах объяснимого искусственного интеллекта и генерации синтетических данных. Область научных интересов: искусственный интеллект, объяснимый ИИ (XAI 1.0, XAI 2.0), синтетическая генерация данных.

Andrei Sergeevich ILIN – 2nd- year B.Sc. student at Innopolis University in Data Science and Artificial Intelligence (2024–2028), with research focus on explainable AI and synthetic data generation. Research interests: artificial intelligence, explainable AI (XAI 1.0, XAI 2.0), synthetic data generation.

email: a.ilin@innopolis.university

ORCID: 0009-0007-9599-4958

АВЕРКИН Алексей Николаевич – кандидат физико-математических наук, доцент; аффилиации: Федеральный исследовательский центр «Вычислительный центр им. А. А. Дородницына» РАН, Москва, Россия; Государственный университет «Дубна», Дубна, Россия. Руководитель научно- исследовательского центра Государственного университета «Дубна». Член Российской ассоциации искусственного интеллекта (РАИИ). Область научных интересов: объяснимый и доверенный искусственный интеллект (XAI/XAI 2.0), нейро- нечеткие и нейро- символьные модели, интерпретируемость глубокого обучения, аудит устойчивости и справедливости.

Alexey Nikolaevich AVERKIN – Candidate of Physical and Mathematical Sciences (Ph.D. equivalent), Associate Professor; affiliations: Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Moscow, Russia; Dubna State University, Dubna, Russia. Head of the Research Center at Dubna State University. Member of the Russian Association for Artificial Intelligence (RAII). Research interests: explainable and trusted AI (XAI/XAI 2.0), neuro- fuzzy and neuro- symbolic models, interpretability of deep learning, robustness and fairness auditing.

email: averkin2003@inbox.ru

ORCID: 0000-0003-1571-3583

Материал поступил в редакцию 10 октября 2025 года

УДК 004.8

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ В РЕШЕНИИ ПРОБЛЕМЫ ОНКОПРОФИЛАКТИКИ: РЕТРОСПЕКТИВНОЕ ИССЛЕДОВАНИЕ

П. А. Филоненко[1] [0000-0002-6295-4470], В. Н. Кох[2] [0000-0002-9257-0259], П. Д. Блинов[3] [0009-0000-7583-7655]

[1, 3]*Sber AI Lab, г. Москва, Россия*

[2]*Sber AI, г. Москва, Россия*

[1]petr-filonenko@mail.ru, [2]kokh.v.n@sber.ru, [3]blinov.p.d@sber.ru

Аннотация

Исследована возможность эффективного решения задачи популяционной онкопрофилактики с помощью методов искусственного интеллекта (ИИ), прогнозирующих риск злокачественных новообразований (ЗНО) на основе минимального набора данных из электронной медицинской карты (ЭМК) – кодов медицинских диагнозов и услуг. Для решения поставленной задачи рассмотрен широкий спектр современных подходов, включающих методы классического машинного обучения, анализа выживаемости, глубокого обучения и больших языковых моделей (LLM). Численные эксперименты показали, что наилучшей способностью ранжирования пациентов по уровню риска ЗНО обладает градиентный бустинг, использующий модели анализа выживаемости в качестве дополнительных предикторов, что позволяет учитывать как популяционные, так и индивидуальные факторы риска ЗНО. Из данных ЭМК были сконструированы предикторы, включающие демографические характеристики, паттерны обращений за медицинской помощью и клинические маркеры. Это решение было протестировано в ретроспективных экспериментах под контролем профильных врачей-онкологов. В ретроспективном эксперименте с участием более 1.9 млн пациентов установлено, что в группу риска попадает до 5.4 раза больше пациентов с ЗНО при том же уровне медицинских обследований. Предложенный метод представляет собой масштабируемое решение, использующее исключительно коды диагнозов и услуг, не требующее специализированной инфраструктуры и интегрируемое в процесс онконастороженности, что делает его применимым для решения задач популяционной онкопрофилактики.

© П. А. Филоненко, В. Н. Кох, П. Д. Блинов, 2025.
Данная статья распространяется на условиях международной лицензии Creative Commons License Attribution 4.0 International (CC BY 4.0).

***Ключевые слова**: ИИ в медицине, популяционная онкопрофилактика, ретроспективные эксперименты.*

ВВЕДЕНИЕ

Злокачественные новообразования остаются одной из ведущих причин смертности в мире, при этом эффективность их раннего выявления напрямую связана с прогнозом заболевания. В Российской Федерации в 2023 г. заболеваемость ЗНО составила около 461 новых случая на каждые 100 тыс. населения [1], что подчеркивает критическую важность развития эффективных методов популяционной онкопрофилактики.

Действующие программы профилактики демонстрируют ограниченную эффективность [2], а традиционные методы остаются дорогостоящими, трудозатратными и практически неприменимыми для масштабного популяционного применения [3], что создает разрыв между потребностью в раннем выявлении ЗНО и возможностями системы здравоохранения. Внедрение электронных медицинских карт (ЭМК) в сочетании с развитием методов искусственного интеллекта (ИИ) открывает новые возможности для автоматизированного анализа медицинских данных. Однако такие ИИ-решения для прогнозирования риска ЗНО требуют либо специализированных данных (например, биомаркеров [4], семейного анамнеза [5], генетических данных [6] и др.), либо специализированной инфраструктуры для развертывания вычислительно сложных решений, что существенно ограничивает их практическое применение в задачах массовой профилактики.

Исследование направлено на поиск, разработку и валидацию метода, способного эффективно решать задачу популяционной онкопрофилактики, основанной исключительно на кодах медицинских диагнозов и услуг, доступных в любой медицинской организации. Для этого: 1) проведен сравнительный анализ широкого спектра различных ИИ-решений и проверена их эффективность в ранжировании пациентов по уровню риска ЗНО; 2) под контролем профильных врачей-онкологов лучшее ИИ-решение прошло ретроспективную валидацию на предмет эффективного применения в задаче популяционной онкопрофилактики.

МАТЕРИАЛЫ И МЕТОДЫ

1. Постановка задачи

Задача прогнозирования риска ЗНО формулируется как бинарная классификация (рис. 1): в момент времени t_{pred} требуется оценить вероятность ЗНО (C00-C97 по МКБ-10) в следующие 12 месяцев, использовав данные ЭМК за предшествующие N месяцев. Значение N выбрано равным 24, чтобы большинство ЭМК не были пустыми, а решение было доступно для массового применения. Такая постановка задачи является консистентной целям федерального проекта «Борьба с онкологическими заболеваниями» на 2025-2030 гг. Целевая переменная определяется как:

- **target = 1**, если выявлено ЗНО в период [t_{pred}, t_{pred} + 12M];
- **target = 0** в противном случае.

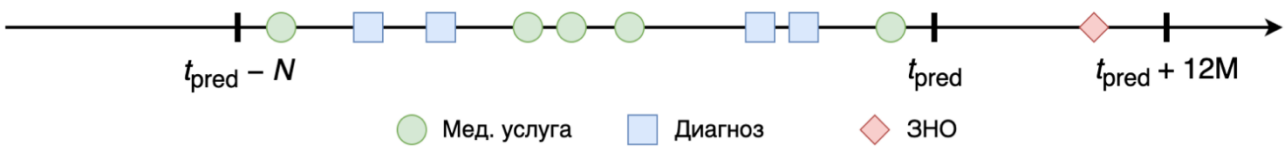


Рис. 1. Постановка задачи

Такая постановка позволяет решать задачу ранжирования пациентов, формируя группы риска для приоритетного прохождения обследований. Для этого используем метрику Average Precision (AP) [7], которая максимизирует долю верных ответов в верхней части списка (Precision@TOP → MAX) и стабильна даже при экстремальном дисбалансе классов, что критически важно в задаче прогнозирования риска ЗНО. Для сравнения с другими решениями, известными в литературе, приведем значения ROC AUC.

2. Методы решения

Для решения поставленной задачи применимы самые разные методы, каждый из которых имеет схожий пайплайн: извлечение признаков из ЭМК → ИИ-модель → оценка вероятности P(ЗНО|ЭМК) или функции выживаемости $S(t|$ЭМК).

В данной работе мы рассматриваем следующие виды решений:

1) **Методы машинного обучения:** логистическая регрессия, случайный лес, градиентный бустинг (GBM);
2) **Модели выживаемости:** AFT-модель, случайные леса выживаемости, DeepHit [8], Deep Survival Machines [9];
3) **Глубокое обучение:** CoLES (*дообучение*) [10], BERT (*претрейн на медицинских текстах*) [11], Longformer (*претрейн на медицинских текстах*) [12];
4) **LLM-энкодеры:** DeepSeek-R1-Distill-Qwen-1.5B (*эмбеддинг последнего скрытого слоя*), Qwen3-Embedding-0.6B, GigaChat-Embeddings;
5) **LLM-конвейер:** LLM-суммаризация (*DeepSeek-R1*) → LLM-энкодер (*Qwen3-Emb*) → *ИИ*-модель (*GBM, DeepHit, LoRA-адаптер*);
6) **Ансамбль GBM и моделей выживаемости**[1] (*популяционные риски – оценки Каплана-Мейера для каждого пола, индивидуальные риски – AFT-модель*): P(ЗНО|ЭМК)=GBM(ML-предикторы $\oplus$ Предикторы-выживаемости), где $\oplus$ – это конкатенация.

ЭКСПЕРИМЕНТЫ

1. Сравнение методов

Методы сравнивались на амбулаторных данных 175411 пациентов (18+) за период 2017-2021 гг. Для корректности выводов была проведена стратификация пациентов по полу и возрасту с проверкой однородности выборок [13]. Проверялись многомерная гипотеза однородности демографических характеристик H_0: $F_{Train}(x)=F_{Validate}(x)=F_{Test}(x)$ и одномерная гипотеза однородности времени до ЗНО H_0: $S_i(t)=S_j(t)$ $\forall i\neq j$, $i,j\in$\{Train,Validate,Test\}. Минимальное p-value > 0.05, что подтверждает отсутствие систематических различий между выборками.

Модели были обучены на выборке Train (54%), гиперпараметры оптимизированы с помощью Optuna на выборке Validate (23%). В табл. 1 показаны результаты сравнения методов на выборке Test (23%) с указанием 95%-х доверительных интервалов.

[1] https://github.com/sb-ai-lab/Can-SAVE

Из полученных результатов видно, что ансамбль градиентного бустинга и моделей выживаемости превосходит другие решения с AP 22.8%. Даже несмотря на то что у других решений значения ROC AUC выше, ансамбль обладает лучшей способностью ранжировать пациентов по уровню риска ЗНО, чем остальные методы. Покажем далее, на какие факторы опирается найденное ИИ-решение.

Табл. 1. Сравнительный анализ методов на тестовой выборке (95% ДИ)

Метод	Average Precision, %	ROC AUC, %
Логистическая регрессия	10.4 ± 1.3	83.4 ± 0.7
Случайный лес	10.2 ± 0.5	83.3 ± 0.6
Градиентный бустинг (GBM)	16.0 ± 1.8	78.6 ± 1.3
Ансамбль GBM и моделей выживаемости	**22.8 ± 2.7**	83.7 ± 1.7
AFT-модель	11.7 ± 1.7	84.8 ± 2.2
Случайные леса выживаемости	7.4 ± 0.3	78.6 ± 0.5
DeepHit	10.2 ± 2.5	86.4 ± 1.6
Deep Survival Machines	10.1 ± 0.5	82.3 ± 0.6
CoLES (fine-tuned)	10.3 ± 0.2	81.3 ± 0.2
BERT → GRU	15.1 ± 2.6	84.9 ± 0.8
Longformer → GBM	9.3 ± 0.2	77.7 ± 0.5
Qwen3-Emb → GBM	15.1 ± 0.9	86.9 ± 0.3
Qwen3-Emb → DeepHit	18.6 ± 0.7	88.5 ± 0.3
DeepSeek-R1 → GBM	16.4 ± 1.0	87.3 ± 0.5
GigaChat → GBM	18.5 ± 0.2	89.6 ± 0.1
DeepSeek-R1 → Qwen3-Emb → GBM	17.6 ± 1.0	88.1 ± 0.5
DeepSeek-R1 → Qwen3-Emb → DeepHit	17.4 ± 0.4	89.5 ± 0.2
DeepSeek-R1 → Qwen3-Emb → LoRA	19.3 ± 0.4	**90.1 ± 0.2**

2. Важность признаков

Для найденного ансамбля градиентного бустинга и моделей выживаемости проанализируем важность входящих предикторов. Для этого вычислим Feature Importance (*как часто предикторы используются при построении деревьев решений*) и Permutation Importance (*как сильно случайные перестановки значений предикторов влияют на целевую метрику*). В табл. 2 суммарные значения показателей важности предикторов по каждой группе признаков. Из полученных результатов очевидно, что модели выживаемости вносят ключевой вклад в прогностическую силу найденного ИИ-решения, поскольку их суммарные значения Feature Importance (39.692) и Permutation Importance (6.594) максимальны.

Табл. 2. Важность групп предикторов найденного ИИ-решения

Группа предикторов	Feature Importance	Permutation Importance
Социально-демографические признаки (*пол, возраст*)	Σ = 21.792	Σ = 2.290
Паттерны визитов (*время с первого визита; месяц визита; доля диагнозов от числа визитов*)	Σ = 21.562	Σ = 6.322
Клинические маркеры (*частотность диагнозов D37-D48, O20-O29; время с первого диагноза D00-D48, I00-I99, Q00-Q99; частотность медицинских услуг по иммунной системе*)	Σ = 16.954	Σ = 1.542
Модели выживаемости (*оценки Каплана-Мейера для мужчин (М), женщин (Ж) и М+Ж; AFT-модель; приращение риска оценок Каплана-Мейера через 12 месяцев; приращение риска AFT-модели через 12 месяцев*)	**Σ = 39.692**	**Σ = 6.594**

3. Ретроспективные эксперименты

Предлагаемое ИИ-решение было протестировано в ретроспективных экспериментах под контролем профильных врачей-онкологов в условиях, приближенных к реальным. Для этого под руководством профильных врачей онкологов в 5 регионах Российской Федерации был проведен эксперимент, демонстрирующий способность формировать группы риска пациентов с ЗНО в сравнении с контрольной группой на основе диспансеризации взрослого населения РФ.

Эксперимент состоял из следующих действий: 1) оценить риск ЗНО каждого пациента в выборке с помощью ИИ-решения; 2) сформировать группы риска 1%, 3%, 5% от численности выборки (такой дополнительный поток пациентов не перегрузит систему здравоохранения); 3) передать списки групп риска контролирующим врачам-онкологам для верификации числа верных ЗНО в каждой группе; 4) сравнить с аналогичными результатами контрольной группы. В ретро-эксперименте анализировались ЭМК из 5 регионов РФ численностью более 1.9 млн пациентов (мужчин 43%, женщин 57%), охватывающих периоды прогнозирования 2018-2024 гг. Пациенты были включены в исследование, если были не младше 18 лет на момент t_{pred} и у них отсутствовало ЗНО в анамнезе жизни. Результаты данного эксперимента представлены в табл. 3.

Табл. 3. Результаты сравнения случаев ЗНО в группах риска 1%-5%

Группа риска, количество пациентов		Контрольная группа, ЗНО	ИИ-решение, ЗНО	Прирост
%	чел.			
Регион 1 (численность: 92 985)				
1%	930	9	41	4,4x
3%	2 790	28	94	3,4x
5%	4 649	46	133	2,9x
Регион 2 (численность: 112 620)				
1%	1 126	11	60	5,3x
3%	3 378	34	117	3,5x
5%	5 631	56	178	3,2x

Регион 3 (численность: 165 355)				
1%	1 653	15	74	4,9x
3%	4 960	35	97	2,8x
5%	8 267	58	117	2,0x
Регион 4 (численность: 651 697)				
1%	6 516	85	315	3,7x
3%	19 550	254	658	2,6x
5%	32 584	424	933	2,2x
Регион 5 (численность: 889 293)				
1%	8 893	80	434	5,4x
3%	26 679	240	781	3,3x
5%	44 465	400	1103	2,8x

Как видно из представленных результатов, ИИ-решение способно эффективно формировать группы риска ЗНО, превышая от 2.0 до 5.4 раза пациентов с ЗНО в каждой группе риска (1%-5%) в сравнении с текущим состоянием процесса выявления ЗНО. Это значит, что если выполнить ИИ-сканирование пациентов целого региона, то работа с группой риска размером от 1% до 5% от численности региона способна повысить показатели выявляемости ЗНО, не перегрузив систему здравоохранения. Таким образом, полученные результаты ретроспективного эксперимента подтверждают, что методы ИИ даже на минимально доступных данных способны повысить качество результатов популяционной онкопрофилактики.

ЗАКЛЮЧЕНИЕ

Найденное решение на основе ИИ доказывает возможность эффективного решения задачи популяционной онкопрофилактики с помощью ИИ-методов с использованием исключительно кодов медицинских диагнозов и услуг. Ансамбль моделей выживаемости и градиентного бустинга превосходит другие рассмотренные подходы при минимальных требованиях к данным и вычислительным ресурсам.

Результаты ретроспективной валидации на более чем 1.9 млн пациентов под контролем профильных онкологов подтвердили клиническую значимость

найденного решения и его способность существенно повысить эффективность популяционной онкопрофилактики. Представленное решение естественным образом встраивается в существующий медицинский процесс онконастороженности, направляя пациентов из группы риска к специалистам первичного звена для принятия решения о дополнительном обследовании и направлении к врачу-онкологу.

Минимальные требования к инфраструктуре делают решение доступным для широкого внедрения в различных системах здравоохранения, что открывает новые возможности для раннего выявления онкологических заболеваний и снижения смертности от ЗНО.

СПИСОК ЛИТЕРАТУРЫ

1. *Каприн А. Д., Старинский В. В., Шахзадова А. О.* Злокачественные новообразования в России в 2023 году (заболеваемость и смертность) / Под ред. А. Д. Каприна, В. В. Старинского, А. О. Шахзадовой. М.: МНИОИ им. П. А. Герцена — филиал ФГБУ «НМИЦ радиологии» Минздрава России, 2024. 276 с. ISBN 978-5-85502-298-8.

2. *Cenin D. R., Tinmouth J., Naber S. K., Khalaf N., Rabeneck L., Tinmouth J. M., Earle C. C., Hilsden R. J., Leddin D., Rostom A., Issaka R. B., Heitman S. J., Lansdorp-Vogelaar I.* Calculation of stop ages for colorectal cancer screening based on comorbidities and screening history // Clinical Gastroenterology and Hepatology. 2021. Vol. 19, No. 3. P. 547–555. https://doi.org/10.1016/j.cgh.2020.05.038

3. *Ratushnyak S., Hoogendoorn M., van Baal P. H. M.* Cost-effectiveness of cancer screening: health and costs in life years gained // American Journal of Preventive Medicine. 2019. Vol. 57, No. 6. P. 792–799. https://doi.org/10.1016/j.amepre.2019.07.027

4. *Alexander M., Burbury K.* A systematic review of biomarkers for the prediction of thromboembolism in lung cancer — Results, practical issues and proposed strategies for future risk prediction models // Thrombosis Research. 2016. Vol. 148. P. 63–69. https://doi.org/10.1016/j.thromres.2016.10.020

5. *Jacobs M. F.* Predicting cancer risk based on family history // eLife. 2021. Vol. 10. e73380. https://doi.org/10.7554/eLife.73380

6. *Wang X., Oldani M. J., Zhao X., Huang X., Qian D.* A review of cancer risk prediction models with genetic variants // Cancer Informatics. 2014. Vol. 13, Suppl. 2. P. 19–28. https://doi.org/10.4137/CIN.S13788

7. *Zhu M.* Recall, precision and average precision // Technical Report. Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, 2004. 6 p.

8. *Lee C., Zame W. R., Yoon J., van der Schaar M.* DeepHit: A deep learning approach to survival analysis with competing risks // Proceedings of the AAAI Conference on Artificial Intelligence. 2018. Vol. 32, No. 1. P. 2314–2321. https://doi.org/10.1609/aaai.v32i1.11842

9. *Nagpal C., Li X., Dubrawski A.* Deep survival machines: Fully parametric survival regression and representation learning for censored data with competing risks // IEEE Journal of Biomedical and Health Informatics. 2021. Vol. 25, No. 8. P. 3163–3175. https://doi.org/10.1109/JBHI.2021.3052441

10. *Babaev D., Ovsov N., Kireev I., Ivanova M., Gusev G., Nazarov I., Tuzhilin A.* CoLES: Contrastive learning for event sequences with self-supervision // Proceedings of the 2022 International Conference on Management of Data (SIGMOD '22). New York, NY, USA: ACM, 2022. P. 1190–1199. https://doi.org/10.1145/3514221.3526129

11. *Blinov P., Kokh V.* Medical profile model: scientific and practical applications in healthcare // IEEE Journal of Biomedical and Health Informatics. 2023. Vol. 28, No. 1. P. 450–458. https://doi.org/10.1109/JBHI.2023.3295631

12. *Yalunin A., Nesterov A., Umerenkov D.* RuBioRoBERTa: a pre-trained biomedical language model for Russian language biomedical text mining // arXiv preprint. 2022. arXiv:2204.03951. https://doi.org/10.48550/arXiv.2204.03951

13. *Philonenko P., Postovalov S.* The new robust two-sample test for randomly right-censored data // Journal of Statistical Computation and Simulation. 2019. Vol. 89, No. 8. P. 1357–1375. https://doi.org/10.1080/00949655.2019.1577858

AI IN CANCER PREVENTION: A RETROSPECTIVE STUDY

P. A. Philonenko[1] [0000-0002-6295-4470], V. N. Kokh[2] [0000-0002-9257-0259],
P. D. Blinov[3] [0009-0000-7583-7655]

[1, 3]*Sber AI Lab, Moscow, Russia*
[2]*Sber AI, Moscow, Russia*

[1]petr-filonenko@mail.ru, [2]kokh.v.n@sber.ru, [3]blinov.p.d@sber.ru

Abstract

This study investigates the feasibility of effectively solving population-scale cancer screening problems using artificial intelligence (AI) methods that predict malignant neoplasm risk based on minimal electronic health record (EHR) data – medical diagnosis and service codes. To address the formulated problem, we considered a broad spectrum of modern approaches, including classical machine learning methods, survival analysis, deep learning, and large language models (LLMs). Numerical experiments demonstrated that gradient boosting using survival analysis models as additional predictors possesses the best ability to rank patients by cancer risk level, enabling consideration of both population-level and individual risk factors for malignant neoplasms. Predictors constructed from EHR data include demographic characteristics, healthcare utilization patterns, and clinical markers. This solution was tested in retrospective experiments under the supervision of specialized oncologists. In the retrospective experiment involving more than 1.9 million patients, we established that the risk group captures up to 5.4 times more patients with cancer at the same level of medical examinations. The investigated method represents a scalable solution using exclusively diagnosis and service codes, requiring no specialized infrastructure and integrable into oncological vigilance processes, making it applicable for population-scale cancer screening.

Keywords*: AI in medicine, cancer prevention, retrospective experiments.*

REFERENCES

1. *Kaprin A. D., Starinskiy V. V., Shakhzadova A. O.* Malignant neoplasms in Russia in 2023 (incidence and mortality) / Ed. by A. D. Kaprin, V. V. Starinskiy, A. O. Shakhzadova. Moscow: P. A. Herzen Moscow Oncology Research Institute — Branch of

the National Medical Research Radiological Centre of the Ministry of Health of Russia, 2024. 276 p. ISBN 978-5-85502-298-8. (In Russian).

2. *Cenin D. R., Tinmouth J., Naber S. K., Khalaf N., Rabeneck L., Tinmouth J. M., Earle C. C., Hilsden R. J., Leddin D., Rostom A., Issaka R. B., Heitman S. J., Lansdorp-Vogelaar I.* Calculation of stop ages for colorectal cancer screening based on comorbidities and screening history. Clinical Gastroenterology and Hepatology, 2021, vol. 19, no. 3, pp. 547–555. https://doi.org/10.1016/j.cgh.2020.05.038

3. *Ratushnyak S., Hoogendoorn M., van Baal P. H. M.* Cost-effectiveness of cancer screening: health and costs in life years gained. American Journal of Preventive Medicine, 2019, vol. 57, no. 6, pp. 792–799. https://doi.org/10.1016/j.amepre.2019.07.027

4. *Alexander M., Burbury K.* A systematic review of biomarkers for the prediction of thromboembolism in lung cancer — Results, practical issues and proposed strategies for future risk prediction models. Thrombosis Research, 2016, vol. 148, pp. 63–69. https://doi.org/10.1016/j.thromres.2016.10.020

5. *Jacobs M. F.* Predicting cancer risk based on family history. eLife, 2021, vol. 10, e73380. https://doi.org/10.7554/eLife.73380

6. *Wang X., Oldani M. J., Zhao X., Huang X., Qian D.* A review of cancer risk prediction models with genetic variants. Cancer Informatics, 2014, vol. 13, suppl. 2, pp. 19–28. https://doi.org/10.4137/CIN.S13788

7. *Zhu M.* Recall, precision and average precision. Technical Report, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, 2004, 6 p.

8. *Lee C., Zame W. R., Yoon J., van der Schaar M.* DeepHit: A deep learning approach to survival analysis with competing risks. Proceedings of the AAAI Conference on Artificial Intelligence, 2018, vol. 32, no. 1, pp. 2314–2321. https://doi.org/10.1609/aaai.v32i1.11842

9. *Nagpal C., Li X., Dubrawski A.* Deep survival machines: Fully parametric survival regression and representation learning for censored data with competing risks. IEEE Journal of Biomedical and Health Informatics, 2021, vol. 25, no. 8, pp. 3163–3175. https://doi.org/10.1109/JBHI.2021.3052441

10. *Babaev D., Ovsov N., Kireev I., Ivanova M., Gusev G., Nazarov I., Tuzhilin A.* CoLES: Contrastive learning for event sequences with self-supervision. Proceedings of

the 2022 International Conference on Management of Data (SIGMOD '22), New York, NY, USA, ACM, 2022, pp. 1190–1199. https://doi.org/10.1145/3514221.3526129

11. *Blinov P., Kokh V.* Medical profile model: scientific and practical applications in healthcare. IEEE Journal of Biomedical and Health Informatics, 2023, vol. 28, no. 1, pp. 450–458. https://doi.org/10.1109/JBHI.2023.3295631

12. *Yalunin A., Nesterov A., Umerenkov D.* RuBioRoBERTa: a pre-trained biomedical language model for Russian language biomedical text mining. arXiv preprint, 2022, arXiv:2204.03951. https://doi.org/10.48550/arXiv.2204.03951

13. *Philonenko P., Postovalov S.* The new robust two-sample test for randomly right-censored data. Journal of Statistical Computation and Simulation, 2019, vol. 89, no. 8, pp. 1357–1375. https://doi.org/10.1080/00949655.2019.1577858

СВЕДЕНИЯ ОБ АВТОРАХ

ФИЛОНЕНКО Петр Александрович – окончил Новосибирский государственный технический университет (ныне – НГТУ-НЭТИ). В 2018 году защитил диссертацию на соискание кандидата технических наук по специальности 05.13.17 – Теоретические основы информатики. В настоящее время работает руководителем направления по исследованию данных в Sber AI Lab. Область научных интересов: прогнозирование риска заболеваний, теория выживаемости, проверка статистических гипотез, машинное обучение.

Petr Aleksandrovich PHILONENKO – graduated from Novosibirsk State Technical University (now NSTU-NETI). In 2018, he defended his dissertation for the degree of Candidate of Technical Sciences in specialty 05.13.17 Theoretical foundations of computer science. Currently, he works as a Senior Data Scientist at the Sber AI Lab. Research interests: risk prediction, survival analysis, hypothesis testing, machine learning.

email: petr-filonenko@mail.ru
ORCID: 0000-0002-6295-4470

КОХ Владимир Николаевич – окончил Сибирский государственный медицинский университет. В настоящее время работает исполнительным директором по анализу данных в Sber AI. Область научных интересов: ИИ в медицине, R&D на основе искусственного интеллекта в здравоохранении.

Vladimir Nikolaevich KOKH – graduated from Siberian State Medical University. Currently, he works as an Executive Director of Data Analysis at Sber AI. Research interests: AI in medicine, artificial intelligence–based R&D in healthcare.

email: kokh.v.n@sber.ru

ORCID: 0000-0002-9257-0259

БЛИНОВ Павел Дмитриевич – окончил Вятский государственный университет по специальности прикладная математика и информатика, закончил в 2011 году. В 2016 году защитил кандидатскую диссертацию на соискание учёной степени кандидата технических наук по специальности 05.13.17 "Теоретические основы информатики". В настоящее время работает исполнительным директором по исследованию данных в Sber AI Lab. Область научных интересов: обработка естественного языка, методы машинного обучения, интеллектуальный анализ медицинских данных.

Pavel Dmitrievich BLINOV – graduated with a degree in Applied Mathematics and Computer Science from Vyatka State University in 2011. In 2016, he defended his dissertation and was awarded the degree of Candidate of Sciences (Ph.D. equivalent) in "Theoretical Foundations of Computer Science" (specialty code 05.13.17). He is currently the Executive Director of Data Science at Sber AI Lab. His research interests include natural language processing, machine learning methods, and medical knowledge mining.

email: blinov.p.d@sber.ru

ORCID: 0009-0000-7583-7655

Материал поступил в редакцию 10 октября 2025 года

УДК 004.891.3

СТИЛОМЕТРИЧЕСКИЙ АНАЛИЗ В ЗАДАЧЕ ПОИСКА ЗАИМСТВОВАНИЙ ТЕКСТОВ НА ТАТАРСКОМ ЯЗЫКЕ

И. З. Хаялеева[1] [0009-0007-5837-7010], М. М. Абрамский[2] [0000-0003-3063-8948]

[1, 2] *Казанский (Приволжский) федеральный университет, г. Казань, Россия*

[1]izidakh@yandex.ru, [2]mabramsk@kpfu.ru

Аннотация

Рассмотрена возможность применения методов стилометрического анализа для поиска заимствований в текстах на татарском языке. Разработаны соответствующие инструменты, в которых использованы алгоритмы машинного обучения, включая кластеризацию (метод *k*-средних), классификацию (метод случайного леса, метод опорных векторов, наивный байесовский классификатор) и гибридный подход (модель FastText + логистическая регрессия). Особое внимание уделено адаптации лингвистических метрик для татарского языка.

Ключевые слова*: поиск заимствований, обработка естественного языка, стилометрический анализ, татарский язык.*

ВВЕДЕНИЕ

В современном мире, где информация играет ключевую роль, анализ текстов и определение их авторства становятся все более актуальными задачами. Особенно это касается малоресурсных языков, чьи носители стремятся к сохранению и развитию своего культурного наследия. Одним из таких языков является татарский. В государственной программе «Сохранение, изучение и развитие государственных языков Республики Татарстан и других языков в Республике Татарстан на 2023–2030 годы», принятой Постановлением № 821 Кабинета Министров Республики Татарстан в 2020 г. [1], отмечено, что использование татарского языка в сфере науки, в том числе при написания квалификационных работ для присуждения академической или ученой степеней, сохраняет свою актуальность. А это, в свою очередь, требует современных и точных средств определения уникальности текста.

© И. З. Хаялеева, М. М. Абрамский, 2025.
Данная статья распространяется на условиях международной лицензии Creative Commons License Attribution 4.0 International (CC BY 4.0).

Целью настоящей работы являются исследование и разработка подходов поиска заимствований в текстах на татарском языке, анализирующих исходный документ с помощью стилометрических[1] методов. Для их достижения были поставлены следующие задачи:

– провести исследование стилометрических методов поиска заимствований;

– применить эти методы в задаче поиска заимствований на татарском языке;

– протестировать и оценить корректность применения стилометрических методов для поиска заимствований на татарском языке.

ОПРЕДЕЛЕНИЕ АВТОРСТВА ТЕКСТА

Одной из задач, решаемых стилометрическим анализом, является определение авторства текста. Для ее решения был реализован инструмент, основанный на алгоритме кластеризации *k*-средних.

Алгоритм *k*-средних – один из методов кластерного анализа, позволяющий разделить произвольный набор данных на заданное количество кластеров таким образом, чтобы объекты внутри одного кластера находились достаточно близко друг к другу, а объекты из разных кластеров не пересекались [2].

В настоящей работе алгоритм *k*-средних был использован для определения *k* различных центроидов в тексте, имеющем разные стили написания. Каждый центроид охватывает такие фрагменты, которые имеют одинаковый стиль написания. Следовательно, количество центроидов соответствует различному количеству стилей написания, присутствующих в документе. На основе предположения о том, что каждый отдельный стиль принадлежит каждому отдельному автору, можно получить оценку авторства каждой части текста. Схема работы созданного инструмента представлена на рис. 1.

[1] Стилометрия – система средств и приемов количественного измерения стилистических характеристик текста.

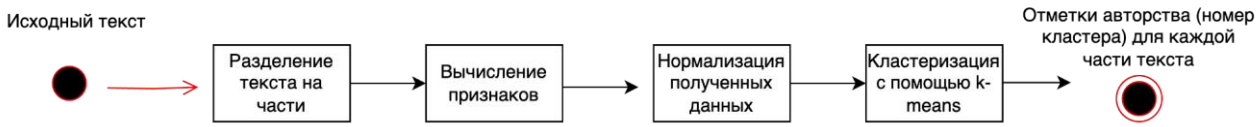


Рис. 1. Схема работы инструмента определения авторства

Для применения алгоритма кластеризации исходный текст был разделен на фрагменты определенной длины, каждый из которых был представлен в виде вектора следующих характеристик:

- сложность чтения текста;
- разнообразие используемых в тексте слов;
- лексические особенности текста.

Для определения метрик измерения перечисленных выше характеристик в тексте на татарском языке были проанализированы работы [3–5]. Стоит отметить, что некоторые подходы к оценке должны рассчитываться с учетом возраста, образования или уровня развития читателя. Соответствующие метрики не были включены в рассмотрение из-за отсутствия данных о пользователях.

Выявленные стилометрические метрики, используемые для векторизации текста, можно отнести к трем группам: лексические, вычисляющие разнообразие используемой лексики и вычисляющие сложность чтения. К лексическим метрикам относятся:

- средняя длина слова;
- среднее количество символов в предложении;
- среднее количество слов в предложении;
- среднее количество слогов в слове;
- количество пунктуационных символов;
- частота специальных символов;
- частота служебных частей речи.

Описанные выше метрики опираются на устоявшиеся понятия в области лингвистики и языкознания. Гораздо больший интерес представляют две другие группы метрик. К метрикам, вычисляющим разнообразие используемой лексики, относятся:

- количество слов *hapax legomenon* – слов, которые встречаются в тексте только один раз. Этот термин часто используют для изучения уникальных слов, которые могут содержать важную информацию о тексте или культуре, в которой был написан текст;
- количество слов *dis legomenon* – таких слов, которые встречаются в тексте только два раза [5];
- мера Оноре – мера, зависящая от количества hapax legomenon и вычисляемая по формуле $H = 100 \log N/(1 - l/d)$, где N – количество слов в тексте, l – количество hapax legomena, d – количество уникальных слов в тексте [6];
- мера Сичела – мера, зависящая от количества dis legomenon и вычисляемая по формуле

$$S = \frac{\mathrm{dis}}{d},$$

где dis – количество dis legomenon, d – количество уникальных слов в тексте [5].

- мера Брюнета – мера, опирающаяся на количество hapax legomenon и вычисляемая по формуле

$$W = N^{d^{-0.17}},$$

где N – количество слов в тексте, d – количество уникальных слов в тексте [6];

- соотношение количества уникальных слов к общему количеству;
- энтропия Шеннона – мера количества информации, которую несет текст, вычисляемая по формуле

$$E = \sum_{i=0}^{N-1} P_i \log P_i,$$

где P_i – вероятность того, что слово под номером i встретится в тексте, а *N* – количество слов в тексте [7].

В качестве метрики сложности текста был отобран индекс удобочитаемости Флеша, оценивающий сложность текста по следующей формуле:

$$\text{УФ} = 206.835 - (1.015\ a) - (84.6\ b),$$

где *a* – средняя длина предложения в словах, *b* – среднее число слогов в слове [3].

Для валидации предложенного подхода был проведен эксперимент на синтетических данных, где в один текст искусственно объединялись фрагменты

от двух разных авторов. Алгоритм показал точность сегментации (accuracy) – 0.78, precision – 0.81 и recall – 0.75 при обнаружении границ стилей.

ОПРЕДЕЛЕНИЕ ЖАНРА ТЕКСТА

Еще одной задачей, решаемой с применением стилометрического анализа, является определение жанра текста. В настоящей работе для ее решения была разработана модель классификации, основанная на векторном представлении текста в виде словаря известных слов.

Для обучения модели были использованы 3450 текстов из разных татароязычных источников, имеющих научный, новостной или художественный жанры. Распределение жанров текстов приведено на рис. 2.

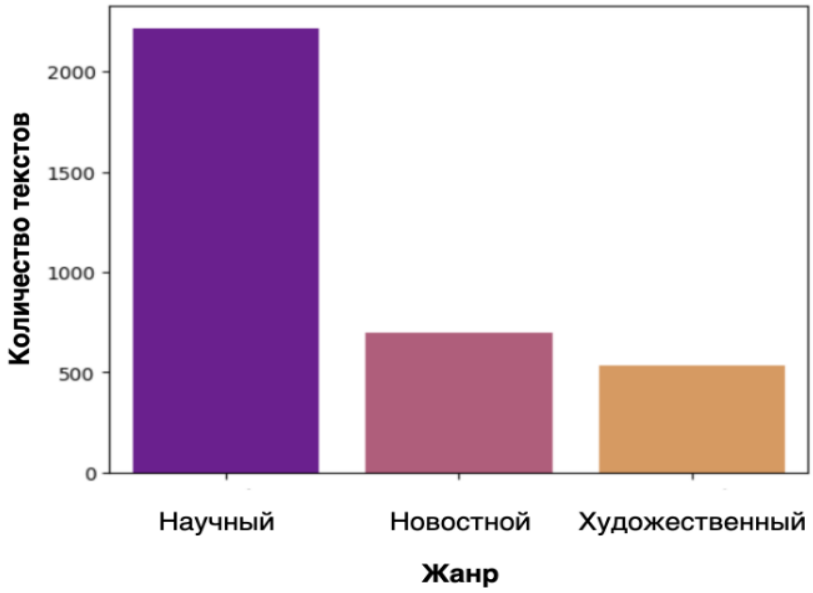


Рис. 2. Распределение текстов по жанрам в обучающем наборе данных (количество текстов – 3450)

Опираясь на работу [9], для данной задачи были реализованы и протестированы три алгоритма классификации: метод случайного леса [10], метод опорных векторов [11] и мультиномиальный наивный байесовский классификатор [12]. Наибольшую точность показал метод случайного леса. Результаты тестирования алгоритмов представлены в табл. 1.

Табл. 1. Сравнительные результаты оценки методов классификации текстов по жанрам

	Метод случайного леса	Метод опорных векторов	Мультиномиальный байесовский классификатор
Доля правильных ответов алгоритма	0.982	0.976	0.852
Точность	0.983	0.977	0.854
Полнота	0.982	0.976	0.852
F1-мера	0.982	0.976	0.844

ОПРЕДЕЛЕНИЕ ЭМОЦИОНАЛЬНОГО ТОНА ТЕКСТА

Определение эмоционального тона текста представляет собой еще одну важную задачу, решаемую в рамках стилометрического анализа. Эта задача заключается в автоматическом определении общего настроения текста: положительного, отрицательного или нейтрального. В более детализированных постановках задачи можно также говорить о классификации по типу эмоций (радость, гнев, печаль и др.). В рамках настоящей работы была рассмотрена базовая модель с трехклассовой классификацией.

Эта задача особенно актуальна для татароязычных текстов, представленных в социальных сетях, комментариях, форумах и пользовательских отзывах. В условиях отсутствия готовых корпусов и моделей на татарском языке разработка инструментов анализа тональности позволяет расширить возможности автоматической обработки текстов и способствует применению языка в современных цифровых сервисах.

Для решения этой задачи был использован гибридный подход, сочетающий методы векторизации текста при помощи предобученной модели FastText и классического машинного обучения. FastText включает в себя модель, обученную на татарском языке (cc.tt.300.vec) [13]. Она позволяет представить любой текст в виде вектора фиксированной размерности, основанного на усреднении векторов слов, входящих в текст.

Алгоритм определения эмоционального тона включал следующие этапы:

- предобработка текста (приведение к нижнему регистру, удаление пунктуации, токенизация);
- получение векторного представления текста;
- обучение классификатора на размеченном корпусе.

Для обучения модели был собран корпус текстов, размеченных вручную по трем категориям: положительный, отрицательный, нейтральный. Каждый текст представлял собой короткое высказывание (1–3 предложения), имитирующее типичные фрагменты отзывов, комментариев или пользовательских мнений. Примеры таких текстов приведены в табл. 2.

Табл. 2. Примеры размеченных текстов

Текст на татарском	Перевод на русский язык	Метка
Бу фильм бик күңелле иде	Этот фильм был очень интересным	положительный
Мин бу китапны яратмадым	Мне эта книга не понравилась	отрицательный
Кичә яңгыр яуды, һава салкын	Вчера дождь шел, было холодно	нейтральный

Для классификации была использована логистическая регрессия, реализованная с помощью библиотеки Scikit-learn [14] и обученная на векторах, полученных с помощью FastText. Модель показала удовлетворительные результаты на тестовом множестве (точность – 0.89, F1-мера – 0.88). Несмотря на небольшой размер корпуса, уже на этой стадии система способна различать базовые эмоциональные категории в татарских текстах.

Таким образом, определение эмоционального тона представляет собой перспективное направление в рамках стилометрического анализа текстов на татарском языке и может быть полезным как в задаче оценки субъективной окраски

текста, так и в качестве дополнительной информации при обнаружении заимствований и определении авторства.

ЗАКЛЮЧЕНИЕ

Успешно применены методы стилометрического анализа для решения задач определения авторства, жанра и эмоционального тона текстов. Результаты тестирования созданных инструментов показали хорошие результаты, однако исследование имеет ряд ограничений, таких, например, как размер корпуса для анализа тональности и отсутствие валидации на реальных данных для задачи определения авторства. Дальнейшие направления исследований включают:

- увеличение корпуса с привлечением пользовательских данных из открытых источников;
- дообучение предобученных многоязычных трансформеров (например, XLM-RoBERTa) [15] на татароязычных текстах;
- внедрение более тонкой классификации (радость, грусть, тревога и др.);
- обнаружение иронии, сарказма и других сложных эмоциональных проявлений.

СПИСОК ЛИТЕРАТУРЫ

1. Постановление кабинета министров Республики Татарстан «Об утверждении государственной программы Сохранение, изучение и развитие государственных языков Республики Татарстан и других языков в Республике Татарстан на 2023 – 2030 годы» // Официальный портал правовой информации Республики Татарстан. Казань, 2020.
URL: https://pravo.tatarstan.ru/npa_kabmin/post/?npa_id=625356 (Дата обращения: 19.08.2025)

2. *Каримов К.Х., Василий Е.А.* Теоретические основы кластеризации данных // Актуальные вопросы фундаментальных и прикладных научных исследований. 2023. С. 242–247.

3. *Балясова И.И.* Параметры сложности текста в татарском языке // Вызовы и тренды мировой лингвистики. 2020. Т. 16. С. 302.

4. *Солнышкина М.И., Макнамара Д.С., Замалетдинов Р.Р.* Обработка естественного языка и изучение сложности дискурса // Russian Journal of Linguistics.

2022. Т. 26. № 2. С. 317–341.

5. *Scott M., Tribble C.* Textual patterns: Key words and corpus analysis in language education. Амстердам: John Benjamins Publishing, 2006. 203 с.

6. *Honoré A. et al.* Some simple measures of richness of vocabulary // Association for literary and linguistic computing bulletin. 1979. Vol. 7, No. 2. P. 172–177.

7. *Flesch R.* A new readability yardstick // Journal of applied psychology. 1948. Vol. 32, No. 3. P. 221–233.

8. *Kincaid J.P., Fishburne Jr R.P., Rogers R.L., Chissom B.S.* Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel // Institute for Simulation and Training. 1975. 49 p.

9. *Kuzman T., Ljubešić N.* Automatic genre identification: a survey // Language Resources and Evaluation. 2025. Vol. 59, No. 1. P. 537–570.

10. *Salman H.A., Kalakech A., Steiti A.* Random forest algorithm overview // Babylonian Journal of Machine Learning. 2024. Vol. 2024. P. 69–79.

11. *Bansal M., Goyal A., Choudhary A.* A comparative analysis of K-nearest neighbor, genetic, support vector machine, decision tree, and long short-term memory algorithms in machine learning // Decision Analytics Journal. 2022. Vol. 3. P. 100071.

12. *Rastogi S., Sambyal R., Tyagi P., Kushwaha R.* Multinomial Naive Bayes Classification Algorithm Based Robust Spam Detection System // 2024 OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 4.0. IEEE, 2024. P. 1–5.

13. *Khusainova A., Khan A., Rivera A.R.* Sart-similarity, analogies, and relatedness for tatar language: New benchmark datasets for word embeddings evaluation // International Conference on Computational Linguistics and Intelligent Text Processing. Cham: Springer Nature Switzerland, 2019. P. 380–390.

14. *Pedregosa F. et al.* Scikit-learn: Machine Learning in Python // Journal of Machine Learning Research. 2020. Vol. 12. P. 2825–2830.

15. *Conneau A., Khandelwal K., Goyal N., Chaudhary V., Wenzek G., Guzmán F., 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.

STYLOMETRIC ANALYSIS IN THE TASK OF SEARCHING FOR BORROWINGS OF TEXTS IN THE TATAR LANGUAGE

I. Z. Khayaleeva[1] [0009-0007-5837-7010], M. M. Abramskiy[2] [0000-0003-3063-8948]

[1, 2]*Kazan (Volga Region) Federal University Kazan, Russia*

[1]izidakh@yandex.ru, [2]mabramsk@kpfu.ru

Abstract

This article discusses the use of stylometric analysis in searching for borrowings of text in the Tatar language. Relevant tools have been developed, utilizing machine learn-ing algorithms, including clustering (k-means method), classification (random forest method, support vector machine method, naive Bayes classifier), and a hybrid approach (FastText model + logistic regression). Special attention is paid to the adaptation of lin-guistic metrics for the Tatar language.

Keywords*: plagiarism detection, natural language processing, stylometric analysis, Tatar language.*

REFERENCES

1. Postanovlenie Kabineta Ministrov Respubliki Tatarstan "Ob Utverzhdenii Gosudarstvennoy Programmy Sokhranenie, Izucheniye i Razvitie Gosudarstvennykh Yazykov Respubliki Tatarstan i Drugikh Yazykov v Respublike Tatarstan na 2023–2030 Gody" // Official Portal of Juridical Information of Republic of Tatarstan. Kazan, 2020. URL: https://pravo.tatarstan.ru/npa_kabmin/post/?npa_id=625356 (access date: 19.08.2025).

2. *Karimov K.Kh., Vasily E.A.* Teoreticheskie osnovy klasterizacii dannyh // Aktual'nye voprosy fundamental'nyh i prikladnyh nauchnyh issledovanij. 2023. P. 242–247.

3. *Balyasova I.I.* Parametry Slozhnosti Teksta v Tatarskom Yazyke // Vyzovy i Trendy Mirovoy Lingvistiki. 2020. Vol. 16. P. 302.

4. *Solnyshkina M.I. McNamara D.S., Zamaletdinov R.R.* Obrabotka Yestestvennogo Yazyka i Izucheniye Slozhnosti Diskursa // Russian Journal of Linguistics. 2022. Vol. 26, No. 2. P. 317–341.

5. *Scott M., Tribble C.* Textual patterns: Key words and corpus analysis in language education. Amsterdam: John Benjamins Publishing, 2006. 203 c.

6. *Honoré A. et al.* Some simple measures of richness of vocabulary // Association for literary and linguistic computing bulletin. 1979. Vol. 7, No. 2. P. 172–177.

7. *Flesch R.* A new readability yardstick // Journal of applied psychology. 1948. Vol. 32, No. 3. P. 221–233.

8. *Kincaid J.P., Fishburne Jr R.P., Rogers R.L., Chissom B.S.* Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel // Institute for Simulation and Training. 1975. 49 p.

9. *Kuzman T., Ljubešić N.* Automatic genre identification: a survey // Language Resources and Evaluation. 2025. Vol. 59, No. 1. P. 537–570.

10. *Salman H.A., Kalakech A., Steiti A.* Random forest algorithm overview // Babylonian Journal of Machine Learning. 2024. Vol. 2024. P. 69–79.

11. *Bansal M., Goyal A., Choudhary A.* A comparative analysis of K-nearest neighbor, genetic, support vector machine, decision tree, and long short-term memory algorithms in machine learning // Decision Analytics Journal. 2022. Vol. 3. P. 100071.

12. *Rastogi S., Sambyal R., Tyagi P., Kushwaha R.* Multinomial Naive Bayes Classification Algorithm Based Robust Spam Detection System // 2024 OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 4.0. IEEE, 2024. P. 1–5.

13. *Khusainova A., Khan A., Rivera A.R.* Sart-similarity, analogies, and relatedness for tatar language: New benchmark datasets for word embeddings evaluation // International Conference on Computational Linguistics and Intelligent Text Processing. Cham: Springer Nature Switzerland, 2019. P. 380–390.

14. *Pedregosa F. et al.* Scikit-learn: Machine Learning in Python // Journal of Machine Learning Research. 2020. Vol. 12. P. 2825–2830.

15. *Conneau A., Khandelwal K., Goyal N., Chaudhary V., Wenzek G., Guzmán F., 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

ХАЯЛЕЕВА Изида Зуфаровна – аспирант Института информационных технологий и интеллектуальных систем Казанского (Приволжского) федерального университета

Izida Zufarovna KHAYALEEVA – PhD student at the Institute of Information Technology and Intelligent Systems, Kazan Federal University.

email: izidakh@yandex.ru

ORCID: 0009-0007-5837-7010

АБРАМСКИЙ Михаил Михайлович – директор Института информационных технологий и интеллектуальных систем Казанского (Приволжского) федерального университета, кандидат технических наук.

Mikhail Mikhailovich ABRAMSKIY – director of the Institute of Information Technology and Intelligent Systems, Kazan Federal University, PhD (Cand Sci. – Tech.)

email: mabramsk@kpfu.ru

ORCID: 0000-0003-3063-8948

Материал поступил в редакцию 15 октября 2025 года