

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

К. С. Клокова¹ [0000-0002-6650-1946], **М. А. Кронгауз**² [0000-0003-2956-0565],
В. А. Шульгинов³ [0000-0003-2217-3677], **Т. А. Юдина**⁴ [0000-0001-6631-2128]

^{1, 3, 4}*Московский физико-технический институт, г. Долгопрудный, Россия*

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

¹kloкова.ks@mipt.ru, ²mkronhaus@yandex.ru, ³shulginov.va@mipt.ru,

⁴yudina.tatiana.a@mipt.ru

Аннотация

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

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

ВВЕДЕНИЕ

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

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

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

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

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

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

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

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

1. ОПИСАНИЕ КОРПУСА И ОБНОВЛЕНИЯ

1.1. Обзор корпуса

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

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

2.2. Недавние обновления и дополнения

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

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

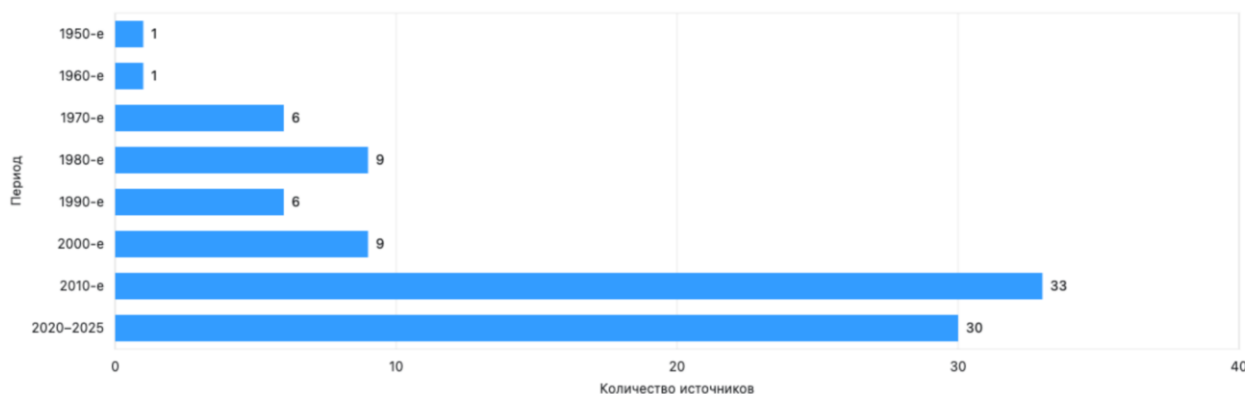


Рис. 1. Распределение количества фильмов/сериалов в корпусе по десятилетиям.

1.3 Веб-интерфейс и возможности поиска

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

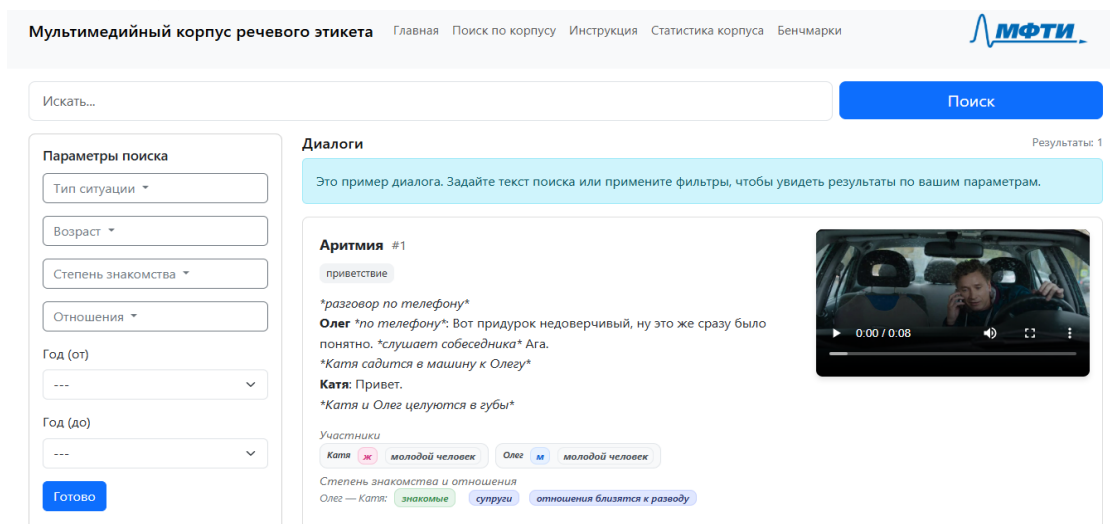


Рис. 2. Страница поиска по корпусу.

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

Табл. 1. Параметры поиска.

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

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

1.4. «Бенчмарки на материале корпуса»

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

2. ОПРЕДЕЛЕНИЕ ЗАДАЧИ И НАБОР ДАННЫХ

2.1. Постановка задачи

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

2.1.1. Идентификация и извлечение последовательностей

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

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

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

2.1.2. Классификация диалогов

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

2.2. Набор данных

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

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

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

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

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

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

3. ЭКСПЕРИМЕНТАЛЬНАЯ УСТАНОВКА

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

3.1. Дизайн промпта и генерация

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

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

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

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

3.2. Стратегии голосования

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

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

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

3.3. Метрики оценки

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

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

$\{p_1, p_2, \dots, p_n\}$ вычисляем

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

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

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

4. РЕЗУЛЬТАТЫ

4.1. Классификация типа диалога

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

Табл. 3. Эффективность классификации типа диалога.

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

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

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

4.2. Извлечение последовательностей приветствий/прощаний

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

Табл. 4. Результаты извлечения последовательностей по типам (Т1: точное совпадение, Т2: нечеткое совпадение, Т3: уровень токенов).

Модель	Тип последовательности	Т1 (P/R/F1)	Т2 (P/R/F1)	Т3 (P/R/F1)
YandexGPT				

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

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

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

4.3. Анализ ошибок

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

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

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

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

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

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

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

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

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

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

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

ЗАКЛЮЧЕНИЕ

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

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

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

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

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

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

Благодарности

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

СПИСОК ЛИТЕРАТУРЫ

1. Brown P., Levinson S. C. Politeness: Some Universals in Language. Cambridge University Press, 1987.
2. Kádár D.Z., Haugh M. Understanding politeness. Cambridge University Press, 2013.
3. Schegloff E.A. Sequence Organization in Interaction: A Primer in Conversation Analysis I. Cambridge University Press, 2007.
4. Ruis L. et al. The goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by LLMs // arXiv:2210.14986. 2024. <https://doi.org/10.48550/arXiv.2210.14986>
5. Jian M., Narayanaswamy S. Are LLMs good pragmatic speakers? // arXiv:2411.01562. 2024. <https://doi.org/10.48550/arXiv.2411.01562>
6. Hu J., Floyd S., Jouravlev O., Fedorenko E., Gibson E. A fine-grained comparison of pragmatic language understanding in humans and language models // Proc. of the 61st Annual Meeting of the Association for Computational Linguistics. 2023. Vol. 1. P. 4194–4213.
7. Lipkin B., Wong L., Grand G., Tenenbaum J.B. Evaluating statistical language models as pragmatic reasoners // arXiv:2305.01020. 2023. <https://doi.org/10.48550/arXiv.2305.01020>
8. Iida A., Okuoka K., Omori T., Nakashima R., Osawa M. Llm-based evaluation of utterances with implicature understanding: A preliminary study // Proc. of the 13th International Conference on Human-Agent Interaction. 2025. P. 485–487. <https://doi.org/10.1145/3765766.3765841>
9. Yerukola A., Vaduguru S., Fried D., Sap M. Is the pope catholic? Yes, the pope is catholic. Generative evaluation of non-literal intent resolution in LLMs // Proc. of the 62nd Annual Meeting of the Association for Computational Linguistics. 2024. Vol. 2. P. 265–275.

10. *Liu R., Sumers T. R., Dasgupta I., Griffiths T.L.* How do large language models navigate conflicts between honesty and helpfulness? // Proc. of the 41st International Conference on Machine Learning. 2024. P. 31844–31865.
 11. *Chen Z., Yang R., Zhao Z., Cai D., He X.* Dialogue act recognition via CRF-attentive structured network. // Proc. of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. 2018. P. 225–234.
 12. *Qamar A., Tong J., Huang R.* Do LLMs understand dialogues? A case study on dialogue acts. // Proc. of the 63rd Annual Meeting of the Association for Computational Linguistics. 2025. Vol. 1. P. 26219–26237.
 13. *Schegloff E., Sacks H.* Opening up closings // *Semiotica*. 1973. Vol. 8, No. 4. P. 289–327.
 14. *Klokoza K., Krongauz M., Shulginov V., Yudina T.* Towards a russian multimedia politeness corpus. // Proc. of the International Conference “Dialogue 2023”. 2023. P. 233–244.
 15. *Terkourafi M.* An argument for a frame-based approach to politeness: Evidence from the use of the imperative in Cypriot Greek // *Pragmatics and beyond*. New series. 2005. Vol. 139. P. 99–116.
 16. *Laver J.* Communicative Functions of Phatic Communion // *Organization of Behavior in Face-to-face Interaction*. 1975. P. 215–238.
 17. *Laver J.* Linguistic Routines and Politeness in Greeting and Parting // *Conversational routines*. 1981. P. 289–305.
 18. *Brown T.B. et al.* Language models are fewshot learners // *arXiv:2005.14165*. 2020. <https://doi.org/10.48550/arXiv.2005.14165>
 19. *Winkler W. E.* String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage // Proc. of the Section on Survey Research. 1990. P. 354–359.
 20. *Cahyono S.C.* Comparison of document similarity measurements in scientific writing using jaro-winkler distance method and paragraph vector method // *IOP Conference Series: Materials Science and Engineering*. 2019. Vol. 662, No. 5. <https://doi.org/10.1088/1757-899X/662/5/052016>
-

ПРИЛОЖЕНИЕ А. ПРИМЕРЫ РАЗМЕЧЕННЫХ ДИАЛОГОВ И ПОСЛЕДОВАТЕЛЬНОСТЕЙ, ИДЕНТИФИЦИРОВАННЫХ МОДЕЛЯМИ

(1) Приветствие

Катя входит в комнату и включает свет

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

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

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

(2) Прощание

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

Максим Васильев: Тебе идти надо?

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

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

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

(3) Приветствие + прощание

У Саяпина звонит телефон

разговор по телефону

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

Подчиненный Саяпина: Ну что, и куда теперь?

...

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

Подчиненный Саяпина: Угу.

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

разговор по телефону завершён

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

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

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

K. S. Klokova¹ [0000-0002-6650-1946], M. A. Krongauz² [0000-0003-2956-0565],

V. A. Shulginov³ [0000-0003-2217-3677], T. A. Yudina⁴ [0000-0001-6631-2128]

^{1, 3, 4}*Moscow Institute of Physics and Technology, Dolgoprudny, Russia*

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

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

Abstract

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

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

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

REFERENCES

1. *Brown P., Levinson S.C.* Politeness: Some Universals in Language. Cambridge University Press, 1987.
2. *Kádár D.Z., Haugh M.* Understanding politeness. Cambridge University Press, 2013.
3. *Schegloff E.A.* Sequence Organization in Interaction: A Primer in Conversation Analysis I. Cambridge University Press, 2007.
4. *Ruis L. et al.* The goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by LLMs // arXiv:2210.14986. 2024. <https://doi.org/10.48550/arXiv.2210.14986>
5. *Jian M., Narayanaswamy S.* Are LLMs good pragmatic speakers? // arXiv:2411.01562. 2024. <https://doi.org/10.48550/arXiv.2411.01562>
6. *Hu J., Floyd S., Jouravlev O., Fedorenko E., Gibson E.* A fine-grained comparison of pragmatic language understanding in humans and language models // Proc. of the 61st Annual Meeting of the Association for Computational Linguistics. 2023. Vol. 1. P. 4194–4213.
7. *Lipkin B., Wong L., Grand G., Tenenbaum J.B.* Evaluating statistical language models as pragmatic reasoners. // arXiv:2305.01020. 2023. <https://doi.org/10.48550/arXiv.2305.01020>
8. *Iida A., Okuoka K., Omori T., Nakashima R., Osawa M.* LLM-based evaluation of utterances with implicature understanding: A preliminary study // Proc. of the 13th International Conference on Human-Agent Interaction. 2025. P. 485–487. <https://doi.org/10.1145/3765766.3765841>
9. *Yerukola A., Vaduguru S., Fried D., Sap M.* Is the pope catholic? Yes, the pope is catholic. Generative evaluation of non-literal intent resolution in LLMs // Proc. of the 62nd Annual Meeting of the Association for Computational Linguistics. 2024. Vol. 2. P. 265–275.
10. *Liu R., Sumers T. R., Dasgupta I., Griffiths T.L.* How do large language models navigate conflicts between honesty and helpfulness? // Proc. of the 41st International Conference on Machine Learning. 2024. P. 31844–31865.

11. *Chen Z., Yang R., Zhao Z., Cai D., He X.* Dialogue act recognition via CRF-at-
tentive structured network // Proc. of the 41st International ACM SIGIR Conference
on Research & Development in Information Retrieval. 2018. P. 225–234.
12. *Qamar A., Tong J., Huang R.* Do LLMs understand dialogues? A case study on
dialogue acts. // Proc. of the 63rd Annual Meeting of the Association for Computa-
tional Linguistics. 2025. Vol. 1. P. 26219–26237.
13. *Schegloff E., Sacks H.* Opening up closings // *Semiotica*. 1973. Vol. 8, No. 4.
P. 289–327.
14. *Klokoval K., Krongauz M., Shulginov V., Yudina T.* Towards a russian multime-
dia politeness corpus. // Proc. of the International Conference “Dialogue 2023”. 2023.
P. 233–244.
15. *Terkourafi M.* An argument for a frame-based approach to politeness: Evi-
dence from the use of the imperative in Cypriot Greek // *Pragmatics and beyond*. New
series. 2005. Vol. 139. P. 99–116.
16. *Laver J.* Communicative Functions of Phatic Communion // *Organisation of
Behaviour in Face-to-face Interaction*. 1975. P. 215–238.
17. *Laver J.* Linguistic Routines and Politeness in Greeting and Parting // *Conver-
sational routines*. 1981. P. 289–305.
18. *Brown T.B. et al.* Language models are few-shot learners // arXiv:2005.14165.
2020. <https://doi.org/10.48550/arXiv.2005.14165>
19. *Winkler W. E.* String comparator metrics and enhanced decision rules in the
felligi-sunter model of record linkage // Proc. of the Section on Survey Research.
1990. P. 354–359.
20. *Cahyono S.C.* Comparison of document similarity measurements in scientific
writing using jaro-winkler distance method and paragraph vector method // *IOP Con-
ference Series: Materials Science and Engineering*. 2019. Vol. 662, No. 5.
<https://doi.org/10.1088/1757-899X/662/5/052016>

СВЕДЕНИЯ ОБ АВТОРАХ



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

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

email: klokovala.ks@mipt.ru

ORCID: 0000-0002-6650-1946



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

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

email: mkronhaus@yandex.ru

ORCID: 0000-0003-2956-0565



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

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

email: shulginov.va@mipt.ru

ORCID:0000-0003-2217-3677



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

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

email: yudina.tatiana.a@mipt.ru

ORCID: 0000-0001-6631-2128

Материал поступил в редакцию 20 августа 2026 года