Научная статья
УДК 159.9.07
https://doi.org/10.21702/rpj.2026.2.4

Большие языковые модели как инструмент прогнозирования типов поведения личности

Александр А. Агафонов, Ахат Р. Хусаинов, Павел Н. Устин*, Леонид М. Попов

Казанский (Приволжский) федеральный университет, Казань, Российская Федерация

Почта ответственного автора: pavust@mail.ru

Аннотация

Введение. Интенсивное развитие инструментов и технологий искусственного интеллекта открывает широкие междисциплинарные возможности для исследования различных типов поведения личности. К таким инструментам относятся большие языковые модели (LLM) – нейронные сети глубокого обучения со множеством параметров. Одним из перспективных вариантов их использования выступает прогнозирование поведения человека через интерпретацию содержания текстовых сообщений пользователей различных онлайн платформ в цифровой среде.

Методы. Исследование осуществлено на выборке, включающей 541 текстовое сообщение, которые были извлечены из онлайн-платформ. Тексты были размечены экспертами-психологами на основе стандартизированных психолингвистических критериев по трем типам поведения: деструктивное, конструктивное, информативное. В качестве инструмента интерпретации текстов был выбрана большая языковая модель «Claude 3.5», адаптированная под задачи многоклассовой классификации через конфигурируемые запросы. Для автоматической классификации текстов был применен API модели Claude 3.5 с базовым запросом, формализующим критерии экспертов-психологов.

Результаты. Впервые предложена технология прогнозирования (интерпретации) конструктивного, деструктивного и информативного поведения пользователей социальных сетей с точностью классификации в пределах 62-76%. Прогнозирование осуществляется через квалитативный анализ текстов любого объема на базе больших языковых моделей. Технология разработана на базе междисциплинарного

взаимодействия психологов, IT-специалистов и систем искусственного интеллекта с использованием возможностей больших языковых моделей к рефлексии через итеративный диалог.

Обсуждение результатов. Технология расширяет возможности психологов прогнозировать поведение на выборках, включающих миллионы пользователей цифровой среды и показывает векторы разработки инструментов интерпретации различных психологических феноменов через анализ текстов. Также показана необходимость дальнейшей адаптации технологии под культурные и лингвистические особенности для минимизации ошибок и противоречий в классификации.

Ключевые слова

большие языковые модели, цифровое поведение, прогнозирование поведения, цифровая среда, рефлексия больших языковых моделей, интерпретируемый искусственный интеллект, цифровая коммуникация, онлайн-платформы

Финансирование

Исследование выполнено за счет гранта Российского научного фонда № 24-18-01125, «Психологическая модель поведения субъекта в цифровой среде», https://rscf.ru/project/24-18-01125/

Для цитирования

Агафонов А.А., Хусаинов А.Р., Устин П.Н, Попов Л.М. (2026). Большие языковые модели как инструмент прогнозирования типов поведения личности. *Российский психологический журнал, 23*(2), 54–73, https://doi.org/10.21702/rpj.2026.2.4

Введение

Цифровая среда, как неотъемлемый атрибут современной жизни, выступает одним из значимых и информативных ресурсов для анализа и прогнозирования поведения личности. Разнообразные онлайн-платформы ежедневно фиксируют миллионы текстовых коммуникаций пользователей социальных сетей, которые отражают эмоции, ценности и социальные динамики конструктивного и деструктивного характера (Ермаков и др., 2022). Современные большие языковые модели (LLM), такие как Claude-3, GPT-4 и YandexGPT, начинают активно применяться для анализа качественных данных, включая интервью, открытые опросы и публичные дискуссии (Kublik & Saboo, 2022; Markowitz, 2024; Ke et al., 2024). Их эмерджентные способности – качественные скачки в решении задач, которые отсутствуют у меньших моделей (Wei et al., 2022) – делают

их перспективным инструментом анализа и прогнозирования поведения личности в психологических исследованиях. Однако эта задача сталкивается с двумя серьезными проблемами. Во-первых, *контекстная зависимость*: одно и то же высказывание может нести разный прагматический смысл. Например, фраза «Власти бездействуют» может быть деструктивной (если цель – нагнетание паники) или конструктивной (если далее следует предложение создать общественный совет), что требует глубокого анализа контекста (Linegar et al., 2023). Во-вторых, *интерпретируемость*: сложные модели машинного обучения, включая LLM, часто работают как «черные ящики» (Dillion et al., 2023; Bender et al., 2021), что ограничивает их применение в психологических исследованиях, где объяснимость решений существенно важна.

Данное исследование предлагает подход, преодолевающий эти ограничения через *рефлексию LLM* – способность моделей анализировать собственные ошибки и корректировать критерии классификации. В отличие от традиционных способов, требующих ручного кодирования и триангуляции (Turner & Reynolds, 2012), наша технология автоматизирует процесс, сохраняя прозрачность решений. Современные LLM, основанные на архитектуре трансформеров с механизмами self-attention (Vaswani et al., 2017), способны не только предсказывать классы поведения, но и генерировать объяснения своих решений через интерактивный *промптинг (инструктирование или запросы данных при взаимодействии с нейросетью)*. Например, запрос к модели может включать два этапа: определение типа поведения в тексте и объяснение лингвистических маркеров, повлиявших на выбор. Это позволяет выявлять ключевые паттерны, такие как использование слов «предлагаю» для конструктивности или «запретить» для деструктивности, и корректировать критерии через итеративный диалог с моделью, аналогично подходам в симуляции человеческого доверия (Xie et al., 2024).

Пример 1 «Двухэтапный запрос к большой языковой модели»

«1. Определи тип поведения текста: деструктивный, конструктивный, информативный.

2. Объясни, какие фразы повлияли на твой выбор.

Текст: [Пример]»

Методологической основой исследования выступает типологический анализ, предполагающий поиск общих признаков в данных через анализ сочетаний существенных характеристик (Татарова и др., 2023). Данный подход согласуется с теорией коммуникативного действия (Habermas, 1985), разделяющей поведение на стратегическое (манипулятивное) и консенсусное (конструктивное), а также с теорией временных масштабов Ньюэлла (Newell, 1994), которая связывает уровни поведения (от биологического до социального) с возможностями LLM имитировать когнитивные процессы (Abdurahman et al., 2024; Almeida et al., 2024; Binz & Schulz,

2023; Hutson, 2023; Ke et al., 2024; Parra, 2024). Большие языковые модели все чаще используются в психологических исследованиях (Sartori & Orrù, 2023; Stokel-Walker, 2022; Rathje et al., 2024). Ранее применение LLM в социальных науках, например, в идентификации популистских нарративов, демонстрировало высокую точность, но не объясняло логику принятия решений. Предлагаемая технология устраняет этот пробел, используя *самооптимизацию запросов*: модель анализирует ошибки классификации, уточняет определения классов и добавляет примеры-эталоны, что повышает точность и адаптивность (Abramski et al., 2023; Liu et al., 2023; Park et al., 2024).

Культурная адаптивность подхода проявляется в учете языковых норм, таких как русскоязычная ирония, через контекстные пояснения в запросах. Например, фраза «Давайте не будем спешить» в русскоязычном контексте часто сигнализирует о конструктивном диалоге, но может быть ошибочно классифицирована как нейтральная из-за отсутствия явных эмоциональных маркеров. Этот аспект согласуется с исследованиями культурных различий в мышлении LLM (Jin et al., 2023). Этическая рамка исследования включает открытое обсуждение ограничений LLM, таких как риски предвзятости (Hagendorff, 2023), необходимость верификации результатов экспертами (Dehouche, 2021; Harding et al., 2023) и опасность антропоморфизации моделей (Stella et al., 2023). Например, анализ религиозных текстов позволяет выявлять радикальные формы поведения авторов, демонстрирующих призывы под видом догматов, а изучение эволюции дискурса в онлайн-общинах – фиксировать сдвиги от конструктивного диалога к поляризации (Demszky et al., 2023; Zhong et al., 2023).

Практическая значимость предлагаемого подхода связана с возможностями его примененияя в междисциплинарных исследованиях. В социальной психологии LLM помогают анализировать групповую динамику и культурные различия (Grossmann et al., 2023; Ziems et al., 2023), в психологической практике – распознавать эмоциональные состояния (Patel & Fan, 2023), а в образовании – разрабатывать персонализированные обучающие стратегии (Frank, 2023). Интеграция рефлексии LLM в классификацию поведения не только снижает зависимость от ручного кодирования, но и задает новые стандарты прозрачности, сочетая преимущества искусственного интеллекта с глубиной гуманитарного анализа (Tamkin et al., 2021; Ke et al., 2024).

Методы

Тексты, выступая продуктами психической активности субъекта, в силу своей материальной природы открывают возможности проведения объективного анализа (Кузнецова, 2018). Опираясь на положения современной когнитивистики, мы, вслед за утверждениями ряда авторов (Мишланов и др., 2020) также придерживаемся тех позиций, что одним из отражений интенциональности поведения человека является его речевая активность. То есть тексты, включающие коммуникативные интенции, жанровую форму и содержание, наиболее эксплицитно воплощают заключенную в

них программу поведения человека. Следовательно, текст, порождаемый субъектом в связи с возможными сценариями его поведения, является одним из источников прогнозирования его действий.

В рамках исследования была собрана выборка из 541 текстовых сообщений, извлеченных из открытых источников интернета (публичные форумы, социальные сети, блоги). Каждый текст предварительно был размечен командой профессиональных психологов по трем типам поведения: *деструктивное* (агрессия, провокации, токсичность), *конструктивное* (эмпатия, сотрудничество, решение проблем), *информативное* (нейтральная передача фактов без эмоциональной окраски). Разметка была основана на стандартизированных психолингвистических критериях, включая анализ тональности, частотности ключевых маркеров (напр., императивы, оценочные прилагательные) и контекстуальной согласованности. Ниже приведены примеры текстов, отражающих типы поведения.

Пример 2 «Текст, отражающий деструктивный тип поведения»

«Подмосковье, таким образом, стало первым регионом, где местные власти сами приняли такое решение, не дожидаясь, пока аналогичный документ будет выработан на федеральном уровне. Член Совета при президенте РФ по межнациональным отношениям Владимир Волох и вовсе надеется, что опыт Московской области начнут перенимать и другие регионы. Надо понимать, что указанная инициатива принадлежит губернатору Воробьёву и мэру столицы Собянину как ответ на отключения зимой тепла, электричества и воды? Чтоб следующей зимой было еще больше аварий ЖКХ из-за перегруза? Пахнет какой-то диверсией...

Идея строительства кварталов для приезжих не нова - в своё время во Франции подобные кварталы были построены во многих крупнейших городах, они же стали и ядром формирующихся гетто, ныне превративших Пятую республику в арабско-африканский бандустан. Наибольшее количество таких кварталов было построено в Марселе, где на данный момент 40% населения (в основном проживающего в пригородах) составляют мигранты или их потомки, большинство из них не работает и живут на выплачиваемую социалку (у нас вместо социалки чужакам выплачивают маткапитал и детские выплаты). Сам Марсель в СМИ именуется криминальной столицей Европы, где преступлений на 1000 человек населения совершается в 4.5 раза больше, чем в среднем по всей, погибающей из-за нашествия (по протекции Макрона) мигрантов, Франции. Подобную судьбу москвичам готовят нынешние региональные власти московского региона». (https://vk.com/wall-54012242_2789421)

Пример 3 «Текст, отражающий информативный тип поведения»

«Национальный стандарт Р 56939 Защита информации. Разработка безопасного программного обеспечения» 2016 года уже признали устаревшим и на замену ему

разработали проект нового ГОСТа Документ опубликован на сайте Федеральной службы по техническому и экспортному контролю (ФСТЭК), его публичное обсуждение продлится до 15 февраля 2024 года.

С 2016 года сменился набор языков программирования. Сейчас превалирующие языки совсем другие, чем были семь лет назад. Сменился и набор технологий, например, повысилась доля опенсорса с открытым исходным кодом. В таком ПО больше «дырок» для хакеров, что увеличивает долю рисков, которые несколько лет назад были не так актуальны

Документ, находящийся сейчас на обсуждении, предназначен для разработчиков и производителей софта, а также для организаций, контролирующих соответствие результатов их деятельности национальному стандарту. Как говорится в пояснительной записке к проекту, принятие нового ГОСТа позволит повысить эффективность и качество разработки безопасного программного обеспечения». (https://pikabu.ru/story/otechestvennyiy_soft_zashchityat_po_novomu_gostu_11009643)

Пример 4 «Текст, отражающий конструктивный тип поведения»

«В России законодательно «продлили» молодежный возраст до 35 лет, в Беларуси он официально – до 31-го. Опрос тут в рунете провели ко Дню молодежи. Спрашивали, до какого возраста россияне чувствуют себя молодыми. Почти 27 тысяч человек приняли участие, 86% опрошенных – люди старше 35 лет. Так вот, треть из них считают, что каждый человек вправе сам решать – причислять себя к молодежи или нет. Независимо от того, какой возраст указан у него в паспорте.

И тут же мнение 39-летней женщины приводят, которая считает День молодежи и своим праздником, потому что «выглядит на 10 лет моложе», и еще одной, которая в 52 тоже молода и обижается на знакомую за ссылку на статью, как надо одеваться в 50+, чтобы не выглядеть смешной. Неприятны такие «толстые намеки» вечным девочкам. Они хотят, чтобы в 55 их принимали за 25-летних и это правда смешно.

Мне вот совсем не хотелось быть частью молодежи в 40 лет. Потому что это прекрасный возраст, идеальный просто. Дети почти выросли, а ты еще полон сил, но уже имеешь опыт. Такие краски жизни открываешь для себя, которые в молодости просто невозможно заметить, потому что бежишь как наскипидаренный. А в 40 неинтересно уже просто бежать, хочется понять куда, зачем, кого взять с собой. Ты начинаешь говорить «нет» лишним людям, меняешь привычки, понимаешь, что можно жить и работать только с теми, с кем тебе действительно интересно, ты многому научился и снова возвращаешься к самому себе...». (https://dzen.ru/a/YNt_tLsyHgFXgK25)

Для решения задачи классификации поведения были протестированы 5 крупных языковых моделей (LLM), доступных через публичные API: ChatGPT 4.0 (OpenAI), Gemini (Google), Claude 3.5 (Anthropic), YandexGPT (Яндекс), GigaChat (Сбер). Были выявлены

ключевые ограничения: GigaChat и YandexGPT продемонстрировали системные отказы в обработке текстов с политическим контекстом, что связано с внутренними фильтрами безопасности; YandexGPT оказался непригодным для длинных текстов из-за ограничения на 4,000 токенов (≈3,000 символов), тогда как 27% данных выборки превышали этот лимит; Gemini был исключен из-за возникших сложностей в получении доступа к платному API. После исключения неподходящих вариантов, Claude 3.5 (Anthropic) и ChatGPT 4.0 показали сопоставимую точность на тестовой выборке. Решающим фактором стала экономическая эффективность: средняя стоимость обработки одного текста в Claude 3.5 составила $0.006, а ChatGPT 4.0 – $0.026 (разница в 4.3 раза). Для минимизации затрат без потери качества финальным инструментом был выбран *Claude 3.5*, адаптированный под задачи многоклассовой классификации через кастомные запросы.

Для автоматической классификации текстов был применен API модели Claude 3.5 с базовым запросом, формализующим критерии экспертов-психологов (Пример 5).

Пример 5 «Запрос к большой языковой модели с критериями экспертов-психологов»

«Поведение автора текста определяется тем, как текст влияет на читателя. Есть три типа поведения:

- Деструктивный: неконструктивная критика, призывы к разрушительным действиям, оскорбления, нагнетание паники.

- Информативный: аналитика, статистика, отсутствие призывов к действиям.

- Конструктивный: формирование спокойного отношения к событиям, созидательные призывы, успокаивающая лексика или ирония.

Определи тип поведения для текста ниже и ответь одним словом: [ТЕКСТ]»

Таблица 1

Результаты первичной классификации (матрица неточностей)

Экспертная разметка	Информативный (0)	Деструктивный (1)	Конструктивный (2)
Предсказания LLM:			
Информативный (0)	168	28	113
Деструктивный (1)	10	144	28
Конструктивный (2)	3	8	39

Точность (Accuracy) по классам: информативный – 93.3%, деструктивный – 80.0%, конструктивный – 21.7%. Низкое качество классификации конструктивного поведения объясняется следующими факторами:

1. Семантическая неоднозначность: конструктивные тексты часто содержат элементы аналитики (признак информативности) или иронию, которую модель интерпретирует как негативную (признак деструктивности).
2. Культурно-лингвистические особенности: успокаивающая лексика на русском языке (напр., «не спешите», «взвесим варианты») слабо различима для LLM, обученных на мультиязычных корпусах.
3. Ограничения запроса: например, ирония, указанная в критериях конструктивности, часто трактовалась моделью как сарказм (деструктивный класс).

Результаты исследования

Корректировка индивидуальных запросов методом рефлексии LLM

Для устранения низкой точности распознавания конструктивных текстов (21.7%) был применен метод рефлексии LLM. Алгоритм включал:

1. Анализ ошибок: модели предоставлялся исходный текст, её ошибочный ответ и корректная экспертная метка.
2. Генерация улучшенных запросов: Claude 3.5 получал задачу скорректировать исходный запрос, чтобы избежать повторения ошибки (табл. 2).
3. Повторная классификация: модифицированные запросы применялись для повторной обработки текстов, при этом результаты сравнивались с экспертными разметками для оценки эффективности изменений.

Таблица 2
Примеры коррекции критериев конструктивности

Исходный критерий	Модифицированный критерий
«...в отдельных случаях ирония»	Удалено из-за двусмысленности
«призывающие к созидательным действиям»	Дополнено: «предлагающие решения с примерами: «Предлагаю создать рабочую группу»
«формирование спокойного отношения»	Уточнено: «нейтральный тон с фокусировкой на диалог, напр.: "Давайте обсудим варианты"»

Результаты показали значительное улучшение точности классификации после внедрения метода рефлексии LLM. Финальные показатели демонстрируют наибольший прирост в категории конструктивного поведения, где точность возросла с исходных 21.7% до 71.1% (табл. 3).

Таблица 3

Динамика изменения точности по классам в ходе оптимизации запросов

Класс поведения	Исходный запрос	После оптимизации
Информативный	93.3%	97.2%
Деструктивный	80.0%	96.7%
Конструктивный	21.7%	71.1%
Средняя точность	65.0%	88.3%

В процессе рефлексии LLM были выявлены ключевые лингвистические маркеры, характерные для каждого типа поведения. Интересным эффектом итеративной оптимизации запросов стало обнаружение феномена *«смещения точности»*. При целенаправленном улучшении классификации конструктивного поведения (точность повысилась до 71.1%) наблюдался небольшой, но статистически значимый спад в классификации информативных текстов. Этот феномен можно объяснить размыванием границ между категориями при повышении специфичности критериев. Например, добавление в запрос критерия «предложение конкретных решений» для конструктивного класса привело к тому, что некоторые информативные тексты с перечислением возможных вариантов действий («можно сделать X, Y или Z») стали ошибочно классифицироваться как конструктивные.

Метод продемонстрировал способность языковых моделей (LLM) к итеративной самокоррекции за счет выявления ключевых аспектов, требующих внимания при классификации конкретных текстов. Важно отметить, что корректировка запросов осуществлялась индивидуально для каждого текста с учетом экспертной разметки. Однако основная методологическая задача заключалась в последовательной оптимизации универсального базового запроса таким образом, чтобы он обеспечивал максимальную точность классификации на всей совокупности текстов без необходимости индивидуальной адаптации под каждый конкретный случай.

Автоматическое обобщение запросов

Для унификации индивидуальных корректировок конструктивного поведения был применен *итеративный алгоритм слияния запросов*, заключающийся в попарном объединении скорректированных запросов к Claude 3.5 (Пример 6):

Пример 6 «Использование итеративного алгоритма слияния запросов»

«Объедини два запроса, удалив повторы и сохранив ключевые критерии:

1) [Промпт 1]

2) [Промпт 2]»

Результирующий запрос содержал расширенную таксономию маркеров для каждого типа поведения с уточненными дефинициями. Применение автоматически обобщенного запроса к набору текстов дало результаты, приведенные в табл. 4.

Таблица 4

Результаты повторной классификации текстов с помощью автоматически обобщенного запроса

Экспертная разметка	Информативный (0)	Деструктивный (1)	Конструктивный (2)
Предсказания LLM:			
Информативный (0)	133	3	42
Деструктивный (1)	37	169	67
Конструктивный (2)	11	8	71

Точность по классам: информативный –73.5% (Δ -23.8% от предыдущего этапа), деструктивный – 93.9% (Δ +0.1%), конструктивный – 39.4% (Δ -43.4%).

Наблюдается значительная регрессия точности для классов информативного и конструктивного поведения по сравнению с предыдущим этапом оптимизации. Такой результат противоречит ожиданиям, согласно которым агрегация индивидуальных улучшений должна была привести к повышению общей точности.

Анализ ошибок выявил ключевые проблемы автоматического обобщения:

1. Конфликт критериев: включение «аналитики с авторской позицией» в конструктивный класс привело к путанице с информативным классом.
2. Семантическое перекрытие: критерий «использование разговорного стиля» для конструктивности совпал с деструктивными текстами, содержащими сленг.
3. Избыточная детализация: добавленные уточнения сузили охват деструктивного класса, повысив его точность за счет других категорий. Это создало «узкие места» в классификации, где тексты, не подпадающие под детализированные критерии, были ошибочно отнесены к другим классам.

4. Смещение в сторону детекции деструктивности: обобщенный запрос приобрел уклон в сторону идентификации деструктивного поведения (93.9% точности), что привело к «перехвату» текстов из других категорий. Количество текстов, ошибочно классифицированных как деструктивные, увеличилось на 42% по сравнению с предыдущим этапом.

Таким образом, исследование определило ключевое ограничение метода: автоматическое обобщение не столько интегрирует улучшения, сколько размывает исходную концептуальную таксономию.

Разработка маркерного подхода к классификации

После выявления ограничений автоматического обобщения запросов был разработан альтернативный *маркерный подход*, основанный на точной идентификации лингвистических индикаторов каждого типа поведения. Для построения такой модели был применен итеративный алгоритм, включающий: выделение наиболее часто встречающихся фраз и конструкций для каждого класса; запрос к LLM для группировки семантически близких маркеров; создание специализированного запроса с инструкцией по применению весовой схемы.

Ключевым отличием данного подхода является *квантификация лингвистических признаков* – вместо целостного семантического анализа модель фокусируется на обнаружении и подсчете конкретных маркеров в тексте. Применение маркерного подхода к набору текстов дало результаты, приведенные в табл. 5.

Таблица 5

Результат классификации текстов после применения маркерного подхода

Экспертная разметка	Информативный (0)	Деструктивный (1)	Конструктивный (2)
Предсказания LLM:			
Информативный (0)	104	0	18
Деструктивный (1)	46	168	74
Конструктивный (2)	31	12	88

Точность по классам: информативный – 57.5% (Δ -35.8% от максимального результата), деструктивный – 93.3% (Δ -0.6%), конструктивный – 48.9% (Δ +27.2% от исходного значения).

Результаты демонстрируют противоречивую динамику: значительное улучшение классификации конструктивного поведения по сравнению с исходной точностью

(21.7% → 48.9%), но при этом существенное снижение точности для информативного класса (93.3% → 57.5%).

Анализ ошибок классификации выявил системные проблемы маркерного подхода:

1. Усиление конструктивных критериев: включение маркера «стимулирование к размышлению» привело к массовой классификации философских и рефлексивных текстов как конструктивных.
2. Сужение критериев информативности: требование «полного отсутствия оценочных суждений» исключило из информативного класса значительное количество научно-популярных текстов.
3. Культурные артефакты в классификации: маркер «описание технологических улучшений» привел к систематической ошибке классификации рекламных и маркетинговых текстов как конструктивных.
4. Эффект «перетекания» текстов в деструктивный класс: деструктивные маркеры имеют наибольшую дискриминативную силу, что привело к смещению классификации в сторону обнаружения негативных паттернов даже в нейтральных текстах. 46 информативных текстов были ошибочно классифицированы как деструктивные из-за наличия единичных критических замечаний.

Интересно отметить, что автоматизированные методы (обобщение и маркерный подход) продемонстрировали схожие средние результаты (68.9% и 66.6% соответственно), но с существенно различной структурой ошибок: обобщение запросов максимизирует точность деструктивного класса за счет информативного, маркерный подход повышает точность конструктивного класса за счет информативного. Это указывает на проблему в определении понятия «информативное поведение», границы которого оказываются наиболее размытыми в процессе автоматизированной оптимизации.

Таргетированная оптимизация запроса для конструктивного типа поведения через рефлексивную коррекцию

Анализ результатов предшествующих этапов исследования выявил, что одной из ключевых проблем классификации является семантическое пересечение между конструктивным и информативным типами поведения. В целях устранения данной проблемы был реализован контекстно-ориентированный подход, акцентирующий внимание на систематических ошибках классификации. Подобный методологический выбор позволил сфокусироваться на уточнении критериев разграничения указанных классов с учетом их контекстуальной специфики.

Была выделена подгруппа из 113 текстов, которые были ошибочно классифицированы как информативные, но получили экспертную разметку как конструктивные. Каждый текст был подвергнут детальному анализу с привлечением LLM для объяснения причин ошибочной классификации. Далее LLM было предложено сформулировать правки для улучшения классификации конструктивных текстов, которые были агрегированы в иерархический список критериев.

На основе анализа ошибок был разработан уточненный критерий для идентификации конструктивного типа поведения. Ключевые изменения в критериях включают: *расширение понятия конструктивности* (включение неявных призывов к действию через модальные конструкции), *признание аналитических текстов* с авторской позицией как потенциально конструктивных, *учет стилистического разнообразия* (распознавание иронии и юмора как инструментов снижения социальной напряженности, учет культурно-специфичных форм выражения конструктивности), *структурный анализ* (выявление конструктивных паттернов в композиции текста, учет баланса между критическими и позитивными элементами сообщения). В результате был получен обобщенный запрос (Пример 7):

Пример 7 «Обобщенный запрос к большой языковой модели»

«Поведение автора текста определяется тем, как текст влияет на читателя. Есть три типа поведения: деструктивный, конструктивный, информативный.

Деструктивный тип: тексты, содержащие неконструктивную критику, явно и неявно призывающие к необдуманным разрушительным или незаконным действиям, оскорбляющие социальные группы, нагнетающие панику и беспокойство. Включают пропаганду дискриминации, ограничения прав и свобод. Тексты с негативной оценкой действий властей без предложения решений. Могут использовать сарказм, иронию, вызывающие недоверие к властям и институтам. Тексты с негативной эмоциональной окраской, выражающие раздражение. Включают прямые призывы к насилию, агрессивные высказывания, дезинформацию, манипулятивные техники, искажающие факты, неэтичные высказывания. Могут использовать провокационные заголовки или выражения. Также сюда относятся тексты, подрывающие доверие к властям, содержащие намеки на нелояльность региональных элит, предполагающие возможность государственного переворота или распада страны. Сюда также относятся тексты, содержащие радикальные взгляды, односторонние обвинения и агрессивную риторику в отношении определенных групп людей, даже если они преподносятся под видом борьбы за справедливость. Тексты, выражающие сильные негативные эмоции или описывающие негативный личный опыт без конструктивных выводов. Также включают грубую лексику, оскорбления, уничижительные высказывания о людях или группах людей, пропаганду негативных стереотипов. Информативный тип: тексты, содержащие преимущественно фактическую информацию, объективную аналитику, статистические данные, описание процессов или явлений, исторические факты без эмоциональной окраски, субъективных оценок и призывов к действиям. НЕ выражают личное мнение автора или явную авторскую позицию. Конструктивный тип: тексты, направленные на формирование спокойного или позитивного отношения к событиям или людям, призывающие к созидательным действиям, содержащие успокаивающе-ободряющую лексику, одобрительные оценки, похвалу, уважение, критику с предложением решений, иронию или юмор. Могут включать личный опыт автора,

аргументированные суждения, практические советы, направленные на помощь другим людям. Также аналитические тексты, предлагающие конкретные решения проблем и новые интерпретации фактов. Поддерживают равенство и уважение ко всем людям. Вдохновляют на позитивные изменения без агрессии или раздражения. Такие тексты могут сочетать информативность с позитивным настроем и стремлением помочь читателю, а также иметь четко выраженную авторскую позицию. Могут включать аналитику и статистику, но подаются в более эмоциональной и личной манере, часто с использованием разговорного стиля. Способствуют личностному росту и самопознанию, поощряют творчество и саморазвитие. Определи тип поведения для следующего текста и ответь одним словом: [ТЕКСТ]».

Внедрение таргетированной оптимизации конструктивного типа поведения привело к существенному изменению в распределении классификаций, отраженному в табл. 6.

Таблица 6

Результат классификации текстов после внедрения таргетированной оптимизации запроса для конструктивного типа поведения

Экспертная разметка	Информативный (0)	Деструктивный (1)	Конструктивный (2)
Предсказания LLM:			
Информативный (0)	112	2	25
Деструктивный (1)	12	137	27
Конструктивный (2)	57	41	128

Точность по классам: информативный – 61.9% (Δ +4.4% от предыдущего этапа), деструктивный – 76.1% (Δ -17.2%), конструктивный – 71.1% (Δ +22.2%).

Наблюдается значительное улучшение точности определения конструктивного класса (+22.2% от предыдущего этапа и +49.4% от исходного уровня), что свидетельствует об эффективности таргетированной оптимизации. Однако это улучшение сопровождается заметным снижением точности для деструктивного класса (-17.2%), что указывает на компромиссный характер оптимизации.

Анализ новых ошибок классификации выявил несколько причин наблюдаемого дисбаланса:

1. Перекрытие критериев: расширенное определение конструктивности «поглотило» 19% деструктивных текстов с завуалированной критикой. Формальное присутствие предложений по решению проблем в деструктивных текстах приводило к их ошибочной классификации как конструктивных.
2. Культурная специфика: тексты с патриотическими маркерами часто ошибочно классифицировались как конструктивные, эмоционально насыщенные выражения, характерные для русскоязычного дискурса, снижали точность разграничения конструктивных и деструктивных текстов.
3. Эффект компенсации: рост точности конструктивного класса на 49.4 процентных пункта от исходного уровня достигнут за счет снижения точности других категорий. Перераспределение текстов между классами привело к «перетеканию» части информативных текстов в конструктивный класс (57 случаев).
4. Увеличение ложноположительных срабатываний: расширение критериев конструктивности привело к росту числа ложноположительных результатов, особенно для текстов с манипулятивной риторикой, замаскированной под конструктивные предложения.

Обсуждение результатов

Полученные результаты демонстрируют, что метод рефлексивной оптимизации запросов для LLM открывает новые возможности в создании интерпретируемых систем анализа текста, что особенно актуально как для психологии, так и в целом, для социально-гуманитарных наук. В отличие от традиционных подходов, где улучшение классификаторов происходит через математическую оптимизацию, предложенный метод обеспечивает три ключевые особенности. Во-первых, *критерии классификации выражены на естественном языке*, что позволяет экспертам непосредственно проверять их корректность. Это соответствует требованиям прозрачности в AI-технологиях, подчеркнутым в работах (Dillion et al., 2023; Bender et al., 2021), где акцентируется важность первоисточников для верификации результатов. Во-вторых, *адаптивность к контексту* через модификацию запросов решает проблему «одноразовой» настройки моделей, характерную для традиционных методов. Например, это позволяет анализировать юридические документы с учётом их специфики, как продемонстрировано в исследованиях, где LLM адаптируются к узким доменам через контекстно-зависимые запросы (Savelka & Ashley, 2023). В-третьих, *формализация экспертных определений* в запросах выявляет их противоречия, что важно для гуманитарных наук, где концепции часто носят плюралистический характер. Это соответствует рекомендациям (Hagendorff, 2023), где подчеркивается необходимость не только точности, но и объяснимости в принятии решений.

Практическая значимость предлагаемого метода для психологии проявляется в возможности выявления конструктивных и деструктивных элементов поведения

авторов разнообразных текстов, что может быть актуально, например, для анализа экстремистских материалов (Zhong et al., 2023). Кроме того, он позволяет идентифицировать коммуникативные паттерны поведения в межконфессиональных диалогах пользователей социальных сетей, где ключевую роль играет контекстуальная адаптация (Demszky et al., 2023), а также отслеживать эволюцию коммуникативных стратегий различных сообществ, например, переход от конфликтного дискурса к конструктивному (Grossmann et al., 2023).

Однако метод имеет ограничения, требующие дальнейшего уточнения. Во-первых, необходимо учесть влияние субъективности экспертов при формулировке запросов (Dehouche, 2021). Во-вторых, следует провести сравнительный анализ с методами глубокого обучения, как это предложено в (Ke et al., 2024). Дальнейшие исследования могут направиться на интеграцию промпт-оптимизации с семантическими моделями для улучшения точности в сложных контекстах. Таким образом, предложенный подход не только расширяет методологию анализа текстов, но и задает новые стандарты прозрачности в психологических исследованиях (Tamkin et al., 2021).

Заключение

Исследование продемонстрировало, что применение метода рефлексии больших языковых моделей (LLM) в задачах классификации типов поведения в цифровой среде открывает большие возможности для повышения точности и прозрачности анализа. Таргетированная корректировка запросов через рефлексию LLM позволила достичь роста точности классификации конструктивного типа поведения с 21.7% до 71.1%, что подтверждает гипотезу о способности LLM к самоанализу и адаптации критериев классификации. Метод, основанный на итеративном диалоге с моделью, формализует экспертные критерии в естественном языке и улучшает интерпретируемость решений.

Однако выявлены и некоторые ограничения подхода. Во-первых, эффект «перетягивания одеяла»: улучшение точности одного класса (например, конструктивного) сопровождалось снижением точности других (деструктивный – на 17.2%, информативный – на 23.8%) из-за семантического пересечения категорий. Например, расширение критериев конструктивности привело к ошибочной классификации 19% деструктивных текстов с завуалированной критикой. Во-вторых, культурно-лингвистические артефакты: универсальные критерии не учитывали специфику русскоязычного дискурса, например, иронию вроде «Давайте не будем спешить» часто трактовалась как нейтральная. Это требует адаптации запросов под языковые и культурные особенности. В-третьих, сложность определения «конструктивности»: границы между информативным и конструктивным поведением остаются размытыми. Анализ показал, что 46% информативных текстов с авторской позицией ошибочно классифицировались как конструктивные, что подчеркивает необходимость уточнения теоретической базы.

Методологический вклад исследования включает разработку таксономии психолингвистических маркеров для трех типов поведения личности, применимой для автоматизированного анализа. Например, ключевые паттерны, такие как использование слов «предлагаю» для конструктивности или «запретить» для деструктивности, обеспечивают структурированный подход к классификации. Методология рефлексии LLM открывает возможности для адаптации моделей к специализированным задачам в психологии, например, прогнозированию вероятности конструктивных или деструктивных форм поведения личности на основе анализа содержания текстов пользователей социальных сетей или дискурсов между различными социальными сообществами. Созданный инструментарий снижает зависимость от ручного кодирования и повышает скорость анализа больших данных, что актуально для мониторинга социальных сетей и онлайн-дискуссий.

Для дальнейшего улучшения метода необходимы кросс-культурные исследования для уточнения критериев конструктивности, особенно в контекстах с высоким уровнем иронии или патриотической риторики. Будущие работы должны интегрировать промпт-оптимизацию с семантическими моделями для улучшения точности в сложных контекстах, например, анализе политических дискуссий.

Результаты исследования подтверждают, что LLM могут стать надежным инструментом для анализа и прогнозирования особенностей поведения пользователей социальных сетей, но требуют тщательной настройки и теоретического обоснования критериев. Предложенный метод не только повышает точность классификации, но и задает стандарты прозрачности, что критически важно для психологических исследований. Однако дальнейшие работы должны учитывать культурную специфику и противоречивость категорий, чтобы минимизировать риски предвзятости и обеспечить этически обоснованное применение AI-технологий.

Литература

Ермаков, П. Н., Коленова, А. С., Денисова, Е. Г., Куприянов, И. В. (2022). Психологические предикторы конструктивных и деструктивных форм информационного поведения молодежи. *Российский психологический журнал*, *19*(2), 21–34. https://doi.org/10.21702/rpj.2022.2.2

Кузнецова, Ю. М. Сценарный подход к анализу текстов. (2018). *Труды Института Системного Анализа Российской Академии Наук*, *68*(1), 31–41. https://doi.org/10.14357/20790279180103

Мишланов, В. А., Чуганская, А. А., Смирнов, И. В., Суворова, М. И., Курузов, И. А. (2020). Разработка методов анализа сценариев поведения (на материале инструктивных интернет-текстов). *Медиалингвистика*, *7*(1), 16–28. https://doi.org/10.21638/spbu22.2020.102

Татарова, Г. Г., Бабич, Н. С., Бессокирная, Г. П., Кученкова, А. В. (2023). *Типологический анализ в социологии как диагностическая процедура*. Москва: ФНИСЦ РАН, 2023. https://doi.org/10.19181/monogr.978-5-89697-408-6.2023

Abdurahman, S., Atari, M., Karimi-Malekabadi, F., Xue, M. J., Trager, J., Park, P. S., … Dehghani, M. (2024). Perils and opportunities in using large language models in psychological research. *PNAS Nexus*, *3*(7), 245. https://doi.org/10.1093/pnasnexus/pgae245

Abramski, K., Citraro, S., Lombardi, L., Rossetti, G., & Stella, M. (2023). Cognitive Network Science Reveals Bias in GPT-3, GPT-3.5 Turbo, and GPT-4 Mirroring Math Anxiety in High-School Students. *Big Data and Cognitive Computing*, *7*(3). https://doi.org/10.3390/bdcc7030124

Almeida, G. F. C. F., Nunes, J. L., Engelmann, N., Wiegmann, A., & Araújo, M. de. (2024). Exploring the psychology of LLMs' moral and legal reasoning. *Artificial Intelligence*, 333, 104145. https://doi.org/10.1016/j.artint.2024.104145

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6), e2218523120. https://doi.org/10.1073/pnas.2218523120

Xie, C., Chen, C., Jia, F., Ye, Z., Lai, S., Shu, K., … Li, G. (2024). Can Large Language Model Agents Simulate Human Trust Behavior? *arXiv:2402.04559*. Retrieved from https://arxiv.org/abs/2402.04559

Dehouche, N. (2021). Plagiarism in the age of massive generative pre-trained transformers (GPT-3). *Ethics in Science and Environmental Politics*, 21, 17–23. http://dx.doi.org/10.3354/esep00195

Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., … Pennebaker, J. W. (2023). Using large language models in psychology. *Nature Reviews Psychology*, 2, 688–701. https://doi.org/10.1038/s44159-023-00241-5

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences*, *27*(7), 597–600. https://doi.org/10.1016/j.tics.2023.04.008

Frank, M. C. (2023). Baby steps in evaluating the capacities of large language models. *Nature Reviews Psychology*, *2*(8), 451–452. https://doi.org/10.1038/s44159-023-00211-x

Grossmann, I., Feinberg, M., Parker, D. C., Christakis, N. A., Tetlock, P. E., & Cunningham, W. A. (2023). AI and the transformation of social science research. *Science*, *380*(6650), 1108–1109. https://doi.org/10.1126/science.adi1778

Habermas, J. (1985). *Theory of Communicative Action, Volume One: Reason and the Rationalization of Society*. Boston: Beacon Press.

Hagendorff, T. (2023). Machine Psychology: Investigating Emergent Capabilities and Behavior in Large Language Models Using Psychological Methods. *arXiv:2303.13988*. Retrieved from https://doi.org/10.48550/arXiv.2303.13988

Harding, J., D'Alessandro, W., Laskowski, N. G., & Long, R. (2023). AI language models cannot replace human research participants. *AI & Society*, 39, 2603–2605. https://doi.org/10.1007/s00146-023-01725-x

Hutson, M. (2023). Doing research with human subjects is costly and cumbersome. Can AI chatbots replace them? *Science*, *381*(6654), 121–123. https://doi.org/10.1126/science.adj6791

Jin, C., Zhang, S., Shu, T., & Cui, Z. (2023). The Cultural Psychology of Large Language Models: Is ChatGPT a Holistic or Analytic Thinker? *arXiv:2308.14242*. Retrieved from https://doi.org/10.48550/arXiv.2308.14242

Ke, L., Tong, S. Cheng, P., & Peng, K. (2024). Exploring the Frontiers of LLMs in Psychological Applications: A Comprehensive Review. *arXiv:2401.01519*. Retrieved from https://doi.org/10.48550/arXiv.2401.01519

Kublik, S., & Saboo, S. (2022). *Gpt-3: the ultimate guide to building NLP products with OpenAI API* (1st ed.). Birmingham: Packt Publishing.

Linegar, M., Kocielnik, R., & Alvarez, R. M. (2023). Large language models and political science. *Frontiers in Political Science*, 5. http://dx.doi.org/10.3389/fpos.2023.1257092

Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., … Ge, B. (2023). Summary of ChatGPT-Related research and perspective towards the future of large language models. *Meta-Radiology*, *1*(2), 100017. https://doi.org/10.1016/j.metrad.2023.100017

Markowitz, D. M. (2024). Can generative AI infer thinking style from language? Evaluating the utility of AI as a psychological text analysis tool. *Behavior Research Methods*, 56, 3548–3559. http://dx.doi.org/10.3758/s13428-024-02344-0

Newell, A. (1994). *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

Park, P. S., Schoenegger, P., & Zhu, C. (2024). Diminished diversity-of-thought in a standard large language model. *Behavior Research Methods*, *56*(6), 5754–5770. https://doi.org/10.3758/s13428-023-02307-x

Parra, V., Sureda, P., Corica, A., Schiaffino, S., & Godoy, D. (2024). Can Generative AI Solve Geometry Problems? Strengths and Weaknesses of LLMs for Geometric Reasoning in Spanish. *International Journal of Interactive Multimedia and Artificial Intelligence*, *8*(5), 65. https://doi.org/10.9781/ijimai.2024.02.009

Patel, S. C., & Fan, J. (2023). Identification and Description of Emotions by Current Large Language Models. *bioRxiv.* Retrieved from https://doi.org/10.1101/2023.07.17.549421

Rathje, S., Mirea, D.-M., Sucholutsky, I., Marjieh, R., Robertson, C., & Van Bavel, J. J. (2024). GPT is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences*, *121*(34). https://doi.org/10.1073/pnas.2308950121

Sartori, G., & Orrù, G. (2023). Language models and psychological sciences. *Frontiers in Psychology*, 14. https://doi.org/10.3389/fpsyg.2023.1279317

Stella, M., Hills, T. T., & Kenett, Y. N. (2023). Using cognitive psychology to understand GPT-like models needs to extend beyond human biases. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(43), e2312911120. https://doi.org/10.1073/pnas.2312911120

Savelka, J., & Ashley, K. D. (2023). The unreasonable effectiveness of large language models in zero-shot semantic annotation of legal texts. *Frontiers in Artificial Intelligence*, 6, 1279794. https://doi.org/10.3389/frai.2023.1279794

Stokel-Walker, C. (2022). AI bot ChatGPT writes smart essays – should professors worry? *Nature*. https://doi.org/10.1038/d41586-022-04397-7

Tamkin, A., Brundage, M., Clark, J., & Ganguli, D. (2021). Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models. *arXiv:2102.02503*. Retrieved from https://doi.org/10.48550/arXiv.2102.02503

Turner, J. C., & Reynolds, K. J. (2012). Self-categorization theory. In *Handbook of theories of social psychology*. *Volume 2* (pp. 399–417). New York: SAGE Publications.

Vaswani, A., Shazeer, N. Parmar, N. Uszkoreit, J. Jones, L. Gomez, A. N. … Polosukhin I. (2017). Attention Is All You Need. *arXiv:1706.03762*. Retrieved from https://doi.org/10.48550/arXiv.1706.03762

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S. … Fedus, W. (2022). Emergent Abilities of Large Language Models. *arXiv:2206.07682*. Retrieved from https://doi.org/10.48550/arXiv.2206.07682

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2024). Can Large Language Models Transform Computational Social Science? *Computational Linguistics*, *50*(1), 237–291. https://doi.org/10.1162/coli_a_00502

Zhong, Y., Chen, Y., Zhou, Y., Lyu, Y.-A.-H., Yin, J.-J., & Gao, Y. (2023). The Artificial intelligence large language models and neuropsychiatry practice and research ethic. *Asian Journal of Psychiatry*, 84, 103577. https://doi.org/10.1016/j.ajp.2023.103577

Поступила в редакцию: 08.12.2025

Поступила после рецензирования: 12.01.2026

Принята к публикации: 03.04.2026

Заявленный вклад авторов

Александр Алексеевич Агафонов – работа с литературными источниками, организация и проведение эмпирического исследования с использование методов искусственного интеллекта, подготовка и написание текста статьи.

Ахат Рустемович Хусаинов – сбор материалов и обработка результатов исследования, техническая поддержка работ, связанных с использованием искусственного интеллекта.

Павел Николаевич Устин – работа с литературными источниками, организация и проведение эмпирического исследования, написание и анализ текста статьи.

Леонид Михайлович Попов – психологическая концептуализация результатов междисциплинарного исследования, критический анализ и редактирование статьи.

Информация об авторах

Александр Алексеевич Агафонов – кандидат физико-математических наук, доцент, доцент кафедры высшей математики и математического моделирования Института математики и механики им. Н. И. Лобачевского, Федеральное государственное автономное образовательное учреждение высшего образования «Казанский (Приволжский) федеральный университет», Казань, Россия; Researcher ID: L-4818-2015, Scopus ID: 37015137600, Author ID: 825443, ORCID ID: https://orcid.org/0000-0001-6924-6701; e-mail: a.a.agathonov@gmail.com

Ахат Рустемович Хусаинов – лаборант, Федеральное государственное автономное образовательное учреждение высшего образования «Казанский (Приволжский) федеральный университет», Казань, Россия; e-mail: ahat2182@gmail.com

Павел Николаевич Устин – кандидат психологических наук, доцент, заведующий кафедрой общей психологии Института психологии и образования, Федеральное государственное автономное образовательное учреждение высшего образования «Казанский (Приволжский) федеральный университет», Казань, Россия; Researcher ID: AAU-5199-2020, Scopus ID: 56766208200, Author ID: 458473, ORCID ID: https://orcid.org/0000-0003-3950-7434; e-mail: pavust@mail.ru

Леонид Михайлович Попов – доктор психологических наук, профессор, профессор кафедры клинической психологии и психологии личности Института психологии и образования, Федеральное государственное автономное образовательное учреждение высшего образования «Казанский (Приволжский) федеральный университет», Казань, Россия; Researcher ID: N-1069-2013, Scopus ID: 56765399100, Author ID: 493043, ORCID ID: https://orcid.org/0000-0002-1866-2411; e-mail: leonid.popov@inbox.ru

Информация о конфликте интересов

Авторы заявляют об отсутствии конфликта интересов.