БУКИНА Т.В.,
НИИ прикладного искусственного интеллекта и цифровых решений, РЭУ им. Г.В. Плеханова, г. Москва, Россия; e-mail: bukinatatyanav@gmail.com. ORCID: 0000-0003-2658-4545

СМИРНОВ Н.М.,
НИИ прикладного искусственного интеллекта и цифровых решений, РЭУ им. Г.В. Плеханова, г. Москва, Россия; e-mail: nikita.smirnov.m@gmail.com. ORCID: 0009-0002-7593-4345

МАКСИМЕНКО В.А.,
д.ф.-м.н., НИИ прикладного искусственного интеллекта и цифровых решений, РЭУ им. Г.В. Плеханова, г. Москва, Россия; Национальный университет Сингапура, г. Сингапур, Республика Сингапур; e-mail: maximenkovl@gmail.com. ORCID: 0000-0002-4632-6896

КАРПОВ К.О.,
НИИ прикладного искусственного интеллекта и цифровых решений, РЭУ им. Г.В. Плеханова, г. Москва, Россия; e-mail: flype586@gmail.com

СКАЛИН Ю.Е.,
к.м.н., профессор, БФУ им. И. Канта, г. Калининград, Россия; e-mail: ISkalin@kantiana.ru. ORCID: 0000-0001-5504-0307

ХРАМОВА М.В.,
к.пед.н., Высшая школа кибертехнологий, математики и статистики, РЭУ им. Г.В. Плеханова, г. Москва, Россия; e-mail: mhramova@gmail.com. ORCID: 0000-0002-6392-4580

КУРКИН С.А.,
д. ф.-м.н., НИИ прикладного искусственного интеллекта и цифровых решений, РЭУ им. Г.В. Плеханова, г. Москва, Россия; e-mail: kurkinsa@gmail.com. ORCID: 0000-0002-3438-5717

ХРАМОВ А.Е.,
член-корреспондент РАН, профессор, д.ф.-м.н., НИИ прикладного искусственного интеллекта и цифровых решений, РЭУ им. Г.В. Плеханова, г. Москва, Россия; ФГБУ «НМХЦ им. Н.И. Пирогова» Минздрава России, г. Москва, Россия; e-mail: hramov.ae@rea.ru. ORCID: 0000-0003-2787-2530

ОЦЕНКА ЛОКАЛИЗАЦИИ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ ДЛЯ ЗАДАЧ КЛИНИЧЕСКОЙ ПСИХОЛОГИИ: ПРОГНОЗИРОВАНИЕ ПОКАЗАТЕЛЕЙ ТРЕВОГИ И ДЕПРЕССИИ НА ВЫБОРКЕ РОССИЙСКИХ СТУДЕНТОВ

DOI: 10.25881/18110193_2026_3_50

Аннотация. *Цель. Оценить пригодность больших языковых моделей (Large Language Models, LLM; далее – БЯМ), работающих в закрытом контуре, для прогнозирования показателей тревоги и депрессии по Госпитальной шкале тревоги и депрессии (Hospital Anxiety and Depression Scale, HADS) на выборке российских студентов.*

Материалы и методы. Эмпирическую базу составили деперсонализированные ответы 946 студентов-первокурсников (669 женщин, 277 мужчин) российского вуза на 14 вопросов HADS. Для 11 моделей (DeepSeek-R1-32B, Qwen3-32B, T-pro-2.0/2.1, YandexGPT-8B, GigaChat-20B и др.), развернутых локально на GPU, распределения вероятностей ответов извлекались из логарифмических вероятностей токенов. Экспериментальный дизайн включал 5 групп промптов (текстовых инструкций модели), 2 формата подачи (структурированный диалог и неструктурированный текст) и 4 варианта формулировок. Модельные и эмпирические распределения сравнивались по расстоянию Вассерштейна (W); дополнительно рассчитывались ожидаемые баллы субшкал HADS.

Результаты. Лучшую точность воспроизведения распределений показала DeepSeek-R1-32B (W = 0,293), однако она систематически занижала ожидаемый балл тревожности (5,6 против 7,7 у людей). Наиболее точные ожидаемые баллы тревожности продемонстрировали YandexGPT-pretrain (7,2) и T-pro-2.0 (7,1). Выбор модели объяснял 47% дисперсии метрики; детализация описания персоны и пол не влияли на результат ($p > 0,05$). Субшкала депрессии воспроизводилась хуже субшкалы тревожности.

Выводы. Оценки устойчивы к вариациям промпта: детализация описания персоны и формулировка не влияют на результат – достаточно минимального описания респондента. Модели с англоязычным обучением предпочитают англоязычные промпты; для русскоязычных моделей языковые предпочтения определяются базовой архитектурой. Гипотеза о систематическом преимуществе русскоязычных моделей не подтвердилась: качество определяется доменной адаптацией, а не языковой принадлежностью. Для практического применения необходима валидация на эксперименте с людьми. Модели не заменяют клиническую оценку специалиста.

Ключевые слова: *большие языковые модели; HADS; тревога; депрессия; локализация; скрининг.*

Для цитирования: *Букина Т.В., Смирнов Н.М., Максименко В.А., Карпов К.О., Скалин Ю.Е., Храмова М.В., Куркин С.А., Храмов А.Е. Оценка локализации больших языковых моделей для задач клинической психологии: прогнозирование показателей тревоги и депрессии на выборке российских студентов. Врач и информационные технологии. 2026; 3: 50-69. DOI: 10.25881/18110193_2026_3_50.*

BUKINA T.V.,
Research Institute of Applied Artificial Intelligence and Digital Solutions, Plekhanov Russian University of Economics, Moscow, Russia; e-mail: bukinatatyanav@gmail.com. ORCID: 0000-0003-2658-4545

SMIRNOV N.M.,
Research Institute of Applied Artificial Intelligence and Digital Solutions, Plekhanov Russian University of Economics, Moscow, Russia; e-mail: nikita.smirnov.m@gmail.com. ORCID: 0009-0002-7593-4345

MAKSIMENKO V.A.,
DSc, Research Institute of Applied Artificial Intelligence and Digital Solutions, Plekhanov Russian University of Economics, Moscow, Russia; National University of Singapore, Singapore, Republic of Singapore; e-mail: maximenkovl@gmail.com. ORCID: 0000-0002-4632-6896

KARPOV K.O.,
Research Institute of Applied Artificial Intelligence and Digital Solutions, Plekhanov Russian University of Economics, Moscow, Russia; e-mail: flype586@gmail.com

SKALIN Y.E.,
PhD, Professor, Immanuel Kant Baltic Federal University, Kaliningrad, Russia; e-mail: ISkalin@kantiana.ru. ORCID: 0000-0001-5504-0307

KHRAMOVA M.V.,
PhD, Higher School of Cyber Technologies, Mathematics and Statistics, Plekhanov Russian University of Economics, Moscow, Russia; e-mail: mhramova@gmail.com. ORCID: 0000-0002-6392-4580

KURKIN S.A.,
DSc, Research Institute of Applied Artificial Intelligence and Digital Solutions, Plekhanov Russian University of Economics, Moscow, Russia; e-mail: kurkinsa@gmail.com. ORCID: 0000-0002-3438-5717

HRAMOV A.E.,
Corresponding Member of the Russian Academy of Sciences, Professor, DSc, Research Institute of Applied Artificial Intelligence and Digital Solutions, Plekhanov Russian University of Economics, Moscow, Russia; Pirogov National Medical and Surgical Center, Moscow, Russia; e-mail: hramov.ae@rea.ru. ORCID: 0000-0003-2787-2530

EVALUATING LARGE LANGUAGE MODEL LOCALIZATION FOR CLINICAL PSYCHOLOGY: PREDICTING ANXIETY AND DEPRESSION SCORES IN A RUSSIAN STUDENT SAMPLE

DOI: 10.25881/18110193_2026_3_50

Abstract. *Aim. To evaluate the suitability of large language models (LLMs), deployed in a closed-loop environment, for predicting anxiety and depression scores on the Hospital Anxiety and Depression Scale (HADS) in a sample of Russian students.*

Materials and methods. The empirical dataset consisted of de-identified responses from 946 first-year university students (669 female, 277 male) in Russia to 14 HADS items. For 11 models (DeepSeek-R1-32B, Qwen3-32B, T-pro-2.0/2.1, YandexGPT-8B, GigaChat-20B, and others) deployed locally on GPUs, response probability distributions were extracted from token log probabilities. The experimental design included 5 prompt groups (textual instructions for the model), 2 presentation formats (structured dialogue and unstructured text), and 4 wording variations. Model-generated and empirical distributions were compared using Wasserstein distance (W); expected HADS subscale scores were additionally calculated.

Results. T DeepSeek-R1-32B demonstrated the best accuracy in reproducing distributions (W = 0.293); however, it consistently underestimated expected anxiety scores (5.6 vs. 7.7). The most accurate expected anxiety scores were produced by YandexGPT-pretrain (7.2) and T-pro-2.0 (7.1). Model selection explained 47% of metric variance; persona detail and respondent sex had no significant effect ($p > 0.05$). The depression subscale was reproduced less accurately than the anxiety subscale.

Conclusions. Scores are robust to prompt variations: the level of detail in the persona description and the specific wording do not affect the outcome—a minimal description of the respondent suffices. Models trained in English favor English-language prompts, whereas for Russian-language models, language preferences are determined by the underlying architecture. The hypothesis regarding a systematic advantage for Russian-language models was not confirmed: quality is determined by domain adaptation rather than language. Practical application requires validation through experiments involving human subjects. The models do not replace clinical assessment by a specialist.

Keywords*: large language models; HADS; anxiety; depression; localization; screening.*

For citation: *Bukina T.V., Smirnov N.M., Maksimenko V.A., Karpov K.O., Skalin Y.E., Khramova M.V., Kurkin S.A., Hramov A.E. Evaluating Large Language Model Localization for Clinical Psychology: Predicting Anxiety and Depression Scores in a Russian Student Sample. Medical doctor and information technology. 2026; 3: 50-69. DOI: 10.25881/18110193_2026_3_50.*

ВВЕДЕНИЕ

Актуальность проблемы. Проблема своевременного выявления тревожных и депрессивных расстройств сохраняет свою значимость в популяции студентов высших учебных заведений в связи с высокими психоэмоциональными нагрузками, характерными для этого периода жизни. Госпитальная шкала тревоги и депрессии (Hospital Anxiety and Depression Scale, HADS), разработанная Zigmond и Snaith в 1983 году [1], является одним из наиболее широко применяемых инструментов для скрининга тревоги и депрессии в различных популяциях и культурных контекстах. За последние годы психометрические свойства HADS были подтверждены в многочисленных международных исследованиях, включая испанский, индонезийский, греческий, китайский, урду языки [2, 3, 4, 5, 6]. На русскоязычной выборке студентов подтверждена двухфакторная структура опросника, соответствующая авторской модели: пункты группируются в субшкалы тревоги и депрессии так же, как в оригинальной версии, что свидетельствует в пользу конструктной валидности русскоязычной версии [7]. Внутренняя согласованность, оцененная по альфе Кронбаха, составила 0,73 для субшкалы «Тревога» и 0,62 для субшкалы «Депрессия». Значение по субшкале депрессии ниже общепринятого порога (≥ 0,70), что следует учитывать при интерпретации результатов по этой субшкале.

Потенциал языковых моделей в клинической психологии. Параллельно с развитием традиционных психодиагностических методов в последние годы активно исследуются возможности применения больших языковых моделей (Large Language Models, LLM; далее – БЯМ) для задач психического здоровья. Согласно систематическому обзору [8], наибольшая доля исследований БЯМ в медицине приходится именно на психические расстройства (86%), при этом ключевыми областями применения выступают выявление и классификация депрессии (38%), обнаружение суицидальных мыслей (7%) и классификация эмоций (5%). Другой масштабный обзор [9] демонстрирует, что применение БЯМ позволяет повысить точность ранней диагностики психических расстройств на 33%, эффективность персонализированных планов лечения на 27% и вовлеченность в психообразовательные программы на 24%. Вместе с тем возможности моделей в этой области неравномерны. Kuzmin et al. [10] сравнили традиционные методы машинного обучения, энкодерные модели и БЯМ на пяти русскоязычных наборах данных для выявления депрессии и тревоги: БЯМ превзошли традиционные подходы на зашумленных и малых выборках, однако на клинически подтвержденных случаях депрессии психолингвистические признаки и энкодерные модели показали сопоставимое качество. Также языковые модели используются для генерирования виртуальных пациентов, которые имитируют реалистичные симптомы, поведение и ответы в диалоге, позволяя моделировать клинические сценарии без участия реальных людей [9]. Такие симуляции применяются для обучения специалистов, тестирования терапевтических подходов и оценки качества психологической помощи в контролируемых условиях. Потенциал применения БЯМ в этих областях несомненно высок. Однако их способность достоверно связывать психоэмоциональное состояние с лингвистическими паттернами требует дополнительной валидации, особенно при работе с популяциями, отличными от западных и не использующими английский язык в качестве основного средства коммуникации (данная проблема известна в литературе как проблема локализации).

Проблема локализации языковых моделей. Ключевой вызов при применении БЯМ для анализа текстов на языках, отличных от английского – проблема кросс-языкового переноса и культурной адаптации. Данная проблема проистекает из глубокой асимметрии в обучающих данных: подавляющая доля корпусов предварительного обучения приходится на английский язык, тогда как большинство языков мира представлены в них существенно хуже [11]. Это приводит к систематическому снижению качества работы моделей на таких языках, особенно в задачах, требующих тонкого семантического и прагматического понимания. Для преодоления этого разрыва необходима разработка подходов, учитывающих не только лингвистическую специфику, но и культурные особенности, релевантные для конечных пользователей [12]. Для того чтобы интеллектуальные системы стали полезными, они должны обеспечивать взаимодействие на предпочитаемых пользователями языках и учитывать культурные особенности их окружения.

Редакционная статья Nature Machine Intelligence [13] подчеркивает, что модели, обученные преимущественно на англо-центричных и западно-ориентированных данных, демонстрируют низкую производительность в не-западных контекстах. Исследователи указывают на необходимость разработки экономически эффективных моделей, адаптированных к локальным контекстам, обученных на конкретных национальных языках и способных работать с минимальными инфраструктурными требованиями.

Для русского языка проблема культурной адаптации БЯМ также остается актуальной. В работе [14] культурная осведомленность GPT-4o, Claude 3.5 Sonnet и Gemini 1.5 Pro оценивалась на 400 вопросах о российском социокультурном контексте, включая крылатые выражения, мемы и культурные типажи. При удовлетворительном знании фактологии модели плохо улавливали тонкие культурные нюансы: ошибались при выборе между синонимичными, но контекстуально различными ответами и не всегда распознавали культурно-специфическую логику; при этом закрытые коммерческие модели превосходили открытые. Кроме того, работа [15] представила набор данных RuBia для выявления социальных и культурных предубеждений в БЯМ применительно к русскому языку, охватывающий 4 домена и 19 поддоменов с почти 2000 уникальных пар предложений. Было показано, что все 9 протестированных моно- и мультиязычных БЯМ воспроизводят стереотипы, причем моноязычные русские модели предвзяты сильнее мультиязычных. Наличие таких стереотипов может проявляться в систематических смещениях при оценке психологического состояния конкретной популяции, что обосновывает необходимость сравнительного анализа моделей с разной культурно-языковой направленностью обучения на российской выборке. Обзор применения БЯМ в российском здравоохранении [16] систематизирует возможности ряда моделей (ChatGPT, ClinicalGPT, BioBERT, GigaChat, Yandex GPT) для автоматизации диагностики и поддержки клинических решений; апробация на ответах ChatGPT по реальным данным пациентов показала приемлемую точность, однако выявила сохраняющиеся риски «галлюцинаций» – правдоподобных, но фактически ошибочных ответов модели. Авторы обосновывают необходимость разработки специализированных медицинских БЯМ, адаптированных к российской клинической практике.

Постановка проблемы и цель исследования. Несмотря на растущий интерес к применению БЯМ в задачах ментального здоровья и признание важности культурной адаптации моделей, вопрос о том, какие модели – российские или зарубежные – лучше описывают русскую популяцию и, соответственно, лучше локализованы для решения задач клинической психологии, остается эмпирически не исследованным. Особый интерес представляет сравнение российских БЯМ, изначально обучавшихся на больших корпусах русскоязычных текстов и потенциально лучше учитывающих локальный культурный контекст, с зарубежными аналогами. Учитывая фокус работы на медицинской диагностике и чувствительность медицинских данных, основное внимание сосредоточено на БЯМ, работающих в закрытом контуре – в изолированной среде без доступа к внешним данным, для обеспечения безопасности обработки данных. В первую очередь рассматриваются языковые модели с малым числом параметров, которые больше подходят для локального развертывания, однако с некоторым снижением качества.

Целью настоящего исследования является сравнительная оценка способности российских и зарубежных языковых моделей воспроизводить эмпирические распределения ответов на вопросы Госпитальной шкалы тревоги и депрессии (HADS) на выборке российских студентов-первокурсников как инструмент измерения культурной локализации модели. В данной работе под локализацией понимается способность модели воспроизводить эмпирические распределения ответов конкретной культурно-языковой выборки на стандартизированный психологический опросник. Отметим, что это индикатор, а не исчерпывающая оценка культурной адаптации в широком смысле. Симуляция ответов выступает методом исследования, а не самоцелью. Мы исходим из гипотезы, что российские модели, обученные на больших объемах русскоязычных текстов и лучше адаптированные к локальному культурному контексту, продемонстрируют более высокую точность в воспроизведении показателей тревоги и депрессии по сравнению с зарубежными моделями.

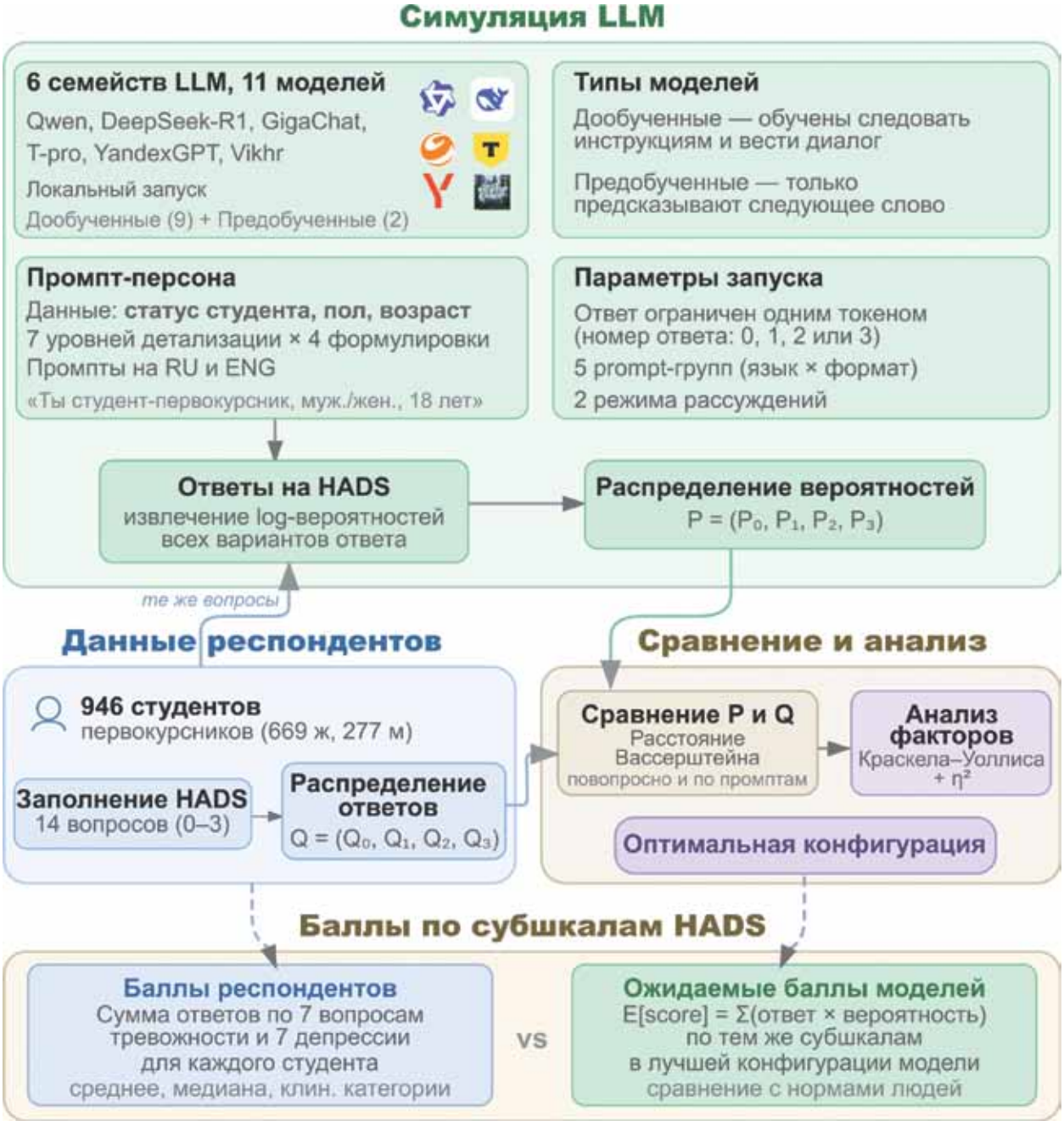


Рисунок 1 – Схема исследования. Верхний блок — симуляция ответов БЯМ: 6 семейств моделей (11 моделей) получают промпт-персону и вопросы HADS, генерируя распределения вероятностей ответов P. Нижний-левый блок — данные респондентов: ответы 946 студентов преобразуются в эталонные распределения Q. Центральный блок — сравнение распределений P и Q по расстоянию Вассерштейна повопросно и по промптам с анализом факторов. Нижний блок — сравнение ожидаемых баллов по субшкалам тревожности и депрессии.

Материалы и методы

Общая схема исследования представлена на рисунке 1. Исследование включало три этапа: (1) симуляция ответов на HADS с помощью 11 языковых моделей при различных конфигурациях промптов, (2) формирование эталонных распределений ответов на основе данных 946 студентов-первокурсников, (3) сравнение модельных и эмпирических распределений по расстоянию Вассерштейна и ожидаемым баллам субшкал.

Опросник HADS

HADS – стандартизированный самоопросник из 14 вопросов[17], разработанный для скрининга тревоги и депрессии у соматических пациентов. Принципиальное отличие HADS от других шкал – исключение соматических симптомов (утомляемость, нарушения сна, аппетита), что позволяет избежать ложноположительных результатов у пациентов с физическими заболеваниями. Опросник включает субшкалы тревожности (HADS-A, вопросы 1, 3, 5, 7, 9, 11, 13) и депрессии (HADS-D, вопросы 2, 4, 6, 8, 10, 12, 14). Каждый вопрос имеет 4 варианта ответа (0-3 балла). Сумма баллов по субшкале: 0-7 – норма, 8-10 – субклинический уровень, 11-21 – клинический уровень. Русскоязычная версия HADS валидирована и широко применяется в российской клинической практике.

Выборка

Эмпирическую базу составили деперсонализированные ответы 946 студентов-первокурсников БФУ им. И. Канта (669 женщин, 277 мужчин) [18] на 14 вопросов HADS. Для каждой комбинации (пол, вопрос, вариант ответа) рассчитывалась доля респондентов; из этих долей формировались эталонные распределения $Q = (Q_0, Q_1, Q_2, Q_3)$, где Q_0, Q_1, Q_2, Q_3 рассчитываются как доля респондентов, выбравших вариант ответа 0, 1, 2, 3, соответственно.

Модели

В исследование были включены 11 моделей, развернутых на 4 GPU NVIDIA V100-SXM3-32GB (таблица 1). Критерием отбора моделей послужило число параметров (до 33 млрд), позволяющее развернуть их на локальном оборудовании и обеспечить работу в закрытом контуре без передачи данных внешним сервисам.

В исследованиях по симуляции респондентов описание персоны, как правило, передается модели через шаблон структурированного диалога: системное сообщение задает роль (например, «Ты – студент первого курса российского университета»), а пользовательское содержит текст вопроса и варианты ответов. Такой формат поддерживают дообученные модели, которые в процессе дополнительного обучения научились различать роли собеседников (система,

Таблица 1 – Модели, включенные в исследование

№	Модель (сокращение)	Тип	Число параметров	Язык
1	Qwen3-0.6B	дообученная	0,6B	ENG
2	Qwen3-32B	дообученная	32B	ENG
3	Qwen2.5-32B-Instruct (Qwen2.5-32B)	дообученная	32B	ENG
4	Vikhr-YandexGPT-5-Lite-8B-it (Vikhr-YaGPT-8B)	дообученная	8B	RU
5	DeepSeek-R1-Distill-Qwen-32B (DS-R1-32B)	дообученная	32B	ENG
6	GigaChat-20B-A3B-instruct (GigaChat-instr)	дообученная	20B	RU+ENG
7	GigaChat-20B-A3B-base (GigaChat-base)	предобученная	20B	RU+ENG
8	T-pro-it-2.0 (T-pro-2.0)	дообученная	33B	RU
9	T-pro-it-2.1 (T-pro-2.1)	дообученная	33B	RU
10	YandexGPT-5-Lite-8B-pretrain (YaGPT-pretr)	предобученная	8B	RU+ENG
11	YandexGPT-5-Lite-8B-instruct (YaGPT-instr)	дообученная	8B	RU+ENG

Примечания: столбец «Язык» указывает преимущественную языковую ориентацию обучающих данных модели, т.е. на каком языке представлена основная доля корпуса предварительного обучения (RU – русскоязычный корпус, ENG – англоязычный, RU+ENG – смешанный). Языковая ориентация определялась на основании опубликованных открытых данных (технические отчеты, карточки моделей).

пользователь, ассистент) и следовать инструкциям. Предобученные модели обучены только продолжать текст – они не различают роли и не могут корректно обработать структурированный диалог. Поэтому для них описание персоны и вопрос объединялись в единый неструктурированный текст.

При отборе моделей также были выявлены особенности, повлиявшие на дизайн эксперимента:

- Ряд моделей построен на базе других БЯМ с дообучением: DS-R1-32B – на Qwen2.5-32B, T-pro – на семействе Qwen3, Vikhr-YaGPT-8B – на YandexGPT-5-Lite-8B-pretrain, YaGPT-5-Lite-8B – на Qwen2.5. Это позволяет при анализе результатов сопоставить модель-потомка с базовой архитектурой и оценить влияние дообучения на качество симуляции (см. раздел «Обсуждение»);
- YaGPT-instr формально имеет шаблон диалога, однако при его использовании модель демонстрировала патологическую сверхуверенность: вероятность одного варианта ответа достигала 0,999, что не соответствует реалистичному распределению и свидетельствует об отказе модели формировать информативное распределение ответов. Поэтому для YaGPT-instr результаты в формате структурированного диалога были исключены, и модель оценивалась только по результатам в неструктурированном тексте;
- Часть моделей (Qwen3, DS-R1-32B, T-pro-2.0) перед формированием ответа генерирует так называемые токены рассуждений – скрытый от пользователя текст, в котором модель «обдумывает» ответ. Наличие этого промежуточного текста могло влиять на итоговые вероятности ответов, поэтому в эксперименте для каждой модели сравнивались два режима: стандартная генерация (с рассуждениями) и подавление рассуждений.

Экспериментальный дизайн

Во время эксперимента изучалось влияние следующих факторов:

Группа промптов. Для исследования влияния языка на результаты работы модели мы выделили 5 групп промптов (текстовых инструкций модели), которые различались языком системного промпта с описанием персоны, языком вопроса и форматом подачи:

- Группа I: системный промпт на русском, вопрос на русском, шаблон диалога (далее – RU→RU);
- Группа II: системный промпт на английском, вопрос на русском, шаблон диалога (далее – ENG→RU);
- Группа III: системный промпт на английском, вопрос на английском, шаблон диалога (далее – ENG→ENG);
- Группа IV: системный промпт на русском, вопрос на русском, неструктурированный текст (далее – RU→RU (н.т.));
- Группа V: системный промпт на английском, вопрос на английском, неструктурированный текст (далее – ENG→ENG (н.т.)).

Группы I–III использовали шаблон структурированного диалога с ролями. Группы IV и V объединяли описание персоны и вопрос в единый неструктурированный текст, что позволяло работать с предобученными моделями, не поддерживающими шаблон диалога.

Режим рассуждений. Генерируемые некоторыми моделями токены рассуждений могли повлиять на эффективность вычисления логарифмических вероятностей. В связи с этим каждая модель тестировалась в двух режимах: стандартная генерация и подавление рассуждений модели – к началу ответа добавлялся пустой блок токенов рассуждений.

Пол респондента, формулировка промпта, детализация промпта. Промпты варьировались по: полу (мужской и женский); 4 вариантам формулировки, различающимся способом указания пола респондента в системном промпте (в начале или в конце предложения, с явным или косвенным указанием пола); 7 уровням детализации описания персоны – от минимального описания до расширенного с указанием возраста, учебного заведения и дополнительных характеристик.

Пример минимального промпта: «Ты – студент первого курса российского университета. Насколько ты согласен с данным утверждением:...», после чего следовал текст вопроса HADS с вариантами ответов.

Извлечение логарифмических вероятностей

Генерация выполнялась с помощью vLLM [19] – открытой платформы для эффективной работы

языковых моделей с использованием аппаратного ускорения. Модель генерировала ровно один символ (токен). Для этого применялся механизм управляемого декодирования [20] – метод, ограничивающий допустимые ответы модели заданным набором токенов: в данном случае цифрами 0–3, соответствующими вариантам ответа опросника. Из результатов извлекались логарифмические вероятности вариантов – числа, отражающие уверенность модели в каждом из возможных ответов (чем больше значение, тем выше уверенность). Затем они переводились в нормализованный вектор $P = (P_0, P_1, P_2, P_3)$ с суммой 1 – распределение ответов модели на один вопрос.

Отмечаемый в литературе риск «галлюцинаций» в настоящем дизайне устраняется конструктивно: управляемое декодирование исключает свободную генерацию, ограничивая ответ четырьмя допустимыми токенами. Вопрос о корректности самого распределения вероятностей при этом остается открытым и составляет предмет настоящего исследования.

Метрики сравнения

Расстояние Вассерштейна. Модельное распределение P сравнивалось с эталонным распределением людей Q по расстоянию Вассерштейна (W) – метрике, учитывающей порядковую природу ответов: перепутать ответы 0 и 3 «дороже», чем 0 и 1 [21]. Диапазон значений – от 0 (полное совпадение) до 3, т.е. чем меньше значение, тем лучше результат. Выбор обусловлен упорядоченным характером вариантов ответа HADS [22, 23].

Ожидаемые баллы субшкал. Для каждого вопроса рассчитывалось математическое ожидание ответа: $E = \Sigma(a \cdot P_a) = 0 \cdot P_0 + 1 \cdot P_1 + 2 \cdot P_2 + 3 \cdot P_3$, где a – номер варианта ответа (от 0 до 3), P_a – вероятность выбора варианта a моделью (или доля респондентов Q_a для эмпирических данных). Таким образом, чем больше вероятностной массы приходится на вариант 0 (чем больше P_0), тем меньше остается на варианты 1–3, и тем ниже ожидаемый балл. Например, при распределении P = (0,5; 0,3; 0,15; 0,05) ожидаемый балл $E = 0 \cdot 0{,}5 + 1 \cdot 0{,}3 + 2 \cdot 0{,}15 + 3 \cdot 0{,}05 = 0{,}75$, тогда как при P = (0,1; 0,3; 0,4; 0,2) E = 1,70. Суммируя E по 7 вопросам субшкалы, получали ожидаемый балл тревожности или депрессии. Для людей ожидаемые баллы вычислялись аналогично из долей Q. Данный подход корректен для сравнения средних, однако не позволяет оценить распределение суммарных баллов или провести классификацию по порогам HADS, поскольку модель отвечает на каждый вопрос независимо.

Статистический анализ

Расстояние Вассерштейна рассчитывалось для каждого вопроса, затем усреднялось по 14 вопросам на промпт-уровне. Усреднение допустимо, поскольку все вопросы HADS используют единую порядковую шкалу 0–3, что обеспечивает сопоставимость значений W между вопросами; тем не менее, данная агрегация предполагает равнозначность всех вопросов [24]. В пункте «Анализ по субшкалам» проиллюстрированы усредненные показатели по вопросам. Для каждой модели перебирались все 10 возможных комбинаций настроек (5 групп промптов × 2 режима рассуждений); лучшей конфигурацией считалась та, которая давала минимальное среднее W по 14 вопросам. Стандартное отклонение (SD) рассчитывалось для каждого промпта как разброс W между 14 вопросами, а затем усреднялось по всем промптам в рамках оптимальной конфигурации модели. Таким образом, SD отражает среднюю неравномерность воспроизведения разных вопросов для данной модели. Влияние факторов анализировалось тестом Краскела-Уоллиса с расчетом размера эффекта η^2 [25]. Тест применялся для оценки величины эффектов; каждый фактор тестировался без последующих попарных сравнений. В связи с разведочным характером анализа коррекция на множественность проверок не применялась.

РЕЗУЛЬТАТЫ

Описательная статистика выборки

Средний балл тревожности по HADS у людей составил 7,72 ± 4,31 (у женщин 8,07, у мужчин 6,87). Средний балл депрессии – 5,22 ± 3,55. По субшкале тревожности 51,4% респондентов находились в диапазоне нормы (0–7), 24,2% – субклинического уровня (8–10), 24,4% – клинического (11+). По субшкале депрессии в норме находились 76,2% респондентов.

Лучшие конфигурации моделей

В таблице 2 представлены оптимальные конфигурации для каждой модели, обеспечивающие

Таблица 2 – Лучшая конфигурация каждой модели

Модель	Режим	Группа
DS-R1-32B	станд.	RU→RU
YaGPT-pretr	подавл.	ENG→ENG (н.т.)
Qwen3-0.6B	подавл.	ENG→ENG
Qwen3-32B	станд.	ENG→ENG
Vikhr-YaGPT-8B	подавл.	RU→RU (н.т.)
T-pro-2.0	станд.	ENG→RU
GigaChat-base	подавл.	RU→RU (н.т.)
YaGPT-instr*	подавл.	RU→RU (н.т.)
GigaChat-instr	подавл.	ENG→RU
Qwen2.5-32B	станд.	ENG→ENG (н.т.)
T-pro-2.1	станд.	RU→RU

Примечания: (н.т.) – неструктурированный текст.. YaGPT-instr в формате структурированного диалога демонстрировала патологическую сверхуверенность (P ≈ 0,999), что соответствует отказу модели формировать информативное распределение; результаты в этом формате исключены, поскольку включение вырожденных распределений исказило бы рейтинг.

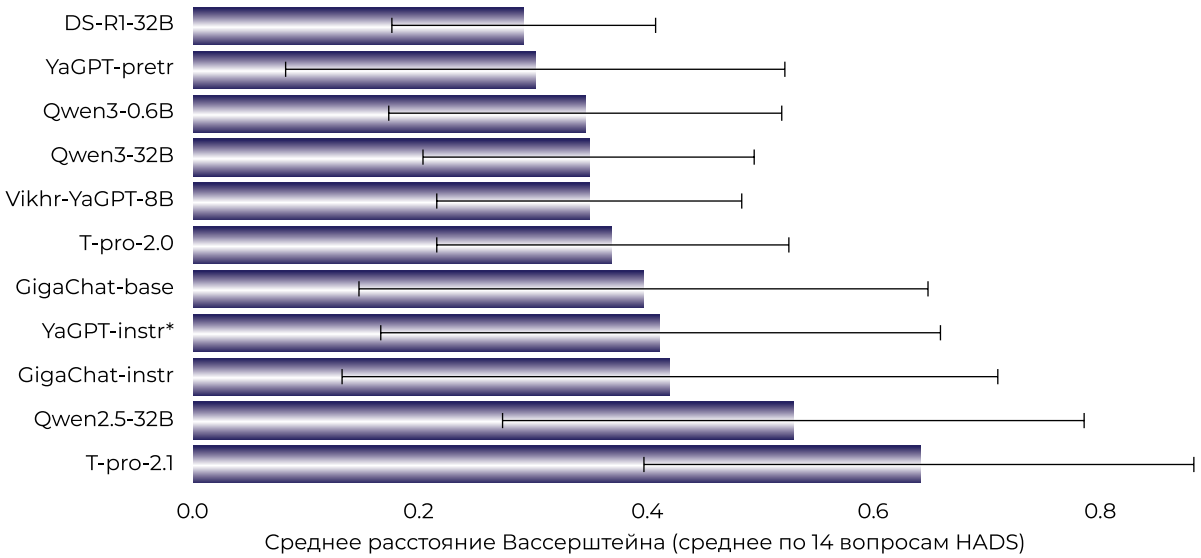


Рисунок 2 – Среднее расстояние Вассерштейна (W) для каждой модели в лучшей конфигурации. Столбцы отсортированы по возрастанию W; планки погрешностей – стандартное отклонение. Меньшие значения W соответствуют лучшему воспроизведению распределений ответов респондентов.

минимальное среднее расстояние Вассерштейна.

На рисунке 2 представлено среднее расстояние Вассерштейна для каждой модели в лучшей конфигурации, при этом лучший результат по распределениям показала DS-R1-32B (W = 0,293), худший – T-pro-2.1 (W = 0,642). DS-R1-32B также продемонстрировала наименьший разброс между вопросами (SD = 0,117), что свидетельствует о равномерном воспроизведении всех 14 вопросов.

Сравнение ожидаемых баллов HADS

На рисунке 3 можно заметить, что анализ ожидаемых баллов субшкал выявил расхождение

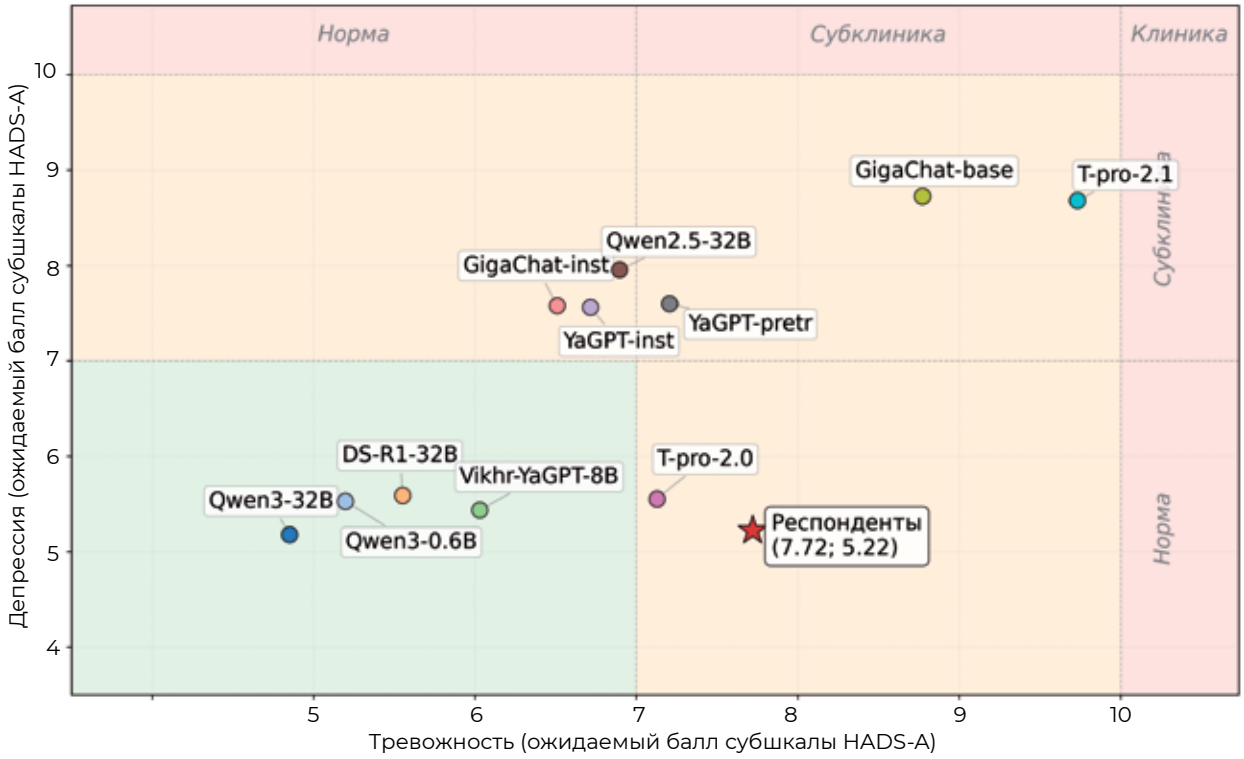


Рисунок 3 – Ожидаемые баллы тревожности (A) и депрессии (D) для каждой модели в лучшей конфигурации в сравнении с реальными средними баллами респондентов.

между двумя метриками. DS-R1-32B, лучшая по расстоянию Вассерштейна, систематически занижала ожидаемый балл тревожности: 5,55 против 7,72 у людей (разница 2,17 балла). Наиболее точные ожидаемые баллы тревожности (с наименьшим средним абсолютным отклонением от эмпирических значений) показали YaGPT-pretr (7,20; Δ = 0,52) и T-pro-2.0 (7,13; Δ = 0,59), где Δ – абсолютное отклонение от среднего балла респондентов. По расстоянию Вассерштейна обе модели занимают средние позиции.

По субшкале депрессии (среднее у людей 5,22) наименьшее отклонение от эмпирических значений выборки продемонстрировали Qwen3-32B (5,18; Δ = 0,04), Vikhr-YaGPT-8B (5,44; Δ = 0,22) и T-pro-2.0 (5,55; Δ = 0,33).

Примечательно, что T-pro-2.0 – русскоязычная модель, построенная на архитектуре Qwen3, – продемонстрировала наименьшее среднее абсолютное отклонение ожидаемых баллов от эмпирических значений выборки по совокупности обеих субшкал: ожидаемый балл тревожности 7,13 (Δ = 0,59) и депрессии 5,55 (Δ = 0,33). На рисунке 3 точка T-pro-2.0 расположена ближе всего к точке респондентов, что делает ее наиболее сбалансированной моделью для задачи воспроизведения ожидаемых баллов HADS.

Две модели систематически завышали баллы обеих субшкал: GigaChat-base (A = 8,77; D = 8,73) и T-pro-2.1 (A = 9,73; D = 8,68), – предсказывая субклинический и клинический уровни тревожности и депрессии при реальном среднем уровне на границе нормы.

Таким образом, расстояние Вассерштейна оценивает близость формы повопросных распределений, но не контролирует направление систематического смещения. Модель может хорошо воспроизводить форму ответов, при этом систематически занижая или завышая уровень симптоматики.

Интегральная оценка моделей

Для объединения двух аспектов оценки – качества воспроизведения формы

распределений (W) и точности ожидаемых баллов (близость к реальным средним A = 7,72 и D = 5,22) – была рассчитана композитная метрика C (таблица 3). Среднее абсолютное отклонение ожидаемых баллов от реальных (Score_dev) определялось как (|ΔA| + |ΔD|) / 2. Оба компонента нормализовались к диапазону [0, 1] методом min-max по наблюденным значениям среди 11 моделей, после чего суммировались с равными весами: C = 0,5 · W_norm + 0,5 · ScoreDev_norm. Меньшие значения C соответствуют лучшему результату. Равные веса выбраны как нейтральный вариант в отсутствие клинических оснований для приоритизации одного из компонентов: точность воспроизведения формы повопросных распределений и точность уровня симптоматики одинаково важны для оцениваемой задачи. Корреляция между W и Score_dev в наблюдаемых данных невысока (r ≈ 0,28), что свидетельствует о том, что компоненты отражают разные аспекты поведения модели и объединение их в одну метрику несет дополнительную информацию. Состав и параметры метрики зафиксированы до завершения анализа данных; применение min-max нормализации постфактум технически необходимо (компоненты имеют разные диапазоны значений) и не использовалось для целенаправленного изменения рейтинга.

По композитной метрике лучшей моделью оказалась T-pro-2.0 (C = 0,112), сочетающая умеренное расстояние Вассерштейна (W = 0,371) с наименьшим отклонением ожидаемых баллов от реальных (Score_dev = 0,46). DS-R1-32B (C = 0,178), лучшая по W, заняла второе место из-за существенного занижения балла тревожности. Третье и четвертое места заняли Vikhr-YaGPT-8B (C = 0,192) и YaGPT-pretr (C = 0,232). Наихудшие результаты показали T-pro-2.1 (C = 1,000) и Qwen2.5-32B (C = 0,631).

Влияние факторов

Для оценки вклада каждого фактора в качество воспроизведения распределений все 5488 значений расстояния Вассерштейна (полученных при разных комбинациях модели, промпта, формата, режима и т.д.) были проанализированы тестом Краскела-Уоллиса. Тест проверяет, отличаются ли значения W между уровнями каждого фактора (например, между 11 моделями или между 5 группами промптов). Размер эффекта η^2 используется для оценки величины влияния фактора: малый ($\eta^2 < 0,01$), средний (0,01–0,06) и большой ($\eta^2 > 0,14$) [25]. Результаты представлены в таблице 4.

Выбор модели являлся доминирующим фактором, объясняя 46,6% дисперсии расстояния Вассерштейна ($\eta^2 = 0,466$). Группа промптов и формат подачи также оказывали значимое влияние ($\eta^2 = 0,036$ и 0,027 соответственно). Пол, способ указания пола и детализация описания

Таблица 3 – Рейтинг моделей по композитной метрике

№	Модель	W	Ожид. A	Ожид. D	Score_dev	C
1	T-pro-2.0	0,371	7,13	5,55	0,46	0,112
2	DS-R1-32B	0,293	5,55	5,59	1,27	0,178
3	Vikhr-YaGPT-8B	0,351	6,03	5,44	0,96	0,192
4	YaGPT-pretr	0,303	7,20	7,60	1,45	0,232
5	Qwen3-0.6B	0,347	5,20	5,53	1,42	0,287
6	Qwen3-32B	0,350	4,85	5,18	1,46	0,300
7	YaGPT-instr	0,413	6,71	7,56	1,67	0,439
8	GigaChat-instr	0,421	6,51	7,58	1,79	0,475
9	GigaChat-base	0,398	8,77	8,73	2,28	0,550
10	Qwen2.5-32B	0,531	6,90	7,96	1,78	0,631
11	T-pro-2.1	0,642	9,73	8,68	2,74	1,000

Примечания: W – расстояние Вассерштейна; Ожид. A и D – ожидаемые баллы тревожности и депрессии (у респондентов: A = 7,72, D = 5,22); Score_dev – среднее абсолютное отклонение ожидаемых баллов от реальных ((|ΔA| + |ΔD|) / 2); C – композитная метрика (0,5 · W_norm + 0,5 · ScoreDev_norm, нормализация min-max); меньшие значения C соответствуют лучшему результату.

Таблица 4 – Влияние факторов на расстояние Вассерштейна

Фактор	H	p	η^2	k
Модель	2559,79	< 0,001	0,466	11
Группа промптов	202,26	< 0,001	0,036	5
Формат (структурированный диалог / неструктурированный текст)	147,26	< 0,001	0,027	2
Режим рассуждений	41,02	< 0,001	0,007	2
Пол (модели)	1,18	0,278	< 0,001	2
Формулировка промпта	1,46	0,691	< 0,001	4
Детализация промпта	1,41	0,965	< 0,001	7

Примечания: H – статистика Краскела-Уоллиса; k – число уровней фактора; η^2 (eta-квадрат) – размер эффекта; N = 5488.

персоны у моделей не влияли на результат ($p > 0,05$). При этом, для респондентов пол оказался статистически значимым, однако только по субшкале тревожности, а размеры эффекта оказались достаточно малыми – ($H = 17,16$; $p < 0,001$; $\eta^2 = 0,017$).

Языковые предпочтения моделей

Группа промптов являлась вторым по значимости фактором после выбора модели ($\eta^2 = 0,036$, $p < 0,001$). Анализ по комбинации модель-группа промптов (рисунок 4) выявил различие в языковых предпочтениях моделей.

Модели семейства Qwen с преимущественно англоязычным обучением показывали лучшие результаты при англоязычных промптах. Qwen2.5-32B достигла наименьшего W = 0,531 в неструктурированном тексте с англоязычным промптом – лучше, чем в формате структурированного диалога (W = 0,686–0,748), несмотря на то что модель является дообученной версией. Для Qwen3-32B: W = 0,350 при ENG→ENG против W = 0,550 при RU→RU – различие в 1,6 раза.

Среди русскоязычно-адаптированных моделей картина была неоднородной. YaGPT-instr показала лучший результат при RU→RU в неструктурированном тексте (W = 0,413), а GigaChat-base – при RU→RU (н.т.) (W = 0,398). Vikhr-YaGPT-8B формально лучше всего работала при RU→RU (н.т.) (W = 0,351), однако разница с форматом структурированного диалога была минимальной (ENG→ENG: W = 0,371), что делает эту модель практически нечувствительной к выбору формата. В то же время YaGPT-pretr, несмотря на русскоязычную ориентацию обучающих данных, достигла лучшего результата при англоязычных промптах (ENG→ENG (н.т.), W = 0,303). Возможное объяснение – качественные англоязычные представления базовой архитектуры Qwen2.5, на которой построена модель.

DS-R1-32B, также построенная на Qwen2.5, показала лучший результат при русскоязычном промпте в формате структурированного диалога (W = 0,293), что, вероятно, связано с приобретением русскоязычных представлений при дистилляции из DeepSeek-R1. Примечательно, что DS-R1-32B продемонстрировала минимальную чувствительность к выбору группы промптов: разброс W между пятью группами составил всего 0,045 (от 0,293 до 0,337), тогда как у других моделей он достигал 0,20–0,44.

GigaChat-instr и T-pro-2.0 (двуязычная и англоязычная базовые архитектуры, соответственно) показали лучшие результаты при формате ENG→RU – англоязычном системном промпте с русскоязычным вопросом опросника (W = 0,421 и 0,371). T-pro-2.1, напротив, показала неудовлетворительные результаты во всех языковых группах (W = 0,642–0,773).

Неструктурированный текст уступал шаблону диалога для большинства дообученных моделей: например, GigaChat-instr показала W = 0,569–0,659 в неструктурированном тексте против W = 0,421–0,477 в формате структурированного диалога. Исключениями стали Qwen2.5-32B (лучший результат в неструктурированном тексте) и Vikhr-YaGPT-8B, для которой разница между форматами была минимальной (W = 0,351 в неструктурированном тексте против W = 0,371 в структурированном диалоге).

Прослеживается определенная тенденция: модели с преимущественно англоязычным обучением (семейство Qwen) стабильно показывают лучшие результаты при англоязычных промптах.

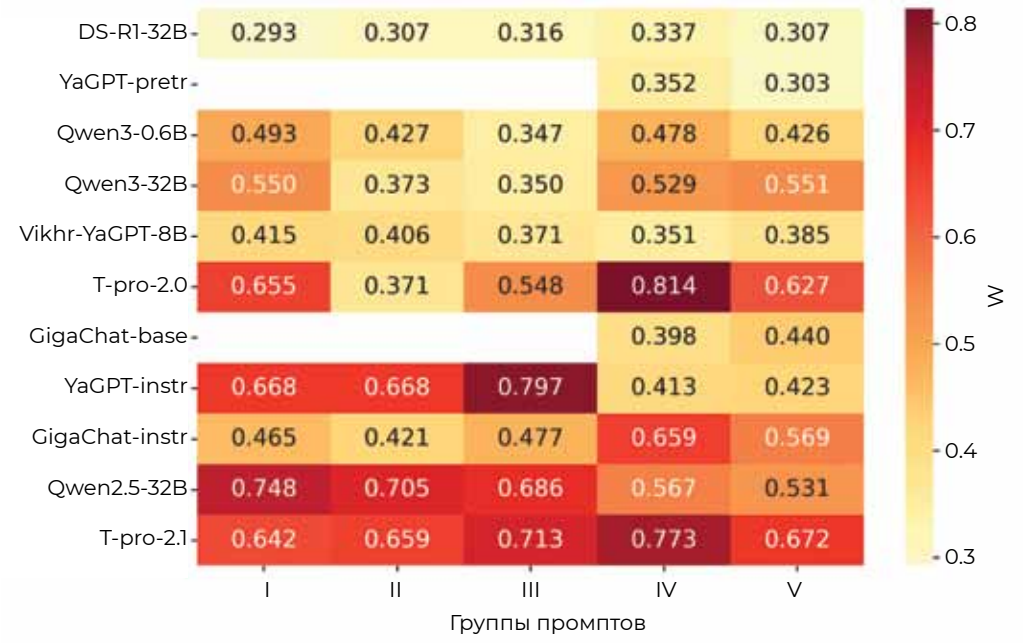


Рисунок 4 – Среднее расстояние Вассерштейна по комбинации модель-группа промптов. Группы промптов: I – системный промпт и вопрос на русском, шаблон диалога (RU→RU); II – системный промпт на английском, вопрос на русском, шаблон диалога (ENG→RU); III – системный промпт и вопрос на английском, шаблон диалога (ENG→ENG); IV – системный промпт и вопрос на русском, неструктурированный текст (RU→RU (н.т.)); V – системный промпт и вопрос на английском, неструктурированный текст (ENG→ENG (н.т.)). Для каждой комбинации модель-группа выбран лучший режим рассуждений. Темные ячейки – высокое W (худшее совпадение); пустые ячейки – предобученные модели, не поддерживающие шаблон диалога (группы I–III).

Для русскоязычных и двуязычных моделей картина неоднородна: YaGPT-instr и GigaChat-base предпочитают русскоязычные промпты, тогда как YaGPT-pretr достигла лучшего результата при англоязычном промпте, а T-pro-2.0 и GigaChat-instr – при комбинации ENG→RU (англоязычный системный промпт с русскоязычным вопросом). Это свидетельствует о том, что языковые предпочтения моделей определяются не только языком дообучения, но и базовой архитектурой, на которой построена модель.

Анализ по субшкалам

Среднее W для субшкалы депрессии составило 0,450, для тревожности – 0,354. Среди наиболее проблемных вопросов (в среднем по всем моделям) – Q12 «Удовлетворение от дел» (W = 0,595), Q14 «Удовольствие от книги/ТВ» (W = 0,595), Q10 «Внешность» (W = 0,518). Наиболее точно воспроизводились Q2 «Удовольствие от привычного» (W = 0,274) и Q13 «Паника» (W = 0,292) (рисунок 5).

ОБСУЖДЕНИЕ

В исследованном диапазоне (0,6–33 млрд параметров) для задач воспроизведения опросных распределений архитектура и состав обучающих данных модели оказались более значимыми предикторами качества симуляции, чем число параметров; полученные результаты не позволяют исключить значимость размера модели за пределами данного диапазона. Выбор модели являлся доминирующим фактором качества прогнозирования. Этот результат согласуется с ранее установленными в литературе закономерностями: наблюдениями Zhao et al. [26], зафиксировавших значительную меж-модельную

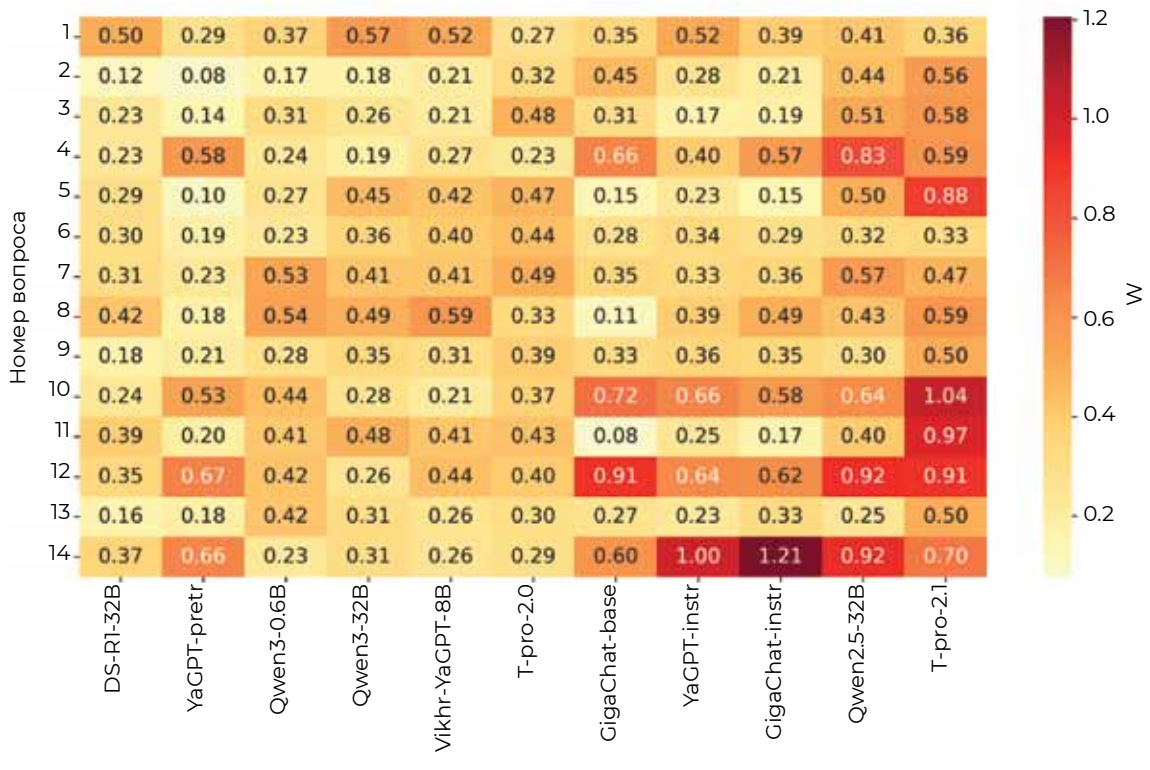


Рисунок 5 – Тепловая карта среднего расстояния Вассерштейна по 14 вопросам HADS x 11 моделей (каждая в лучшей конфигурации). Номера вопросов обозначены на оси ординат. Темные ячейки – высокое W (худшее совпадение).

вариативность при анализе данных социологических исследований. Закономерность «больше параметров – лучше симуляция» не подтвердилась в исследованном диапазоне: модели с $8x10^9$ параметров (YaGPT-pretr, Vikhr-YaGPT-8B) опередили ряд моделей с $32x10^9$ параметров. Аналогичный эффект описан в работах Sukiennik et al. [27] и Lutz et al. [28].

Гипотеза о преимуществе русскоязычных моделей не подтвердилась. В данной работе сопоставлены 11 конкретных моделей на русскоязычной HADS-выборке студентов: по расстоянию Вассерштейна лучшей оказалась DS-R1-32B – англоязычная модель, дистиллированная из DeepSeek-R1. По ожидаемым баллам тревожности наименьшее отклонение от эмпирических значений выборки продемонстрировали YaGPT-pretr и T-pro-2.0 – модели с русскоязычной адаптацией. При этом T-pro-2.0 оказалась наиболее сбалансированной моделью по совокупности обеих субшкал: она ближе всего к распределению баллов респондентов. Однако некоторые другие русскоязычные модели (T-pro-2.1, GigaChat-base) показали наихудшие результаты, что не позволяет говорить о систематическом преимуществе русскоязычных моделей. Определяющим фактором представляется не языковая принадлежность модели как таковая, а качество ее доменной адаптации – представленность целевой культуры и языка в обучающих данных и процедуры дообучения. Аналогичный вывод подтверждается и в смежных задачах: в работе [29] тонкая настройка RuBERT на русскоязычных медицинских текстах повысила метрики семантического поиска на 40–55% по сравнению с базовой моделью.

Сопоставление моделей-потомков с их базовыми архитектурами подтверждает эту интерпретацию. DS-R1-32B, дообученная на основе Qwen2.5-32B, существенно превзошла базовую модель, что указывает на приобретение русскоязычных представлений при дистилляции из DeepSeek-R1. T-pro-2, построенная на Qwen3, также продемонстрировала значительное

улучшение по сравнению с Qwen3-32B по ожидаемым баллам. Vikhr-YaGPT-8B и YaGPT-pretr, построенные на общей базе Qwen2.5, показали сопоставимые результаты (при этом Vikhr-YaGPT-8B показал более близкий результат в депрессии). Эти наблюдения свидетельствуют о том, что дообучение может как существенно улучшить (DS-R1-32B, T-pro-2.0), так и ухудшить (T-pro-2.1) качество симуляции по сравнению с базовой архитектурой.

Наши результаты свидетельствуют о том, что несоответствие языка обучения и языка промпта также является самостоятельным источником систематической ошибки, не сводимым к культурному смещению. Ряд исследований фиксирует смещение БЯМ в сторону англоязычного (преимущественно американского) контекста: Sukiennik et al. [27] показали минимальное отклонение от реальных данных для США, а Qi et al. [30] и Kabir et al. [31] обнаружили, что модели существенно хуже воспроизводят ответы неанглоязычных популяций. Rupprecht et al. [32], работавшие с немецким ALLBUS и открытыми моделями (Llama, Mistral, Qwen, Gemma), столкнулись с аналогичной проблемой недостаточной представленности немецкоязычных данных.

Полученные результаты выявляют расхождение между метриками W и ожидаемыми баллами, что имеет практическое значение для интерпретации. DS-R1-32B лучше всех воспроизводила форму повопросных распределений, но систематически занижала тревожность на 2,17 балла (28%), что при клиническом применении привело бы к ложноотрицательной оценке. T-pro-2.1 и GigaChat-base, напротив, завышали баллы обеих субшкал, предсказывая субклинический-клинический уровень при фактической норме. Для задач скрининга направленность ошибки принципиальна: занижение тревожности означает пропуск пациентов группы риска, а завышение – избыточное направление к специалистам [33].

Худшее воспроизведение субшкалы депрессии (W = 0,450 против 0,354 для тревожности) отражает ограничение БЯМ в моделировании субъективного гедонистического опыта. Дополнительным фактором может выступать более низкая внутренняя согласованность субшкалы депрессии в эталонной выборке (α = 0,62), указывающая на большую разнородность ее пунктов. Вопросы тревожности описывают психофизиологические состояния (напряжение, паника), широко представленные в текстовых корпусах. Вопросы депрессии (удовольствие от чтения, забота о внешности) отражают культурно-обусловленные нормы переживания удовольствия, для которых БЯМ не располагают достаточным контекстом [34, 35].

Полученные результаты показывают, что детализация описания персоны, пол респондента и способ указания пола не влияли на качество воспроизведения распределений ($p > 0,05$) – для корректной симуляции ответов на HADS достаточно минимального описания респондента, содержащего лишь статус студента (например, «Ты – студент первого курса российского университета»). Этот результат согласуется с ранее установленным в литературе: Hu и Collier [36], количественно показавшими, что переменные персоны объясняют менее 10% дисперсии в задачах NLP. Исключение – задачи с высокой групповой согласованностью (например, электоральные предпочтения), где вклад персоны достигает 71%. Расхождение с Giorgi et al. [37], зафиксировавшими значимое влияние явного указания демографии, вероятно, объясняется спецификой опросника: HADS измеряет клинические состояния (тревога, депрессия), менее зависимые от социально-демографического профиля, чем политические установки или потребительские предпочтения.

Проблема скошенности логарифмических вероятностей у дообученных моделей [22, 28] в настоящем исследовании решалась применением управляемого декодирования, что обеспечивало информативные распределения.

Практическое значение результатов состоит в возможности использования локально развернутых БЯМ для пилотирования скрининговых инструментов в российском здравоохранении. Работа в закрытом контуре обеспечивает конфиденциальность данных, что критично для медицинских учреждений. Предполагаемым пользователем может являться исследователь или клинический психолог, адаптирующий стандартизированный опросник к новой популяции (иная возрастная группа, профессиональная или региональная выборка). С помощью симуляции БЯМ специалист может предварительно оценить повопросные распределения ответов до проведения полноценного эмпирического исследования. Это позволяет выявить потенциально проблемные пункты и обоснованно сузить объем пилотного

исследования. Преимущество перед традиционными процедурами – возможность получить ориентировочные оценки без сбора первичных данных, что особенно ценно при ограниченных ресурсах или при необходимости быстрой проверки гипотезы о применимости инструмента. Предложенный подход дополняет, но не заменяет традиционную психометрическую валидацию, которая остается обязательной для клинического использования.

Следует подчеркнуть, что модели не ставят диагноз и не заменяют клиническую оценку специалиста. Настоящее исследование оценивает прогностические возможности БЯМ как инструмента сравнительного анализа, а не их диагностическую точность.

ОГРАНИЧЕНИЯ ИССЛЕДОВАНИЯ

Эталонная выборка ограничена одной группой (студенты-первокурсники), что не позволяет оценить обобщаемость на другие популяции. Невозможно исключить контаминацию: HADS и его результаты могут присутствовать в обучающих корпусах моделей, что затрудняет разграничение между «пониманием» и «запоминанием» опросника. Модели отвечают на каждый вопрос независимо, не воспроизводя корреляционную структуру между вопросами, что исключает классификационный анализ по порогам HADS на индивидуальном уровне. Исследование ограничено моделями с открытыми весами; результаты могут отличаться для закрытых коммерческих моделей (GPT-4, Claude и др.).

Сравнение дообученных и предобученных моделей частично смешано с форматом подачи инструкции: дообученные модели оценивались преимущественно в формате структурированного диалога, предобученные – в формате неструктурированного текста. Полностью разграничить эффект архитектуры и эффект формата в рамках данного дизайна невозможно; внутри каждого класса моделей (дообученные и предобученные) сравнения остаются сопоставимыми [38].

ВЫВОДЫ

Результаты исследования позволяют сформулировать два основных практических вывода.

Во-первых, стабильность оценок относительно вариаций промпта. Детализация описания персоны, формулировка промпта и указание пола респондента не влияют на качество воспроизведения распределений. Для корректной симуляции ответов на HADS достаточно минимального описания, содержащего лишь статус студента. Это упрощает практическое применение метода, поскольку не требует тщательного конструирования персоны.

Во-вторых, влияние языка обучения модели на результат. Модели с преимущественно англоязычным обучением (семейство Qwen) стабильно показывают лучшие результаты при англоязычных промптах. Для русскоязычных и двуязычных моделей языковые предпочтения определяются не только языком дообучения, но и базовой архитектурой: так, YaGPT-pretr достигла лучшего результата при англоязычном промпте, а T-pro-2.0 – при комбинации англоязычного системного промпта с русскоязычным вопросом. При этом T-pro-2.0 оказалась наиболее сбалансированной моделью по совокупности ожидаемых баллов обеих субшкал.

Гипотеза о систематическом преимуществе русскоязычных моделей не подтвердилась: лучшая модель по расстоянию Вассерштейна (DS-R1-32B) является англоязычной, а среди русскоязычных моделей наряду с успешными (T-pro-2.0, YaGPT-pretr) присутствуют модели с наихудшими результатами (T-pro-2.1, GigaChat-base). Качество симуляции определяется доменной адаптацией, а не языковой принадлежностью модели как таковой.

Субшкала депрессии воспроизводится хуже субшкалы тревожности. Расхождение между метриками (расстояние Вассерштейна и ожидаемые баллы) указывает на необходимость многомерной оценки: модель может точно воспроизводить форму распределений, но систематически искажать уровень симптоматики.

Метод анализа логарифмических вероятностей потенциально применим к другим стандартизированным опросникам с фиксированными вариантами ответов (PHQ-9, GAD-7, SF-36 и др.), однако данное обобщение в настоящем исследовании непосредственно не проверялось и требует самостоятельной валидации. Локальное развертывание моделей с открытыми весами обеспечивает конфиденциальность данных, что перспективно для российского здравоохранения. Вместе с тем для практического использования синтетических респондентов необходима

обязательная валидация результатов на эксперименте с реальными людьми. Модели не заменяют клиническую оценку специалиста и могут рассматриваться только как вспомогательный инструмент предварительного анализа.

Конфликт интересов

Авторы заявляют об отсутствии конфликта интересов.

Источники финансирования

Статья подготовлена в рамках выполнения научно-исследовательской работы № FSSW-2026-0008 «Развитие фундаментальных основ построения цифровых двойников инфраструктурных систем в экономике данных» за счет средств субсидии из федерального бюджета в рамках государственного задания Минобрнауки России.

ЛИТЕРАТУРА/REFERENCES

1. Snaith RP, Zigmond AS. The Hospital Anxiety and Depression Scale. Br Med J (Clin Res Ed). 1986; 292(6516): 344.
2. Herrero MJ, Blanch J, Peri JM, et al. A validation study of the Hospital Anxiety and Depression Scale (HADS) in a Spanish population. Gen Hosp Psychiatry. 2003; 25(4): 277-283. doi: 10.1016/s0163-8343(03)00043-4.
3. Tiksnadi BB, Triani N, Fihaya, FY, et al. Validation of Hospital Anxiety and Depression Scale in an Indonesian population: a scale adaptation study. Family Medicine and Community Health. 2023; 11; e001775. doi: 10.1136/fmch-2022-001775.
4. Michopoulos I, Douzenis A, Kalkavoura C, et al. Hospital Anxiety and Depression Scale (HADS): validation in a Greek general hospital sample. Ann Gen Psychiatry. 2008; 7: 4. doi: 10.1186/1744-859X-7-4.
5. Chan YF, Leung DY, Fong DY, et al. Psychometric evaluation of the Hospital Anxiety and Depression Scale in a large community sample of adolescents in Hong Kong. Qual Life Res. 2010; 19(6): 865-873. doi: 10.1007/s11136-010-9645-1.
6. Mumford DB, Tareen IA, Bajwa MA, Bhatti MR, Karim R. The translation and evaluation of an Urdu version of the Hospital Anxiety and Depression Scale. Acta Psychiatr Scand. 1991; 83(2): 81-85. doi: 10.1111/j.1600-0447.1991.tb07370.x.
7. Marakshina JA, Pavlova AA, Ismatullina VI, et al. Psychometric properties analysis of the Hospital Anxiety and Depression Scale (HADS) on a sample of Russian-speaking students. Acta Biomed Sci. 2025; 10(5): 182-193. doi: 10.29413/ABS.2025-10.5.20.
8. Triantafyllidis A, Segkouli S, Kokkas S, et al. Large language models for cardiovascular disease, cancer, and mental disorders: a review of systematic reviews. Healthcare (Basel). 2026; 14(1): 45. doi: 10.3390/healthcare14010045.
9. Ghorbian M, Ghobaei-Arani M. Large language models for mental health diagnosis and treatment: a survey. Artif Intell Rev. 2026; 59: 9. doi: 10.1007/s10462-025-11418-0.
10. Kuzmin G, Strepetov P, Stankevich M, et al. Exploring large language models for detecting mental disorders. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. 2025: 34535-34559. doi: 10.18653/v1/2025.emnlp-main.1752.
11. Joshi P, Santy S, Budhiraja A, et al. The state and fate of linguistic diversity and inclusion in the NLP world. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 6282-6293. doi: 10.18653/v1/2020.acl-main.560.
12. Hedderich MA, Lange L, Adel H, Strotgen J, Klakow D. A survey on recent approaches for natural language processing in low-resource scenarios. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021: 2545-2568. doi: 10.18653/v1/2021.naacl-main.201.
13. Localizing AI in the Global South. Nat Mach Intell. 2025; 7: 675. doi: 0.1038/s42256-025-01057-z.
14. Gromenko E, Kalacheva D, Klokova K, et al. Cultural evaluation of LLMs in Russian: catchphrases and cultural types. In: Proceedings of the International Conference "Dialogue". 2025.
15. Grigoreva V, Ivanova A, Alimova I, et al. RuBia: a Russian language bias detection dataset. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). 2024: 14227-14239.
16. Назаров Д.М., Бадаев Ф.И. Применение больших языковых моделей в сфере здравоохранения // Менеджер здравоохранения. – 2025. – №5. – С.142-154. [Nazarov DM, Badaev FI. Primenenie bol'shikh yazykovykh modelei v sfere zdravookhraneniya. Menedzher zdravookhraneniya. 2025; 5: 142-154. (In Russ.)] doi: 10.21045/1811-0185-2025-5-142-154.
17. Zigmond AS, Snaith RP. The Hospital Anxiety and Depression Scale. Acta Psychiatr Scand. 1983; 67(6): 361-370. doi: 10.1111/j.1600-0447.1983.tb09716.x.

18. Агапов М.А., Скалин Ю.Е., Бочарова Н.Н. Скрининговая диагностика психоэмоционального состояния обучающихся в системе высшего образования // Вестник Балтийского федерального университета им. И. Канта. Серия: Филология, педагогика, психология. – 2023. – №1. – С.90-101. [Agapov MA, Skalin YuE, Bocharova NN. Skriningovaya diagnostika psikhoemotsional'nogo sostoyaniya obuchayushchikhsya v sisteme vysshego obrazovaniya. Vestnik Baltiyskogo federal'nogo universiteta im. I. Kanta. Seriya: Filologiya, pedagogika, psikhologiya. 2023; 1: 90-101. (In Russ.)] doi: 10.5922/pikbfu-2023-1-9.
19. Kwon W, Li Z, Zhuang S, et al. Efficient memory management for large language model serving with PagedAttention. In: Proceedings of the 29th Symposium on Operating Systems Principles. 2023. doi: 10.1145/3600006.3613165.
20. Structured outputs. vLLM documentation v0.8.2. 2025. Available from: https://docs.vllm.ai/en/v0.8.2/features/structured_outputs.html. Cited February 25, 2026.
21. Givens CR, Shortt RM. A class of Wasserstein metrics for probability distributions. Michigan Math J. 1984; 31(2): 231-240. doi: 10.1307/mmj/1029003026.
22. Kim S, Jeong J, Han JS, et al. LLM-Mirror: a generated-persona approach for survey pre-testing. Preprint. 2024. doi: 10.48550/arXiv.2412.03162.
23. Moon S, Abdulhai M, Kang M, et al. Virtual personas for language models via an anthology of backstories. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024: 19864-19897. doi: 10.18653/v1/2024.emnlp-main.1110.
24. Santurkar S, Durmus E, Ladhak F, et al. Whose opinions do language models reflect? In: Proceedings of the 40th International Conference on Machine Learning (ICML). 2023. doi: 10.48550/arXiv.2303.17548.
25. Tomczak M, Tomczak E. The need to report effect size estimates revisited: an overview of some recommended measures of effect size. Trends Sport Sci. 2014; 1(21): 19-25.
26. Zhao J, Yuan C, Luo W, et al. Large language models as virtual survey respondents: evaluating sociodemographic response generation. Preprint. 2025. doi: 10.48550/arXiv.2509.06337.
27. Sukiennik N, Gao C, Xu F, Li Y. An evaluation of cultural value alignment in LLM. Preprint. 2025. doi: 10.48550/arXiv.2504.08863.
28. Lutz M, Sen I, Ahnert G, Rogers E, Strohmaier M. The prompt makes the person(a): a systematic evaluation of sociodemographic persona prompting for large language models. In: Findings of the Association for Computational Linguistics: EMNLP 2025. 2025: 23212-23237. doi: 10.18653/v1/2025.findings-emnlp.1261.
29. Каширина И.Л., Старичкова Ю.В., Ле Ч.К. Тонкая настройка языковой модели RuBERT для повышения точности анализа медицинских запросов // Врач и информационные технологии. – 2026. – №1. – С.64-73. [Kashirina IL, Starichkova YuV, Le ChK. Tonkaya nastroika yazykovoi modeli RuBERT dlya povysheniya tochnosti analiza meditsinskikh zaprosov. Vrach i informatsionnye tekhnologii. 2026; 1: 64-73. (In Russ.)] doi: 10.25881/18110193_2026_1_64.
30. Qi W, Huang F, An J, Kwak H. A cross-cultural comparison of LLM-based public opinion simulation. Preprint. 2025. doi: 10.48550/arXiv.2506.21587.
31. Kabir M, Abrar A, Ananiadou S. Break the checkbox: challenging closed-style evaluations of cultural alignment in LLMs. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. 2025: 24-51. doi: 10.18653/v1/2025.emnlp-main.2.
32. Rupprecht J, Froehling L, Wagner C, Strohmaier M. German General Social Survey personas: a survey-derived persona prompt collection for population-aligned LLM studies. In: Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026). 2026: 1761-1780. doi: 10.63317/2sod6uekicbg.
33. Bjelland I, Dahl AA, Haug TT, et al. The validity of the Hospital Anxiety and Depression Scale: an updated literature review. J Psychosom Res. 2002; 52(2): 69-77. doi: 10.1016/S0022-3999(01)00296-3.
34. Petrov NB, Serapio-Garcia G, Rentfrow J. Limited ability of LLMs to simulate human psychological behaviours: a psychometric analysis. Preprint. 2024. doi: 10.48550/arXiv.2405.07248.
35. Teutloff JK. Synthetic founders: AI-generated social simulations for startup validation research in computational social science. Preprint. 2025. doi: 10.48550/arXiv.2509.02605.
36. Hu T, Collier N. Quantifying the persona effect in LLM simulations. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024: 10289-10307. doi: 10.18653/v1/2024.acl-long.554.
37. Giorgi S, Liu T, Aich A, et al. Modeling human subjectivity in LLMs using explicit and implicit human factors in personas. In: Findings of the Association for Computational Linguistics: EMNLP 2024. 2024: 7174-7188. doi: 10.18653/v1/2024.findings-emnlp.420.
38. Sclar M, Choi Y, Tsvetkov Y, Suhr A. Quantifying language models' sensitivity to spurious features in prompt design or: how I learned to start worrying about prompt formatting. In: Proceedings of the Twelfth International Conference on Learning Representations (ICLR). 2024. doi: 10.48550/arXiv.2310.11324.