Экспертная оценка ответов больших генеративных моделей на запросы врачей в клинической практике*

И.А. Тыров[I], Ю.А. Васильев[II], А.В. Владзимирский[II,III], О.В. Омелянская[II,IV], Т.А. Березнер[II,V], Ю.В. Соболева[II], А.А. Яфарова[II], М.Д. Варюхина[II], Ю.Ф. Шумская[II], Д.А. Ахмедзянова[II], К.Ю. Беженова[II], И.А. Блохин[II], Р.В. Решетников[II]

[I] Департамент здравоохранения г. Москвы, Россия
[II] Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения г. Москвы, Россия
[III] Первый Московский государственный медицинский университет им. И. М. Сеченова, г. Москва, Россия
[IV] Российский технический университет «МИРЭА», г. Москва, Россия
[V] Национальный исследовательский университет «Высшая школа экономики», г. Москва, Россия

Аннотация. Использование больших генеративных моделей (БГМ) в здравоохранении в настоящее время вызывает значительный интерес. Проведено сравнение БГМ GigaChat и YandexGPT. Выборка 1200 ответов БГМ на запросы врачей из медицинских организаций, оказывающих помощь в амбулаторных и стационарных условиях, по семи тематикам оценивалась 13 врачами-экспертами по пяти пятибалльным критериям: релевантность, правильность, безопасность, полнота, языковая грамотность. Анализ выполнен с использованием парного t-критерия с оценкой размера эффекта. Наблюдались разнонаправленные различия между моделями в качестве ответов, в зависимости от критерия оценки и тематики запроса. Отмечено неудовлетворительное качество ответов по критериям безопасности и правильности по ряду тематик (GigaChat «Лекарственные средства» 3,1±1,5 правильность, 3,3±1,6 безопасность, YandexGPT «Параметры и нормы методов исследования» 3,3±1,5 для обоих критериев). Таким образом, БГМ в настоящее время не готовы выполнять функцию ассистента врача.

Ключевые слова: *искусственный интеллект, медицинский чат-бот, большие генеративные модели, GigaChat, YandexGPT, экспертная оценка LLM.*
DOI: 10.14357/20790279260310 **EDN:** EUAPVH

Введение

В настоящее время наблюдается очевидный тренд на внедрение технологий искусственного интеллекта (ИИ) в системы здравоохранения во всем мире [1]. Эта тенденция справедлива в том числе для Российской Федерации [2-4]. По состоянию на апрель 2026 года в РФ зарегистрировано 56 медицинских изделий с технологиями ИИ, включающих ИИ-решения для анализа медицинских изображений, анализа электронных медицинских карт, в том числе сервисы поддержки принятия врачебных решений[1].

Использование технологий генеративного ИИ (большие генеративные модели, БГМ) в медицине в настоящее время вызывает повышенный интерес. Zhihong и соавторы проанализировали 415 исследований и показали, что чаще всего в здравоохранении используются БГМ ChatGPT, GPT-3.5 и GPT-4, в основном для помощи в принятии врачебных решений и для предоставления информации пациенту [5]. Однако авторы осторожны в своих выводах: они отмечают большой потенциал в развитии БГМ в медицине, но демонстрируют и существенные ограничения. В частности, в большинстве работ оцени-

* Исследование выполнено в рамках научно-исследовательской работы «Тестирование и контроль качества больших языковых моделей, интегрированных в ЕМИАС в условиях практического здравоохранения».

[1] Перечень https://portal.egisz.rosminzdrav.ru/materials/4873.

вают точность и правильность ответов, уделяя мало внимания оценке их безопасности.

Схожей оценки преждевременности внедрения придерживаются также в работе [6], в которой исследователи рассмотрели более сотни научных статей и пришли к выводу, что в подавляющем большинстве случаев БГМ успешно используют только для извлечения данных из различных источников (почти 70 % случаев). В то же время подготовка с помощью БГМ клинических рекомендаций или, например, внедрение в чат-бот для общения с пациентами изучаются крайне редко (11,5 и 7%, соответственно). Что касается готовности к внедрению БГМ и технологий на их основе в практику врача, то лишь в 25 % случаев модели были признаны пригодными к внедрению, в то время как в 70 % случаев требуется их доработка или проверка работоспособности.

Если говорить о российской действительности, то отмечается недостаточность медицинских данных для обучения моделей на русском языке и отсутствие четкой нормативно-правовой базы, регулирующей применение БГМ [7]. При этом Щетинин и соавторы видят ряд перспектив для внедрения БГМ, в том числе интеграцию в Единую медицинскую информационно-аналитическую систему (ЕМИАС), разработку собственных нейросетей, а также, что особенно важно, специальных методов оценки качества работы БГМ в клинических условиях.

В 2025 году Департамент здравоохранения г. Москвы (ДЗМ) инициировал пилотный проект, который был предназначен для оценки возможности реализации этих перспектив. В контур ЕМИАС ряда медицинских организаций ДЗМ для оказания поддержки медицинским сотрудникам был интегрирован чат-бот на основе БГМ, включающий в себя решения от ведущих российских разработчиков. Была разработана и апробирована методология экспертной оценки подобных ассистентов врача [8]. В рамках пилота врачи могли пользоваться услугами чат-бота по запросам на любые темы, в том числе касающиеся их профессиональной деятельности.

Цель исследования: представить промежуточные результаты пилотного проекта, проведя сравнительную оценку качества работы БГМ-ассистентов врача в клинических условиях.

1. Методология

В настоящем исследовании использовали проприетарные БГМ с закрытым кодом GigaChat (версия 4, Сбер, Россия) и YandexGPT (версия 5.1 Pro, Яндекс, Россия). Программное обеспечение (ПО) GigaChat было предварительно дообучено разработчиками для оказания поддержки врачам при решении профессиональных задач. ПО YandexGPT предварительного дообучения не проходило, но имело доступ к коллекции медицинской документации, актуальной для ДЗМ. Взаимодействие с БГМ-ассистентами осуществлялось посредством ПО ЕМИАС.Бот (версия 1.2.0., информационно-аналитический центр ДЗМ, Россия), встроенного в интерфейс ЕМИАС для подключенных к пилотному проекту медицинских организаций (МО) в виде отдельного окна.

Запросы к БГМ-ассистентам, выступавшим в роли медицинских справочников, создавались врачами подключенных к пилотному проекту медицинских организаций (Сопроводительная информация, Таблица S1[2]) в рамках выполнения своей профессиональной деятельности в индивидуальных сессиях взаимодействия с ПО ЕМИАС. Бот. Пользователи БГМ-ассистентов не были ограничены в том, как именно формулировать свои запросы. При составлении выборки исследования сохранялись орфография и пунктуация оригинальных запросов.

Критерии включения для запросов:

- дата выполнения соответствует целевому временному периоду;
- формулировка запроса позволяет однозначно определить содержание ожидаемого ответа БГМ-ассистентов.

Критерии невключения для запросов:

1) запросы с неоднозначной трактовкой (например, для запроса «отривин», ввиду его неопределенности, будет правильным любой из вариантов ответа: действующее вещество, производитель, фармакологическое действие, показания, противопоказания, а также любая из комбинаций этих вариантов);
2) индивидуальные вопросы из цепочек запросов (prompt chains), а также сами цепочки запросов;
3) запросы от сотрудников медицинских организаций ДЗМ, оказывающих помощь детскому населению;
4) запросы с ответами БГМ-ассистентов, которые свидетельствовали об ошибке ПО при обработке данных (например, «Сервис недоступен, повторите запрос позже»).

В настоящем исследовании не рассматривали сценарий работы БГМ-ассистентов непосредственно с ЭМК-пациентов.

[2] Здесь и далее для таблиц S см. сопроводительные материалы по ссылке https://disk.yandex.ru/i/c1PgPHuHCZ-aPg

Запросы пользователей, вошедшие в выборку исследования, были сделаны в период с 16 мая по 20 июля 2025 года. Из общей выгрузки по организациям, оказывающим медицинскую помощь взрослому населению, методом случайной выборки отобрали 90 000 запросов, из которых формировали финальную выборку исследования методом стратифицированной случайной выборки для обеспечения соответствия количественного распределения тем запросов распределению в генеральной совокупности [9]. При формировании финальной выборки создавали отдельные подвыборки для запросов сотрудников МО, работающих в амбулаторных и стационарных условиях.

Финальную выборку формировали консенсусным пересмотром каждого запроса с определением темы (Таблица S2) и соответствия критериям включения и невключения. Пересмотр производили 4 врача-исследователя разных медицинских специальностей (гастроэнтерология, терапия, нейрохирургия, рентгенология) с опытом работы по медицинской специальности более трех лет, а также опытом научной работы в области медицинского ИИ более трех лет.

Оценку ответов моделей проводили методом экспертной оценки посредством опросника Э-5 (Сопроводительные материалы, Таблица S3) [10]. Каждую пару «запрос пользователя – ответ БГМ» независимо оценивали пять врачей-специалистов с опытом работы более 5 лет. Всего в оценке приняли участие 13 врачей, среди которых было 2 кандидата медицинских наук и 1 доктор медицинских наук, средний стаж работы составил 13 лет. Медицинские специальности врачей, проводивших оценку: хирургия, оториноларингология, аллергология и иммунология, кардиология, функциональная диагностика, рентгенология, нефрология, анестезиология-реаниматология, ревматология и терапия. Отбор запросов и оценку ответов проводили разные специалисты для обеспечения максимально возможной объективной оценки. В тех же целях специалистам, проводившим оценку ответов БГМ, не раскрывали название модели.

Длину запроса в словах определяли с помощью формулы Excel, подсчитывающей число пробелов в запросе и добавляющей к этому значению единицу. Перед подсчетом текст запроса очищали от лишних пробелов.

Размер выборки исследования был определен с использованием ПО G Power 3.1.9.7 [11]. Исходя из предварительных оценок, согласно которым для детекции статистически значимого различия между двумя оценками, не превышающего 0,2, при условии среднеквадратичного отклонения обеих оценок, не превышающего 0,8, при уровне статистической значимости 0,05 и статистической мощности 0,8, минимально приемлемым значением является 253 пары «запрос врача – ответ БГМ». Минимально приемлемое количество независимых оценок каждой пары «запрос врача – ответ БГМ» определяли по методологии [12], согласно которой для оценки ответов БГМ в здравоохранении в исследовательском или образовательном контексте должно быть не менее 4 эксперта.

Для описательной статистики показатели представлены в виде абсолютных и относительных значений, среднего значения (среднеквадратичного отклонения) либо медианы (95 % доверительного интервала), в зависимости от типа данных. Значения по пятибалльной шкале Ликерта были интерпретированы как переменные непрерывного типа [13]. Согласие между экспертными оценками вычисляли по методологии, предложенной Kottner J. и соавторы [14]. Значения 95 % доверительного интервала для согласия между экспертами вычисляли в соответствии с методологией [15]. Сравнение показателей проводили проверкой парного t-критерия с поправкой Бенджамини – Хохберга для множественных сравнений. Проверка допущения о нормальности распределения выборочных средних значений не проводилась исходя из накопленных доказательств об устойчивости t-критерия к нарушению допущений о типе и характере распределения данных [16,17]. Для каждого сравнения определяли размер эффекта с вычислением показателя d Коэна. Использовали граничные значения для интерпретации размера эффекта $d \leq 0{,}2$ (малый размер), $0{,}2 < d \leq 0{,}5$ (средний размер), $0{,}5 < d \leq 0{,}8$ (большой размер), $d > 0{,}8$ (очень большой размер). Все вычисления проводили в среде R версии 4.5.0 [18].

2. Результаты

Исходная выборка состояла из 90 000 запросов, совершенных 737 уникальными пользователями (сотрудниками МО, имеющими свой отдельный аккаунт в системе) 14 МО, для каждого из которых были получены ответы обеих моделей. Запросы в исходной выборке совершали врачи 22 разных специальностей, большая часть специалистов представлена врачами-терапевтами участковыми и врачами общей практики (Таблица S4).

Итоговая выборка составила 1200 пар «запрос врача – ответ БГМ-ассистентов» (по 600 пар от организаций, оказывающих медицинскую помощь в амбулаторных или стационарных услови-

Табл. 1

Распределение вопросов по темам в сценарии «Поиск по базе данных», %

Тема	МО, оказывающие медицинскую помощь в амбулаторных условиях	МО, оказывающие медицинскую помощь в стационарных условиях
Нормативно-правовая и техническая информация	19 (6,3)	89 (29,7)
Лекарственные средства	87 (29)	39 (13)
Терминология диагноза	35 (11,7)	29 (9,7)
Информация о заболевании	87 (29)	85 (28,3)
Общемедицинские рекомендации	20 (6,7)	17 (5,7)
Параметры и нормы методов исследования	39 (13)	11 (3,7)
Определения и общая информация	13 (4,3)	30 (10)

*Указано число уникальных запросов. Для каждого запроса были получены ответы двух БГМ, таким образом, число оцениваемых пар составило 1200.

ях, с равным распределением ответов от моделей GigaChat и YandexGPT). Наиболее частой темой запросов была «Информация о заболевании» (29 и 28,3% запросов для сотрудников МО, оказывающих медицинскую помощь в амбулаторных и в стационарных условиях, соответственно), наиболее редкой – «Общемедицинские рекомендации» (6,7 и 5,7 %, соответственно) (табл. 1).

Общее сравнение. Сравнение по общей оценке ПО компаний Сбер (4,19±1,26) и Яндекс (4,13±1,29) показало, что между ними нет значимых различий в общей оценке их работы (t = 0,82; p-value = 0,41; d Коэна = 0,047 – малый размер эффекта).

При сравнении работы моделей по пяти доменам опросника Э-5 статистически значимые различия наблюдались только для домена «Языковая грамотность» (табл. 2). Отметим, что более низкие оценки ПО YandexGPT по этому домену ассоциированы с его регулярно повторяемыми рекомендациями обратиться к врачу: эксперты «штрафовали» модель, поскольку БГМ-ассистент предназначен именно для врачей.

Если рассматривать различия между ПО на уровне тем запросов (см. табл. S2), то обнаруживаются значимые различия для темы «Параметры и нормы методов исследования» в пользу ответов ПО GigaChat со средним или большим размером эффекта по доменам «Правильность» (4,11±1,10 против 3,29±1,47, p<0,01, d=0,63), «Безопасность» (4,29±1,07 против 3,34±1,47, p<0,01, d=0,74), «Полнота» (4,18±1,20 против 3,30±1,49, p<0,01, d=0,65) и «Языковая грамотность» (4,80±0,63 против 4,24±1,21, p=0,03, d=0,63). При этом для ПО YandexGPT обнаружено значимое преимущество по домену «Безопасность» для темы «Лекарственные средства» (4,06±1,38 против 3,28±1,57, p<0,01, d=0,53) (Сопроводительная информация, табл. S5).

Сравнение качества ответов модели на запросы врачей МО, оказывающих медицинскую помощь в различных условиях, для ПО GigaChat. Для ПО GigaChat Сбер при сравнении ответов на запросы работников МО, оказывающих медицинскую помощь в различных условиях, обнаружены значимые различия по общей оценке и по всем доменам опросника, кроме релевантности. Во всех случаях работа БГМ-ассистента на вопросах стационарного кластера была оценена значимо лучше, чем на вопросах амбулаторного кластера. При этом для общей оценки и для всех доменов, кроме полноты, размер эффекта является малым (~0,2 d Коэна). В то же время для полноты обнаружен большой размер эффекта (0,7 d Коэна).

При сравнении ответов модели по индивидуальным темам статистически значимыми, после применения поправки на множественные сравнения, оказались различия по доменам правильности

Табл. 2

Сравнение работы моделей по доменам. (*) выделены статистически значимые различия

Показатель	GigaChat, среднее значение±СКО	YandexGPT, среднее значение±СКО	p-value	p-value с поправкой	d Коэна
Релевантность	4,82±0,63	4,82±0,62	0,890	0,949	-0,008
Правильность	3,61±1,41	3,58±1,44	0,761	0,949	0,018
Безопасность	3,84±1,37	3,84±1,35	0,949	0,949	-0,004
Полнота	3,94±1,35	3,87±1,38	0,375	0,937	0,051
Языковая грамотность	4,76±0,72	4,61±0,92	0,001*	0,006*	0,188

и полноты по темам «Лекарственные средства» и «Информация о заболевании». Во всех случаях работа модели с вопросами стационарного кластера была значимо лучше (Сопроводительная информация, табл. S6).

Сравнение качества ответов модели на запросы врачей МО, оказывающих медицинскую помощь в различных условиях, для ПО YandexGPT. По общей оценке ответы БГМ-ассистента на запросы врачей МО, оказывающих медицинскую помощь в стационарных условиях, вновь были оценены выше. Статистически значимые различия наблюдались для доменов полноты и языковой грамотности (Сопроводительная информация, табл. S7). Ни по одному домену, в том числе при рассмотрении на уровне тем, не наблюдали статистически значимых различий в пользу ответов на запросы врачей МО, оказывающих медицинскую помощь в амбулаторных условиях.

Сравнение запросов врачей МО, оказывающих медицинскую помощь в различных условиях. По результатам анализа длины запросов в словах было выявлено статистически значимое различие. Среднее значение длины запроса врачей МО, оказывающих медицинскую помощь в амбулаторных условиях, составляет 4,7±2,4 слов против 10,1±9,9 слов для запросов врачей МО, оказывающих медицинскую помощь в стационарных условиях (p<0,001, d=-0,745).

Сравнение ответов БГМ-ассистентов на запросы врачей МО, оказывающих медицинскую помощь в стационарных условиях. При сравнении оценок по доменам опросника единственное статистически значимое различие обнаруживается в пользу GigaChat Сбер по домену полноты (4,39±0,99 против 4,05±1,17 для GigaChat и YandexGPT соответственно), размер эффекта средний (d Коэна = 0,314) (Сопроводительная информация, таблица S8). Статистически значимые различия между оценками ответов ПО отсутствуют при рассмотрении на уровне тем, за исключением темы «Нормативно-правовая и техническая информация» по домену правильности – ПО YandexGPT отработал значимо лучше (3,81±1,13 против 3,25±1,44 у GigaChat), размер эффекта средний (d Коэна 0,433).

Сравнение ответов моделей на запросы врачей МО, оказывающих медицинскую помощь в амбулаторных условиях. Наблюдались статистически значимо более высоко оцениваемые экспертами ответы GigaChat по домену «языковая грамотность» (Сопроводительная информация, табл. S9). При рассмотрении отдельных тем запросов ответы ПО YandexGPT были оценены статистически значимо выше для темы «Лекарственные средства» (домен «Безопасность»: 4,23±1,36 против 3,06±1,65 для GigaChat, домен «Полнота»: 3,68±1,66 против 2,92±1,59 для GigaChat). Обратная ситуация наблюдалась для темы «Параметры и нормы методов исследования» (домен «Безопасность»: 4,21±1,14 против 3,29±1,56 для GigaChat и YandexGPT, соответственно; домен «Полнота»: 4,04±1,27 против 3,16±1,55 для GigaChat и YandexGPT, соответственно).

3. Обсуждение

При сравнении ответов двух БГМ-ассистентов (GigaChat и YandexGPT) на запросы врачей МО ДЗМ значимых различий по общей оценке между моделями не обнаруживается, за исключением критерия языковой грамотности. Различие по этому домену в пользу GigaChat можно объяснить более адекватной для условий применения стилистики ответов. В отличие от GigaChat, YandexGPT дополняла свои ответы формулировками о необходимости проконсультироваться с врачом перед приемом препаратов, при обнаружении симптомов и для интерпретации результатов исследований лабораторной диагностики. Учитывая, что настоящие ПО создавались как ассистенты врача, подобные рекомендации являются излишними и эксперты снижали за них оценку.

На уровне тем были выявлены различия в оценках работы ПО по некоторым доменам: в теме «Параметры и нормы методов исследований» – в пользу GigaChat, в теме «Лекарственные средства» – в пользу YandexGPT. Отметим, что при разработке БГМ-ассистентов были использованы разные подходы: GigaChat был дообучен для оказания поддержки врачам, в то время как YandexGPT не проходил дообучения, но имел доступ к коллекции релевантной медицинской документации. Таким образом, ни один из этих подходов индивидуально не является достаточным решением проблемы повышения качества ответов БГМ в области здравоохранения.

Обе БГМ демонстрировали значимо лучшее качество ответов для вопросов врачей стационарного кластера, нежели амбулаторного. Объяснить данный результат можно спецификой промптирования: для врачей МО, оказывающих медицинскую помощь в стационарных условиях, было характерно более сфокусированное и конкретное формулирование запроса (средняя длина запроса 10,1±9,9 слов против 4,7±2,4 слов для запросов врачей МО, оказывающих медицинскую помощь в амбулаторных условиях). В то же время для вра-

чей амбулаторных МО была характерна большая неопределенность в формулировках и более общие запросы, на которые ответ БГМ может с большей вероятностью оказаться неполным или не совсем верным.

Характерным примером запроса врача МО, оказывающих медицинскую помощь в стационарных условиях, является «распиши антигипертензивную терапию пациенту с давлением 190/110 с диагнозом ГБ 3ст, 3ст, риск ССО 4, тХПН, цирроз печени». Для сравнения, примеры запросов врачей МО, оказывающих медицинскую помощь в амбулаторных условиях: «холецистит симптомы», «лечение рефлюкс эзофагита». В первом случае пользователь действительно предоставил БГМ достаточный контекст для того, чтобы она могла составить полный, точный и релевантный ответ. Примеры запросов второго типа демонстрируют, что БГМ-ассистент находится в условиях большей свободы при формулировке ответа, что, по нашим результатам, ассоциируется с более низким качеством таких ответов.

В настоящей работе продемонстрировано, что две БГМ с разной архитектурой и разными подходами к обучению показывают крайне схожие результаты по метрикам исследования. Это согласуется с результатами Tiller и соавторы, проводивших сравнение моделей ChatGPT, Gemini, Grok, Meta AI и DeepSeek [19]. Моделям задавались вопросы из пяти категорий: онкологические заболевания, вакцины, стволовые клетки, питание и спорт. По общей оценке между моделями не было статистически значимых различий, а по показателю числа проблемных ответов только Grok проявила себя хуже остальных. Tiller и соавторы объясняют это спецификой данных, на которых она обучалась: они были взяты из социальных сетей, где стоит ожидать значительной доли недостоверного и ненаучного контента. При этом у всех пяти БГМ доля проблемных ответов превышала 50 %, значительное число ответов содержали галлюцинации и ссылались на источники низкого качества.

Аналогичные результаты были продемонстрированы Blasingame и соавторы [20], проводившими сравнение ответов моделей ChatGPT-4o, Gemini 2.0 Flash, Copilot, DeepSeek R1, Grok-3 друг с другом, а также с «золотым стандартом», подготовленным врачами-экспертами. Вопросы, задаваемые БГМ, касались лечения, прогноза и этиологии заболеваний, статистически значимых различий между качеством ответов обнаружено не было.

Выявленное в настоящей работе локальное преимущество БГМ-ассистентов друг перед другом по определенным метрикам было также ранее показано в работе Sumbullu и соавторы [21]. Их исследование было посвящено сравнению ChatGPT-5.2, ChatGPT-5.2 Plus, Gemini 3.0, Gemini 3.0 Plus и DeepSeek V3.2 при ответах на вопросы в области эндодонтии. Было продемонстрировано, что модель ChatGPT 5.2 показывает меньшую точность, но большую читаемость ответов, тогда как модели Gemini 3 Plus и DeepSeek V3.2 характеризовались большей согласованностью. Схожие результаты были получены также в работе Yüceer-Çetiner и соавторы [22].

Настоящее исследование имеет ограничения. В него были включены только два решения, оба от российских разработчиков. Ответы БГМ не сравнивали с результатами работы лидирующих решений в области, таких как ChatGPT, Gemini и DeepSeek, что не позволяет включать GigaChat и YandexGPT в общий пул научных результатов в области оценки качества работы БГМ в здравоохранении. С другой стороны, условия проведения пилотного проекта не предполагают возможности использования зарубежных моделей. Тем не менее в будущем мы планируем провести дополнительное исследование, в котором сравним ответы отечественных и зарубежных БГМ на запросы врачей, не содержащие рисков для персональных данных пациентов.

Заключение

Результаты настоящего исследования показывают, что БГМ, обладая высоким потенциалом для интеграции в клиническую практику в качестве ассистента врача, не готовы выполнять эту функцию на текущем уровне развития. Ключевые проблемы заключаются в правильности и безопасности ответов таких моделей, которые по ряду тематик могут опускаться до оценки на уровне 3,0-3,3 по пятибалльной шкале, создавая существенные риски как для пациентов, так и для врачей. Возможным решением этих проблем является объединение подходов к дообучению моделей, проработка запросов по каждой из тем, а также внедрение постоянного мониторинга таких инструментов с предоставлением обратной связи разработчикам.

Литература

1. Aravazhi P.S. et al. The integration of artificial intelligence into clinical medicine: trends, challenges, and future directions. // Disease-a-Month. 2025. Vol. 71. No. 6. P. 101882.
2. Указ Президента РФ от 10.10.2019 № 490 (ред. от 15.02.2024) «О развитии искусственного

интеллекта в Российской Федерации» (вместе с «Национальной стратегией развития искусственного интеллекта на период до 2030 года»).
3. Указ Президента РФ от 08.12.2025 № 896 «О Стратегии развития здравоохранения в Российской Федерации на период до 2030 года».
4. Распоряжение Правительства Российской Федерации от 17.04.2024 г. № 959-р об утверждении стратегического направления в области цифровой трансформации здравоохранения.
5. Zhihong Z. et al. Advancing healthcare with large language models: A scoping review of applications and future directions. // International Journal of Medical Informatics, 2025. P. 106231.
6. Liang E.N. et al. Clinical applications of large language models in medicine and surgery: A scoping review. // Journal of International Medical Research. 2025. Vol. 53. No. 7. P. 03000605251347556.
7. Shchetinin E.Y. et al. Methods for developing and implementing large language models in healthcare: challenges and prospects in Russia. // Discrete and Continuous Models and Applied Computational Science. 2025. Vol. 33. No. 3. P. 327–344.
8. Решетников Р.В. и др. Методики оценки качества больших генеративных моделей для базовых сценариев применения в здравоохранении. // Врач и информационные технологии. 2025. № 3. С. 64–76.
9. Suresh K., Thomas S.V., Suresh G. Design, data analysis and sampling techniques for clinical research. // Annals of Indian Academy of Neurology. 2011. Vol. 14. No. 4. P. 287–290.
10. Васильев Ю.А. и др. Создание и валидация опросника ERA-GMA для оценки качества работы больших генеративных моделей в качестве сервиса предоставления ответа на вопросы врача-клинициста. // Национальное здравоохранение. 2026. В печати.
11. Faul F. et al. Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. // Behavior Research Methods. 2009. Vol. 41. No. 4. P. 1149–1160.
12. Tam T.Y.C. et al. A framework for human evaluation of large language models in healthcare derived from literature review. // NPJ Digital Medicine. 2024. Vol. 7. No. 1. P. 258.
13. Huh I., Gim J. Exploration of Likert scale in terms of continuous variable with parametric statistical methods. // BMC Medical Research Methodology. 2025. Vol. 25. No. 1. P. 218.
14. Kottner J. et al. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. // International Journal of Nursing Studies. 2011. Vol. 48. No. 6. P. 661–671.
15. de Vet H.C.W., Dikmans R.E., Eekhout I. Specific agreement on dichotomous outcomes can be calculated for more than two raters. // Journal of Clinical Epidemiology. 2017. Vol. 83. P. 85–89.
16. Heeren T., D'Agostino R. Robustness of the two independent samples t-test when applied to ordinal scaled data. // Statistics in Medicine. 1987. Vol. 6. No. 1. P. 79–90.
17. Posten H.O. Robustness of the two-sample t-test. // Robustness of statistical methods and nonparametric statistics. Dordrecht: Springer Netherlands, 1984. P. 92–99.
18. R Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2020.
19. Tiller N.B. et al. Generative artificial intelligence-driven chatbots and medical misinformation: an accuracy, referencing and readability audit. // BMJ Open. 2026. Vol. 16. No. 4. P. e112695.
20. Blasingame M.N. et al. Comparing five generative AI chatbots' answers to LLM-generated clinical questions with medical information scientists' evidence summaries. // Journal of the Medical Library Association: JMLA. 2026. Vol. 114. No. 2. P. 94.
21. Sümbüllü M. et al. Large language models' performances regarding common patient questions about broken endodontic instruments: a comparative analysis of ChatGPT-5.2, Gemini 3 and DeepSeek V3.2 in accuracy, consistency and readability. // BMC Oral Health. 2026.
22. Yüceer-Çetiner E. et al. Evaluating the competence of AI chatbots in answering patient-oriented frequently asked questions on orthognathic surgery. // Healthcare. 2025. Vol. 13. No. 17. P. 2114.

Тыров Илья Александрович. Департамент здравоохранения г. Москвы. Заместитель руководителя. Область научных интересов: цифровизация медицины, инновации в медицине, медицинская информатика. E-mail: zdrav@mos.ru

Васильев Юрий Александрович. Государственное бюджетное учреждение здравоохранения города Москвы «Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения г. Москвы». Главный врач. Доктор медицинских наук, доцент. Область научных интересов: рентгенология, радиология, цифровая диагностика, компьютерное зрение, искусственный интеллект в медицине. E-mail: VasilevYA1@zdrav.mos.ru

Владзимирский Антон Вячеславович. Государственное бюджетное учреждение здравоохранения города Москвы «Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения города Москвы». Заместитель директора. Доктор медицинских наук, доктор исторических наук, доцент. Федеральное государственное автономное образовательное учреждение высшего образования Первый Московский государственный медицинский университет им. И.М. Сеченова (Сеченовский Университет). Профессор. Область научных интересов: лучевая диагностика, искусственный интеллект в медицине, телемедицина, цифровые инновации в здравоохранении. E-mail: VladzimirskijAV@zdrav.mos.ru

Омелянская Ольга Васильевна. Государственное бюджетное учреждение здравоохранения города Москвы «Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения города Москвы». Заместитель директора. Федеральное государственное бюджетное образовательное учреждение высшего образования «МИРЭА – Российский технологический университет». Преподаватель. Область научных интересов: инновации в медицине, организация здравоохранения, менеджмент в здравоохранении. E-mail: OmelyanskayaOV@zdrav.mos.ru

Березнер Тимофей Александрович. Государственное бюджетное учреждение здравоохранения города Москвы «Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения города Москвы». Эксперт. Национальный исследовательский университет «Высшая школа экономики». Преподаватель. Область научных интересов: искусственный интеллект, когнитивные науки, цифровизация. E-mail: BereznerTA@zdrav.mos.ru (Ответственный за переписку).

Соболева Юлия Владимировна. Государственное бюджетное учреждение здравоохранения города Москвы «Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения города Москвы». Эксперт. Область научных интересов: организация здравоохранения, искусственный интеллект в медицине. E-mail: SobolevaYV4@zdrav.mos.ru

Яфарова Адель Айратовна. Государственное бюджетное учреждение здравоохранения города Москвы «Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения города Москвы». Эксперт, кандидат медицинских наук. Область научных интересов: инновации в медицине, искусственный интеллект в медицине, кардиология. E-mail: YafarovaAA@zdrav.mos.ru

Варюхина Мария Дмитриевна. Государственное бюджетное учреждение здравоохранения города Москвы «Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения города Москвы». Заведующий сектором. Кандидат медицинских наук, ведущий научный сотрудник. Область научных интересов: цифровизация здравоохранения, искусственный интеллект в медицине. E-mail: VaryukhinaMD@zdrav.mos.ru

Шумская Юлия Федоровна. Государственное бюджетное учреждение здравоохранения города Москвы «Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения города Москвы». Заведующий сектором, научный сотрудник. Область научных интересов: телемедицина, инновации в медицине. E-mail: ShumskayaYF@zdrav.mos.ru

Ахмедзянова Дина Альфредовна. Государственное бюджетное учреждение здравоохранения города Москвы «Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения города Москвы». Младший научный сотрудник. Область научных интересов: телемедицина, цифровизация здравоохранения. E-mail: AkhmedzyanovaDA@zdrav.mos.ru

Беженова Карина Юрьевна. Государственное бюджетное учреждение здравоохранения города Москвы «Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения города Москвы». Младший научный сотрудник. Область научных интересов: телемедицина, цифровизация здравоохранения. E-mail: BezhenovaKY@zdrav.mos.ru

Блохин Иван Андреевич. Государственное бюджетное учреждение здравоохранения г. Москвы «Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения города Москвы». Ведущий научный сотрудник. Кандидат медицинских наук. Область научных интересов: лучевая диагностика, радиология, рентгенология, компьютерное зрение, искусственный интеллект. E-mail: BlokhinIA@zdrav.mos.ru

Решетников Роман Владимирович. Государственное бюджетное учреждение здравоохранения города Москвы «Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения города Москвы». Руководитель отдела. Кандидат физико-математических наук. Область научных интересов: искусственный интеллект, медицинская информатика, цифровые технологии в медицине, компьютерное зрение. E-mail: ReshetnikovRV1@zdrav.mos.ru

Expert Evaluation of Large Generative Model Responses to Physicians' Queries in Clinical Practice

I.A. Tyrov[I], Yu.A. Vasilev[II], A.V. Vladzymyrskyy[II,III], O.V. Omelyanskaya[II,IV], T.A. Berezner[II,V], J.V. Soboleva[II], A.A. Yafarova[II], M.D. Varyukhina[II], Yu.F. Shumskaya[II], D.A. Akhmedzyanova[II], K.Yu. Bezhenova[II], I.A. Blokhin[II], R.V. Reshetnikov[II]

[I] Department of Healthcare of Moscow, Moscow, Russia

[II] Moscow Center for Diagnostics and Telemedicine, Moscow, Russia

[III] Sechenov University, Moscow, Russia

[IV] MIREA, Moscow, Russia

[V] HSE University, Moscow, Russia

Abstract. Large language models (LLMs) in healthcare currently attract significant interest. This study compared GigaChat and YandexGPT by evaluating 1,200 LLM-generated responses to physician queries from outpatient and inpatient settings across 7 topics. Thirteen expert physicians assessed the responses using a 5-point scale across five criteria: relevance, correctness, safety, completeness, and language literacy. We performed paired t-tests and calculated effect size. Heterogeneous differences between LLM scores were observed depending on assessment criteria and topic. Safety and correctness were unsatisfactory on several topics, including "Medications" for GigaChat (3.1±1.5 correctness, 3.3±1.6 safety) and "Diagnostic parameters, reference ranges, and their interpretation" for YandexGPT (3.3±1.5 for both). Overall, our findings suggest that LLMs are not yet ready to function as physician assistants in clinical practice.

Keywords: *artificial intelligence, medical chatbot, large language models, GigaChat, YandexGPT, expert evaluation of LLMs*.

DOI: 10.14357/20790279260310 **EDN:** EUAPVH

References

1. Aravazhi P.S. et al. The integration of artificial intelligence into clinical medicine: trends, challenges, and future directions. Disease-a-Month. 2025;71(6):101882.
2. Ukaz Prezidenta Rossiyskoy Federatsii No. 490 ot 10 oktyabrya 2019 g. (red. ot 15 fevralya 2024 g.) "O razvitii iskusstvennogo intellekta v Rossiyskoy Federatsii" (vmeste s "Natsional'noy strategiey razvitiya iskusstvennogo intellekta na period do 2030 goda"). (In Russ.)
3. Ukaz Prezidenta Rossiyskoy Federatsii No. 896 ot 8 dekabrya 2025 g. "O Strategii razvitiya zdravookhraneniya v Rossiyskoy Federatsii na period do 2030 goda". (In Russ.)
4. Rasporyazheniye Pravitel'stva Rossiyskoy Federatsii No. 959-r ot 17 aprelya 2024 g. o strategicheskom napravlenii v oblasti tsifrovoy transformatsii zdravookhraneniya. (In Russ.)
5. Zhihong Z. et al. Advancing healthcare with large language models: A scoping review of applications and future directions. International Journal of Medical Informatics. 2025;106231.
6. Liang E.N. et al. Clinical applications of large language models in medicine and surgery: A scoping review. Journal of International Medical Research. 2025;53(7):03000605251347556.
7. Shchetinin E.Y. et al. Methods for developing and implementing large language models in healthcare: challenges and prospects in Russia. Discrete and Continuous Models and Applied Computational Science. 2025;33(3):327-344. (In Russ.)
8. Reshetnikov R.V. et al. Assessing the quality of large generative models for basic healthcare applications. Vrach i informatsionnye tekhnologii. 2025;(3):64-76. (In Russ.)

9. Suresh K., Thomas S.V., Suresh G. Design, data analysis and sampling techniques for clinical research. Annals of Indian Academy of Neurology. 2011;14(4):287-290.
10. Vasilev Y.A. et al. Sozdanie i validatsiya oprosnika ERA-GMA dlya otsenki kachestva raboty bol'shikh generativnykh modeley v kachestve servisa predostavleniya otveta na voprosy vracha-klinitsista. Natsional'noye zdravookhraneniye (Rossiya). 2026. In press. (In Russ.)
11. Faul F. et al. Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. Behavior Research Methods. 2009;41(4):1149-1160.
12. Tam T.Y.C. et al. A framework for human evaluation of large language models in healthcare derived from literature review. NPJ Digital Medicine. 2024;7(1):258.
13. Huh I., Gim J. Exploration of Likert scale in terms of continuous variable with parametric statistical methods. BMC Medical Research Methodology. 2025;25(1):218.
14. Kottner J. et al. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. International Journal of Nursing Studies. 2011;48(6):661-671.
15. de Vet H.C.W., Dikmans R.E., Eekhout I. Specific agreement on dichotomous outcomes can be calculated for more than two raters. Journal of Clinical Epidemiology. 2017;83:85-89.
16. Heeren T., D'Agostino R. Robustness of the two independent samples t-test when applied to ordinal scaled data. Statistics in Medicine. 1987;6(1):79-90.
17. Posten H.O. Robustness of the two-sample t-test. In: Robustness of statistical methods and nonparametric statistics. Dordrecht: Springer Netherlands, 1984:92-99.
18. R Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2020.
19. Tiller N.B. et al. Generative artificial intelligence-driven chatbots and medical misinformation: an accuracy, referencing and readability audit. BMJ Open. 2026;16(4):e112695.
20. Blasingame M.N. et al. Comparing five generative AI chatbots' answers to LLM-generated clinical questions with medical information scientists' evidence summaries. Journal of the Medical Library Association: JMLA. 2026;114(2):94.
21. Sümbüllü M. et al. Large language models' performances regarding common patient questions about broken endodontic instruments: a comparative analysis of ChatGPT-5.2, Gemini 3 and DeepSeek V3.2 in accuracy, consistency and readability. BMC Oral Health, 2026.
22. Yüceer-Çetiner E. et al. Evaluating the competence of AI chatbots in answering patient-oriented frequently asked questions on orthognathic surgery. Healthcare. 2025;13(17):2114.

Tyrov Ilya A. Deputy Head, Moscow Health Care Department, 43 Oruzheyny pereulok, Moscow, 127006. E-mail: zdrav@mos.ru

Vasilev Yuriy A. Assoc. Prof., Ph.D. in Medicine, D.Sc. Medical director, State Budget-Funded Health Care Institution of the City of Moscow "Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Health Care Department", 24 Petrovka str., bld. 1, Moscow, 127051. E-mail: VasilevYA1@zdrav.mos.ru

Vladzymyrskyy Anton V. Assoc. Prof., Ph.D. in Medicine, D.Sc., Ph.D. in History, D.Sc. Deputy Director of R&D, State Budget-Funded Health Care Institution of the City of Moscow "Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Health Care Department", 24 Petrovka str., bld. 1, Moscow, 127051. Professor, Department of Information Technologies and Medical Data Processing, I.M. Sechenov First Moscow State Medical University of the Ministry of Health of the Russian Federation, 8-2 Trubetskaya str., Moscow, 119991. E-mail: VladzimirskijAV@zdrav.mos.ru

Omelyanskaya Olga V. Chief Innovation Officer, State Budget-Funded Health Care Institution of the City of Moscow "Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Health Care Department", 24 Petrovka str., bld. 1, Moscow, 127051. Lecturer, Department of Artificial Intelligence Technologies, Ministry of Science and Higher Education, Moscow Technical University — MIREA, 78 Vernadsky Avenue, Moscow, 119454. E-mail: OmelyanskayaOV@zdrav.mos.ru

Berezner Timofei A. Expert, State Budget-Funded Health Care Institution of the City of Moscow "Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Health Care Department", 24 Petrovka str., bld. 1, Moscow, 127051. Lecturer, School of Psychology, Faculty of Social Sciences, HSE University, 20 Myasnitskaya str., Moscow, 101000. E-mail: BereznerTA@zdrav.mos.ru

Soboleva Julia V. Expert, State Budget-Funded Health Care Institution of the City of Moscow "Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Health Care Department", 24 Petrovka str., bld. 1, Moscow, 127051. E-mail: SobolevaYV4@zdrav.mos.ru

Yafarova Adel A. Ph.D. in Medicine, Expert, State Budget-Funded Health Care Institution of the City of Moscow "Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Health Care Department", 24 Petrovka str., bld. 1, Moscow, 127051. E-mail: YafarovaAA@zdrav.mos.ru

Varyukhina Maria D. Ph.D. in Medicine, Lead Researcher, Head of Lab, Digital Medicine Lab, State Budget-Funded Health Care Institution of the City of Moscow "Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Health Care Department", 24 Petrovka str., bld. 1, Moscow, 127051. E-mail: VaryukhinaMD@zdrav.mos.ru

Shumskaya Yulia F. Researcher, Head of Lab, Telemedicine Research Lab, State Budget-Funded Health Care Institution of the City of Moscow "Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Health Care Department", 24 Petrovka str., bld. 1, Moscow, 127051. E-mail: ShumskayaYF@zdrav.mos.ru

Akhmedzyanova Dina A. Junior Researcher, Telemedicine Research Lab, State Budget-Funded Health Care Institution of the City of Moscow "Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Health Care Department", 24 Petrovka str., bld. 1, Moscow, 127051. E-mail: AkhmedzyanovaDA@zdrav.mos.ru

Bezhenova Karina Yu. Junior Researcher, Telemedicine Research Lab, State Budget-Funded Health Care Institution of the City of Moscow "Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Health Care Department", 24 Petrovka str., bld. 1, Moscow, 127051. E-mail: BezhenovaKY@zdrav.mos.ru

Blokhin Ivan A. Ph.D. in Medicine, Lead Researcher, Medical Research Department, State Budget-Funded Health Care Institution of the City of Moscow "Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Health Care Department", 24 Petrovka str., bld. 1, Moscow, 127051. E-mail: BlokhinIA@zdrav.mos.ru

Reshetnikov Roman V. Ph.D. in Physics and Mathematics, Head of Department, Medical Research Department, State Budget-Funded Health Care Institution of the City of Moscow "Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Health Care Department", 24 Petrovka str., bld. 1, Moscow, 127051. E-mail: ReshetnikovRV1@zdrav.mos.ru