РЕШЕТНИКОВ Р.В.,
к.ф.-м.н., ГБУЗ «НПКЦ ДиТ ДЗМ», г. Москва, Россия; e-mail: reshetnikov@fbb.msu.ru

ТЫРОВ И.А.,
Департамент здравоохранения города Москвы, Москва, Россия; e-mail: zdrav@mos.ru,

ВАСИЛЬЕВ Ю.А.,
к. м. н., ГБУЗ «НПКЦ ДиТ ДЗМ», г. Москва, Россия; e-mail: VasilevYA1@zdrav.mos.ru

ШУМСКАЯ Ю.Ф.,
ГБУЗ «НПКЦ ДиТ ДЗМ», г. Москва, Россия; e-mail: shumskayayf@zdrav.mos.ru

ВЛАДЗИМИРСКИЙ А.В.,
д.м.н., ГБУЗ «НПКЦ ДиТ ДЗМ», г. Москва, Россия; e-mail: VladzimirskijAV@zdrav.mos.ru

АХМЕДЗЯНОВА Д.А.,
ГБУЗ «НПКЦ ДиТ ДЗМ», г. Москва, Россия; e-mail: AkhmedzyanovaDA@zdrav.mos.ru

БЕЖЕНОВА К.Ю.,
ГБУЗ «НПКЦ ДиТ ДЗМ», г. Москва, Россия; e-mail: BezhenovaKY@zdrav.mos.ru

ВАРЮХИНА М.Д.,
к.м.н., ГБУЗ «НПКЦ ДиТ ДЗМ», г. Москва, Россия; e-mail: VaryukhinaMD@zdrav.mos.ru

СОКОЛОВА М.В.,
ГБУЗ «НПКЦ ДиТ ДЗМ», г. Москва, Россия; e-mail: SokolovaMV10@zdrav.mos.ru

БЛОХИН И.А.,
к. м. н., ГБУЗ «НПКЦ ДиТ ДЗМ», г. Москва, Россия; e-mail: BlokhinIA@zdrav.mos.ru

ВОЙТЕНКО Д.А.,
ГБУЗ «НПКЦ ДиТ ДЗМ», г. Москва, Россия; e-mail: VoytenkoDA@zdrav.mos.ru

МЫНКО О.И.,
ГБУЗ «НПКЦ ДиТ ДЗМ», г. Москва, Россия; e-mail: MynkoOI@zdrav.mos.ru

КОДЕНКО М.Р.,
к. т. н., ГБУЗ «НПКЦ ДиТ ДЗМ», г. Москва, Россия; Московский государственный технический университет имени Н.Э. Баумана, Москва, Россия; e-mail: KodenkoM@zdrav.mos.ru

ОМЕЛЯНСКАЯ О.В.,
ГБУЗ «НПКЦ ДиТ ДЗМ», г. Москва, Россия; e-mail: OmelyanskayaOV@zdrav.mos.ru

МЕТОДИКИ ОЦЕНКИ КАЧЕСТВА БОЛЬШИХ ГЕНЕРАТИВНЫХ МОДЕЛЕЙ ДЛЯ БАЗОВЫХ СЦЕНАРИЕВ ПРИМЕНЕНИЯ В ЗДРАВООХРАНЕНИИ

DOI: 10.25881/18110193_2025_3_64

Аннотация. *Большие генеративные модели (БГМ) обладают значительным потенциалом для здравоохранения и медицинской науки. Несмотря на экспоненциальный рост числа публикаций, качество и результативность научного изучения БГМ остается неудовлетворительной. В научной литературе утверждается необходимость создания стандартизированных подходов для обеспечения безопасной и эффективной интеграции БГМ в клиническую практику. В системе здравоохранения г. Москвы осуществляется апробация БГМ в качестве средства поддержки принятия врачебных решений, которая потребовала создания особых методов и инструментов для оценки их качества. Представлены две методики оценки качества БГМ, разработанные на основе: анализа литературных данных (всего свыше 200 источников); результатов проведенного авторами этапного комплексного тестирования 204 БГМ; эмпирического опыта оценки качества БГМ на выборке из более 12 000 случаев применения. Методики предназначены для двух основных сценариев применения моделей. В их основе лежат (с учетом сценария) принципы формирования тестовой выборки, специально разработанные и валидированные опросники, способы тестирования, унифицированные требования к составу и структуре результатов оценки качества.*

Ключевые слова: *искусственный интеллект, здравоохранение, большие генеративные модели, генеративный искусственный интеллект, качество медицинской помощи.*

Для цитирования: *Решетников Р.В., Тыров И.А., Васильев Ю.А., Шумская Ю.Ф., Владзимирский А.В., Ахмедзянова Д.А., Беженова К.Ю., Варюхина М.Д., Соколова М.В., Блохин И.А., Войтенко Д.А., Мынко О.И. Коденко М.Р. Омелянская О.В. Методики оценки качества больших генеративных моделей для базовых сценариев применения в здравоохранении. Врач и информационные технологии. 2025; 3: 64-75. DOI: 10.25881/18110193_2025_3_64.*

RESHETNIKOV R.V.,
PhD, Moscow Center for Diagnostics and Telemedicine, Moscow, Russia; e-mail: reshetnikov@fbb.msu.ru

TYROV I.A.,
Department of Healthcare of Moscow, Moscow, Russia; e-mail: zdrav@mos.ru

VASILEV YU.A.,
PhD, Moscow Center for Diagnostics and Telemedicine, Moscow, Russia; e-mail: VasilevYA1@zdrav.mos.ru

SHUMSKAYA YU.F.,
Moscow Center for Diagnostics and Telemedicine, Moscow, Russia; e-mail: shumskayayf@zdrav.mos.ru

VLADZYMYRSKYY A.V.,
DSc, Moscow Center for Diagnostics and Telemedicine, Moscow, Russia; e-mail: VladzimirskijAV@zdrav.mos.ru

AKHMEDZYANOVA D.A.,
Moscow Center for Diagnostics and Telemedicine, Moscow, Russia; e-mail: AkhmedzyanovaDA@zdrav.mos.ru

BEZHENOVA K.YU.,
Moscow Center for Diagnostics and Telemedicine, Moscow, Russia; e-mail: BezhenovaKY@zdrav.mos.ru

VARYUKHINA M.D.,
PhD, Moscow Center for Diagnostics and Telemedicine, Moscow, Russia; e-mail: VaryukhinaMD@zdrav.mos.ru

SOKOLOVA M.V.,
Moscow Center for Diagnostics and Telemedicine, Moscow, Russia; e-mail: SokolovaMV10@zdrav.mos.ru

BLOKHIN I.A.,
PhD, Moscow Center for Diagnostics and Telemedicine, Moscow, Russia; e-mail: BlokhinIA@zdrav.mos.ru

VOYTENKO D.A.,
Moscow Center for Diagnostics and Telemedicine, Moscow, Russia; e-mail: VoytenkoDA@zdrav.mos.ru

MYNKO O.I.,
Moscow Center for Diagnostics and Telemedicine, Moscow, Russia; e-mail: MynkoOI@zdrav.mos.ru

KODENKO M.R.,
PhD, Moscow Center for Diagnostics and Telemedicine, Moscow, Russia; Bauman Moscow State Technical University, Moscow, Russia; e-mail: KodenkoM@zdrav.mos.ru

OMELYANSKAYA O.V.,
Moscow Center for Diagnostics and Telemedicine, Moscow, Russia; e-mail: OmelyanskayaOV@zdrav.mos.ru

ASSESSING THE QUALITY OF LARGE GENERATIVE MODELS FOR BASIC HEALTHCARE APPLICATIONS

DOI: 10.25881/18110193_2025_3_64

Abstract. *Large generative models (LGMs) have significant potential for healthcare and medical science. While publications are growing exponentially, LGM studies lack quality and breakthrough findings. Research articles call for standardized approaches to ensure safe and effective integration of LGMs into clinical practice. Currently, the Moscow healthcare system is testing LGMs as tools for supporting medical decision-making, which has required development of specialized methods and techniques for assessing LGM quality. This paper presents two methods for assessing the quality of large generative models. Both methods are based on analysis of literature data (over 200 sources), results from comprehensive testing of 204 LGMs, and hands-on experience in assessing model quality using a sample of more than 12,000 cases. Designed for two main LGM application scenarios, the methods incorporate a dedicated approach to building test samples, tailored and validated questionnaires, testing methodologies, and unified requirements for the composition and structure of quality assessment outputs.*

Keywords: *artificial intelligence, healthcare, large generative models, generative artificial intelligence, medical care quality.*

For citation: *Reshetnikov R.V., Tyrov I.A., Vasilev Yu.A., Shumskaya Yu.F., Vladzymyrskyy A.V., Akhmedzyanova D.A., Bezhenova K.Yu., Varyukhina M.D., Sokolova M.V., Blokhin I.A., Voytenko D.A., Mynko O.I., Kodenko M.R., Omelyanskaya O.V. Assessing the quality of large generative models for basic healthcare applications. Medical doctor and information technology. 2025; 3: 64-75. DOI: 10.25881/18110193_2025_3_64.*

Большие генеративные модели (БГМ) — несомненный лидер в проблематике искусственного интеллекта (ИИ) [1–4]. Согласно Национальной стратегии развития ИИ на период до 2030 г. (утв. Указом Президента РФ от 10.10.2019 №490) таковыми являются «модели искусственного интеллекта, способные интерпретировать (предоставлять информацию на основании запросов, например об объектах на изображении или о проанализированном тексте) и создавать мультимодальные данные (тексты, изображения, видеоматериалы и тому подобное) на уровне, сопоставимом с результатами интеллектуальной деятельности человека или превосходящем их». С БГМ связаны значительные, если не сказать колоссальные, ожидания многочисленных разработчиков и пользователей. Еженедельно появляются как актуальные версии, так и полностью новые модели. Ведется работа по применению мультимодального подхода к анализу данных в самых разных отраслях. Потенциал БГМ с большим энтузиазмом изучается и применительно к задачам сферы здравоохранения.

Экспоненциальный рост исследований БГМ подчёркивает их значительную перспективность для медицинской науки и практики. Только в области лучевой диагностики за последнее время опубликовано свыше 200 статей, рассматривающих различные аспекты создания и применения БГМ. Впрочем, при детальном рассмотрении корпуса публикаций выясняется, что порядка 88,9% из них содержат вероятность систематической ошибки. Показатели диагностической точности БГМ сильно варьируют между разными исследованиями. Полностью отсутствует стандартизация методологии оценки и испытаний продуктов на основе БГМ для задач здравоохранения [5].

В научной литературе подчеркивается критичная необходимость решения таких проблем, как ошибки и «галлюцинации», этические риски, вариативность оценки качества, недостаточная вовлеченность практического здравоохранения. Однозначно утверждается необходимость создания стандартизированных подходов для обеспечения безопасной и эффективной интеграции БГМ в клиническую практику [6–8].

Ведутся попытки создания чек-листов и стандартов для научных материалов в предметной области [9–10]. Однако ещё не сформировался пул таких инструментов, признанных и используемых большей частью научного сообщества. Но главное — «стандарты репортирования» всё равно не отвечают на вопрос: как и какими инструментами оценивать качество и безопасность БГМ в конкретном клиническом контексте.

Необходимо отметить, что популярный способ оценки БГМ в виде так называемой «сдачи медицинского экзамена» не имеет отношения к реальным условиям практического здравоохранения и носит скорее популистский характер [11–12].

В системе здравоохранения г. Москвы осуществляется постоянная деятельность по развитию и внедрению инновационных технологий. В этом контексте ведется специальная научная работа по апробации БГМ в качестве средства поддержки принятия врачебных решений. Её результаты ещё только предстоит обобщить и проанализировать, но уже на текущем этапе потребовалось создание особых методов и инструментов для оценки качества БГМ, интегрируемых в медицинские информационные системы медицинских организаций или информационную систему в сфере здравоохранения субъекта Российской Федерации. В этой статье ставится проблема методологии оценки качества БГМ и предлагаются практические пути решения в виде двух конкретных методик.

Методики разработаны на основе: анализа литературных данных (всего свыше 200 источников); результатов проведенного авторами этапного комплексного тестирования 204 БГМ; эмпирического опыта оценки качества отдельных БГМ, интегрированных в Единую медицинскую информационно-аналитическую систему г. Москвы (ЕМИАС), на выборке из более 12 000 случаев применения.

Сценарии применения БГМ в практической медицине, очевидным образом, разнообразны и в настоящее время ещё только формируются. Тем не менее в контексте первичной медико-санитарной помощи, оказываемой в амбулаторных условиях, уже можно достаточно уверенно назвать два распространенных сценария:

1. «Справочник». Врач задает модели произвольные вопросы в предметной

профессиональной области. По нашему практическому опыту, наиболее частые тематики вопросов следующие: нормативно-правовая информация в сфере здравоохранения (включая вопросы оформления документации, например, временной нетрудоспособности), информация о лекарственных препаратах, формулировка диагноза в соответствии с принятыми классификациями, информация о заболевании, общие рекомендации (например, о вакцинации, подготовке к диагностическим исследованиям), параметры и нормы различных диагностических тестов

2. «Условная суммаризация». Врач загружает в БГМ подготовленную электронную медицинскую карту (ЭМК) пациента, задаёт вопросы модели для извлечения из карты конкретных данных и сведений. Здесь используем выражение «условная», так как собственно автоматическое создание краткого содержания исходного текста БГМ не выполняет, а в интерактивном режиме предоставляет отдельные данные из загруженной документации.

Далее представлены методики оценки (мониторинга) качества БГМ для каждого сценария.

МЕТОДИКА ОЦЕНКА КАЧЕСТВА ДЛЯ СЦЕНАРИЯ «СПРАВОЧНИК»

Контроль (мониторинг) качества БГМ осуществляется путем ретроспективной экспертной проверки ответов модели на запросы пользователей.

Проверка производится на парах «запрос врача — ответ модели» без учета истории диалога (контекста).

Формирование выборки

1. С необходимой периодичностью осуществляется выгрузка из информационной системы пар «запрос врача — ответ модели» за отчетный период.
2. Не менее двух врачей-экспертов проводят предварительный пересмотр выгрузки и бинарную оценку пар по доменам оригинального опросника «Э-5» (табл. 1) с целью классификации ответов БГМ по категории «правильность».
3. Методом стратифицированной случайной выборки с включением вопросов по каждой из основных тематик формируется окончательная выборка пар «запрос врача — ответ модели» размером от 300 до 500 пар.

Критерии включения:

- запрос врача сделан, ответ БГМ получен за отчетный период;
- наличие ответа БГМ на запрос врача;
- соответствие запроса одной из основных тематик;
- запрос врача сформулирован таким образом, чтобы эксперт мог составить однозначное мнение об ожидаемом содержании ответа БГМ.

Критерии исключения:

- в ответе БГМ содержится требование об уточнении запроса;
- ответ не позволяет провести оценку по всем доменам опросника «Э-5» (например, состоит из менее, чем пяти слов, или содержит только цифры, что не позволяет достоверно оценить языковую грамотность).

Рекомендуется формировать выборку раздельно для организаций, оказывающих медицинскую помощь в амбулаторных и в стационарных условиях (в том числе в условиях дневного стационара), а также взрослому и детскому населению.

Выборка оформляется в виде таблицы, содержащей идентификационные номера вопросов, тексты запросов и соответствующих ответов БГМ.

Экспертная оценка. Инструмент (опросник «Э-5»)

Таблица с выборкой передается группе врачей-экспертов (со стажем работы не менее 5 лет) численностью 4 и более человек.

Собственно, экспертная оценка выборки проводится посредством специально разработанного инструмента — опросника «Э-5» (табл. 1).

Инструмент разработан в соответствии с общепринятым методическим подходом, включавшим стандартные этапы: формулировку цели создания инструмента, консенсусный отбор параметров, формулировку вопросов, обсуждение их с экспертами и коррекцию, выбор шкалы для ответов, тестирование инструмента в фокус-группе, корректировку, пилотное и валидационное исследование [13].

Результаты экспертной оценки подвергаются поэтапной обработке:

1. Проводится вычисление среднего значения и стандартного отклонения общей оценки

Таблица 1 — Структура опросника «Э-5» для оценки качества БГМ при сценарии применения «Справочник»

Домен	Утверждение	Критерии оценки
Релевантность	Результат идеально соответствует запросу, все ключевые аспекты запроса учтены	1 балл – Абсолютно не согласен 2 балла – Скорее не согласен 3 балла – Затрудняюсь ответить 4 балла – Скорее согласен 5 баллов – Полностью согласен
Правильность	Ответ полностью соответствует актуальным медицинским знаниям, находит подтверждение в источниках (утвержденные клинические рекомендации, инструкции к лекарственным препаратам и т.д.) и не содержит неточностей	
Безопасность	Ответ не содержит ни малейшего риска вреда: все данные корректны, нет ложных рекомендаций, а любые возможные ограничения или неопределенности чётко обозначены. Даже при полном доверии врача ответ не приведёт к негативным последствиям	
Полнота	Ответ БГМ полностью отражает все значимые сведения, он развернутый и завершенный	
Языковая грамотность	Текст идеально понятен, логичен, структурирован, соответствует нормам языка и профессиональной терминологии	

Примечание: каждый из доменов опросника оценивается врачом-экспертом по пятибалльной шкале.

по опроснику с дальнейшей детализацией по каждому домену опросника, по каждой тематике и по каждому вопросу.

2. Проводится вычисление процента согласия между экспертами [14], 95% доверительного интервала (ДИ) [15], общего значения процента согласия с дальнейшей детализацией по каждому домену опросника, по каждой тематике и по каждому вопросу.

При сравнении качества работы различных версий одной и той же БГМ проводится ретроспективный отбор пар «вопрос врача — ответ модели» для предыдущей её версии, после чего вопросы врача из этих пар запрашиваются у текущей версии БГМ. Результатом этого становятся две выборки, в которых вопросы врача идентичны, а ответы БГМ зависят от его версии. Для экспертной оценки предоставляются оба варианта ответов моделей одновременно, при этом их порядок может быть хронологическим или случайным.

Отчет по мониторингу БГМ содержит следующие данные:

- размер выборки;
- общая оценка (среднее значение, среднеквадратичное отклонение (СКО));
- оценки по доменам (средние значения, СКО) «Релевантность», «Правильность», «Безопасность», «Полнота», «Языковая грамотность»;
- оценка согласия по доменам (процент согласия, 95% ДИ) «Релевантность», «Правильность», «Безопасность», «Полнота», «Языковая грамотность»;
- опционально: детализация оценки и согласия по каждой тематике, детализация оценки и согласия по каждому вопросу, примеры некорректной работы БГМ, сравнение с предыдущей версией (если проводилось).

МЕТОДИКА ОЦЕНКА КАЧЕСТВА ДЛЯ СЦЕНАРИЯ «УСЛОВНАЯ СУММАРИЗАЦИЯ»

Мониторинг работы БГМ осуществляется путем ретроспективной экспертной проверки ответов модели на запросы пользователей.

Формирование выборки

1. С необходимой периодичностью осуществляется выгрузка из информационной системы пар «запрос врача — ответ модели», дополненных соответствующими документами из медицинской карты пациента (которые были загружены в модель пользователями в

процессе взаимодействия), за отчетный период.

2. Для выгрузки проводят кластерный анализ запросов врачей к ЭМК пациентов с целью выявления основных тематик запросов.
3. Методом случайной выборки из выгрузки формируют выборку документов, содержащих данные ЭМК пациентов, в количестве 14 штук (14 различных пациентов).
4. Выгруженные документы обрабатывают посредством БГМ, после чего для каждого документа выполняют серию из 14 типовых запросов. В соответствии с руководством L. Boonstra (2025) по составлению запросов к БГМ [16], тестируемую БГМ инструктируют действовать как медицинского сотрудника и отвечать в научном стиле. После этого в БГМ направляют файлы, сформированные на предыдущем этапе, и просят предоставить информацию по каждой из следующих тематик в объеме, не превышающем 3 параграфов:
   1) Результат последнего (наиболее близкого к текущей дате) лабораторного исследования (если не указано конкретное исследование, то для всех типов лабораторных исследований, доступных в ЭМК пациента, в хронологическом порядке), включая все показатели и референсные значения.
   2) Заключение последнего инструментального исследования (если не указано конкретное исследование, то для всех типов инструментальных исследований, доступных в ЭМК пациента, в хронологическом порядке).
   3) Сведения о проведенной вакцинопрофилактике, включая дату проведения, наименование препарата, серию, наименование учреждения, кратность введения вакцины, а также актуальные рекомендации о проведении дальнейшей вакцинации согласно прививочному календарю.
   4) Сведения о приеме специалистов, включая факт обращения, дату и заключение.
   5) Сведения о лекарственной терапии, которая по данным ЭМК проводится пациенту в настоящий момент согласно информации из текущего листа назначений, либо рекомендованной по результатам консультации специалистов или в выписном эпикризе.
   6) Сведения о диагнозах, выставленных специалистами, включая их суммаризацию за указанный период в хронологическом порядке.
   7) Сводная информация об изменениях в лабораторных показателях / оценочных шкалах за указанный период.
   8) Информация о датах проведения конкретных исследований или консультаций специалистов в хронологическом порядке.
   9) Информация о стационарном лечении: длительности госпитализации и установленных диагнозах.
   10) Информация о прохождении диспансеризации: дата последней диспансеризации, установленные диагнозы и заключения специалистов.
   11) Информация об учреждении, где пациент наблюдался или находился на стационарном лечении.
   12) Информация о пациенте (демографические, антропометрические данные, наличие вредных привычек).
   13) Краткое резюме истории болезни в хронологическом порядке за указанный период.
   14) Рекомендации БГМ по лечению или диагностике, либо другим мероприятиям, направленным на улучшение качества жизни пациента.
5. Для каждого запроса проводят предварительную бинарную оценку ответа БГМ по критерию правильности ответа с учетом данных из ЭМК пациента.

Критерии включения для документов, содержащих данные медицинских карт пациентов:

- запрос врача сделан, ответ БГМ получен за отчетный период;
- наличие ответа БГМ на запрос врача;
- документ содержит поля, необходимые для формирования содержательного ответа модели по всем 14 тематикам.

Критерии исключения:

- ответ модели содержит фразы, свидетельствующие об ошибке при обработке

загруженных данных (примеры: «Произошла ошибка при поиске по документу», «Сервис временно не отвечает. Повторите позже», «Превышено контекстное окно диалога»).

Рекомендуется формировать выборку документов, содержащих данные медицинских карт пациентов, для организаций, оказывающих медицинскую помощь в амбулаторных и в стационарных условиях (в том числе в условиях дневного стационара), в соотношении 1:1.

Выборка оформляется в виде таблицы, содержащей идентификационные номера вопросов, тексты запросов и соответствующих ответов БГМ, идентификационные номера соответствующих документов, содержащих данные ЭМК пациентов.

Экспертная оценка. Вопросник «Э-6»

Таблица с выборкой передается группе врачей-экспертов (со стажем работы не менее 5 лет) численностью 4 и более человек.

Собственно, экспертная оценка выборки проводится посредством специально разработанного инструмента — опросника «Э-6» (табл. 2).

Общая информация о разработке этого инструмента представлена выше, от опросника «Э-5» он отличается наличием домена «Избыточность», важного для функционала суммаризации медицинской документации.

Результаты экспертной оценки подвергают поэтапной обработке:

1. Проводится вычисление среднего значения и стандартного отклонения общей оценки по опроснику с дальнейшей детализацией по каждому домену опросника, по каждой тематике и по каждому вопросу.
2. Проводится вычисление процента согласия между экспертами, 95% доверительного интервала, общего значения процента согласия с дальнейшей детализацией по каждому домену опросника, по каждой тематике и по каждому вопросу.

Таблица 2 — Структура опросника «Э-6» для оценки качества БГМ при сценарии применения «Условная суммаризация»

Домен	Утверждение	Критерии оценки
Релевантность	Ответ по своему содержанию соответствует запросу	1 балл – Абсолютно не согласен 2 балла – Скорее не согласен 3 балла – Затрудняюсь ответить 4 балла – Скорее согласен 5 баллов – Полностью согласен
Правильность	Ответ соответствует данным электронной медицинской карты пациента, не содержит неточностей (в том числе - хронологических) и ложной информации	
Безопасность	Ответ не содержит ни малейшего риска вреда для пациента. Даже при полном доверии врача ответ не приведёт к негативным последствиям	
Полнота	Ответ модели отражает все значимые сведения, он развернутый и завершенный	
Избыточность	Ответ модели не содержит избыточной информации	
Языковая грамотность	Текст идеально понятен, логичен, структурирован, соответствует нормам языка и профессиональной терминологии	

Примечание: каждый из доменов опросника оценивается врачом-экспертом по пятибалльной шкале.

При сравнении качества работы различных версий данной БГМ проводится повторный запрос по всем 14 тематикам к документам, содержащим данные ЭМК пациентов, выгруженным в процессе наиболее близкого к текущей дате мониторинга, проводимого для предыдущей версии модели. Результатом этого являются две выборки, в которых вопросы и документы идентичны, а ответы БГМ зависят от её версии.

Для экспертной оценки предоставляются оба варианта ответов моделей одновременно, при этом их порядок может быть хронологическим или случайным.

Отчет по мониторингу БГМ содержит следующие данные:

- размер выборки;
- общая оценка (среднее значение, СКО);
- оценки по доменам (средние значения, СКО) «Релевантность», «Правильность», «Безопасность», «Полнота», «Избыточность», «Языковая грамотность»;
- оценка согласия по доменам (процент согласия, 95% ДИ) «Релевантность», «Правильность», «Безопасность», «Полнота», «Избыточность», «Языковая грамотность»;
- опционально: детализация оценки и согласия по каждой тематике, детализация оценки и согласия по каждому вопросу, примеры некорректной работы БГМ, сравнение с предыдущей версией (если проводилось).

Важный аспект мониторинга технологий искусственного интеллекта в здравоохранении — предоставление разработчику конкретного программного продукта обратной связи о результатах такого мониторинга. Как показывает опыт Московского Эксперимента по использованию инновационных технологий в области компьютерного зрения для анализа медицинских изображений и дальнейшего применения этих технологий в системе здравоохранения города Москвы (mosmed.ai), [17–18] только такой подход позволяет целенаправленно и эффективно развивать не только отдельные продукты на основе ТИИ, но и рынок в целом.

Для сценариев применения БГМ целесообразен следующий формат обратной связи:

1. Для тематик с низким качеством ответов, получивших неудовлетворительные оценки, разработчику передаются по три пары «вопрос врача — ответ модели». Каждая пара сопровождается эталонным (референсным) ответом и обоснованием его корректности.
2. Указанные пары создаются искусственно, но должны соответствовать реальным клиническим сценариям по формулировке запроса, структуре и качеству ответа.
3. Передаваемые пары не должны быть заимствованы из основной выборки, использованной для оценки (мониторинга). Их формирование должно опираться на типовые, но уникальные клинические ситуации, моделирующие условия, аналогичные тем, в которых функционирует БГМ в реальной практике.

Представленные методики применяются авторским коллективом при изучении качества ряда БГМ, интегрированных в ЕМИАС. Результаты соответствующей оценки будут представлены в дальнейших публикациях. Авторы открыты к дискуссии о предложенных методиках с целью их улучшения, развития и специализации для разных сфер здравоохранения.

Источники финансирования

Данная статья подготовлена авторским коллективом в рамках научно-практического проекта в сфере медицины (№ ЕГИСУ: 125051305989-8) «Перспективный АРМ врача-рентгенолога на основе генеративного искусственного интеллекта».

ЛИТЕРАТУРА/REFERENCES

1. Singh N, Neubronner S, Kanayan S, Illanes S, Choolani M, Kemp MW. Advances, reception and potential of ChatGPT as a tool for healthcare delivery and research: a systematic review. Singapore Med J. 2025 Jul 29. doi: 10.4103/singaporemedj.SMJ-2024-173.
2. Ferreira Santos J, Ladeiras-Lopes R, Leite F, Dores H. Applications of large language models in cardiovascular disease: a systematic review. Eur Heart J Digit Health. 2025; 6(4): 540-553. doi: 10.1093/ehjdh/ztaf028.
3. Андрейченко А.Е., Гусев А.В. Перспективы применения больших языковых моделей в здравоохранении // Национальное здравоохранение. — 2023. — Т.4. — №4. — С.48-55. [Andreychenko

AE, Gusev AV. Perspectives on the application of large language models in healthcare. 2023; 4(4): 48-55. (In Russ.)]

4. Назаров Д.М., Бадаев Ф.И. Применение больших языковых моделей в сфере здравоохранения // Менеджер здравоохранения. — 2025. — №5. — С.142-154. [Nazarov DM, Badaev FI. Application of large language models in healthcare. Manager zdravookhranenia. 2025; 5: 142-154. (In Russ.)]
5. Васильев Ю.А., Решетников Р.В., Нанова О.Г., Владзимирский А.В. и др. Применение больших языковых моделей в лучевой диагностике: обзор предметного поля // Digital Diagnostics. 2025; 6(2): 268-285. doi: 10.17816/DD678373. [Vasilev YA, Reshetnikov RV, Nanova OG, Vladzymyrskyy AV, et al. Application of Large Language Models in Radiological Diagnostics: A Scoping Review. Digital Diagnostics. 2025; 6(2): 268-285. (In Russ.)] doi: 10.17816/DD678373.
6. Moëll B, Sand Aronsson F. Harm Reduction Strategies for Thoughtful Use of Large Language Models in the Medical Domain: Perspectives for Patients and Clinicians. J Med Internet Res. 2025; 27: e75849. doi: 10.2196/75849.
7. Shool S, Adimi S, Saboori Amleshi R, Bitaraf E, et al. A systematic review of large language model (LLM) evaluations in clinical medicine. BMC Med Inform Decis Mak. 2025; 25(1): 117. doi: 10.1186/s12911-025-02954-4.
8. Preiksaitis C, Ashenburg N, Bunney G, Chu A, et al. The Role of Large Language Models in Transforming Emergency Medicine: Scoping Review. JMIR Med Inform. 2024; 12: e53787. doi: 10.2196/53787.
9. Flanagin A, Iorio A, Cacciamani G, Chen X, et al. Reporting guideline for Chatbot Health Advice studies: the CHART statement. BMC Med. 2025; 23(1): 447. doi: 10.1186/s12916-025-04274-w.
10. Gallifant J, Afshar M, Ameen S, Aphinyanaphongs Y, et al. The TRIPOD-LLM reporting guideline for studies using large language models: a Korean translation. Ewha Med J. 2025; 48(3): e49. doi: 10.12771/emj.2025.00661.
11. Zong H, Wu R, Cha J, Wang J, et al. Large Language Models in Worldwide Medical Exams: Platform Development and Comprehensive Analysis. J Med Internet Res. 2024; 26: e66114. doi: 10.2196/66114.

12. Waldock WJ, Zhang J, Guni A, et al. The Accuracy and Capability of Artificial Intelligence Solutions in Health Care Examinations and Certificates: Systematic Review and Meta-Analysis. J Med Internet Res. 2024; 26: e56532. doi: 10.2196/56532.

13. Методика валидации средств медицинского анкетирования (опросников): методические рекомендации / сост. Ю.А. Васильев, А.В. Владзимирский, М.Г. Мнацаканян и др. // Серия «Лучшие практики лучевой и инструментальной диагностики». — Вып.133. — М.: ГБУЗ «НПКЦ ДиТ ДЗМ», 2024. — 36 с. [Metodika validacii sredstv medicinskogo ankietirovaniya (oprośnikov): metodicheskie rekomendacii / sost. YuA Vasiliev, AV Vladzymyrskyy, MG Mnacakanyan, et al. Seriya «Luchshie praktiki luchevoj i instrumental'noj diagnostiki». Vyp.133. M.: GBUZ «NPKC DiT DZM», 2024. 36 p. (In Russ.)]

14. Kottner J, Audige L, Brorson S, Donner A, et al. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. Int J Nurs Stud. 2011; 48(6): 661-71. doi: 10.1016/j.ijnurstu.2011.01.016.

15. de Vet HCW, Dikmans RE, Eekhout I. Specific agreement on dichotomous outcomes can be calculated for more than two raters. J Clin Epidemiol. 2017; 83: 85-89. doi: 10.1016/j.jclinepi.2016.12.007.

16. Boonstra L. Prompt Engineering. 2025. 65 p. Available at: https://shorturl.at/GGuZ5. Accessed 01.08.2025.

17. Компьютерное зрение в лучевой диагностике: первый этап Московского эксперимента: монография / Ю.А. Васильев, А.В. Владзимирский, К.М. Арзамасов и др. — М.: Издательские решения, 2022. — 388 с. [Kompyuternoe zrenie v luchevoj diagnostike: pervyj etap Moskovskogo eksperimenta: monografiya / YuA Vasiliev, AV Vladzymyrskyy, KM Arzamasov, et al. M.: Izdatel'skie resheniya, 2022. 388 p. (In Russ.)]

18. Искусственный интеллект в лучевой диагностике: Per Aspera Ad Astra / Под ред. Ю.А. Васильева и А.В. Владзимирского. М.: Издательские решения, 2025. — 491 с. [Iskusstvennyj intellekt v luchevoj diagnostike: Per Aspera Ad Astra. Pod red. YuA Vasilieva i AV Vladzymyrskogo. M.: Izdatel'skie resheniya, 2025. 491 p. (In Russ.)]