БОЛЬШИЕ ЯЗЫКОВЫЕ МОДЕЛИ В МЕДИЦИНЕ: АКТУАЛЬНЫЕ ЭТИЧЕСКИЕ ВЫЗОВЫ

С. А. Костров ✉, М. П. Потапов

Ярославский государственный медицинский университет, Ярославль, Россия

Статья посвящена анализу актуальных этических вызовов, связанных с внедрением больших языковых моделей (LLM) в сферу медицины и здравоохранения. Рассматриваются различные архитектуры LLM, этапы их обучения (предобучение, донастройка, обучение с подкреплением на основе обратной связи от человека) и критерии качества обучающих данных. Основное внимание уделяется комплексу этических проблем: вопросам авторского права на контент, сгенерированный искусственным интеллектом (ИИ); систематической предвзятости алгоритмов и риску генерации недостоверной информации; необходимости обеспечения прозрачности и объяснимости ИИ (XAI); проблемам конфиденциальности и защиты персональных медицинских данных, включая сложности анонимизации и получения информированного согласия. Также анализируются аспекты юридической ответственности за применение LLM в клинической практике и обсуждаются технологические решения (федеративное обучение, гомоморфное шифрование) для минимизации рисков. Подчеркивается необходимость комплексного подхода, сочетающего технологическое совершенствование, разработку этических стандартов, адаптацию законодательства и критический надзор медицинского сообщества для безопасной и эффективной интеграции LLM в клиническую практику.

Ключевые слова: искусственный интеллект в медицине, большие языковые модели, авторское право генеративного текста, объяснимый ИИ, федеративное обучение ИИ, предвзятость, кибербезопасность

Вклад авторов: М. П. Потапов — планирование исследования, анализ, редактирование; С. А. Костров — сбор, анализ, интерпретация данных, подготовка черновика рукописи.

✉ **Для корреспонденции:** Сергей Александрович Костров
ул. Революцилнная, д. 5, г. Ярославль, 150000, Россия; kosea@ysmu.ru

Статья поступила: 06.05.2025 **Статья принята к печати:** 20.05.2025 **Опубликована онлайн:** 29.06.2025

DOI: 10.24075/medet.2025.008

LARGE LANGUAGE MODELS IN MEDICINE: CURRENT ETHICAL CHALLENGES

Kostrov SA ✉, Potapov MP

Yaroslavl State Medical University, Yaroslavl, Russia

The article analyzes the latest ethical challenges associated with introduction of large language models (LLMs) in medicine and healthcare. Various LLM architectures, stages of their training (pretraining, pretuning, reinforcement learning from human feedback) and criteria for quality of training data are reviewed. The emphasis is on a range of ethical issues such as copyright for AI-generated content; systematic bias in algorithms and risk of generating false information; a need to ensure transparency and explainability of AI (XAI); issues of confidentiality and protection of personal medical data, including difficulties with anonymization and obtaining informed consent. Aspects of legal responsibility for using LLMs in clinical practice are also analyzed and technological solutions (federated learning, homomorphic encryption) to minimize risks are discussed. The need for an integrated approach combining technological improvement, development of ethical standards, adaptation of legislation and critical supervision of the medical community is emphasized to ensure safe and effective integration of LLMs into clinical practice.

Keywords: artificial intelligence in medicine, large language models, generative text authorship, explainable AI, federated learning AI, bias, cybersecurity

Author contribution: Potapov MP — research planning, analysis, editing; Kostrov SA — collection, analysis, interpretation of data, preparation of a draft manuscript.

✉ **Correspondence should be addressed:** Sergey A. Kostrov
Revolutsionnaya str., 5., Yaroslavl, 150000, Russia; kosea@ysmu.ru

Received: 06.05.2025 **Accepted:** 20.05.2025 **Published online:** 29.06.2025

DOI: 10.24075/medet.2025.008

В течение последнего пятилетия искусственный интеллект (ИИ) утвердился в качестве одной из фундаментальных технологий, инициирующих трансформацию базовых парадигм функционирования медицины и системы здравоохранения [1, 2]. Признавая потенциально завышенные ожидания, связанные с данной технологией, необходимо уточнить используемую далее терминологию. В научном и профессиональном дискурсе принято различать две основные концепции ИИ: сильный (общий) искусственный интеллект (Artificial General Intelligence, AGI) — гипотетическая форма ИИ, которая обладает способностью к универсальному обучению и решению задач, аналогично человеческому интеллекту, остающаяся только теоретической и пока не реализованной на практике; слабый (узкоспециализированный) ИИ (Artificial Narrow Intelligence, ANI) — существующие программные системы, помогающие человеку решать конкретные, четко ограниченные задачи, например, диагностика заболеваний по медицинским изображениям или автоматизация рутинных операционных процессов. Именно ANI, применяемый на сегодняшний день в практической медицине, следует рассматривать под общим термином ИИ.

В рамках слабого ИИ выделяют два крупных класса: дескриптивный (описательный) и генеративный ИИ. Дескриптивные системы осуществляют анализ и интерпретацию данных (включая числовые, текстовые, графические, аудио- и видеоматериалы), обеспечивая классификацию, прогнозирование и выявление скрытых закономерностей. Генеративный ИИ, напротив, способен создавать (компилировать) новые тексты, изображения или иные форматы данных на основе обучающих выборок, что открывает новые возможности для поддержки принятия клинических решений, автоматизации

процессов документооборота и коммуникации в здравоохранении [3–5].

Особое положение в современной парадигме ИИ занимает обработка естественного языка (Natural Language Processing, NLP), позволяющая анализировать, интерпретировать и генерировать текстовую информацию на человеческом языке. Развитие и практическое значение получили большие языковые модели (Large Language Models, LLM, БЯМ) — специализированные архитектуры ИИ, способные оперировать сверхбольшими массивами текстовых данных. Данная публикации будут посвящена рассмотрению некоторых этических аспектов, связанных с применением больших языковых моделей в медицине.

За последние несколько лет наблюдается значительный рост исследований, посвященных применению LLM в медицинской сфере [6, 7]. Этические аспекты становятся центральными в дискуссии о безопасной и эффективной имплементации данных технологий в клиническую практику [8, 9]. Систематические исследования выявляют как потенциальные преимущества LLM в анализе медицинских данных, информационном обеспечении и поддержке принятия решений, так и существенные этические вызовы, связанные с алгоритмической предвзятостью, недостаточной прозрачностью и рисками нарушения конфиденциальности. Особую озабоченность вызывает способность LLM генерировать контент, обладающий высокой степенью убедительности, но потенциально содержащий неточности, что требует обеспечения человеческого надзора и разработки строгих этических руководств.

МАТЕРИАЛЫ И МЕТОДЫ

В рамках подготовки данной обзорной публикации был применен комплексный подход к поиску, анализу и отбору релевантной информации, в том числе с применением LLM. Информационный поиск осуществлялся в отечественных и международных библиографических базах данных: eLibrary, Scopus и PubMed, а также с использованием специализированных платформ поиска научных публикаций и аналитических инструментов, таких как Consensus, Semantic Scholar и Elicit, использующих в том числе LLM в своих алгоритмах. Поисковая стратегия, направленная на обеспечение полноты и актуальности охвата исследуемой тематики, включала ключевые термины и их англоязычные эквиваленты: большие языковые модели, медицина, здравоохранение, этика, биоэтика, риски, предвзятость, достоверность и др. Критериями включения источников была доступность полнотекстовой версии, опубликованной на русском или английском языках, публикации датировались интервалом 2015–2025 гг. Оценка релевантности отобранных публикаций по данным аннотации проводилась по следующим параметрам: соответствие тематике больших языковых моделей в медицине и здравоохранении, анализ этических аспектов, описание рисков внедрения, а также степень новизны и научной значимости работы. Исключались статьи, не соответствующие заявленным критериям, а также дублирующие источники. Для систематизации, извлечения данных и обобщения отобранных публикаций использовался LLM инструментарий Google NotebookLM и Perplexity. Результирующие подготовленные материалы для обеспечения точности и корректности проверялись коллективом авторов. Подготовка черновика, грамматическая корректура выполнялись с применением OpenAI ChatGPT-4.1 и Google Gemini 2.5.

АРХИТЕКТУРЫ, РАЗВИТИЕ И ОБУЧЕНИЕ LLM

Совершенствование вычислительных мощностей, доступных ресурсов и передовых алгоритмов значительно продвинуло развитие LLM, способствуя их интеграции в различные области человеческой деятельности, в том числе в клиническую практику [1, 4, 5]. Применение LLM можно разделить на три ключевые направления: поддержка принятия клинических решений, автоматизация медицинской документации и отчетности, а также медицинское образование и коммуникация между врачом и пациентом. LLM демонстрируют преимущества в обработке неструктурированных данных [3, 5]. Однако эффективность варьируется в зависимости от конкретной модели и подхода к ее обучению.

Для более полного понимания природы этических вопросов, связанных с LLM, необходимо представление о внутреннем устройстве моделей, определяющих особенности их функционирования. Современные LLM представляют собой результат многолетней эволюции архитектурных подходов в области обработки естественного языка. Хотя сейчас доминирующей архитектурой стали трансформеры, исторически развитие шло через несколько ключевых этапов и архитектур [10–13].

1. Ранние системы NLP основывались на ручном кодировании лингвистических правил (например, система ELIZA, 1966 г.). Статистические языковые модели (SLM) применялись для предсказания слов на основе частотных паттернов (например, IBM Model, 1990 г.), но они не демонстрировали широкого практического применения.
2. Рекуррентные нейронные сети (RNN) — это класс искусственных нейронных сетей, предназначенных для обработки последовательной информации. Они способны запоминать предшествующие элементы последовательности, что делает их эффективными для анализа временных рядов, текстов и биомедицинских сигналов, однако они оперировали ограниченным объемом контекста [11]. Усовершенствованными вариантами являются модели с долгой краткосрочной памятью (Long short-term memory; LSTM), которые анализируют последовательные клинические параметры (например, частоту сердечных сокращений, давление, лабораторные показатели) и выявляют паттерны, предвещающие развитие осложнений. LSTM применяются для анализа ЭКГ, ЭЭГ, данных пульсоксиметрии и других временных сигналов [13].
3. Word2Vec реализует принципы дистрибутивной семантики через векторные представления слов (алгоритмы Skip-gram и CBOW (2013)). В процессе работы текст представляется как последовательность токенов (как правило, отдельные слова или субсловные единицы), которые рассматриваются как минимальные семантические единицы. Для каждого токена Word2Vec строит эмбеддинг — отображает токен в многомерное векторное пространство, где близкие по смыслу слова имеют схожие векторные представления. Эти эмбеддинги используются для анализа семантических и синтаксических отношений в тексте.
4. Сверточные нейронные сети (CNN) — класс глубоких нейросетей, изначально специализированных для обработки данных с пространственной структурой (изображения, 3D-сканы, спектрограммы). Хотя CNN традиционно ассоциируются с анализом изображений, их архитектурные принципы извлечения

локальных признаков с помощью сверточных слоев послужили прообразом для механизмов внимания в трансформерах, став связующим звеном между обработкой локальных паттернов и глобального контекста [11, 13].

5. Трансформеры: революционная архитектура, основанная на механизме внимания. Применяя многослойные энкодеры/декодеры, модель анализирует последовательности токенов и взвешивает значимость каждого токена в контексте всей последовательности. Наиболее известные модели этого класса (например, Generative Pre-trained Transformer, GPT), предварительно обученные на огромных текстовых корпусах, стали широкодоступны и получили доминирующее положение [14].
6. Генерация, дополненная поиском (Retrieval-Augmented Generation, RAG): подход, направленный на преодоление фундаментальных ограничений классических LLM, таких как генерация фактологически некорректной информации («галлюцинации»), устаревание знаний модели и отсутствие ссылок на верифицированные источники. RAG интегрирует LLM с внешними базами знаний, такими как PubMed, UpToDate, базами клинических рекомендаций и другими авторитетными ресурсами.
7. BERT (Bidirectional Encoder Representations from Transformers) — архитектура, основанная на двунаправленных трансформерах, что обеспечивает глубокое понимание семантики и синтаксиса текста за счет учета контекста слева и справа от токена. BERT и его производные широко применяются для извлечения информации из электронных медицинских карт, автоматической классификации медицинских текстов, поиска в биомедицинских базах данных и системах поддержки принятия клинических решений.
8. Гибридные модели: для решения мультимодальных задач разрабатываются системы, комбинирующие механизмы внимания трансформеров со сверточными или рекуррентными слоями, что позволяет обрабатывать разнородные данные — от текстовых медицинских записей до визуализаций (КТ, МРТ) и временных рядов (ЭКГ, показатели мониторинга) [13].
9. Нейросимволические системы интегрируют методы машинного обучения (нейронные сети) с символическими методами представления знаний и рассуждений (формальная логика, экспертные правила, онтологии). Такие системы не только анализируют неструктурированные данные, но и используют формальные знания для повышения интерпретируемости, точности и надежности выводов. Применяются для задач с высокими требованиями к объяснимости решений, например, проверка гипотез, сгенерированных LLM, на соответствие клиническим рекомендациям [15].
10. Рассуждающие (reasoning) модели, предназначенные для решения задач, требующих сложных логических, пространственных или этических выводов, оптимизированные для имитации сложных когнитивных и логических процессов, характерных для медицинской экспертизы. В отличие от классических LLM, ориентированных преимущественно на генерацию текста и выявление паттернов, рассуждающие модели строят цепочки логических выводов, интегрируют разнородные источники знаний и объясняют свои решения на уровне, приближенном к клиническому мышлению специалиста [16].

Подобный путь эволюции технологий от базовых математических алгоритмов, через закрытые нейросетевые модели «черных ящиков» приводит к современным объяснимым (explainable) моделям [16].

ОБУЧЕНИЕ МОДЕЛЕЙ

Эволюция от основанных на жестких лингвистических правилах и статистических моделях к современным трансформерам и гибридным мультимодальным архитектурам позволила значительно расширить спектр применения LLM в клинической практике. Однако качество и надежность функционирования LLM напрямую зависят от методов их обучения и, прежде всего, от характеристик и качества исходного обучающего материала. В клиническом контексте именно исходные данные определяют границы применимости модели, уровень ее достоверности, интерпретируемость результатов и безопасность внедрения LLM в медицинские процессы [17, 18].

Предобучение (Pre-training): начальный этап, на котором модель обучается на больших неструктурированных корпусах текстов общего характера. Целью является формирование универсальных языковых представлений и базовых навыков понимания и генерации текста. Именно этот этап обучения определяет функционирование широкодоступных GPT-моделей общего назначения (YandexGPT, GigaChat, ChatGPT, Gemini, DeepSeek, Grok, Cloude и др.), способных генерировать тексты по широкому кругу тем, в том числе медицинские тексты, но зачастую только общего и поверхностного характера. При обработке запросов, касающихся сложных клинических случаев, такие модели с большой долей вероятности могут допускать ошибки. Во избежание потенциального вреда и юридических претензий, разработчики встраивают в системы модули блокировки ответов на запросы медицинского характера, а также такая LLM должна при ответе сформулировать отказ от ответственности, порекомендовав обратиться к квалифицированному врачу.

Донастройка (Fine-tuning): этап дополнительного обучения модели на специализированных клинических данных с целью адаптации к конкретным задачам, таким как генерация медицинских заключений, поддержка диагностического процесса, анализ клинических диалогов, обработка медицинских изображений и др. Наиболее эффективной считается донастройка на датасетах, размеченных экспертами и отражающих реальные клинические сценарии. Модели, прошедшие такую донастройку (например, BioGPT, BioMedLM, PubMedBERT, ClinicalBERT), применяются, как правило, в профессиональном медицинском сообществе и менее известны широкой публике [17].

Обучение с подкреплением на основе обратной связи от человека (Reinforcement Learning from Human Feedback, RLHF): метод, при котором модель корректирует свое поведение на основе оценок качества и точности сгенерированных ответов, предоставляемых экспертами. Это позволяет минимизировать риск генерации опасных или некорректных медицинских рекомендаций, а также снизить вероятность появления «галлюцинаций». Модели, обученные с применением RLHF (например, GatorTron, Med-PaLM, MetaMedLLM), используются в основном посредством интеграций, обеспечивающих доступ

к контексту в виде персонализированных медицинских записей, электронных историй болезни, совместно с комплексными и телемедицинскими решениями. RLHF утверждается в качестве стандарта для обучения LLM медицинского назначения. Исследования показывают, что LLM, прошедшие RLHF, демонстрируют превосходство по качеству и полноте медицинских консультаций по сравнению как с моделями, донастроенными без RLHF, так и с предобученными LLM. RLHF закрепляется как обязательный этап для создания современных медицинских языковых моделей, обеспечивая их соответствие требованиям клинической практики, безопасности и этики [16].

Критерии качества исходного обучающего материала:

- актуальность и достоверность. В медицинских LLM критически важно использовать только актуальные и верифицированные данные. Использование устаревших или непроверенных источников может привести к распространению ошибочных рекомендаций и создать риски для здоровья пациентов;
- репрезентативность и разнообразие. Для обеспечения справедливости (fairness) и универсальности модели обучающий материал должен охватывать широкий спектр клинических сценариев, демографических групп, языковых и культурных особенностей. Недостаточная репрезентативность приводит к систематическим ошибкам и предвзятости, особенно в отношении малых или уязвимых групп пациентов;
- качество разметки и экспертная валидация. Ошибки в аннотировании данных, неполные или некорректные инструкции ведут к снижению точности и интерпретируемости результатов. Эффективным подходом является комбинированный метод разметки, при котором эксперты формируют ядро датасета, а ИИ-алгоритмы дополняют его вариативными примерами, сочетая масштабируемость и высокое качество аннотаций.

В задачах диагностики, интерпретации медицинских изображений и клинической коммуникации модели, обученные на специализированных, экспертно размеченных данных, демонстрируют существенно более высокую точность и стабильность результатов по сравнению с моделями, обученными на общих или синтетических корпусах [1, 2, 7, 18].

ПРОБЛЕМЫ И ВЫЗОВЫ ВНЕДРЕНИЯ LLM В МЕДИЦИНЕ

Внедрение LLM сопровождается многочисленными этическими вопросами, которые требуют системного подхода к их решению. Комплексный анализ этических вызовов, связанных с LLM, выявил как давно обсуждаемые проблемы, такие как потенциальное нарушение авторских прав, систематическая предвзятость и обеспечение конфиденциальности данных, так и новые дилеммы, включая вопросы правдивости генерируемой информации и ее соответствие социальным нормам [1, 8, 9, 18].

АВТОРСКОЕ ПРАВО

Классическая доктрина авторского права основана на признании автором исключительно физического лица — человека, обладающего творческим замыслом и реализующего его в объективной форме. Появление все более автономных моделей ИИ, способных генерировать тексты, научные гипотезы, диагностические заключения, поднимает вопрос о субъекте авторского права [19–23].

В большинстве национальных правовых систем, включая страны СНГ, ЕС и США, авторское право не признает ИИ в качестве самостоятельного автора (субъекта). Это обусловлено тем, что творческий акт предполагает наличие воли, сознания и субъективного выбора, которыми современные ИИ не обладают. Статья 1228 Гражданского кодекса РФ четко определяет автором произведения исключительно гражданина (физическое лицо), творческим трудом которого создано произведение науки, литературы или искусства. ИИ не обладают правоспособностью и не могут осуществлять творческую деятельность в юридическом смысле.

Тем не менее, растущий объем медицинских текстов, генерируемых LLM, требует пересмотра устоявшихся подходов. Медицинская сфера предъявляет особые требования к качеству, достоверности и юридической чистоте информации. В отличие от художественной или публицистической деятельности, здесь на кону стоят здоровье и жизнь пациентов, а также профессиональная репутация медицинских работников и исследователей [18]. Использование LLM для автоматизированного создания медицинских текстов, протоколов, анализов данных и даже научных статей порождает ряд специфических рисков.

1. Неочевидность источников: LLM обучаются на огромных корпусах текстов, зачастую без четкого разграничения между открытыми и защищенными авторским правом материалами. Это затрудняет идентификацию источников заимствований и может привести к непреднамеренному нарушению прав третьих лиц [20].
2. Проблема плагиата: автоматическая генерация текстов может приводить к созданию производных работ или текстов, частично совпадающих с оригинальными источниками, что создает угрозу обвинений в плагиате со стороны правообладателей.
3. Сложности с атрибуцией: в случае совместного творчества человека и ИИ возникает вопрос об определении вклада каждого участника и порядке распределения авторских прав.

Можно выделить три основных подхода к определению авторства при создании объектов с участием ИИ [23].

Автор — разработчик ИИ. Предполагается, что все права на результаты, созданные с помощью ИИ, принадлежат лицу или организации, разработавшим соответствующую модель. Разработчик вкладывает значительные интеллектуальные усилия и творческий потенциал в создание самой системы ИИ, включая разработку алгоритмов, архитектуры и подготовку данных для обучения, что требует существенных финансовых, временных и человеческих ресурсов со стороны разработчика [23]. Признание авторских прав за разработчиком может служить стимулом для дальнейших инвестиций и инноваций в этой области. Данный вариант дает более простой и предсказуемый механизм определения правообладателя по сравнению с другими. Однако такой подход оправдан, только если пользователь не вносит существенного творческого вклада, а лишь нажимает кнопку для генерации случайного произведения без дальнейшего творческого вмешательства.

Автор — пользователь ИИ. В этом случае автором признается лицо, непосредственно управляющее ИИ

и формирующее запросы. Пользователь выбирает из предложенных вариантов, корректирует и направляет работу ИИ для достижения желаемого результата. Детально сформулированный и креативный запрос может привести к созданию уникального произведения, в то время как общий или стандартный запрос, вероятно, даст более типичный результат. ИИ выступает в роли усовершенствованного инструмента, позволяющего реализовать творческий замысел пользователя, направляя процесс. Эта модель чаще всего применяется в медицинской и юридической практике при условии активного участия пользователя (врача, исследователя) [22–24].

Автор — ИИ (концепция «электронной личности»). В контексте резолюции Европейского парламента с рекомендациями по гражданско-правовым нормам о робототехнике обсуждается возможность признания ИИ самостоятельным субъектом авторских прав [23, 24]. Современные генеративные системы демонстрируют все большую степень автономии в процессе создания произведений. Вклад ИИ может выходить за рамки простого инструментального использования, и система способна генерировать непредвиденные и оригинальные результаты, которые не были напрямую заложены разработчиком или проконтролированы человеком. Однако на практике этот подход не получил признания, поскольку ИИ не обладает ни правосубъектностью, ни способностью к самостоятельной реализации прав и обязанностей. Международная практика свидетельствует о том, что в подавляющем большинстве случаев суды и ведомства по интеллектуальной собственности отказывают в признании авторства за ИИ [22].

Таким образом, мы придерживаемся мнения, что вклад участников в создание произведения (как художественного текста, так и научно-исследовательского, а также любые медицинские записи, сгенерированные LLM) является многоуровневым. В случаях, когда вклад пользователя и работа ИИ (как результат труда разработчика и данных) неразделимы, необходимо применение концепции совместного авторства с компенсацией владельцам авторских прав пропорционально их вкладу в создание контента. Пользователи ИИ в зависимости от выбранного тарифа в определенной мере приобретают ИИ как услугу, усиливая свои авторские позиции.

В то же время в ряде стран обсуждаются варианты введения особых режимов охраны для произведений, созданных с минимальным участием человека, например сокращенного срока действия авторских прав [21], одновременно обеспечив вознаграждение авторам, чьи произведения использовались для обучения ИИ.

Помимо юридических аспектов, использование LLM в медицине порождает ряд научных дилемм. Возможно снижение роли человеческого творчества. Экспоненциальный рост объема контента, генерируемого ИИ, может привести к девальвации человеческого вклада и снижению мотивации к самостоятельному научному поиску. Автоматическая генерация медицинских текстов без должной экспертной валидации чревата распространением недостоверной или даже опасной информации.

Современная правовая система пока не готова к полному учету специфики ИИ-генерируемых объектов, что требует выработки новых подходов к определению авторства, охраноспособности и распределению прав на результаты интеллектуальной деятельности.

С учетом изложенных проблем предлагаются следующие направления развития:

- введение специальных режимов охраны для произведений, созданных с использованием ИИ, например сокращенного срока действия прав;
- обязательное раскрытие степени участия ИИ при публикации медицинских статей, разработке клинических протоколов и иных научных материалов;
- разработка международных стандартов по атрибуции и идентификации источников при использовании LLM;
- создание более совершенных систем отслеживания заимствований и проверки на плагиат на основе токенизированной информации;
- начисление вознаграждений разработчикам и авторам материалов, на основе которых обучаются модели, в том числе посредством систем платных подписок.

ПРЕДВЗЯТОСТЬ (BIAS), ГАЛЛЮЦИНАЦИИ И ОБЪЯСНИМЫЙ ИИ

Несмотря на значительный прогресс в снижении частоты фактологических ошибок («галлюцинаций») в современных LLM, особенно в узкоспециализированных системах, донастроенных с применением RLHF (где достигается релевантность ответов выше 95%), новым серьезным вызовом является проблема систематической предвзятости (systematic bias), которая приводит к ошибкам в медицинских рекомендациях, дискриминации уязвимых групп пациентов, искажению медицинских знаний, обуславливающих снижение доверия к ИИ в здравоохранении [24, 25].

Систематическая предвзятость — это устойчивое искажение результатов работы модели, обусловленное особенностями данных, архитектуры или процессов обучения, приводящее к искаженному или неточному представлению определенных групп, явлений или концепций, а также к искаженной интерпретации клинических данных. Это не случайные сбои, а следствие внутренней логики работы алгоритмов. Алгоритмические системы способны не только воспроизводить, но и амплифицировать существующие предубеждения, создавая потенциально опасный цикл усиления дискриминации [26].

LLM обучаются на корпусах текстов, которые могут содержать исторические, социальные и культурные предубеждения, а также несбалансированное представление медицинских знаний. Ошибки или субъективизм при разметке медицинских данных могут закреплять предвзятость на этапе подготовки датасетов.

Особенности трансформеров, механизмы внимания и способы обработки контекста могут как усиливать, так и ослаблять предвзятость. Как уже обозначалось GPT — это автогрегрессивная трансформерная модель, обученная предсказывать следующий токен на основе статистических закономерностей в обучающих данных. Они склонны воспроизводить наиболее часто встречающиеся паттерны, закрепляя существующие предубеждения и медицинские стереотипы, что может проявляться в диспропорциональном внимании к определенным аспектам информации, коррелирующим с демографическими характеристиками, или в некорректной интерпретации редких или неоднозначных

случаев [25]. GPT не имеет встроенных механизмов проверки фактов или соответствия клиническим стандартам. Увеличение размера модели не всегда гарантирует уменьшение предвзятости; некоторые ее формы могут даже усиливаться [14].

Рассуждающие модели хотя и включают механизмы логического вывода (например, Chain-of-Thought, CoT), они все равно могут воспроизводить предвзятые паттерны рассуждений, если таковые присутствовали в обучающих данных, более того, предвзятость в цепочках рассуждений сложнее обнаружить, так как возможен эффект подтверждения (confirmation bias). Критической проблемой является то, что генерируемые объяснения (рационализации) могут маскировать истинные (возможно, предвзятые) причины предсказания модели, особенно при неверных ответах. Подход к снижению риска — использование выражения неопределенности, когда модель указывает степень уверенности в своем ответе, позволяя клиницистам учитывать это при интерпретации. Когда модели явно выражают свою неопределенность, их прогнозы становятся менее категоричными и менее подверженными систематическим ошибкам [25]. Представления неопределенности могут быть использованы как дополнительный фильтр для выявления случаев, в которых модель потенциально предвзята или не уверена, что позволяет либо отложить решение, либо привлечь эксперта.

Интеграция с внешними базами знаний в моделях RAG потенциально снижает предвзятость за счет доступа к актуальным и доказательным фактам. Однако RAG-модели могут некорректно агрегировать противоречивую информацию из источников или воспроизводить предвзятость, если она содержится во внешних базах данных. Трудно обеспечить воспроизводимость решений, поскольку даже при идентичных запросах модель может ссылаться на разные источники, что затрудняет аудит и коррекцию предвзятости.

В целом, все LLM алгоритмически склонны к генерации наиболее вероятных (частотных) ответов, что приводит к игнорированию редких, но клинически значимых случаев. Использование LLM без экспертной валидации рискует закреплять и распространять предвзятость [14].

Исследования демонстрируют, что большие языковые модели проявляют существенные различия между их «явными убеждениями» (revealed beliefs) и «заявленными ответами» (stated answers), что указывает на наличие множественных предвзятостей и искажений в формируемых ими представлениях [26].

Еще одной из проблем является диссонанс между вероятностной природой алгоритмических выводов и их субъективным восприятием пациентами (а в некоторых случаях и врачами) как детерминированных предсказаний [27].

Исследования в области коммуникации рисков подтверждают, что эффективность передачи медицинской информации существенно зависит от способа представления данных пациенту [27]. Категоричные формулировки прогностических заключений индуцируют выраженные психологические реакции даже в случаях низкой статистической вероятности прогнозируемого исхода. Оптимистичные формулировки создают иллюзию контролируемости, заставляя пациентов недооценивать объективные риски и даже к преждевременному прекращению терапии.

Автоматизационное смещение — тенденция воспринимать алгоритмические выводы как более объективные по сравнению с человеческими суждениями. Цифровые интерфейсы подсознательно вызывают доверие к источникам. Излишнее доверие к алгоритмическим советникам представляет собой комплексный феномен возникновения новых форм зависимости. Многие пользователи склонны приписывать системам ИИ свойства «сверхчеловеческого интеллекта», игнорируя ограничения обучающих данных и архитектурные особенности моделей. Экспериментальные данные показывают, что 68% респондентов готовы следовать советам ИИ даже вопреки мнению лечащего врача [27]. Клинические проявления алгоритмической зависимости включают: компульсивную проверку прогнозов через мобильные приложения, тревожно-фобические реакции при временной недоступности сервиса, отказ от самостоятельного анализа симптомов в пользу автоматизированных диагнозов.

Разработка методологий для количественной оценки предвзятости и степени достоверности ответов в медицинских LLM представляет собой важное направление дальнейших исследований [28, 29].

Несмотря на беспрецедентный потенциал LLM в медицине, их широкому внедрению препятствует непрозрачность механизмов принятия решений для большинства пользователей, что снижает доверие со стороны медицинских специалистов и пациентов. Многие большие языковые модели, такие как GPT-4, представляют собой сложные нейросетевые архитектуры с миллиардами параметров, чье внутреннее функционирование часто остается непостижимым для многих пользователей («черный ящик») [14].

Объяснимый искусственный интеллект (Explainable Artificial Intelligence, XAI) представляет собой направление исследований, ориентированное на разработку методологий и технологий, которые делают процесс принятия решений ИИ-системами понятным для человека, обеспечивая возможность верификации результатов и помогая в преодолении барьера недоверия к ИИ-технологиям [30].

Создание моделей, которые изначально обладают высокой степенью интерпретируемости, является базовым решением (например, линейные модели и деревья решений, которые позволяют явно проследить взаимосвязь между входными данными (вклад каждого признака) и выходными результатами). Хотя эти модели могут уступать в предиктивной точности более сложным архитектурам на некоторых задачах [16].

Генерация промежуточных этапов рассуждения перед выдачей окончательного ответа Chain-of-Thought (CoT) повышает не только точность, но и объяснимость, позволяя проследить логическую цепочку модели. Объяснения могут быть адаптированы для разных аудиторий (врачи, пациенты, регуляторы).

Как уже указывалось ранее, обязательными становятся применение методологии RAG, предоставление моделям доступа к актуальной научной литературе, клиническим рекомендациям и другим верифицируемым источникам, что повышает точность, надежность и прозрачность генерируемой информации. Примером оценки может служить Medical Information Retrieval-Augmented Generation Evaluation (MIRAGE) — первый бенчмарк, включающий 7663 вопроса из пяти медицинских наборов данных для вопросно-ответных систем. Исследования с использованием MIRAGE продемонстрировали, что применение MedRAG по сравнению с методом подсказок на основе цепочки рассуждений улучшает точность ответов различных LLM до 18% [31].

По актуальному состоянию на май 2025 г. бенчмарк MedAgentsBench включает 1453 структурированных клинических случая, охватывающих 13 систем органов и 10 медицинских специальностей. По результатам сравнения в марте 2025 г. лидером являются рассуждающие модели DeepSeek R1 и OpenAI-o3, обеспечивая не только высокую точность, но и оптимальное соотношение между производительностью, стоимостью вычислений и временем вывода, что особенно важно для практического внедрения в медицинских информационных системах, продемонстрировав точность в простых диагностических задачах OpenAI-o3 = 89%, DeepSeek R1 = 93%, однако в сложных сценариях, требующих многоэтапного планирования лечения, показатель снижался до значений OpenAI-o3 = 67%, DeepSeek R1 = 73% [32].

Остро стоит проблема отсутствия стандартизированных метрик и протоколов оценки качества объяснений. Существующие методы XAI генерируют объяснения различного формата и содержания, и на данный момент не существует консенсуса относительно того, какими свойствами должно обладать «хорошее» объяснение и как эти свойства можно объективно измерить [18, 32].

КОНФИДЕНЦИАЛЬНОСТЬ И ЗАЩИТА ПЕРСОНАЛЬНЫХ ДАННЫХ

Использование реальных клинических данных для обучения и применения LLM требует строгого соблюдения стандартов анонимизации и конфиденциальности пациентов, что накладывает дополнительные требования к подготовке обучающих выборок [33, 34].

Эффективность цифровых медицинских технологий напрямую зависит от доверия пациентов. Нарушение конфиденциальности подрывает доверие к системе здравоохранения в целом и может привести к отказу пациентов от предоставления полной и достоверной информации, что негативно скажется на качестве медицинской помощи. Персонализированные LLM повышают качество лечения через учет индивидуальных особенностей, но требуют обработки ультрачувствительных данных (о геноме, образе жизни и психическом статусе пациента) [14].

Медицинские данные характеризуются высокой степенью чувствительности: они содержат сведения о диагнозах, результатах анализов, генетических особенностях, истории болезней и иной информации, способной идентифицировать личность пациента и подлежат строгой правовой и этической защите. LLM обучаются на больших массивах информации, включающих не только открытые источники, но и специализированные медицинские базы данных. Даже после формального обезличивания сохраняется риск восстановления личности пациента на основе косвенных признаков, что особенно актуально для редких заболеваний или уникальных сочетаний клинических признаков.

Вступающий в силу с 1 сентября 2025 г. в России приказ Минздрава от 20 марта 2025 г. № 139н «Об утверждении Порядка обезличивания сведений о лицах, которым оказывается медицинская помощь, а также о лицах, в отношении которых проводятся медицинские экспертизы, медицинские осмотры и медицинские освидетельствования», пришедший на смену приказа № 341н от 14 июня 2018 г., предписывает обезличивать все сведения, позволяющие прямо или косвенно идентифицировать личность пациента, включая ФИО, дату рождения, адрес, контактные данные, индивидуальные номера документов и иные идентификаторы. Процедура должна обеспечивать невозможность восстановления личности пациента без использования дополнительной информации, хранящейся отдельно и защищенной в соответствии с законодательством РФ [35].

Однако даже при удалении прямых идентификаторов (имя, дата рождения, адрес) в медицинских данных сохраняются квази-идентификаторы (например, редкое сочетание симптомов, уникальные схемы лечения), которые могут быть использованы для реидентификации личности пациента. Исследование LLM-Anonymizer продемонстрировало сохранение около 2% идентифицирующей информации после обработки [36]. Исследования демонстрируют, что злоумышленники могут восстанавливать исходные тексты из векторных представлений моделей с точностью до 92% с помощью методов атаки инверсии [37].

Этические стандарты требуют использовать минимально необходимый объем данных для достижения поставленной цели. Однако LLM, обучаясь на огромных датасетах, часто обрабатывают избыточные сведения, что затрудняет контроль за обработкой информации и увеличивает площадь потенциальной утечки.

В большинстве случаев пациенты дают согласие на обработку своих данных для конкретных целей: диагностики, лечения, научных исследований. Классические требования полноты информации, добровольности и компетентности пациента вступают в противоречие с технической сложностью ИИ. Использование LLM, способных к генерации новых знаний и переиспользованию информации в непредвиденных сценариях, выходит за рамки стандартных форм согласия. Пациенты зачастую не осведомлены о том, что их данные могут быть использованы для обучения сложных моделей, которые впоследствии применяются в широком спектре задач. Большинство пациентов не обладают специализированными знаниями, позволяющими оценить архитектуру нейронных сетей, качество обучающих данных или ограничения алгоритмов [18, 34].

LLM непрерывно обновляются, что делает традиционное статичное информирование нерелевантным уже на этапе подписания согласия. Динамическое информированное согласие (dynamic informed consent) — это современная модель взаимодействия между пациентом и медицинской организацией, предполагающая не разовое, а непрерывное, поэтапное информирование пациента и получение его согласия на каждом этапе взаимодействия. Пациент получает информацию не только на этапе начала лечения, но и при каждом существенном изменении в алгоритме ИИ, обновлении программного обеспечения или появлении новых клинических данных, влияющих на принятие решений. Необходимо применение интерактивных цифровых платформ, позволяющих пациенту в реальном времени получать уведомления, разъяснения и давать согласие на новые этапы взаимодействия [38, 39].

В России действует экспериментальный правовой режим для развития и внедрения искусственного интеллекта (ИИ) в здравоохранении, автоматически подразумевая согласие пациентов на передачу обезличенных медицинских данных для обучения искусственного интеллекта [40], по окончании действия которого врачебному сообществу необходимо определиться с формами и способами работы с динамическими согласиями.

Существующие законы (например, HIPAA в США, GDPR в ЕС, ФЗ-152 в РФ) устанавливают требования к защите персональных данных, но не учитывают специфику работы LLM. Требование «права на забвение» сталкивается с технической сложностью выборочного удаления данных в предобученных моделях. Возникают вопросы распределения ответственности за утечки данных (разработчик, медучреждение, пользователь) и соблюдения правил трансграничной передачи данных.

Необходимы комплексные регуляторные меры: проведение обучения персонала по вопросам кибербезопасности и этики работы с медицинскими данными, введение многоуровневой системы контроля доступа к исходным данным и результатам работы моделей, регулярное тестирование моделей на предмет воспроизведения чувствительной информации, внедрение алгоритмов обнаружения и фильтрации персональных данных на этапе генерации ответов моделей, использование методов дифференциальной приватности, позволяющих обучать LLM на агрегированных данных без риска восстановления индивидуальных записей. Требуются актуализация законодательства с учетом специфики работы LLM, введение специальных требований к анонимизации и аудиту моделей, отраслевых стандартов для сертификации алгоритмов обезличивания, обеспечение прозрачности процессов обработки данных и информирования пациентов о возможных рисках.

Технологические решения, такие как добавление гауссовского шума к эмбеддингам, снижают риск инверсии на 60%, но одновременно ухудшают производительность моделей. Федеративное обучение (Federated Learning, FL) и гомоморфное шифрование (Homomorphic Encryption, HE) формируют технологический симбиоз, позволяющий обрабатывать чувствительные медицинские данные без их прямой экспозиции [41].

Федеративное обучение реализует децентрализованный подход, где модели обучаются на локальных наборах данных без их передачи центральному серверу. Это позволяет минимизировать риски утечек при трансграничных исследованиях, объединять знания из разнородных источников (лаборатории, больницы, носимые устройства). Эксперименты с FL-фреймворком Flower демонстрируют высокую точность при значительном снижении рисков конфиденциальности [42].

Гомоморфные схемы шифрования позволяют выполнять вычисления над зашифрованными данными без необходимости их предварительной расшифровки. Суть гомоморфного шифрования заключается в том, что если исходные данные были зашифрованы, то над этим шифротекстом можно производить определенные математические операции (например, сложение, умножение), и результат этих операций также будет находиться в зашифрованном виде. После расшифровки результата врач получает тот же итог, который был бы получен при выполнении аналогичных операций над исходными незашифрованными данными. Однако для оптимизации таких вычислений требуется специализированное дорогостоящее вычислительное оборудование [43].

Прототип MedSecureAI демонстрирует, что такое сочетание FL+HE снижает риск утечек на 99.2% при увеличении времени обучения всего в 2,1 раза по сравнению с базовыми моделями [41]. Это порождает дополнительные технологические вызовы: создание специализированных процессоров для медицинского HE, разработка межгосударственных стандартов обмена зашифрованными моделями, интеграция постквантовых криптографических алгоритмов.

ЮРИДИЧЕСКАЯ ОТВЕТСТВЕННОСТЬ LLM-РЕЗУЛЬТАТОВ

С правовой точки зрения, на сегодняшний день LLM не обладают статусом самостоятельных субъектов права. Они рассматриваются исключительно как инструменты, созданные и используемые физическими или юридическими лицами. Юридическая ответственность за последствия применения LLM возлагается на разработчиков, поставщиков программного обеспечения, а также на медицинских работников и организации, использующие эти технологии [44].

Разработчики и поставщики обязаны обеспечивать соответствие своих продуктов установленным стандартам качества и безопасности, а также информировать пользователей о возможных ограничениях и рисках.

Все медицинские изделия, включая программное обеспечение на основе больших языковых моделей, подлежат обязательной государственной регистрации перед их внедрением в клиническую практику. В зависимости от потенциального вреда при ошибке, ИИ-решения относятся к классам риска: IIa (средний риск — системы предварительной обработки медицинской документации, первичного скрининга), IIb (повышенный риск — системы автоматизированной интерпретации результатов инструментальных исследований, алгоритмы прогнозирования течения заболеваний, программное обеспечение для поддержки принятия клинических решений) или III (высокий — системы ИИ, принимающие самостоятельные клинические решения, формирующие диагностические и лечебные рекомендации и применяемые автономно в имплантируемых медицинских устройствах), поскольку их ошибки могут привести к значимым последствиям для жизни и здоровья пациента. Регистрация требует проведения клинической оценки, подтверждения качества алгоритмов, обеспечения прозрачности и воспроизводимости результатов, а также внедрения механизмов управления рисками и непрерывного мониторинга функционирования [45]. Росздравнадзор осуществляет мониторинг и может приостанавливать применение скомпрометированных решений для принятия корректирующих мероприятий (как, например, это было в 2023–2024 гг. с системой Botkin.AI).

Особое внимание разработчики и эксплуатирующие организации обязаны уделять вопросам информационной безопасности. Информационные системы, обрабатывающие персональные данные пациентов, становятся приоритетной целью для злоумышленников. Современные киберугрозы, включая несанкционированный доступ, атаки на целостность и конфиденциальность данных, а также манипуляции с выводами моделей, способны привести к серьезным последствиям, представляющим угрозы не только здоровью, но жизни пациентов [45, 46]. Данные медицинские информационные системы подпадают под действие Федерального закона от 26.07.2017 № 187-ФЗ «О безопасности критической информационной инфраструктуры Российской Федерации».

Медицинские работники, в свою очередь, несут профессиональную ответственность за принятие клинических решений, даже если они опираются на рекомендации, сформулированные LLM. Врач обязан

критически оценивать полученную информацию и не может полностью делегировать принятие решений искусственному интеллекту [18].

В случае возникновения негативных последствий, связанных с ошибками или недостоверными рекомендациями LLM, ответственность может быть распределена между различными участниками процесса в зависимости от характера и источника ошибки. Если речь идет о дефекте программного обеспечения, ответственность, как правило, возлагается на разработчика. Если же ошибка возникла вследствие некорректного применения технологии или игнорирования врачом профессиональных стандартов и клинических рекомендаций, ответственность несет медицинский работник или организация.

ЗАКЛЮЧЕНИЕ

Таким образом, внедрение больших языковых моделей в здравоохранение требует комплексного подхода, сочетающего дальнейшее технологическое совершенствование моделей, разработку и внедрение строгих этических стандартов, адаптацию нормативно-правовой базы, применение передовых методов информационной безопасности и постоянный критический надзор со стороны экспертного медицинского сообщества.

Одним из ключевых направлений является совершенствование алгоритмов и архитектур. Необходимо отдавать предпочтение современным моделям, сочетающим в себе возможности рассуждений, поиска и объяснения. Переход от прогностических моделей типа «черный ящик» к интерпретируемым системам, способным обосновывать свои выводы, повысит доверие к этим технологиям со стороны медицинских специалистов и пациентов. Важным шагом является развитие нейросимволических методов, интегрирующих машинное обучение с символическими представлениями знаний и логическими рассуждениями. Это позволит не только анализировать неструктурированные данные, но и использовать формальные знания для повышения точности и надежности выводов.

Не менее важным является обеспечение качества и релевантности обучающих данных. LLM должны быть не только предварительно обучены на больших корпусах текстов, но и донастроены на узкоспециализированных предразмеченных клинических данных с участием врачей-экспертов. Обучение с подкреплением на основе обратной связи от эксперта (RLHF) должно стать стандартом для медицинских языковых моделей, подтверждая их соответствие требованиям клинической практики, безопасности и этики. Это позволит обеспечить не только релевантность ответов в общем случае, но и их персонификацию и клиническую доказательность

Обязательным условием успешного внедрения LLM в здравоохранение является адаптация нормативно-правовой базы к технологическим достижениям. Специалистам в области права необходимо учитывать специфику усиливающейся интеграции ИИ во все сферы деятельности и разработать новые подходы к определению авторства, охраноспособности и распределению прав на результаты интеллектуальной деятельности, созданные с участием LLM.

Обеспечение конфиденциальности и защиты персональных данных является необходимым условием. Важно строго соблюдать стандарты анонимизации и конфиденциальности пациентов при использовании реальных клинических данных для обучения и применения LLM. Следует использовать минимально необходимый объем данных для достижения поставленной цели и внедрять технологические решения, такие как федеративное обучение и гомоморфное шифрование, позволяющие обрабатывать чувствительные медицинские данные без их прямой экспозиции. Важно также разработать интерактивные цифровые платформы, позволяющие пациенту в реальном времени получать уведомления, разъяснения и давать согласие на новые этапы взаимодействия (динамическая форма согласия).

Исключение недостоверных ответов, этапный фактчекинг и кросс-проверка необходимы для борьбы с «галлюцинациями», предвзятостью. Необходимо разрабатывать методологии для количественной оценки степени достоверности ответов в медицинских LLM, позволяющие клиницистам учитывать это при интерпретации результатов. Важно учитывать культурные и языковые особенности различных групп пациентов и разрабатывать модели, учитывающие эти различия.

Для формирования у пациентов и врачебного сообщества объективного доверительного отношения к применяемым технологиям ИИ необходимо обеспечить прозрачность и объяснимость работы LLM. Это требует разработки стандартизированных метрик и протоколов оценки качества, а также применения методов XAI, позволяющих проследить логическую цепочку модели и адаптировать объяснения для разных аудиторий. Важно также учитывать психологические аспекты восприятия информации, предоставляемой LLM, и избегать категоричных формулировок, которые могут индуцировать выраженные психологические реакции.

Только при соблюдении этих условий можно добиться качественного повышения уровня здравоохранения за счет использования больших языковых моделей, обеспечив при этом защиту прав и интересов пациентов и медицинских работников.

Литература

1. Meng X, Yan X, Zhang K, et al. The application of large language models in medicine: A scoping review. iScience. 2024; 27(5): 109713.
2. Omiye JA, Gui H, Rezaei SJ, Zou J, Daneshjou R. Large Language Models in Medicine: The Potentials and Pitfalls: A Narrative Review. Ann Intern Med. 2024; 177(2): 210–220. DOI: 10.7326/M23-2772.
3. Chen Y, Esmaeilzadeh P. Generative AI in Medical Practice: In-Depth Exploration of Privacy and Security Challenges. J Med Internet Res. 2024; 26: e53008. DOI: 10.2196/53008.
4. Yim D, Khuntia J, Parameswaran V, Meyers A. Preliminary Evidence of the Use of Generative AI in Health Care Clinical Services: Systematic Narrative Review. JMIR Med Inform. 2024; 12: e52073. DOI: 10.2196/52073.
5. Nógrádi B, Polgár TF, Meszlényi V, et al. ChatGPT M. D. Is there any room for generative AI in neurology? PLoS One. 2024; 19(10): e0310028. DOI: 10.1371/journal.pone.0310028.
6. Wang C, Li M, He J, Wang Z, Darzi E, Chen Z, et al. A Survey for Large Language Models in Biomedicine. ArXiv. 2024; abs/2409.00133.

7. Moglia A, Georgiou K, Cerveri P, et al. Large language models in healthcare: from a systematic review on medical examinations to a comparative analysis on fundamentals of robotic surgery online test. Artif Intell Rev. 2024; 57: 231. DOI: 10.1007/s10462-024-10849-5.
8. Ong JCL, Chang SY, William W, et al. Ethical and regulatory challenges of large language models in medicine. Lancet Digit Health. 2024; 6(6): e428-e432. DOI: 10.1016/S2589-7500(24)00061-X.
9. Zhui LL, Fenghe L, Xuehu W, Qining F, Wei R. Ethical Considerations and Fundamental Principles of Large Language Models in Medical Education: Viewpoint. J Med Internet Res. 2024; 26: e60083. DOI: 10.2196/60083
10. Wei Y, Zhou J, Wang Y, et al. A Review of Algorithm & Hardware Design for AI-Based Biomedical Applications. IEEE Trans Biomed Circuits Syst. 2020; 14(2): 145–163. DOI: 10.1109/TBCAS.2020.2974154
11. Пикалев Я. С. Обзор архитектур систем интеллектуальной обработки естественно-языковых текстов. Проблемы искусственного интеллекта. 2020; 4(19).
12. Jin L, Feng S, Xin Z, Chai Y. Evolution and advancements in deep learning models for Natural Language Processing. Applied and Computational Engineering. 2024.
13. Li K, Ao B, Wu X, Wen Q, Ul Haq E, Yin J. Parkinson's disease detection and classification using EEG based on deep CNN-LSTM model. Biotechnol Genet Eng Rev. 2024; 40(3): 2577–2596. DOI: 10.1080/02648725.2023.2200333
14. Haltaufderheide J, Ranisch R. The ethics of ChatGPT in medicine and healthcare: a systematic review on Large Language Models (LLMs). NPJ Digit Med. 2024; 7(1): 183. Published 2024 Jul 8. DOI: 10.1038/s41746-024-01157-x.
15. Garcez AA, et al. Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. arXiv preprint arXiv:1905.06088. 2019.
16. Li Z, et al. From system 1 to system 2: A survey of reasoning large language models. arXiv preprint arXiv: 2502.17419. 2025.
17. Cascella M, Semeraro F, Montomoli J, Bellini V, Piazza O, Bignami E. The Breakthrough of Large Language Models Release for Medical Applications: 1-Year Timeline and Perspectives. J Med Syst. 2024; 48(1): 22. DOI: 10.1007/s10916-024-02045-3.
18. Андрейченко А. Е., Гусев А. В. Перспективы применения больших языковых моделей в здравоохранении. Национальное здравоохранение. 2023; 4 (4): 48–55. DOI: 10.47093/2713-069X.2023.4.4.48-55.
19. Moffatt B, Hall A. Is AI my co-author? The ethics of using artificial intelligence in scientific publishing. Account Res. 2024. DOI: 10.1080/08989621.2024.2386285.
20. Lee JY. Can an artificial intelligence chatbot be the author of a scholarly article? J Educ Eval Health Prof. 2023; 20: 6. DOI: 10.3352/jeehp.2023.20.6.
21. Johnson A. Generative AI, UK Copyright and Open Licences: considerations for UK HEI copyright advice services. F1000Res. 2024; 13: 134. DOI: 10.12688/f1000research.143131.1.
22. Шайдуров А. С. Авторское право на произведения, созданные искусственным интеллектом в РФ: проблемы и перспективы. В сборнике: Цифровые технологии в научном развитии: новые концептуальные подходы: Сборник статей по итогам Международной научно-практической конференции; 30 апреля 2023 г.; Самара. Стерлитамак: Общество с ограниченной ответственностью «Агентство международных исследований». 2023; 88–92. EDN MDZVBE.
23. Кишкембаев М. Актуальные проблемы защиты авторских прав на объект, созданный искусственным интеллектом. Vestnik Torajgyrov Universiteta. Ûridičeskaâ Seriâ. Март 2024; 75–86. DOI: 10.48081/zusz1934.
24. Avila Negri SMC. Robot as Legal Person: Electronic Personhood in Robotics and Artificial Intelligence. Front Robot AI. 2021;8:789327. Published 2021 Dec 23. DOI: 10.3389/frobt.2021.789327.
25. Schmidgall S, Harris C, Essien I, et al. Evaluation and mitigation of cognitive biases in medical language models. NPJ Digit Med. 2024; 7(1): 295. DOI: 10.1038/s41746-024-01283-6.
26. Mondal M, et al. Do large language models exhibit cognitive dissonance? studying the difference between revealed beliefs and stated answers. Preprint arXiv. 2406.14986. 2024.
27. Goerlandt F, Li J, Reniers G. The Landscape of Risk Communication Research: A Scientometric Analysis. International Journal of Environmental Research and Public Health. 2020; 17(9): 3255. DOI: 10.3390/ijerph17093255.
28. Zarfati M, Soffer S, Nadkarni GN, Klang E. Retrieval-Augmented Generation: Advancing personalized care and research in oncology. Eur J Cancer. 2025; 220: 115341. DOI: 10.1016/j.ejca.2025.115341.
29. Shool S, Adimi S, Saboori Amleshi R, Bitaraf E, Golpira R, Tara M. 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.
30. Шевская Н. В. Объяснимый искусственный интеллект и методы интерпретации результатов. Моделирование, оптимизация и информационные технологии. 2021; 9(2). DOI: 10.26102/2310-6018/2021.33.2.024. EDN VRKUIL.
31. Xiong G, et al. Benchmarking retrieval-augmented generation for medicine. Findings of the Association for Computational Linguistics ACL. 2024. 2024; 6233–6251.
32. Tang X et al. Medagentsbench: Benchmarking thinking models and agent frameworks for complex medical reasoning. arXiv preprint arXiv:2503.07459. 2025.
33. Yadav N, Pandey S, Gupta A, Dudani P, Gupta S, Rangarajan K. Data Privacy in Healthcare: In the Era of Artificial Intelligence. Indian Dermatol Online J. 2023;14(6): 788–792. DOI: 10.4103/idoj.idoj_543_23.
34. Алтынников М. С., Кузнецова Н. О. Особенности организации защиты персональных данных в медицинском учреждении. Инновации. Наука. Образование. 2021; 36: 1479–1486. EDN ZISIGQ.
35. Приказ Минздрава России от 20.03.2025 № 139н «Об утверждении Порядка обезличивания сведений о лицах, которым оказывается медицинская помощь, а также о лицах, в отношении которых проводятся медицинские экспертизы, медицинские осмотры и медицинские освидетельствования». Официальный интернет-портал правовой информации. 2025.
36. Wiest IC, et al. Anonymizing medical documents with local, privacy preserving large language models: The LLM-Anonymizer. medRxiv. 2024. C. 2024.06. 11.24308355.
37. Morris JX, et al. Text embeddings reveal (almost) as much as text. arXiv preprint arXiv:2310.06816. 2023.
38. Mascalzoni D, Melotti R, Pattaro C, et al. Ten years of dynamic consent in the CHRIS study: informed consent as a dynamic process. Eur J Hum Genet. 2022;30: 1391–1397 DOI: 10.1038/s41431-022-01160-4. 2022.
39. Harishbhai Tilala M, Kumar Chenchala P, Choppadandi A, et al. Ethical Considerations in the Use of Artificial Intelligence and Machine Learning in Health Care: A Comprehensive Review. Cureus. 2024; 16(6): e62443. Published 2024 Jun 15. DOI: 10.7759/cureus.62443.
40. Постановление Правительства РФ от 18.07.2023 № 1164 (ред. от 01.02.2025) «Об установлении экспериментального правового режима в сфере цифровых инноваций и утверждении Программы экспериментального правового режима в сфере цифровых инноваций по направлению медицинской деятельности, в том числе с применением телемедицинских технологий и технологий сбора и обработки сведений о состоянии здоровья и диагнозах граждан». Официальный интернет-портал правовой информации. 2025.
41. Lessage X, Collier L, Van Ouytsel C-HB, Legay A, Mahmoudi S and Massonet P. Secure federated learning applied to medical imaging with fully homomorphic encryption. 2024 IEEE 3rd International Conference on AI in Cybersecurity (ICAIC). Houston. TX. USA. 2024; 1–12. DOI: 10.1109/ICAIC60265.2024.10433836.
42. Walskaar, I, Tran, MC, & Catak, FO. A Practical Implementation of Medical Privacy-Preserving Federated Learning Using Multi-Key Homomorphic Encryption and Flower Framework. Cryptography. 2023; 7(4): 48. DOI: 10.3390/cryptography7040048.
43. Mohandas R, Veena S, Kirubasri G. Thusnavis Bella Mary I & Udayakumar, R. Federated Learning with Homomorphic Encryption for Ensuring Privacy in Medical Data. Indian Journal of Information Sources and Services. 2024; 14(2): 17–23. DOI: 10.51983/ijiss-2024.14.2.03.

44. Shumway DO, Hartman HJ. Medical malpractice liability in large language model artificial intelligence: legal review and policy recommendations. J Osteopath Med. 2024; 124(7): 287–290. DOI: 10.1515/jom-2023-0229.
45. Третьякова Е. П. Использование искусственного интеллекта в здравоохранении: распределение ответственности и рисков. Цифровое право. 2021; 2(4): 51–60. DOI: 10.38044/2686-9136-2021-2-4-51-60. EDN NGHUTA.
46. Мажуга Е. Ю., Петрова А. О., Гордеев Я. И. Правовое регулирование систем искусственного интеллекта в медицинской сфере. В сборнике: Актуальные проблемы современной России: психология, педагогика, экономика, управление и право. Сборник научных трудов международных научно-практических конференций; 07–24 апреля 2023 г.; Москва. Москва: Московский психолого-социальный университет. 2023; 1281–1285. EDN IAGUFA.

References

1. Meng X, Yan X, Zhang K, et al. The application of large language models in medicine: A scoping review. iScience. 2024; 27(5): 109713.
2. Omiye JA, Gui H, Rezaei SJ, Zou J, Daneshjou R. Large Language Models in Medicine: The Potentials and Pitfalls: A Narrative Review. Ann Intern Med. 2024; 177(2): 210–220. DOI: 10.7326/M23-2772.
3. Chen Y, Esmaeilzadeh P. Generative AI in Medical Practice: In-Depth Exploration of Privacy and Security Challenges. J Med Internet Res. 2024; 26: e53008. DOI: 10.2196/53008.
4. Yim D, Khuntia J, Parameswaran V, Meyers A. Preliminary Evidence of the Use of Generative AI in Health Care Clinical Services: Systematic Narrative Review. JMIR Med Inform. 2024; 12: e52073. DOI: 10.2196/52073.
5. Nógrádi B, Polgár TF, Meszlényi V, et al. ChatGPT M. D. Is there any room for generative AI in neurology? PLoS One. 2024; 19(10): e0310028. DOI: 10.1371/journal.pone.0310028.
6. Wang C, Li M, He J, Wang Z, Darzi E, Chen Z, et al. A Survey for Large Language Models in Biomedicine. ArXiv. 2024; abs/2409.00133.
7. Moglia A, Georgiou K, Cerveri P, et al. Large language models in healthcare: from a systematic review on medical examinations to a comparative analysis on fundamentals of robotic surgery online test. Artif Intell Rev. 2024; 57: 231. DOI: 10.1007/s10462-024-10849-5.
8. Ong JCL, Chang SY, William W, et al. Ethical and regulatory challenges of large language models in medicine. Lancet Digit Health. 2024; 6(6): e428-e432. DOI: 10.1016/S2589-7500(24)00061-X.
9. Zhui LL, Fenghe L, Xuehu W, Qining F, Wei R. Ethical Considerations and Fundamental Principles of Large Language Models in Medical Education: Viewpoint. J Med Internet Res. 2024; 26: e60083. DOI: 10.2196/60083
10. Wei Y, Zhou J, Wang Y, et al. A Review of Algorithm & Hardware Design for AI-Based Biomedical Applications. IEEE Trans Biomed Circuits Syst. 2020; 14(2): 145–163. DOI: 10.1109/TBCAS.2020.2974154
11. Pikalov Ya S. Obzor arkhitektur sistem intellektual'noy obrabotki yestestvenno-yazykovykh tekstov. Problemy iskusstvennogo intellekta. 2020; 4(19). Russian.
12. Jin L, Feng S, Xin Z, Chai Y. Evolution and advancements in deep learning models for Natural Language Processing. Applied and Computational Engineering. 2024.
13. Li K, Ao B, Wu X, Wen Q, Ul Haq E, Yin J. Parkinson's disease detection and classification using EEG based on deep CNN-LSTM model. Biotechnol Genet Eng Rev. 2024; 40(3): 2577–2596. DOI: 10.1080/02648725.2023.2200333
14. Haltaufderheide J, Ranisch R. The ethics of ChatGPT in medicine and healthcare: a systematic review on Large Language Models (LLMs). NPJ Digit Med. 2024; 7(1): 183. Published 2024 Jul 8. DOI: 10.1038/s41746-024-01157-x.
15. Garcez AA, et al. Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. arXiv preprint arXiv:1905.06088. 2019.
16. Li Z, et al. From system 1 to system 2: A survey of reasoning large language models. arXiv preprint arXiv: 2502.17419. 2025.
17. Cascella M, Semeraro F, Montomoli J, Bellini V, Piazza O, Bignami E. The Breakthrough of Large Language Models Release for Medical Applications: 1-Year Timeline and Perspectives. J Med Syst. 2024; 48(1): 22. DOI: 10.1007/s10916-024-02045-3.
18. Andreychenko AYe., Gusev AV. Perspektivy primeneniya bol'shikh yazykovykh modeley v zdravookhranenii. Natsional'noye zdravookhraneniye. 2023; 4 (4): 48–55. DOI: 10.47093/2713-069X.2023.4.4.48-55. Russian.
19. Moffatt B, Hall A. Is AI my co-author? The ethics of using artificial intelligence in scientific publishing. Account Res. 2024. DOI: 10.1080/08989621.2024.2386285.
20. Lee JY. Can an artificial intelligence chatbot be the author of a scholarly article? J Educ Eval Health Prof. 2023; 20: 6. DOI: 10.3352/jeehp.2023.20.6.
21. Johnson A. Generative AI, UK Copyright and Open Licences: considerations for UK HEI copyright advice services. F1000Res. 2024; 13: 134. DOI: 10.12688/f1000research.143131.1.
22. Shaydurov AS. Avtorskoye pravo na proizvedeniya, sozdannyye iskusstvennym intellektom v RF: problemy i perspektivy. V sbornike: Tsifrovyye tekhnologii v nauchnom razvitii: novyye kontseptual'nyye podkhody: Sbornik statey po itogam Mezhdunarodnoy nauchno-prakticheskoy konferentsii; 30 aprelya 2023 g.; Samara. Sterlitamak: Obshchestvo s ogranichennoy otvetstvennost'yu «Agentstvo mezhdunarodnykh issledovaniy». 2023; 88–92. EDN MDZVBE. Russian.
23. Kishkembayev M. Aktual'nyye problemy zashchity avtorskikh prav na ob"yekt, sozdannyy iskusstvennym intellektom. Vestnik Torajgyrov Universiteta. Ûridičeskaâ Seriâ. Mart 2024; 75–86. DOI: 10.48081/zusz1934. Russian.
24. Avila Negri SMC. Robot as Legal Person: Electronic Personhood in Robotics and Artificial Intelligence. Front Robot AI. 2021;8:789327. Published 2021 Dec 23. DOI: 10.3389/frobt.2021.789327.
25. Schmidgall S, Harris C, Essien I, et al. Evaluation and mitigation of cognitive biases in medical language models. NPJ Digit Med. 2024; 7(1): 295. DOI: 10.1038/s41746-024-01283-6.
26. Mondal M, et al. Do large language models exhibit cognitive dissonance? studying the difference between revealed beliefs and stated answers. Preprint arXiv. 2406.14986. 2024.
27. Goerlandt F, Li J, Reniers G. The Landscape of Risk Communication Research: A Scientometric Analysis. International Journal of Environmental Research and Public Health. 2020; 17(9): 3255. DOI: 10.3390/ijerph17093255.
28. Zarfati M, Soffer S, Nadkarni GN, Klang E. Retrieval-Augmented Generation: Advancing personalized care and research in oncology. Eur J Cancer. 2025; 220: 115341. DOI: 10.1016/j.ejca.2025.115341.
29. Shool S, Adimi S, Saboori Amleshi R, Bitaraf E, Golpira R, Tara M. 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.
30. Shevskaya NV. Ob"yasnimyy iskusstvennyy intellekt i metody interpretatsii rezul'tatov. Modelirovaniye, optimizatsiya i informatsionnyye tekhnologii. 2021; 9(2). DOI: 10.26102/2310-6018/2021.33.2.024. EDN VRKUIL. Russian.
31. Xiong G, et al. Benchmarking retrieval-augmented generation for medicine. Findings of the Association for Computational Linguistics ACL. 2024. 2024; 6233–6251.
32. Tang X et al. Medagentsbench: Benchmarking thinking models and agent frameworks for complex medical reasoning. arXiv preprint arXiv:2503.07459. 2025.
33. Yadav N, Pandey S, Gupta A, Dudani P, Gupta S, Rangarajan K. Data Privacy in Healthcare: In the Era of Artificial Intelligence. Indian Dermatol Online J. 2023;14(6): 788–792. DOI: 10.4103/idoj.idoj_543_23.

34. Altynnikov MS, Kuznetsova NO. Osobennosti organizatsii zashchity personal'nykh dannykh v meditsinskom uchrezhdenii. Innovatsii. Nauka. Obrazovaniye. 2021; 36: 1479–1486. EDN ZISIGQ. Russian.
35. Prikaz Minzdrava Rossii ot 20.03.2025 № 139n «Ob utverzhdenii Poryadka obezlichivaniya svedeniy o litsakh, kotorym okazyvayetsya meditsinskaya pomoshch', a takzhe o litsakh, v otnoshenii kotorykh provodyatsya meditsinskiye ekspertizy, meditsinskiye osmotry i meditsinskiye osvidetel'stvovaniya». Ofitsial'nyy internet-portal pravovoy informatsii. 2025. Russian.
36. Wiest IC, et al. Anonymizing medical documents with local, privacy preserving large language models: The LLM-Anonymizer. medRxiv. 2024. C. 2024.06. 11.24308355.
37. Morris JX, et al. Text embeddings reveal (almost) as much as text. arXiv preprint arXiv:2310.06816. 2023.
38. Mascalzoni D, Melotti R, Pattaro C, et al. Ten years of dynamic consent in the CHRIS study: informed consent as a dynamic process. Eur J Hum Genet. 2022;30: 1391–1397 DOI: 10.1038/s41431-022-01160-4. 2022.
39. Harishbhai Tilala M, Kumar Chenchala P, Choppadandi A, et al. Ethical Considerations in the Use of Artificial Intelligence and Machine Learning in Health Care: A Comprehensive Review. Cureus. 2024; 16(6): e62443. Published 2024 Jun 15. DOI: 10.7759/cureus.62443.
40. Postanovleniye Pravitel'stva RF ot 18.07.2023 № 1164 (red. ot 01.02.2025) «Ob ustanovlenii eksperimental'nogo pravovogo rezhima v sfere tsifrovykh innovatsiy i utverzhdenii Programmy eksperimental'nogo pravovogo rezhima v sfere tsifrovykh innovatsiy po napravleniyu meditsinskoy deyatel'nosti, v tom chisle s primeneniyem telemeditsinskikh tekhnologiy i tekhnologiy sbora i obrabotki svedeniy o sostoyanii zdorov'ya i diagnozakh grazhdan». Ofitsial'nyy internet-portal pravovoy informatsii. 2025. Russian.
41. Lessage X, Collier L, Van Ouytsel C-HB, Legay A, Mahmoudi S and Massonet P. Secure federated learning applied to medical imaging with fully homomorphic encryption. 2024 IEEE 3rd International Conference on AI in Cybersecurity (ICAIC). Houston. TX. USA. 2024; 1–12. DOI: 10.1109/ICAIC60265.2024.10433836.
42. Walskaar, I, Tran, MC, & Catak, FO. A Practical Implementation of Medical Privacy-Preserving Federated Learning Using Multi-Key Homomorphic Encryption and Flower Framework. Cryptography. 2023; 7(4): 48. DOI: 10.3390/cryptography7040048.
43. Mohandas R, Veena S, Kirubasri G. Thusnavis Bella Mary I & Udayakumar, R. Federated Learning with Homomorphic Encryption for Ensuring Privacy in Medical Data. Indian Journal of Information Sources and Services. 2024; 14(2): 17–23. DOI: 10.51983/ijiss-2024.14.2.03.
44. Shumway DO, Hartman HJ. Medical malpractice liability in large language model artificial intelligence: legal review and policy recommendations. J Osteopath Med. 2024; 124(7): 287–290. DOI: 10.1515/jom-2023-0229.
45. Tret'yakova Ye P. Ispol'zovaniye iskusstvennogo intellekta v zdravookhranenii: raspredeleniye otvetstvennosti i riskov. Tsifrovoye pravo. 2021; 2(4): 51–60. DOI: 10.38044/2686-9136-2021-2-4-51-60. EDN NGHUTA. Russian.
46. Mazhuga YeYu, Petrova AO, Gordeyev Ya I. Pravovoye regulirovaniye sistem iskusstvennogo intellekta v meditsinskoy sfere. V sbornike: Aktual'nyye problemy sovremennoy Rossii: psikhologiya, pedagogika, ekonomika, upravleniye i pravo. Sbornik nauchnykh trudov mezhdunarodnykh nauchno-prakticheskikh konferentsiy; 07–24 aprelya 2023 g.; Moskva. Moskva: Moskovskiy psikhologo-sotsial'nyy universitet. 2023; 1281–1285. EDN IAGUFA. Russian.