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

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

Over the past five years, artificial intelligence (AI) has become one of the fundamental technologies launching transformation of the basic paradigms of medicine and healthcare system [1, 2]. Recognizing potentially inflated expectations associated with this technology, it is necessary to clarify the terminology used below. In scientific and professional discourse, it is natural to distinguish between two main concepts of AI. The first one is artificial general intelligence (AGI), also known as strong artificial intelligence (AI), a hypothetical form of AI that can learn universally and solve problems like a human, which is only theoretical and has not been implemented in practice yet; the second one is artificial narrow intelligence (ANI), also referred to as weak AI, an existing software system that helps a person solve specific, clearly limited tasks, such as diagnosing diseases using medical images or automatization of routine operational processes.

The general term AI denotes ANI, which is used in practical medicine today.

Two major classes of weak AI are distinguished: descriptive and generative AI. Descriptive systems analyze and interpret data (including numerical, textual, graphical, audio, and video materials), providing classification, prediction, and identification of hidden patterns. On the contrary, generative AI can create (compile) new texts, images, or other data formats based on training samples, which opens up new opportunities to support clinical decision-making and automate workflow and communication processes in healthcare [3–5].

Natural Language Processing (NLP) holds a special place in the modern AI paradigm. It allows to analyze, interpret and generate textual information in a human language. Large Language Models (LLM), specialized AI architectures capable of operating with ultra-large arrays of textual data, have gained development and practical significance. This publication will be

devoted to review of some ethical aspects related to the use of large language models in medicine.

There has been a significant increase in research on the use of LLMs in medical field over the last few years [6,7]. Ethical aspects occupy a central position in discussion about safe and effective implementation of these technologies in clinical practice [8, 9]. Systematic research reveals both the potential advantages of LLMs in medical data analysis, information support, and decision support, significant ethical challenges related to algorithmic bias, lack of transparency, and risks of privacy violations. The ability of LLMs to generate highly persuasive but potentially inaccurate content, which requires human control and development of strict ethical guidelines, is of particular concern.

MATERIALS AND METHODS

While preparing this review publication, an integrated approach was applied to search, analysis and selection of relevant information, including using LLMs. Information search was carried out in domestic and international bibliographic databases: eLibrary, Scopus and PubMed, specialized platforms for searching scientific publications and analytical tools such as Consensus, Semantic Scholar and Elicit, which use LLMs in their algorithms. The search strategy that ensured complete and relevant coverage of the topic under study included key terms and their English-language equivalents such as large language models, medicine, healthcare, ethics, bioethics, risks, bias, reliability, and others. To include sources, a full-text version published in Russian or English from 2015–2025 was required. Relevance of the selected publications according to the abstract was assessed using the following parameters: relevance to the topic of large language models in medicine and healthcare, analysis of ethical aspects, description of implementation risks, novelty and scientific significance of the work. Articles that did not meet the stated criteria, as well as duplicate sources, were excluded. Google's NotebookLM and Perplexity LLM tools were used to systematize, extract data, and summarize selected publications. The resulting prepared materials were checked by a team of authors to ensure accuracy and correctness. The draft was prepared and grammatical proofreading was performed using OpenAI ChatGPT-4.1 and Google Gemini 2.5.

LLM ARCHITECTURE, DEVELOPMENT AND TRAINING

Improvement of computing power, available resources, and advanced algorithms has significantly promoted LLM development, facilitating their integration into various fields of human activity, including clinical practice [1, 4, 5]. LLMs can be used in three main areas such as clinical decision support, automation of medical documentation and reporting, as well as medical education and doctor-patient communication. LLMs have advantages of processing unstructured data [3, 5]. However, effectiveness varies depending on the specific model and approach to training.

To ensure a better understanding of the nature of LLM-related ethical issues, it is necessary to get an idea about the internal structure of the models that shape their functioning specifics. Modern LLMs represent the result of a long-term evolution of architectural approaches in natural language processing. Although transformers have now become a dominant architecture, historically they have gone through several key stages and architectures [10–13]:

1. Early NLP systems were based on manual coding of linguistic rules (for example, the ELIZA system, 1966). Statistical language models (SLM) have been used to predict words based on frequency patterns (for example, IBM Model, 1990). However, they haven't been widely used in practice.
2. Recurrent neural networks (RNNs) include a class of artificial neural nets designed to process sequential information. They can memorize the preceding elements of a sequence. Thus, they can effectively analyze time series, texts, and biomedical signals, operating however with a limited amount of context [11]. Advanced variants with long short-term memory (LSTM) analyze consistent clinical parameters (for example, heart rate, blood pressure, laboratory parameters) and identify patterns that predict complications. LSTMs are used to analyze ECG, EEG, pulse oximetry data, and other time signals [13].
3. Word2Vec implements principles of distributive semantics through vector representations of words (Skip-gram and CBOW algorithms (2013)). In the working process, the text is seen as a sequence of tokens (usually individual words or sub-word units), which are considered as minimal semantic units. For each token, Word2Vec creates an embedding: it maps the token into a multidimensional vector space where words that are similar in meaning have similar vector representations. These embeddings are used to analyze semantic and syntactic relationships in a text.
4. Convolutional neural networks (CNNs) are a class of deep neural networks that initially aimed at processing data with a spatial structure (images, 3D scans, spectrograms). Although CNNs are traditionally associated with image analysis, their architectural principles of extracting local features using convolutional layers served as a prototype for attention mechanisms in transformers, becoming a link between processing local patterns and global context [11, 13].
5. Transformers: a revolutionary architecture based on the mechanism of attention. By using multi-layer encoders/decoders, the model analyzes sequences of tokens, weighing the importance of each token in a sequence. The most well-known models of this class (for example, Generative Pre-trained Transformer, GPT), pre-trained on large-scale text data corpora, became widely available and gained a dominant position [14].
6. Retrieval-Augmented Generation (RAG): an approach aimed at overcoming the fundamental limitations of traditional LLMs, such as generation of factually incorrect information ("hallucinations"), obsolescence of model knowledge and lack of references to verified sources. RAG integrates LLMs with external knowledge bases such as PubMed, UpToDate, clinical recommendation databases, and other reputable resources.
7. BERT (Bidirectional Encoder Representations from Transformers) is an architecture based on bidirectional transformers, which provides a deep understanding of semantics and syntax of the text by taking into account the context to the left and right of the token. BERT and its derivatives are widely used to extract information from electronic medical records, automatically classify medical texts, and get access to biomedical databases and clinical decision support systems.
8. Hybrid models: to solve multimodal problems, systems are being developed that combine transformer attention mechanisms with convolutional or recurrent layers, which allows processing heterogeneous data from text-based

medical records to visualizations (CT, MRI) and time series (ECG, monitoring indicators) [13].

9. Neuro-symbolic systems integrate machine learning methods (neural networks) with symbolic methods of knowledge representation and reasoning (formal logic, expert rules, ontologies). Such systems do not only analyze unstructured data, but also use formal knowledge to improve interpretability, accuracy, and reliability of conclusions. They are used to solve tasks with high requirements for explainability of solutions, for example, when it is necessary to test hypotheses generated by LLMs for compliance with clinical recommendations [15].
10. Reasoning models are designed to solve problems that require complex logical, spatial, or ethical conclusions, optimized to simulate complex cognitive and logical processes typical of medical expertise. Unlike traditional LLMs, which focus primarily on generation of texts and identification of patterns, reasoning models build chains of logical conclusions, integrate diverse sources of knowledge, and explain their decisions at the level similar to the clinical thinking of a professional [16].

A similar path of evolution of technology from basic math algorithms through closed neural network models of "black boxes" gives rise to modern explainable models [16].

MODEL TRAINING

Evolution from rigid linguistic rules and statistical models to modern transformers and hybrid multimodal architectures has significantly expanded the range of LLM application in clinical practice. However, quality and reliability of LLMs directly depend on methods of their training as well as characteristics and quality of the starting training material. In clinical context, it is the initial data that determine the boundaries of the model applicability, level of reliability, interpretability of results, and safety of implementing LLMs in medical processes [17, 18].

Pre-training: the initial stage where the model learns patterns from massive unstructured bodies of general texts. The goal is to form universal language concepts and basic skills for text understanding and generating. Functioning of widely available general-purpose GPT models (YandexGPT, GigaChat, ChatGPT, Gemini, DeepSeek, Grok, Cloud, and others) that can generate different texts, including medical ones, which are however often of a general and superficial nature only is commonly determined at this stage of training. Such models can most likely make mistakes while processing queries concerning complex clinical cases. To avoid potential harm and legal claims, developers equip systems with modules that block responses to medical inquiries, and such an LLM must formulate a disclaimer when responding by recommending you to contact a qualified doctor.

Fine-tuning: additional model training based on specialized clinical data in order to adapt to specific tasks such as generating medical reports, supporting the diagnostic process, analyzing clinical dialogues, processing medical images, etc. Customizable datasets marked up by experts that reflected real clinical scenarios are the most effective. Models that went through such a customization (for example, BioGPT, BioMedLM, PubMedBERT, ClinicalBERT) are commonly used by medical professionals and are less known to the general public [17].

Reinforcement Learning based on Human Feedback (RLHF): a method in which a model corrects its behavior assessing quality and accuracy of the generated responses provided by experts. This minimizes the risk of generating dangerous or incorrect medical recommendations and reducing the likelihood of "hallucinations." Models trained with RLHF (for example, GatorTron, Med-PaLM, MetaMedLLM) are used mainly through integrations that provide access to the context in the form of personalized medical records, electronic health records, integrated and telemedicine solutions. RLHF is approved as the standard for medical LLM training. Research shows that LLMs that used RLHF were superior in quality and completeness of medical consultations compared to both models pre-configured without the RLHF and with pre-trained LLMs. RLHF is an obligatory stage for creation of modern medical language models, ensuring their compliance with requirements of clinical practice, safety and ethics [16].

Quality criteria of the starting training material:

- Relevance and reliability. It is critically important to use only up-to-date and verified data in medical LLMs. Use of outdated or unverified sources can lead to distribution of erroneous recommendations and create risks for the health of patients.
- Representativeness and diversity. To ensure fairness and universality of the model, the training material should cover a wide range of clinical scenarios, demographic groups, linguistic and cultural characteristics. Insufficient representation leads to systematic errors and bias, especially in relation to small or vulnerable groups of patients.
- Markup quality and expert validation. Errors in data annotation, incomplete or incorrect instructions lead to decreased accuracy and interpretability of the results. An effective approach is a combined markup method, in which experts form the core of the dataset, and AI algorithms complement it with variable examples, combining scalability and high-quality annotations.

While performing diagnostics, interpretation of medical images, and clinical communication, models trained on specialized, expertly labeled data demonstrate significantly higher accuracy and stability of results compared to those trained on general or synthetic data sets. [1, 2, 7, 18].

PROBLEMS AND CHALLENGES OF LLM IMPLEMENTATION IN MEDICINE

LLM implementation is accompanied by numerous ethical issues that require a systematic approach to their solution. A comprehensive analysis of LLM-associated ethical challenges has revealed both long-discussed issues such as potential copyright infringement, systematic bias, and data privacy, as well as new dilemmas, including verity of the information generated and its compliance with social norms [1, 8, 9, 18].

COPYRIGHT

As per the classical doctrine of copyright, an author, a person who has a creative idea and implements it in an objective form, can be a natural person only. Emergence of increasingly autonomous AI models capable of generating texts, scientific hypotheses, and diagnostic conclusions raises the question of copyright proprietor [19–23].

In most national legal systems, including the CIS countries, the EU and the USA, copyright does not recognize AI as an independent author (subject). It happens because a creative act needs the presence of will, consciousness and subjective choice, which modern AI does not possess. Article 1228 of the Civil Code of the Russian Federation clearly defines that an author of the work is the citizen (natural person) by whose

creative labor such work of literature, science or art has been created. AI does not have legal capacity and cannot carry out creative activities in the legal sense.

However, the growing volume of medical texts generated by LLMs requires a revision of established approaches. The medical field places special demands on quality, reliability and legal purity of information. Health and life of patients, as well as the professional reputation of medical professionals and researchers, are at stake here unlike artistic or journalistic activities [18]. Use of LLMs for automated creation of medical texts, protocols, data analyses, and even scientific articles generates a number of specific risks:

1. Sources are not obvious: training LLMs require vast amounts of text data, often without a clear distinction between open and copyrighted materials. This hinders identification of sources of borrowings and may lead to an unintended violation of the rights of third parties [20].
2. The problem of plagiarism: automatic text generation can lead to derivative works or texts that partially match the original sources, which poses the threat of accusations of plagiarism from copyright holders.
3. Difficulties with attribution: in case of joint human and AI creativity, it is necessary to determine the contribution of each participant and the order of distribution of copyrights.

There are three main approaches to determation of authorship when creating objects with AI participation [23]:

The author develops AI. It is assumed that all rights to the results created using AI belong to the person or organization that developed the corresponding model. The developer invests significant intellectual efforts and creative potential in the AI system, including development of algorithms, architecture and preparation of data for training. It requires significant financial, time and human resources from the developer [23]. Recognition of copyright by the developer can serve as an incentive for further investments and innovations in this area. This option provides a simpler and more predictable mechanism for determining the copyright holder compared to others. However, this approach is justified only if the user does not make a significant creative contribution, but only presses a button to generate a random piece without further creative intervention.

The author uses AI. In this case, the author is the person who directly manages AI and generates requests. The user chooses from the suggested options, corrects and directs AI to achieve the desired result. A detailed and creative query can lead to a unique piece, while a general or standard query is likely to produce a more typical result. AI acts as an improved tool that allows you to implement the user's creative intent by guiding the process. This model is most often used in medical and legal practice provided that the user (doctor, researcher) is engaged in active participation [22–24].

The author is AI (the concept of "electronic personality"). According to the resolution of the European Parliament with recommendations on civil law rules on robotics, the possibility of recognizing AI as an independent subject of copyright is being discussed [23,24]. Modern generative systems show an increasing degree of autonomy in the process of creating works. Contribution of AI can go beyond a simple instrumental use, and the system is able to generate unexpected and original results that were not directly established by the developer or controlled by a human. However, in practice, this approach has not been recognized, since AI has neither legal personality nor ability to exercise rights and obligations independently. International practice shows that in the vast majority of cases, courts and intellectual property offices refuse to recognize authorship of AI [22].

Thus, we believe that contribution of the participants to creation of any work (literary text, scientific text, and medical records generated by LLMs) is multilevel. When contribution of a user and AI (as a result of developer's work and data) is inseparable, it is necessary to apply the concept of joint authorship, providing compensation to copyright holders depending on their contribution to making content. Depending on the chosen tariff, AI users acquire AI as a service, strengthening their copyright positions.

At the same time, a number of countries are discussing options for introducing special protection regimes for works created with minimal human involvement, for example, a shortened copyright term [21], while providing remuneration to those authors whose works were used to teach AI.

Apart from the legal aspects, the use of LLMs in medicine raises a number of scientific dilemmas/ Reducing the role of human creativity is one of them. Exponential growth in the amount of content generated by AI can devaluate human input and decrease motivation for independent scientific research. Automatic generation of medical texts without proper expert validation can result in distribution of unreliable or even dangerous information.

The modern legal system is not yet ready to fully take into account specifics of AI-generated objects, which requires new approaches to determining authorship, protectability and distribution of rights to the results of intellectual activity.

Taking into account the problems outlined, the following directions of development are proposed:

- Introduction of special protection regimes for works created using AI, for example, a shortened term of rights.
- Mandatory disclosure of AI involvement degree in publication of medical articles, development of clinical protocols and other scientific materials.
- Development of international standards on attribution and identification of sources when using LLMs.
- Creation of more advanced systems for tracking borrowings and checking for plagiarism based on tokenized information.
- Accrual of remuneration to developers and authors of materials on the basis of which models are trained, including through paid subscription systems.

BIAS, HALLUCINATIONS, AND EXPLICABLE AI

Despite significant progress in reducing the frequency of factual errors ("hallucinations") in modern LLMs, especially in highly specialized systems configured using RLHF (with relevance of responses above 95%), a new serious challenge is systematic bias, which leads to errors in medical recommendations, discrimination against vulnerable groups of patients, and distortion of medical knowledge, causing a decreased confidence in AI in healthcare [24,25].

Systematic bias is a persistent distortion of the results of a model due to specific data, architecture, or learning processes, leading to a distorted or inaccurate representation of certain groups, phenomena, or concepts, as well as distorted interpretation of clinical data. These failures are not accidental, they constitute a consequence of the internal algorithm logic. Algorithmic systems cannot only reproduce but also amplify existing biases, creating a potentially dangerous cycle of increased discrimination [26].

LLMs are trained on text corpora that may contain historical, social, and cultural biases, as well as an unbalanced medical knowledge. Errors or subjectivity in marking up medical data can consolidate bias at the stage of preparing datasets.

The features of transformers, attention mechanisms, and ways of processing context can both enhance and weaken bias. As it has already been mentioned, GPT is an autogregressive transformer model trained to predict the next token based on statistical patterns in the training data. It tends to reproduce the most common patterns, reinforcing existing biases and medical stereotypes, which may manifest itself in disproportionate attention to certain aspects of information correlating with demographic characteristics, or in incorrect interpretation of rare or ambiguous cases [25]. GPT has no built-in fact-checking or compliance mechanisms for clinical standards. Increasing the size of the model does not always guarantee less biases; some of its forms may even get intensified [14].

Although reasoning models include logical inference mechanisms (for example, Chain-of-Thought, CoT), they can still reproduce biased reasoning patterns if they were present in the training data, moreover, it is more difficult to detect bias in reasoning chains, because the confirmation bias effect is possible. A critical problem is that the explanations (rationalizations) generated can mask the true (possibly biased) reasons for the model's prediction, especially when the answers are incorrect. The risk reduction approach is to use an expression of uncertainty, where the model indicates the degree of confidence in its response, allowing clinicians to take this into account during interpretation. When models explicitly express their uncertainty, their forecasts become less categorical and less prone to systematic errors [25]. Uncertainty representations can be used as an additional filter to identify cases in which the model is potentially biased or uncertain as it allows either to postpone a decision or involve an expert;

Integration with external knowledge bases in RAG models potentially reduces bias through access to relevant and evidence-based facts. However, RAG models may incorrectly aggregate controversial information from sources or reproduce bias if it is contained in external databases. It is difficult to ensure reproducibility of solutions, because the model may refer to different sources even with identical queries, which makes it difficult to audit and correct bias.

In general, all LLMs are algorithmically inclined to generate the most likely (frequent) responses, ignoring rare but clinically significant cases. When a LLM is used without expert validation, it can lead to perpetuating and spreading bias [14].

Research shows that large language models exhibit significant differences between their "revealed beliefs" and "stated answers," indicating the presence of multiple biases and distortions in the representations they form [26].

Another problem is the dissonance between the probabilistic nature of algorithmic conclusions and their subjective perception by patients (and in some cases by doctors) as deterministic predictions [27].

Research in risk communication confirms that effectiveness of transmitting medical information significantly depends on the way the data is presented to the patient [27]. Categorical formulations of prognostic conclusions induce pronounced psychological reactions even in a low statistical probability of the predicted outcome. Optimistic formulations create the illusion of controllability, forcing patients to underestimate the objective risks and even discontinue therapy prematurely.

Automation bias is the tendency to perceive algorithmic inferences as more objective than human judgments. Digital interfaces make us trust sources subconsciously. Excessive trust in algorithmic advisors is a complex phenomenon of emergence of new forms of dependence. Many users tend to attribute the properties of "superhuman intelligence" to AI systems, ignoring limitations of the training data and architectural features of the models. Experimental data show that 68% of respondents are ready to follow the advice of AI, even though their attending physician has a different opinion [27]. Clinical manifestations of algorithmic dependence include compulsive verification of predictions through mobile applications, anxiety-phobic reactions when the service is temporarily unavailable, and refusal to analyze symptoms independently in favor of automated diagnoses.

Development of methodologies for quantifying bias and degree of reliability of responses in medical LLMs is an important area of further research [28, 29].

Despite the unprecedented potential of LLMs in medicine, their widespread adoption is inhibited by the lack of transparency of decision-making mechanisms for most users, which reduces the trust of medical professionals and patients. Many large language models, such as GPT-4, are complex neural network architectures with billions of parameters, with its internal functioning often being incomprehensible to many users (a "black box") [14].

Explicable Artificial Intelligence (XAI) is a research area focused on development of methodologies and technologies that make the decision-making process of AI systems understandable to humans, enable verification of results and help to overcome distrust in AI technologies [30].

Creating models with initially high degree of interpretability are basic solutions (for example, linear models and decision trees that allow you to explicitly trace the relationship between the input data (the contribution of each feature) and output results). However, these models may have inferior predictive accuracy for some tasks as compared to more complex architectures [16].

Generating intermediate stages of reasoning before giving a final Chain-of-Thought (CoT) response increases not only accuracy, but also explainability, allowing to trace the logical chain of the model. Explanations can be adapted for different groups (doctors, patients, regulators).

As mentioned earlier, it becomes mandatory to apply the RAG methodology, provide models with access to relevant scientific literature, clinical recommendations and other verifiable sources, and increase the accuracy, reliability and transparency of the information generated. The Medical Information Retrieval-Augmented Generation Evaluation (MIRAGE), the first benchmark that includes 7,663 questions from five medical datasets for question-and-answer systems, can serve as an example of an assessment. Studies with MIRAGE have demonstrated that the use of MedRAG, compared with the chain-of-reasoning hint method, improves accuracy of responses from various LLMs by up to 18% [31].

As of May 2025, the MedAgentsBench benchmark includes 1,453 structured clinical cases covering 13 organ systems and 10 medical specialties. According to the comparison results, DeepSeek R1 and OpenAI-o3 reasoning models are the leaders in March 2025. They provide not only high accuracy, but also an optimal ratio between performance, cost of calculations and output time, which is especially important for practical implementation in medical information systems with accuracy in simple diagnostic tasks of 89% for OpenAI-o3 and 93% for DeepSeek R1. However, in complex scenarios requiring multi-stage treatment planning, the indicator decreased to 67% for OpenAI-o3 and 73% for DeepSeek R1 [32].

The problem of lack of standardized metrics and protocols for evaluating the quality of explanations is urgent. Existing XAI methods generate explanations of various formats and content. Currently, there is no consensus on what properties a "good" explanation should have and how these properties can be objectively measured [18,32].

CONFIDENTIALITY AND PROTECTION OF PERSONAL DATA

Use of real clinical data for LLM training and application requires strict adherence to patient anonymization and confidentiality standards, which imposes additional requirements on preparation of training samples [33, 34].

Effectiveness of digital medical technologies directly depends on trust of patients. Violation of confidentiality undermines trust in healthcare system as a whole and can lead to refusal of patients to provide complete and reliable information, which will negatively affect the quality of medical care. Personalized LLMs improve treatment quality by paying attention to individual characteristics, but require processing of ultra-sensitive data (regarding genome, lifestyle, and mental status of the patient) [14].

Medical data can be characterized by a high degree of sensitivity: they contain information about diagnoses, test results, genetic characteristics, medical history, and other information that can identify the patient. They are also subject to strict legal and ethical protection. LLMs are trained on a vast amount of text, including not only open sources, but also specialized medical databases. Even formal depersonalization can be followed by a risk of restoring the patient's identity based on indirect signs, which is especially important for rare diseases or unique combinations of clinical signs.

Order No. 139n of the Ministry of Health of the Russian Federation dated March 20, 2025 "On Approval of the Procedure for Depersonalizing Information about persons who receive medical care, as well as about persons for whom medical expertises, medical examinations and medical certifications are conducted", that has been put in force since September 1, 2025 and that replaced Order No. 341n dated June 14, 2018, prescribes depersonalization of all information that allows direct or indirect identification of the patient's identity, including full name, date of birth, address, contact information, individual document numbers and other identifiers. The procedure should ensure that it is impossible to restore the patient's identity without using additional information stored separately and protected in accordance with the legislation of the Russian Federation [35].

However, even when direct identifiers (name, date of birth, address) are deleted, quasi-identifiers (for example, a rare combination of symptoms, unique treatment regimens) are still present in the medical data and can be used to re-identify the patient. The LLM-Anonymizer study demonstrated retention of about 2% of identifying information after processing [36]. Research shows that intruders can restore source texts from vector representations of models with an accuracy of up to 92% using inversion attack methods [37].

According to ethical standards, minimum required amount of data should be used to achieve the goal. However, LLMs, that use huge datasets for training, often process redundant information, which makes it difficult to control information processing and increases the scope of potential leakage.

In most cases, patients consent to processing of their data for specific purposes of diagnosis, treatment, and scientific research. Classical requirements of completeness of information, voluntary nature, and patient competence conflict with the technical complexity of AI. Use of LLMs capable of generating new knowledge and reusing information in unforeseen scenarios goes beyond the standard forms of consent. Patients are often unaware that their data can be used to train complex models that are subsequently used in a wide range of tasks. Most patients do not have specialized knowledge that allows them to evaluate the architecture of neural networks, quality of training data, or limitations of algorithms [18,34].

LLMs are continuously updated. It makes the traditional static provision of information irrelevant already at the stage of signing the consent. Dynamic informed consent is a modern model of interaction between a patient and a medical organization, which involves not a one-time, but continuous, step-by-step informing of the patient and obtaining the patient's consent at each stage of interaction. The patient obtains information not only at the initial stage of treatment, but also with every significant change in AI algorithm, software update, or occurrence of new clinical data that affect decision-making. It is necessary to use interactive digital platforms that allow the patient to receive notifications, clarifications and consent to new stages of interaction in real time [38,39].

In Russia, there is an experimental legal regime for development and implementation of artificial intelligence (AI) in healthcare, automatically implying consent of patients to transfer anonymized medical data for artificial intelligence training [40], after which the medical community needs to determine the forms and methods of working with dynamic consent.

Existing laws (for example, HIPAA in the USA, GDPR in the EU, FZ-152 in the Russian Federation) establish requirements for personal data protection, but do not take into account the specifics of LLM work. The "right to be forgotten" requirement faces the technical difficulty of selective deletion of data in pre-trained models. There are questions about distribution of responsibility for data leakage (developer, medical institution, user) and compliance with the rules of cross-border data transfer.

Comprehensive regulatory measures are needed: staff training on cybersecurity and ethics of working with medical data, introduction of a multi-level system controlling access to source data and model results, regular testing of models for reproducing sensitive information, introduction of algorithms for detecting and filtering personal data at the stage of generating model responses, use of differential privacy methods that allow training LLMs on aggregated data without the risk of restoring individual records. Legislation needs to be updated considering specifics of LLM work, introduction of special requirements for anonymization and audit of models, and industry standards for certification of depersonalization algorithms. Ensuring transparency of data processing processes and informing patients about possible risks is important too.

Technological solutions such as adding Gaussian noise to embeddings reduce the risk of inversion by 60%, but also worsen performance of the models. Federated Learning (FL) and Homomorphic Encryption (HE) form a technological symbiosis that allows processing sensitive medical data without direct exposure [41].

Federated learning implements a decentralized approach where models are trained on local datasets without their transfer to the central server. It can minimize the risks of leaks in cross-border research and combine knowledge from diverse sources (laboratories, hospitals, wearable devices). Experiments with the Flower FL framework demonstrate high accuracy while significantly reducing privacy risks [42].

Homomorphic encryption schemes make it possible to calculate encrypted data without the need to decrypt it first. Homomorphic encryption means that if the source data has been encrypted, then certain mathematical operations can be done with this ciphertext (for example, addition, multiplication), and the result of these operations will also be encrypted. After decryption of the result, the doctor receives the same result that would have been obtained by performing similar operations with original unencrypted data. However, to optimize these calculations, specialized expensive computing equipment is required [43].

The MedSecureAI prototype demonstrates that the FL+HE combination reduces the risk of leaks by 99.2% while increasing the training time by only 2.1 times compared to the basic models [41]. This creates additional technological challenges: creation of specialized processors for medical HE, development of interstate standards for exchange of encrypted models, and integration of post-quantum cryptographic algorithms.

LEGAL LIABILITY OF LLM RESULTS

From a legal point of view, LLMs currently do not have the status of independent legal entities. They are considered exclusively as tools created and used by individuals or legal entities. The legal responsibility for consequences of LLM application lies with developers, software vendors, as well as medical professionals and organizations using these technologies [44].

Developers and suppliers have to ensure that their products comply with established quality and safety standards, and have to inform users about possible limitations and risks.

All medical devices, including large language model software, are subject to mandatory state registration before they are introduced into clinical practice. Depending on the potential harm caused by an error, AI solutions belong to the following risk classes: IIa (medium risk — systems for pre-processing medical documentation, primary screening), IIb (increased risk — systems for automated interpretation of instrumental research results, algorithms for predicting the course of diseases, software for supporting clinical decision-making) or III (high risk — AI systems that make independent clinical decisions, form diagnostic and therapeutic recommendations, and are applied autonomously in implantable medical devices), since their errors can lead to significant consequences for the patient's life and health. Registration requires conducting a clinical assessment, confirming the quality of algorithms, ensuring transparency and reproducibility of results, as well as implementing risk management mechanisms and continuous monitoring of functioning[45]. Roszdravnadzor monitors and may suspend the use of compromised solutions to take corrective action (as it was done in 2023–2024 with Botkin.AI).

Developers and operating organizations should pay special attention to information security issues. Information systems that process personal data of patients are becoming a priority target for intruders. Modern cyber security threats, including unauthorized access, attacking integrity and confidentiality of data, as well as manipulation with model conclusions, can lead to serious consequences that pose threats not only to health, but also to lives of patients [45, 46]. These medical information systems are subject to Federal Law No. 187-FZ dated July 26, 2017 "On Security of Critical Information Infrastructure of the Russian Federation".

Medical professionals, in turn, are professionally responsible for making clinical decisions, even if they rely on recommendations formulated by the LLM. The doctor must critically evaluate the information received and cannot completely delegate decision-making to artificial intelligence [18].

In case of negative consequences related to errors or unreliable LLM recommendations, responsibility can be distributed among various participants of the process, depending on the nature and source of the error. If we are talking about a software defect, responsibility is usually allocated to the developer. The medical professional or organization is claimed responsible if an error occurred due to incorrect use of technology or because a doctor ignored professional standards and clinical recommendations.

CONCLUSION

Thus, introduction of large language models in healthcare requires an integrated approach combining further technological improvement of models, development and implementation of strict ethical standards, adaptation of the regulatory framework, use of advanced information security techniques and constant critical supervision by the expert medical community.

Improvement of algorithms and architectures is one of the key areas. It is necessary to select modern models that combine the possibilities of reasoning, search and explanation. Transition from predictive "black box" models to interpreted systems that can substantiate their conclusions will increase trust of medical professionals and patients in these technologies. Development of neuro-symbolic methods that integrate machine learning with symbolic representations of knowledge and logical reasoning is an important step. It helps not only analyze unstructured data, but also use formal knowledge to improve interpretability, accuracy, and reliability of conclusions.

Quality and relevance of the training data are equally important. LLMs should not only be pre-trained on massive bodies of texts, but also pre-tuned using highly specialized pre-marked clinical data with participation of medical experts. Expert Feedback Reinforcement Learning (RLHF) should become the standard for medical language models, confirming their compliance with requirements of clinical practice, safety and ethics. This will ensure not only the relevance of general answers, but also their personification and clinical evidence.

Adjustment of regulatory framework to technological advances is a prerequisite for successful implementation of LLMs in healthcare. Legal experts need to consider specifics of increasing AI integration into all fields of activity and develop new approaches to authorship identification, protection and distribution of rights to intellectual property results created with LLM participation.

Ensuring confidentiality and protection of personal data is a prerequisite. It is important to strictly adhere to the standards of patient anonymization and confidentiality when using real clinical data for LLM training and application. The minimum required amount of data should be used to achieve the goal set and implement technological solutions such as federated learning and homomorphic encryption that allow to process sensitive medical data without direct exposure. It is also important to develop interactive digital platforms that provide the patient with real-time notifications, clarifications and consent to new stages of interaction (dynamic consent form).

Exclusion of unreliable answers, step-by-step fact-checking and cross-checking are necessary to combat "hallucinations" and bias. It is necessary to develop methodologies to quantify the degree of reliability of responses in medical LLMs, allowing clinicians to take this into account when interpreting the results. It is important to pay attention to cultural and linguistic characteristics of different groups of patients and develop models that take these differences into account.

To develop an objective and trusting attitude towards the applied AI technologies, it is necessary to ensure transparency and explainability of LLM functioning. To do this, it is necessary to develop standardized metrics and protocols assessing quality and use of XAI methods to trace the logical chain of the model and adapt explanations for different audiences. It is also important to take into account psychological aspects of LLM-provided information perception and avoid categorical formulations that can induce pronounced psychological reactions.

Only when these conditions are met, healthcare level can be significantly increased owing to the use of large language models, while protecting the rights and interests of patients and medical professionals.

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. 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.

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. С. 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.