**Review Article**

Healthc Inform Res. 2025 April;31(2):114-124.
https://doi.org/10.4258/hir.2025.31.2.114
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

# Large Language Models in Medicine: Clinical Applications, Technical Challenges, and Ethical Considerations

**Kyu-Hwan Jung**[1,2]

[1]Department of Medical Device Management and Research, Samsung Advanced Institute for Health Sciences and Technology, Sungkyunkwan University, Seoul, Korea
[2]Smart Healthcare Research Institute, Research Institute for Future Medicine, Samsung Medical Center, Seoul, Korea

**Objectives:** This study presents a comprehensive review of the clinical applications, technical challenges, and ethical considerations associated with using large language models (LLMs) in medicine. **Methods:** A literature survey of peer-reviewed articles, technical reports, and expert commentary from relevant medical and artificial intelligence journals was conducted. Key clinical application areas, technical limitations (e.g., accuracy, validation, transparency), and ethical issues (e.g., bias, safety, accountability, privacy) were identified and analyzed. **Results:** LLMs have potential in clinical documentation assistance, decision support, patient communication, and workflow optimization. The level of supporting evidence varies; documentation support applications are relatively mature, whereas autonomous diagnostics continue to face notable limitations regarding accuracy and validation. Key technical challenges include model hallucination, lack of robust clinical validation, integration issues, and limited transparency. Ethical concerns involve algorithmic bias risking health inequities, threats to patient safety from inaccuracies, unclear accountability, data privacy, and impacts on clinician-patient interactions. **Conclusions:** LLMs possess transformative potential for clinical medicine, particularly by augmenting clinician capabilities. However, substantial technical and ethical hurdles necessitate rigorous research, validation, clearly defined guidelines, and human oversight. Existing evidence supports an assistive rather than autonomous role, mandating careful, evidence-based integration that prioritizes patient safety and equity.

**Keywords:** Natural Language Processing, Artificial Intelligence, Clinical Decision Support Systems, Medical Informatics Applications, Medical Ethics

**Submitted:** March 31, 2025
**Accepted:** April 23, 2025

**Corresponding Author**
Kyu-Hwan Jung
Department of Medical Device Management and Research, Samsung Advanced Institute for Health Sciences and Technology, Sungkyunkwan University, 115, Irwon-ro, Gangnam-gu, Seoul 06355, Korea.
Tel: +82-2-3410-3632, E-mail: khwanjung@skku.edu (https://orcid.org/0000-0002-6626-6800)



## I. Introduction

The landscape of artificial intelligence (AI) in medicine is undergoing a profound transformation driven by large language models (LLMs). These advanced AI systems, grounded in deep learning methodologies and commonly based on transformer neural network architectures [1], exhibit remarkable proficiency in processing, understanding, and generating human language. Notable examples at the forefront of this technology include OpenAI's Generative Pre-trained Transformer (GPT) series, Google's Gemini, and Meta's LLaMA [2]. The capabilities of these models derive

from extensive pre-training on massive, diverse textual datasets comprising trillions of words sourced from the internet, books, and other digitized materials. This comprehensive pre-training equips them with broad knowledge of grammar, semantics, general world information, and complex linguistic patterns [2,3].

Adapting these general-purpose models to specialized fields such as medicine requires a process known as fine-tuning [4]. Fine-tuning involves additional training of a pre-trained model using a smaller, domain-specific dataset, such as clinical notes, biomedical literature, or patient communication transcripts. This procedure refines the model's general linguistic skills, aligning them with the specialized vocabulary, subtleties, and reasoning processes unique to medicine. Examples of such tailored models include BioBERT [5], ClinicalBERT [6], GatorTron [7], Med-PaLM [8], and Med-Gemini [9]. These models, fine-tuned or specifically designed using clinical or biomedical texts, typically demonstrate superior performance in medical tasks compared to their general-purpose counterparts [3,10,11].

When introduced into clinical environments, LLMs enter a setting already influenced by earlier AI applications that primarily analyze structured data, such as laboratory results and billing codes, or medical images used in radiology and pathology. These earlier AI applications focused on risk prediction, diagnostic support, and operational optimization [12,13]. However, LLMs represent a paradigm shift due to their unique capacity to analyze unstructured text data, a significant advancement given the abundance of clinically relevant information found in unstructured text formats. Previously, these data sources have largely been inaccessible for computational analysis due to the inherent complexities of natural language. Therefore, LLMs signify a new technological frontier capable of unlocking valuable insights and making practical use of previously underutilized textual information [2,10].

In this article, we survey the current landscape of LLM applications within clinical practice, categorizing these applications by their primary functional roles. We discuss technical issues pertinent to building and effectively implementing these models within clinical settings. Substantial emphasis is also placed on ethical considerations, which are particularly critical in healthcare AI applications. Finally, the review concludes by examining future directions and the potential trajectory of LLM integration into medicine.

## II. Clinical Applications of Large Language Models

LLMs are currently being explored across a spectrum of clinical activities to enhance efficiency, support decision-making, improve communication, and optimize workflows. The evidence supporting these applications varies, with some areas demonstrating greater maturity than others (Table 1).

### 1. Clinical Documentation Assistance

A significant portion of clinician time is dedicated to documentation tasks, contributing notably to clinician burnout. LLMs offer potential solutions by automating or assisting with various documentation-related activities.

#### 1) Summarization

LLMs have demonstrated effectiveness in summarizing lengthy clinical texts. Specific applications include condensing progress notes into problem lists, abstracting findings from radiology reports into concise impressions, generating after-visit summaries from detailed history and physical examination notes, summarizing physician-patient dialogues, and shortening extensive patient-generated health questionnaires [14,15]. Research using models such as GPT-4 indicates that LLM-generated summaries can match or even surpass human-produced summaries in completeness and accuracy for particular tasks, while accomplishing this significantly faster [15].

#### 2) Note generation/drafting

LLMs can assist in drafting various clinical documents, including discharge summaries [14], referral letters, and inter-unit handoff notes [16]. Additionally, LLMs can identify goals of care conversations within electronic health record (EHR) notes and rapidly generate clinically useful summaries. Importantly, the reported incidence of hallucination (incorrectly generated content) for this specific task is notably low [17].

#### 3) Information extraction

These models can extract structured data elements (e.g., diagnoses, medications, symptoms, patient-reported outcomes) from unstructured narrative text within EHRs or other clinical documents [18]. This capability facilitates data aggregation for quality improvement initiatives, research purposes, or populating structured fields in EHRs.

Table 1. Clinical applications of LLMs in medicine

| Application category | Specific examples | Potential benefits | Example studies |
|---|---|---|---|
| Clinical documentation summarization | Summarizing radiology reports, patient questions, progress notes, doctor-patient dialogues; AVS generation. | Reduced documentation time; Improved completeness, correctness, conciseness vs. human summaries for specific tasks. | Clinical reader study comparing GPT-4 vs. human summaries [15]. |
| Clinical documentation generation | Drafting discharge summaries, referral letters; LLM-generated EM-to-IP handoff notes. | Reduced documentation burden; Standardization of notes. | Cohort study comparing LLM vs. physician EM-to-IP handoff notes [16]; Studies on discharge/referral letters [33]. |
| Clinical info extraction | Extracting diagnoses, medications, PROs from unstructured text; Populating structured EHR fields. | Improved data accessibility; Facilitation of quality improvement and research; Reduced manual data entry. | Privacy-preserving medical information retrieval [18]. |
| Diagnostic & clinical decision support | Differential diagnosis suggestions; Answering clinical questions; Potential data interpretation. | Access to synthesized knowledge; Potential to augment clinical reasoning; Speed up information retrieval. | Med-PaLM 2 high accuracy on MedQA [9]; Med-PaLM M outperforms most state-of-the-art on MultiMedBench [23]. |
| Patient communication & engagement | Simplifying medical jargon/documents (e.g., trial info, discharge summaries); Patient Q&A chatbots; Health literacy tools. | Improved patient understanding; Enhanced engagement; Increased accessibility of information; Potential for empathetic responses. | Patient-friendly discharge summaries [33]; Studies on chatbot empathy/utility [22]. |

LLM: large language model, AVS: audio video coding standard, GPT: Generative Pre-trained Transformer, EM: emergency medicine, IP: inpatient, EHR: electronic health record, PRO: patient reported outcome.

#### 4) Documentation improvement

LLMs may also analyze clinical documentation to identify missing information, inconsistencies (e.g., discrepancies between diagnosis and treatment plan), or areas requiring clarification, thereby improving the overall quality and accuracy of medical records [19].

## 2. Diagnostic and Clinical Decision Support

LLMs can analyze extensive medical information and perform inferential functions to support clinical decision-making and reasoning processes.

#### 1) Generation of differential diagnoses

LLMs can evaluate patient symptoms, medical history, and preliminary test results to suggest potential differential diagnoses to clinicians. Studies have shown that LLMs can outperform clinicians in generating differential diagnoses and in the quality of diagnostic and management reasoning, especially when employing a chain-of-thought (CoT) process that mirrors human clinical reasoning [20-22]. Thus, LLMs hold potential to enhance diagnostic accuracy and efficiency, particularly in complex clinical scenarios.

#### 2) Answering clinical questions

LLMs function effectively as advanced information retrieval systems capable of answering specific clinical queries by integrating information from medical knowledge bases, scientific literature, or clinical guidelines [8,23,24]. Systems like Med-PaLM 2 have demonstrated strong performance on clinical knowledge assessments analogous to medical licensing examinations (e.g., the MedQA benchmark), achieving accuracy levels comparable to or exceeding those of human experts in multiple studies [9].

#### 3) Interpretation of complex clinical data

LLMs show emerging capabilities for interpreting complex clinical data, potentially integrating information across multiple modalities (e.g., textual reports and imaging findings). Current research involving multimodal models such as Med-PaLM M and Med-Gemini explores these promising opportunities [23,25]. Such interpretation capabilities could also assist in treatment planning decisions [17,26].


https://doi.org/10.4258/hir.2025.31.2.114

Nevertheless, translating these capabilities into reliable clinical decision support (CDS) remains challenging. Although strong performance on standardized assessments is reassuring, studies simulating real clinical scenarios have identified significant deficiencies [27]. A notable gap persists between performance on knowledge-based tests and the nuanced, context-dependent judgment essential for clinical practice. Current LLMs excel at information retrieval yet lack the sophisticated clinical judgment and integrative capacities of human clinicians [28]. Moreover, seamless integration with clinical workflows and EHR systems, which would be essential for successful CDS, remains difficult to achieve [29].

### 3. Patient Communication and Engagement

LLMs' natural language processing abilities offer potential improvements in clinician-patient communication and patient access to, understanding of, and dissemination of health information.

#### 1) Simplifying of medical information

LLMs can translate complex medical terminology (e.g., informed consent documents) [30,31], research findings (e.g., clinical trial summaries) [32], or clinical reports (e.g., discharge summaries) [6,33] into straightforward language that patients can easily comprehend. This application has significant potential to enhance patient understanding and promote effective shared decision-making.

#### 2) Patient question answering/chatbots

Virtual assistants or chatbots powered by LLMs can respond to patient inquiries regarding health conditions or treatments, offering preliminary information and guidance [34]. Remarkably, LLMs can deliver responses with a significant degree of empathy, even being perceived by patients as more empathetic than human clinicians in certain written communications [22,35].

#### 3) Increasing health literacy

By generating clear, understandable, and personalized health information, LLMs have the potential to enhance patients' overall health literacy [36].

Despite this promise, caution is warranted. LLMs can produce inaccurate or misleading information (misinformation), lack nuanced understanding of individual patient circumstances, and fail to convey critical emotional and psychological dimensions inherent in patient care. The impersonal nature of chatbot-mediated communication cannot replicate the therapeutic benefits of direct human clinician-patient interactions. Thus, employing LLMs for patient communication necessitates careful oversight and rigorous validation of information for accuracy and appropriateness.

## III. Technical Challenges and Ethical Considerations

Despite enthusiasm for using LLMs in medicine, their translation into safe and effective clinical applications is impeded by significant technical hurdles and profound ethical concerns (Table 2). These challenges often intersect and require integrated solutions.

### 1. Technical Challenges

#### 1) Accuracy and reliability

Ensuring the accuracy and reliability of LLM outputs in clinical contexts—where errors carry severe consequences—is arguably the most critical technical barrier. LLMs are known to "hallucinate," generating fluent, plausible-sounding content that is factually incorrect, unsupported by source data, or entirely fabricated [37-40]. If relied upon for clinical decisions, these inaccuracies directly threaten patient safety. Mitigation strategies currently explored include retrieval augmented generation (RAG) to ground responses in external knowledge sources, improving uncertainty estimation and calibration techniques, employing specific prompting strategies such as chain-of-thought, domain-specific fine-tuning, implementing domain-informed safety guardrails, and developing robust hallucination detection methods [41].

Beyond complete fabrication, LLMs can produce outputs that contradict source data or established medical facts or omit critical details from summaries and analyses [16]. These errors also present substantial patient safety risks.

#### 2) Validation and evaluation

Assessing the clinical utility and safety of LLMs is exceptionally challenging. Current validation methodologies often fall short.

Standard natural language processing (NLP) evaluation metrics—such as ROUGE and BLEU for summarization tasks, or accuracy metrics for question-answering—frequently correlate poorly with clinical relevance, factual correctness, or patient safety [42]. High benchmark scores do not necessarily equate to competence in real-world clinical reasoning. Therefore, there is an urgent need for standardized, clinically meaningful benchmarks and evaluation frameworks specifically designed for medical LLMs.

An additional challenge is the rapid evolution of models

Table 2. Technical challenges of LLMs in clinical medicine

| Challenge | Description | Risks for clinical practice | Potential mitigation strategies |
|---|---|---|---|
| Hallucination/ accuracy | LLMs generating plausible but factually incorrect, fabricated, or unsupported information. Includes factual errors & omissions. | Erroneous diagnoses or treatment plans; Patient harm; Erosion of clinician trust; Increased verification workload. | RAG, uncertainty estimation, CoT prompting, fine-tuning, safety guardrails, robust detection methods, human oversight. |
| Validation & evaluation gaps | Lack of clinically relevant benchmarks; Poor correlation of NLP metrics with clinical utility; Paucity of prospective clinical trials. | Difficulty assessing true clinical readiness & safety; Risk of deploying unsafe/ineffective tools; Hinders comparative effectiveness research. | Development of clinically meaningful validation frameworks & benchmarks; Requirement for rigorous prospective trials; Standardized reporting guidelines (e.g., TRIPOD-LLM). |
| Data privacy & security | Risk of exposing sensitive patient data used in training or prompts; Compliance with HIPAA/GDPR; Consent issues. | Breach of patient confidentiality; Identity theft; Erosion of public trust; Legal & regulatory penalties. | Data anonymization/pseudonymization, differential privacy, robust cybersecurity, penetration testing, clear informed consent processes, tiered data access policies. |
| Integration & interoperability | Difficulty integrating LLMs with existing EHRs and clinical workflows. | Limited adoption; Workflow disruption; Reduced efficiency gains; Clinician frustration. | Development of standardized APIs; Collaboration between LLM developers & EHR vendors; User-centered interface design. |
| Transparency & explainability | "Black box" nature of LLMs, making reasoning processes opaque. | Hinders clinician trust & critical appraisal; Complicates error analysis & debugging; Impedes validation; Barrier to informed consent. | Explainable AI (XAI) methods (SHAP, attention maps, KG integration); Transparent reporting of methods & data; Focus on interpretable models where possible. |

LLM: large language model, RAG: retrieval augmented generation, CoT: chain-of-thought, EHR: electronic health record, KG: knowledge graph, NLP: natural language processing, API: application programming interface, HIPAA: Health Insurance Portability and Accountability Act, GDPR: General Data Protection Regulation, SHAP: SHapley Additive exPlanations.

and reporting standards. The swift pace of LLM development complicates timely, rigorous evaluations, as models are continually updated, potentially altering their performance characteristics. Transparent reporting standards, such as the proposed TRIPOD-LLM and MI-CLEAR-LLM guidelines [43,44], are thus essential for proper appraisal and reproducibility of studies.

### 3) Data quality and privacy

Training and fine-tuning LLMs require extensive, high-quality data, the availability of which is often limited [2,3]. Protecting patient privacy is equally critical. Using patient data for model training can result in data memorization, leakage, or risks of re-identification [18]. Mitigation strategies include employing synthetic data, data anonymization, pseudonymization, differential privacy techniques, robust cybersecurity measures, model penetration testing, and implementing transparent patient consent processes [45,46].

### 4) Integration and interoperability

Successful implementation of LLMs depends heavily on seamless integration with existing clinical workflows and healthcare IT infrastructures, particularly EHRs [11]. This presents a considerable technical challenge, requiring extensive customization and interface development. User-friendly interfaces are essential for widespread clinical adoption.

### 5) Transparency and explainability

LLMs generally function as "black boxes," making it challenging to understand how specific outputs or suggestions are derived [47]. This lack of transparency discourages clinician trust, complicates error analysis and validation, and hinders informed acceptance. Developing and applying explainable AI (XAI) methods—such as SHAP [48] attention

 https://doi.org/10.4258/hir.2025.31.2.114

visualization [48], and integration with knowledge graphs [49]—tailored for medical LLMs is an active area of research, along with promoting transparent reporting practices.

## 2. Ethical Considerations

Deploying LLMs in clinical medicine raises profound ethical issues that require careful consideration and proactive management (Table 3).

### 1) Algorithmic bias and equity

A major ethical concern is that LLMs trained on datasets reflecting existing societal inequities may amplify biases in healthcare delivery [36]. This could result in poorer performance or unfair clinical recommendations for underrepresented racial or ethnic groups, gender groups, socioeconomic populations, or linguistic minorities. Linguistic bias is especially relevant, as most models are primarily trained on standard English, resulting in diminished performance for non-English speakers or users of non-standard dialects [50]. Mitigation requires intentional curation of diverse and representative training data, the development of bias detection and mitigation algorithms, validation across diverse demographic groups, and inclusive design and evaluation processes. Addressing bias is both ethically imperative and essential for clinical validity, as a biased tool cannot reliably serve all patients.

Table 3. Ethical considerations of LLMs in clinical medicine

| Consideration | Description | Risks for clinical practice | Potential mitigation strategies |
|---|---|---|---|
| Algorithmic bias & equity | Perpetuation/amplification of societal biases present in training data (racial, gender, linguistic, etc.). | Health disparities; Inequitable quality of care; Unfair or harmful recommendations for certain groups; Non-compliance with anti-discrimination laws. | Diverse & representative training data; Bias detection & mitigation algorithms; Validation across diverse populations; Inclusive design processes; Auditing for fairness. |
| Patient safety & liability | Risk of harm due to inaccurate/hallucinated outputs and unclear responsibility for errors involving LLMs. | Incorrect diagnosis/treatment; Adverse events; Delayed care; Difficulty assigning blame; Legal uncertainty; Potential barrier to adoption; Undermining trust. | Rigorous safety testing; Robust validation; Clear performance limitations disclosure; Mandatory human oversight; Development of clear regulatory guidelines; Defining roles & responsibilities for developers, institutions, clinicians; Establishing legal precedents. |
| Data governance & consent | Ethical sourcing & use of patient data; Ensuring meaningful informed consent; Data ownership rights. | Violation of patient autonomy & privacy rights; Erosion of trust; Legal non-compliance. | Transparent data use policies; Patient education on data rights; Robust consent mechanisms; Strong data governance frameworks within institutions. |
| Transparency & trust (ethical) | Lack of transparency about model capabilities, limitations, and data usage erodes trust. | Clinician reluctance to adopt; Patient skepticism; Difficulty in establishing trustworthy AI systems. | Increased transparency in reporting (data, methods, performance); Clear communication with clinicians & patients about LLM function & limits; Explainability efforts. |
| Impact on clinician-patient relationship | Potential for depersonalization, reduced empathy, or communication breakdown if used improperly. | Weakening of therapeutic alliance; Reduced patient satisfaction; Missed non-verbal cues. | Thoughtful workflow integration; Training clinicians on effective use; Prioritizing human interaction; Using LLMs to augment rather than replace communication. |
| Equity of access | Potential disparities if benefits aren't accessible to all populations/settings. | Widening health inequities; Unequal access to advanced healthcare tools [12]. | Policies promoting equitable deployment; Development of tools for low-resource settings; Ensuring accessibility for diverse linguistic groups & abilities. |

LLM: large language model.

#### 2) Patient safety and liability

LLMs' potential to produce inaccurate information—including hallucinations, factual errors, or critical omissions—constitutes a direct risk to patient safety, potentially causing erroneous diagnoses or treatment delays. Determining responsibility or liability when an LLM contributes to clinical errors is complicated [51,52]. Questions arise about whether liability rests with the clinician using the tool, the developers, the implementing institution, or a combination thereof. The absence of definitive regulatory guidelines and established legal precedents for AI-related medical errors creates uncertainty and may impede implementation. Thus, explicit guidelines, clear accountability mechanisms, and possibly new legal frameworks are essential.

#### 3) Data governance and consent

Using large volumes of patient data for LLM training and operation raises significant governance challenges [29,53]. Ethical data sourcing, patient privacy, data security, and obtaining valid informed consent for data use are critical concerns. Issues around data ownership and patient rights concerning data used by AI systems must be clearly addressed.

#### 4) Transparency and trust

The inherently opaque "black box" nature of most LLMs undermines trust among clinicians and patients. Without the ability to understand or explain an LLM's reasoning behind recommendations, clinicians are less likely to trust or confidently utilize these tools. This mistrust could result in delayed or incorrect clinical decisions, negatively impacting patient care [54]. Advancing explainability and promoting open reporting are crucial steps toward establishing trust.

#### 5) Impact on clinician-patient relationship

Integrating LLMs into clinical interactions carries the risk of depersonalizing care, reducing clinician empathy, or negatively affecting direct communication if poorly implemented. Over-reliance on LLMs could potentially deskill clinicians or diminish their critical-thinking abilities. Conversely, by reducing administrative burdens, LLMs might enable clinicians to spend more meaningful time interacting directly with patients [55]. Careful consideration of workflow integration and human factors is required to ensure technological support enhances rather than detracts from therapeutic relationships.

#### 6) Equity of access

Ensuring equitable access to the benefits of LLM technology across diverse patient populations, socioeconomic groups, and healthcare settings—including low-resource environments—is a significant ethical consideration [29,54]. The digital divide and disparities in technology access could exacerbate existing health inequities unless proactively addressed.

## IV. Discussion

### 1. Risks and Benefits of LLMs in Medicine

LLMs represent a technology with significant potential to profoundly impact clinical medicine, primarily due to their unparalleled ability to process and generate language, granting access to the vast amounts of unstructured textual data available in healthcare [2,11]. Current evidence highlights promising applications, particularly in reducing the burden of clinical documentation through summarization, note generation, and information extraction. Additionally, LLMs show considerable potential for improving patient communication by simplifying complex information, powering informative chatbots, and optimizing various aspects of clinical workflows.

Nevertheless, initial optimism must be tempered by a realistic assessment of the technical and ethical barriers to widespread, secure, and effective clinical deployment. Accuracy concerns, particularly the tendency of these models toward hallucinations, remain critical issues. A notable gap persists due to the lack of rigorous clinical validation, highlighting discrepancies between model performance on standardized benchmarks and real-world clinical decision-making. Furthermore, the "black box" nature of LLMs poses significant transparency and trust challenges, while ethical considerations surrounding bias, equity, privacy, and accountability require careful attention and proactive mitigation.

Therefore, integrating LLMs into clinical practice necessitates carefully balancing their potential benefits against inherent risks. Current technological capabilities and available evidence strongly indicate that LLMs should primarily serve as assistive tools augmenting clinician abilities rather than as autonomous entities intended to replace clinicians. Adopting a cautious, evidence-based approach is crucial, requiring robust human oversight across virtually all clinical applications and embracing a "physician-in-the-loop" paradigm, wherein clinicians retain ultimate responsibility for clinical decisions informed by LLM outputs.

Moving forward requires a collaborative, multi-stakeholder approach involving AI developers, clinicians, ethicists, regulators, healthcare institutions, and patients. Prioritizing pa-


https://doi.org/10.4258/hir.2025.31.2.114

tient safety, ensuring equity, maintaining transparency, and building trust should be guiding principles in the development and integration of these powerful technological tools.

### 2. Future Directions and Research Needs

Fully realizing the potential of LLMs in clinical medicine requires focused efforts in several key areas.

#### 1) Rigorous validation

There is an urgent need for developing and adopting standardized, clinically meaningful validation methodologies and benchmarks. Such frameworks should accurately assess performance, safety, and real-world clinical utility, extending beyond traditional NLP evaluation metrics. Conducting prospective, randomized controlled trials comparing LLM-assisted workflows with conventional care is essential to establishing genuine clinical efficacy.

#### 2) Technical advancement

Continued research is necessary to enhance model accuracy, minimize the frequency and consequences of hallucinations, improve robustness across diverse datasets and scenarios, and develop effective explainability techniques specifically tailored to clinical users. Further investigation into emerging ecosystems—in which multiple specialized AI agents (e.g., diagnostic agents, documentation agents, patient communication agents) collaborate by exchanging information and coordinating tasks within complex clinical workflows—is also warranted [56,57]. Active research areas additionally include integrating multimodal data [58,59] and extending LLM capabilities into physical systems [60].

#### 3) Bias mitigation

Proactive strategies for identifying, measuring, and mitigating biases in training datasets and model outputs are crucial to ensure equitable performance across patient populations and prevent exacerbating existing health disparities.

#### 4) Ethical and regulatory frameworks

Establishing clear ethical guidelines, robust regulatory oversight, and well-defined accountability structures is essential to govern the responsible development, deployment, and clinical use of LLM technologies in healthcare.

#### 5) Human-AI collaboration

Research should prioritize optimizing human-AI interaction models, designing intuitive interfaces, and understanding how to best integrate LLMs into clinical workflows to effectively support clinicians rather than hinder their performance.

## Conflict of Interest

No potential conflict of interest relevant to this article was reported.

## ORCID

Kyu-Hwan Jung (https://orcid.org/0000-0002-6626-6800)

## References

1. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need [Internet]. Ithaca (NY): arXive.org; 2017 [cited at 2025 Apr 15]. Available from: https://arxiv.org/abs/1706.03762.
2. Minaee S, Mikolov T, Nikzad N, Chenaghlu M, Socher R, Amatriain X, et al. Large language models: a survey [Internet]. Ithaca (NY): arXive.org; 2024 [cited at 2025 Apr 15]. Available from: https://arxiv.org/abs/2402.06196v1.
3. Thirunavukarasu AJ, Ting DS, Elangovan K, Gutierrez L, Tan TF, Ting DS. Large language models in medicine. Nat Med 2023;29(8):1930-40. https://doi.org/10.1038/s41591-023-02448-8
4. Wornow M, Xu Y, Thapa R, Patel B, Steinberg E, Fleming S, et al. The shaky foundations of large language models and foundation models for electronic health records. NPJ Digit Med 2023;6(1):135. https://doi.org/10.1038/s41746-023-00879-8
5. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;36(4):1234-40. https://doi.org/10.1093/bioinformatics/btz682
6. Huang K, Altosaar J, Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission [Internet]. Ithaca (NY): arXive.org; 2019 [cited at 2025 Apr 15]. Available from: https://arxiv.org/abs/1904.05342v1.
7. Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, et al. A large language model for electronic health records. NPJ Digit Med 2022;5(1):194. https://doi.org/10.1038/s41746-022-00742-2
8. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023;620(7972):172-80. https://doi.org/10.1038/

s41586-023-06291-2
9. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Amin M, et al. Toward expert-level medical question answering with large language models. Nat Med 2025;31(3):943-50. https://doi.org/10.1038/s41591-024-03423-7
10. The Lancet Digital Health. Large language models: a new chapter in digital health. Lancet Digit Health 2024;6(1):e1. https://doi.org/10.1016/S2589-7500(23)00254-6
11. Meng X, Yan X, Zhang K, Liu D, Cui X, Yang Y, et al. The application of large language models in medicine: a scoping review. iScience 2024;27(5):109713. https://doi.org/10.1016/j.isci.2024.109713
12. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng 2018;2(10):719-31. https://doi.org/10.1038/s41551-018-0305-z
13. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J 2019;6(2):94-8. https://doi.org/10.7861/futurehosp.6-2-94
14. Patel SB, Lam K. ChatGPT: the future of discharge summaries? Lancet Digit Health 2023;5(3):e107-8. https://doi.org/10.1016/S2589-7500(23)00021-3
15. Van Veen D, Van Uden C, Blankemeier L, Delbrouck JB, Aali A, Bluethgen C, et al. Adapted large language models can outperform medical experts in clinical text summarization. Nat Med 2024;30(4):1134-42. https://doi.org/10.1038/s41591-024-02855-5
16. Hartman V, Zhang X, Poddar R, McCarty M, Fortenko A, Sholle E, et al. Developing and evaluating large language model-generated emergency medicine handoff notes. JAMA Netw Open 2024;7(12):e2448723. https://doi.org/10.1001/jamanetworkopen.2024.48723
17. Agaronnik ND, Davis J, Manz CR, Tulsky JA, Lindvall C. Large Language Models to Identify Advance Care Planning in Patients With Advanced Cancer. J Pain Symptom Manage 2025;69(3):243-50.e1. https://doi.org/10.1016/j.jpainsymman.2024.11.016
18. Wiest IC, Ferber D, Zhu J, van Treeck M, Meyer SK, Juglan R, et al. Privacy-preserving large language models for structured medical information retrieval. NPJ Digit Med 2024;7(1):257. https://doi.org/10.1038/s41746-024-01233-2
19. Lyu Q, Tan J, Zapadka ME, Ponnatapura J, Niu C, Myers KJ, et al. Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential. Vis Comput Ind Biomed Art 2023;6(1):9. https://doi.org/10.1186/s42492-023-00136-5
20. Brodeur PG, Buckley TA, Kanjee Z, Goh E, Ling EB, Jain P, et al. Superhuman performance of a large language model on the reasoning tasks of a physician [Internet]. Ithaca (NY): arXive.org; 2024 [cited at 2025 Apr 15]. Available from: https://arxiv.org/abs/2412.10849.
21. McDuff D, Schaekermann M, Tu T, Palepu A, Wang A, Garrison J, et al. Towards accurate differential diagnosis with large language models. Nature 2025 Apr 9 [Epub]. https://doi.org/10.1038/s41586-025-08869-4.
22. Tu T, Schaekermann M, Palepu A, Saab K, Freyberg J, Tanno R, et al. Towards conversational diagnostic artificial intelligence. Nature 2025 Apr 9 [Epub]. https://doi.org/10.1038/s41586-025-08866-7.
23. Tu T, Azizi S, Driess D, Schaekermann M, Amin M, Chang PC, et al. Towards generalist biomedical AI [Internet]. Ithaca (NY): arXive.org; 2023 [cited at 2025 Apr 15]. Available from: https://arxiv.org/abs/2307.14334.
24. Ali R, Tang OY, Connolly ID, Fridley JS, Shin JH, Zadnik Sullivan PL, et al. Performance of ChatGPT, GPT-4, and Google Bard on a neurosurgery oral boards preparation question bank. Neurosurgery 2023;93(5):1090-8. https://doi.org/10.1227/neu.0000000000002551
25. Saab K, Tu T, Weng WH, Tanno R, Stutz D, Wulczyn E, et al. Capabilities of Gemini models in medicine [Internet]. Ithaca (NY): arXive.org; 2024 [cited at 2025 Apr 15]. Available from: https://arxiv.org/abs/2404.18416.
26. Oh Y, Park S, Byun HK, Cho Y, Lee IJ, Kim JS, et al. LLM-driven multimodal target volume contouring in radiation oncology. Nat Commun 2024;15(1):9186. https://doi.org/10.1038/s41467-024-53387-y
27. Hager P, Jungmann F, Holland R, Bhagat K, Hubrecht I, Knauer M, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med 2024;30(9):2613-22. https://doi.org/10.1038/s41591-024-03097-1
28. Chen H, Fang Z, Singla Y, Dredze M. Benchmarking large language models on answering and explaining challenging medical questions [Internet]. Ithaca (NY): arXive.org; 2024 [cited at 2025 Apr 15]. Available from: https://arxiv.org/abs/2402.18060v1.
29. Ong JC, Chang SY, William W, Butte AJ, Shah NH, Chew LS, et al. Ethical and regulatory challenges of large language models in medicine. Lancet Digit Health 2024;6(6):e428-32. https://doi.org/10.1016/S2589-7500(24)00061-X
30. Ali R, Connolly ID, Tang OY, Mirza FN, Johnston B, Abdulrazeq HF, et al. Bridging the literacy gap for surgical consents: an AI-human expert collaborative approach. NPJ Digit Med 2024;7(1):63. https://doi.org/10.1038/s41746-024-01039-2

31. Decker H, Trang K, Ramirez J, Colley A, Pierce L, Coleman M, et al. Large language model-based chatbot vs surgeon-generated informed consent documentation for common procedures. JAMA Netw Open 2023;6(10):e2336997. https://doi.org/10.1001/jamanetworkopen.2023.36997
32. Ghim JL, Ahn S. Transforming clinical trials: the emerging roles of large language models. Transl Clin Pharmacol 2023;31(3):131-8. https://doi.org/10.12793/tcp.2023.31.e16
33. Zaretsky J, Kim JM, Baskharoun S, Zhao Y, Austrian J, Aphinyanaphongs Y, et al. Generative artificial intelligence to transform inpatient discharge summaries to patient-friendly language and format. JAMA Netw Open 2024;7(3):e240357. https://doi.org/10.1001/jamanetworkopen.2024.0357
34. Cho S, Lee M, Yu J, Yoon J, Choi JB, Jung KH, et al. Leveraging large language models for improved understanding of communications with patients with cancer in a call center setting: proof-of-concept study. J Med Internet Res 2024;26:e63892. https://doi.org/10.2196/63892
35. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023;183(6):589-96. https://doi.org/10.1001/jamainternmed.2023.1838
36. Aydin S, Karabacak M, Vlachos V, Margetis K. Large language models in patient education: a scoping review of applications in medicine. Front Med (Lausanne) 2024;11:1477898. https://doi.org/10.3389/fmed.2024.1477898
37. Azamfirei R, Kudchadkar SR, Fackler J. Large language models and the perils of their hallucinations. Crit Care 2023;27(1):120. https://doi.org/10.1186/s13054-023-04393-x
38. Rawte V, Sheth A, Das A. A survey of hallucination in large foundation models [Internet]. Ithaca (NY): arXive.org; 2023 [cited at 2025 Apr 15]. Available from: https://arxiv.org/abs/2309.05922.
39. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation [Internet]. Ithaca (NY): arXive.org; 2022 [cited at 2025 Apr 15]. Available from: https://arxiv.org/abs/2202.03629v1.
40. Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions [Internet]. Ithaca (NY): arXive.org; 2023 [cited at 2025 Apr 15]. Available from: https://arxiv.org/abs/2311.05232v1.
41. Tonmoy SM, Zaman SM, Jain V, Rani A, Rawte V, Chadha A, et al. A comprehensive survey of hallucination mitigation techniques in large language models [Internet]. Ithaca (NY): arXive.org; 2024 [cited at 2025 Apr 15]. Available from: https://arxiv.org/abs/2401.01313.
42. Yu F, Endo M, Krishnan R, Pan I, Tsai A, Reis EP, et al. Evaluating progress in automatic chest X-ray radiology report generation. Patterns (N Y) 2023;4(9):100802. https://doi.org/10.1016/j.patter.2023.100802
43. Gallifant J, Afshar M, Ameen S, Aphinyanaphongs Y, Chen S, Cacciamani G, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med 2025;31(1):60-9. https://doi.org/10.1038/s41591-024-03425-5
44. Park SH, Suh CH, Lee JH, Kahn CE, Moy L. Minimum reporting items for clear evaluation of accuracy reports of large language models in healthcare (MI-CLEAR-LLM). Korean J Radiol 2024;25(10):865-8. https://doi.org/10.3348/kjr.2024.0843
45. Behnia R, Ebrahimi M, Pacheco J, Padmanabhan B. Privately fine-tuning large language models with differential privacy [Internet]. Ithaca (NY): arXive.org; 2022 [cited at 2025 Apr 15]. Available from: https://arxiv.org/abs/2210.15042v1.
46. Charles Z, Ganesh A, McKenna R, McMahan HB, Mitchell N, Pillutla K, et al. Fine-tuning large language models with user-level differential privacy [Internet]. Ithaca (NY): arXive.org; 2024 [cited at 2025 Apr 15]. Available from: https://arxiv.org/abs/2407.07737.
47. Zhao H, Chen H, Yang F, Liu N, Deng H, Cai H, et al. Explainability for large language models: a survey [Internet]. Ithaca (NY): arXive.org; 2023 [cited at 2025 Apr 15]. Available from: https://arxiv.org/abs/2309.01029.
48. Goldshmidt R, Horovicz M. TokenSHAP: interpreting large language models with Monte Carlo Shapley value estimation [Internet]. Ithaca (NY): arXive.org; 2024 [cited at 2025 Apr 15]. Available from: https://arxiv.org/abs/2407.10114.
49. Kau A, He X, Nambissan A, Astudillo A, Yin H, Aryani A. Combining knowledge graphs and large language models [Internet]. Ithaca (NY): arXive.org; 2024 [cited at 2025 Apr 15]. Available from: https://arxiv.org/abs/2407.06564.
50. Li Z, Shi Y, Liu Z, Yang F, Payani A, Liu N, et al. Language ranker: a metric for quantifying LLM performance across high and low-resource languages [Internet]. Ithaca (NY): arXive.org; 2024 [cited at 2025 Apr 15]. Available from: https://arxiv.org/abs/2404.11553.

51. Mesko B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med 2023;6(1):120. https://doi.org/10.1038/s41746-023-00873-0
52. Reddy S, Allan S, Coghlan S, Cooper P. A governance model for the application of AI in health care. J Am Med Inform Assoc 2020;27(3):491-7. https://doi.org/10.1093/jamia/ocz192
53. Harrer S. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. EBioMedicine 2023;90:104512. https://doi.org/10.1016/j.ebiom.2023.104512
54. 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. https://doi.org/10.1038/s41746-024-01157-x
55. Workum JD, van de Sande D, Gommers D, van Genderen ME. Bridging the gap: a practical step-by-step approach to warrant safe implementation of large language models in healthcare. Front Artif Intell 2025;8:1504805. https://doi.org/10.3389/frai.2025.1504805
56. Kim Y, Park C, Jeong H, Grau-Vilchez C, Chan YS, Xu X, et al. A demonstration of adaptive collaboration of large language models for medical decision-making [Internet]. Ithaca (NY): arXive.org; 2024 [cited at 2025 Apr 15]. Available from: https://arxiv.org/abs/2411.00248.
57. Kim Y, Park C, Jeong H, Chan YS, Xu X, McDuff D, et al. MDAgents: an adaptive collaboration of LLMs for medical decision-making [Internet]. Ithaca (NY): arXive.org; 2024 [cited at 2025 Apr 15]. Available from: https://arxiv.org/abs/2404.15155.
58. Lu MY, Chen B, Williamson DF, Chen RJ, Zhao M, Chow AK, et al. A multimodal generative AI copilot for human pathology. Nature 2024;634(8033):466-73. https://doi.org/10.1038/s41586-024-07618-3
59. Hong EK, Roh B, Park B, Jo JB, Bae W, Soung Park J, et al. Value of using a generative AI model in chest radiography reporting: a reader study. Radiology 2025;314(3):e241646. https://doi.org/10.1148/radiol.241646
60. Kim MJ, Pertsch K, Karamcheti S, Xiao T, Balakrishna A, Nair S, et al. OpenVLA: an open-source vision-language-action model [Internet]. Ithaca (NY): arXive.org; 2024 [cited at 2025 Apr 15]. Available from: https://arxiv.org/abs/2406.09246.