AI Medicine
https://www.sciltp.com/journals/aim

Article

Large Language Models in Medicine: Application Status and Challenges

Ziqi Chen [1], Yiwei Lu [1], Yan Zeng [1], Dingcheng Tian [2], Yun Li [1] and Fei Li [1,*]

[1] College of Communication and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210023, China
[2] College of Medicine and Biological Information Engineering, Northeastern University, Shenyang 110016, China
* Correspondence: lifei@njupt.edu.cn

How To Cite: Chen, Z.; Lu, Y.; Zeng, Y.; et al. Large Language Models in Medicine: Application Status and Challenges. *AI Medicine* **2025**, *2*(2), 8. https://doi.org/10.53941/aim.2025.100008

Received: 3 September 2025
Revised: 1 December 2025
Accepted: 2 December 2025
Published: 9 December 2025

Abstract: In recent years, the rapid evolution of large language models (LLMs) has driven revolutionary changes in multiple medical application scenarios, showcasing vast potential. With continuous technological developments, LLMs have made substantial advancements in clinical diagnosis and support, medical education and training, clinical documentation processing, patient interaction and public education, and medical research assistance. This paper first explores the practical applications of LLMs in these medical scenarios, analyzing how LLMs contribute to improving clinical decision-making efficiency, optimizing medical education, improving patient interaction, and advancing medical research. The paper then discusses the key technical elements of LLMs in the medical field, including data and knowledge construction, model training methods, and multi-modal data fusion. We also focus on the challenges faced by LLMs in medical applications, including data limitations, model hallucinations, and insufficient standardization of evaluation. Finally, the paper looks ahead to future research directions, highlighting the improvement of evaluation frameworks, the enhancement of personalized medical capabilities, the development of multi-modal medical LLMs, and the strengthening of ethical and regulatory compliance. Through these analyses, this paper aims to advance the ongoing development and practical application of medical LLMs.

Keywords: large language models; medical applications; clinical diagnosis; multi-modal

1. Introduction

Over the past two years, deep learning-based large language models (LLMs) have rapidly emerged, demonstrating outstanding performance in multiple task scenarios and becoming an important research direction in the field of artificial intelligence [1]. By training on large datasets, LLMs have excelled in tasks such as text generation and complex reasoning, driving the automation of information processing and knowledge discovery [2]. In the medical field, LLMs have broad application prospects, with the potential to enhance clinical work efficiency, refine diagnosis and treatment decisions, and expedite the progress of medical research, making them a primary emphasis in mmmdical AI research [3].

LLMs have made substantial advancements in medical applications, with ChatGPT being one of the key representatives. Although this model has not been specifically fine-tuned for the medical domain, it has still passed and even excelled in the United States Medical Licensing Examination (USMLE) Step 3, and can generate analysis answers deemed valuable by clinical experts [4]. Released in 2024, GPT-4o expanded the context window to 128k tokens and supports multi-modal inputs of text, images, and audio. It allows the model to simultaneously handle different modalities such as medical records and medical images in a single prompt, thus providing new technological possibilities for intelligent processing of complex clinical tasks [5].

At the same time, the open-source community has also introduced some high-quality solutions. Released in early 2025, DeepSeek-R1 achieved performance comparable to that of GPT-4 in mathematical reasoning and Chinese

medical question-and-answer (Q&A) tasks, and fully opened the model weights and training process under the MIT license [6, 7]. With its openness and local deployment capability, the DeepSeek series of models has been applied in multiple medical institutions in China, covering scenarios such as electronic health records (EHRs) system integration, intelligent Q&A, and digital nursing assistants [8].

Current research indicates that LLMs show strong generalization and transfer capabilities in medical applications [9]. On the one hand, general models (such as ChatGPT) have been able to pass standardized medical exams without domain-specific fine-tuning and generate clinically valuable answers [10]. On the other hand, open-source localized models (such as DeepSeek-R1) have achieved performance comparable to that of GPT-4 in Chinese medical contexts, particularly demonstrating practical application potential in tasks such as symptom Q&A and prescription recommendations [11]. These advancements suggest that LLMs are expected to further improve clinical decision-making efficiency, reduce documentation burdens, promote medical education and research innovation, and push the healthcare system toward greater intelligence and precision.

This paper aims to systematically explore the current status, challenges, and future development directions of LLMs in the medical field. We conducted a literature review by searching databases such as Google Scholar, Web of Science, and PubMed, collecting relevant research on LLMs in medicine, and analyzing and discussing five core aspects: clinical diagnosis and support, medical education and training, clinical documentation processing, patient interaction and public education, and medical research assistance.

The structure of the paper is as follows. Section 2 focuses on the application progress of LLMs in the medical field, covering five core scenarios: clinical diagnosis and support, medical education and training, clinical documentation processing, patient interaction and public education, and medical research assistance. Section 3 discusses the key challenges currently faced, including data limitations, hallucination phenomena, and the standardization of evaluation. Section 4 provides a conclusion.

2. Applications of LLMs in Medicine

We searched for existing research works in the medical field based on ChatGPT in the Google Scholar, Web of Science, and PubMed databases. The keywords used were "LLM" and "Medicine". The search period was from 1 January 2023 to 30 June 2025. After retrieving the articles, we categorized them into five aspects: clinical diagnosis and support, medical education and training, clinical documentation, patient interaction and public education, and medical research assistance. This section first introduces several representative large language models and then analyzes the current applications of LLMs in the medical field according to the five scenarios described above.

2.1. Common General-Purpose Large Language Model Backbones Used in Medical Applications

In current research on medical large language models, several advanced general-purpose models, such as ChatGPT [12], DeepSeek [13], LLaMA [14], and Claude [15], have become key foundations for many application studies. ChatGPT, developed by OpenAI based on the GPT series architecture, is one of the most widely used general-purpose large language models to date. It is first pretrained on large-scale general corpora using autoregressive language modeling, and then aligned with human intent through techniques such as instruction tuning and reinforcement learning from human feedback (RLHF), enabling the model to better follow human instructions. This paradigm allows ChatGPT to demonstrate strong robustness and generalization ability in tasks such as open-ended question answering and complex reasoning. In medical settings, ChatGPT has been widely explored for clinical question answering, medical knowledge explanation, case discussions, exam simulation, and clinical document drafting and polishing. To improve the accuracy and trustworthiness of its responses, some studies further combine ChatGPT with retrieval-augmented generation or external medical knowledge bases to trace and verify information sources. Owing to its strong instruction-following capability, multilingual support, and continuous updates, ChatGPT is often used as a baseline model or foundational tool in medical LLM research; however, its closed-source nature also to some extent limits in-depth investigation of its internal mechanisms and training details.

DeepSeek is a rapidly evolving representative of open-source large language models, targeting general language understanding and complex reasoning tasks. It adopts a fully open-source strategy, releasing a series of models with different parameter scales to support flexible deployment from resource-constrained environments to high-performance servers. In terms of training strategies, DeepSeek emphasizes strengthening capabilities in code, mathematical reasoning, and complex problem decomposition, thus markedly enhancing its logical reasoning and tool-use abilities [16]. As the model parameters and weights are fully released under a permissive MIT license, DeepSeek is highly suitable for in-depth instruction tuning and domain adaptation in specialized fields such as medicine, making it an ideal backbone for building domain-specific large models. At present, DeepSeek has already

achieved substantive application progress in the medical domain. In medical applications, researchers often adopt DeepSeek as a foundational model for Chinese or multilingual medical text processing, including tasks such as electronic health record understanding, structured information extraction, clinical document generation, and research writing assistance. Compared with closed-source models, DeepSeek's open-source nature greatly promotes research transparency and experimental reproducibility.

LLaMA is one of the most representative open-source large language model families and has evolved into multiple versions and parameter scales. This series of models is pretrained on large-scale, multi-source corpora, and its architecture and engineering design strike a balance between performance and efficiency, making it well suited as a "backbone" for downstream tasks. In the medical field, the open-source nature of LLaMA has made it a popular choice for constructing specialized medical large language models. Researchers often perform domain-adaptive pretraining or instruction tuning on top of LLaMA to develop models for tasks such as biomedical literature understanding, clinical text processing, specialty-specific question answering, and case analysis. Its good scalability and rich open-source ecosystem enable researchers to flexibly adjust model size to accommodate different computational resource constraints, and to easily integrate medical corpora and knowledge bases for customized optimization, thereby systematically exploring the relationships among model scale, training data, and task performance.

The core feature of Claude lies in its use of the"Constitutional AI"safety-alignment framework. This framework aims to guide the model to perform self-supervision and self-improvement according to a predefined set of principles, thereby substantially reducing harmful, biased, or inappropriate content generation and establishing a foundation for high safety and controllability [15]. In terms of model capabilities, Claude exhibits strong performance in long-context handling and robust reasoning, maintaining semantic coherence and logical consistency over extended contexts. Thanks to these safety-oriented designs and technical strengths, Claude is suitable for medical scenarios with very high safety requirements. Although, compared with fully open-source models, Claude's parameters and complete training details are not publicly available, its engineering practices in safety alignment and long-context reasoning offer a valuable technical pathway for building "safe and controllable medical large language models".

In summary, mainstream large language models such as ChatGPT, DeepSeek, LLaMA, and Claude differ in their training data, degree of openness, core capabilities, and safety strategies, and together constitute the technical foundation for current research and applications of medical large language models. These differences influence model selection and technical adaptation routes for various medical tasks. For example, open-source models (such as DeepSeek and LLaMA) are better suited as backbones for continued pretraining and specialization in the medical domain, whereas models characterized by strong safety and conversational abilities (such as ChatGPT and Claude) are more frequently used directly in clinical question answering, doctor-patient interaction, and medical education scenarios. As will be illustrated in the following sections, many cutting-edge studies on medical LLMs build upon a deep understanding of the characteristics of these models, and, through fine-tuning, adaptation, or application development, continuously extend their boundaries in clinical and research applications.

2.2. Clinical Diagnosis and Support

In the field of clinical diagnosis and support, LLMs, such as ChatGPT and DeepSeek, can provide preliminary diagnostic suggestions by analyzing patients' descriptions of symptoms, signs, and discomfort. When deeply integrated with clinical workflows, these models can significantly improve the accuracy of differential diagnosis and effectively support patient triage management. Current research in this field divides applications into four main tasks: "general reasoning and clinical decision support", "diagnosis and prognosis prediction for specific diseases", "multimodal diagnostic information integration", and "model performance evaluation and comparative analysis". These tasks are interconnected in practice and jointly promote the development of LLMs in medicine.

First, general reasoning and clinical decision support focus on exploring how to utilize LLMs to enhance the efficiency and accuracy of clinical decision-making, optimize the differential diagnosis process, and effectively support triage decisions. Studies have shown that GPT-4 achieves the highest accuracy in medical question-answering tasks. However, it still has certain shortcomings in recognizing rare diseases [17]. To further enhance question-answering and reasoning capabilities, Singhal et al. proposed Med-PaLM 2, which achieved performance comparable to that of experts across multiple datasets and physician evaluations [18]. In more complex differential diagnosis tasks, McDuff et al. developed the AMIE model, which achieved 59.1% and 51.7% accuracy in independent diagnosis and physician-assisted mode, respectively. However, it still did not meet the standards for clinical application [19]. In the emergency department setting, Williams et al. evaluated the performance of GPT-3.5 and GPT-4, and the results showed that the quality of recommendations from both models was significantly lower than that of resident physicians [20]. Wu et al. demonstrated that DeepSeek-R1 significantly improved clinicians' efficiency and accuracy in diagnosing complex critical illnesses, showing its potential in highly complex clinical scenarios [21].

Diagnosis and prognosis prediction for specific diseases focuses on the precise diagnostic and prognostic capabilities of LLMs in specific disease domains, highlighting the advantages of disease-specific models in improving diagnostic accuracy and clinical value. The study by Perlis et al. showed that GPT-4 turbo achieved an accuracy of 50.8% in treatment plan recommendations for bipolar depression, highlighting its application potential in the field of mental health [22]. Shashi et al. proposed COMPOSER-LLM, which performed well in early sepsis prediction, with a sensitivity of 72.1% and an F1 score of 61.0%, providing reliable support for emergency medicine [23]. In oral pathology diagnosis, Ka et al. found that DeepSeek-v3 outperformed ChatGPT-4o in accuracy [24]. Zhang et al. demonstrated that DeepSeek-V3 achieved performance comparable to that of senior radiologists in the LR classification of hepatocellular carcinoma, highlighting its value in liver cancer diagnosis [25]. In addition, Leng et al. pointed out that GPT-4o performed well in cognitive impairment classification and clinical scoring, demonstrating significant clinical potential [26]. These studies demonstrate that disease-specific LLMs can significantly enhance diagnostic accuracy and prognostic prediction capabilities, particularly in the early identification of complex diseases and the development of personalized treatment plans.

In multimodal diagnostic information integration scenarios, by combining images, text, and other multimodal data, LLMs can further improve the accuracy of disease diagnosis and expand their applications in complex clinical settings. For example, Zhou et al. proposed SkinGPT-4, which showed highly reliable diagnostic results in dermatology, highlighting the application prospects of multimodal data in dermatology [27]. The LLMSeg model by Oh et al. outperformed traditional vision AI models in target contour delineation tasks, with stronger generalization ability [28]. Li et al. developed DeepDR-LLM, which significantly improved diagnostic quality in diabetes management and retinopathy screening [29]. Tran et al. proposed HistoGPT, which generated dermatopathology reports of near-expert quality, demonstrating excellent performance [30]. Lu et al. proposed PathChat as a general pathology AI assistant, demonstrating outstanding performance on pathology-related questions and providing new ideas for the intelligentization of pathology [31]. In addition, Wu et al. developed the MMCBM model, which achieved an F1 score of 0.91 in diagnosing choroidal tumors, demonstrating strong potential for application [32]. These studies indicate that multimodal data fusion models can not only improve diagnostic accuracy but also expand their feasibility and effectiveness in different clinical application scenarios.

Finally, the task of model performance evaluation and comparative analysis is devoted to comprehensively evaluating and comparing the performance of mainstream LLMs in clinical diagnostic tasks to reveal their advantages and limitations. Ha et al. evaluated the diagnosis of oral lesions and found that DeepSeek-V3 surpassed ChatGPT-4o in Top-3 accuracy, showing its advantage in oral disease diagnosis [33]. To et al. evaluated the capabilities of DeepSeek-R1, ChatGPT-o1, and Llama 3.1, and showed that DeepSeek-R1 outperformed the other models in diagnostic reasoning [34]. The study by Chan et al. showed that DeepSeek-R1 achieved diagnostic accuracy comparable to GPT-4 in complex cases, but was slightly lower in differential diagnosis [35]. In the performance evaluation of ophthalmology question-answering, Shean et al. found that GPT o1 Pro achieved an accuracy of 83.4% in ophthalmology questions, providing practical support for ophthalmic diagnosis [36]. These studies demonstrate the performance differences among various LLMs in clinical diagnosis and provide valuable data support for the further optimization and development of intelligent healthcare systems.

Current research indicates that applications in the field of clinical diagnosis and support can be categorized into four distinct and interrelated tasks: general reasoning, disease-specific refinement, multimodal integration, and benchmark evaluation. These tasks are intertwined, jointly promoting the application and development of LLMs in clinical diagnosis and support. General reasoning improves decision-making efficiency and accuracy, while disease-specific refinement enhances diagnostic and prognostic prediction capabilities for specific diseases. Multimodal integration expands the application scenarios of the models, and benchmark evaluation provides scientific evidence for comparing the advantages and disadvantages of different models. These research findings provide strong support for optimizing intelligent healthcare systems and enhancing clinical practice. Table 1 summarizes the key applications of LLMs in clinical diagnosis and support, showing their specific performance and outcomes in different tasks.

2.3. Medical Education and Training

With the development of LLMs, their potential in medical education and training is also being explored. The application of LLMs in clinical education, medical exams, automatic grading, and specialized training has gradually become a research hotspot. Through their powerful natural language processing capabilities, LLMs can offer personalized learning experiences and play a crucial role in medical education, thereby enhancing teaching efficiency and assessment accuracy. This section will discuss three application scenarios of LLMs in medical education based on current research findings: applications in residency training and medical exams, in medical specialty education and clinical training, and their role in automatic grading, evaluation, and educational transformation.

Table 1. LLM's application in clinical diagnosis and support.

Task	Characteristics	Object	Foundation Model	Result	Paper
Diagnostic reasoning and clinical recommendations	Enhancing clinical decision-making efficiency and accuracy with LLMs, while optimizing differential diagnosis and triage processes	Evaluate clinical accuracy of large language models on medical Q&A tasks	ChatGPT, Llama 2	GPT-4 best overall; poor on rare-diseases; Llama lightly lower	[17]
		Propose Med-PaLM 2 for expert-level medical question answering	Med-PaLM 2	Med-PaLM 2 performs excellently across multiple datasets and physician evaluations	[18]
		Propose AMIE for accurate differential diagnosis	AMIE	Independent diagnosis accuracy: 59.1%, clinician assistance: 51.7%	[19]
		Providing clinical recommendations in the Emergency Department	GPT-4, GPT-3.5	Significantly lower than resident physicians	[20]
		Diagnosis of complex critical illness cases	DeepSeek-R1	DeepSeek-R1 improved diagnostic accuracy and efficiency	[21]
Disease-specific diagnosis and prediction	Diagnosis and prognosis prediction based on specific diseases, highlighting the precision of disease-specific models	Bipolar depression	GPT4-turbo	50.8% best treatment,	[22]
		Early sepsis prediction	COMPOSER-LLM	Sensitivity:72.1%, Positive predictive: 52.9%, F-1: 61.0%	[23]
		Diagnosis of oral pathologies	ChatGPT-4o, Deepseek-v3	ChatGPT-4o mean score: 3.15 ± 0.41 Deepseek-v3 mean score: 4.02 ± 0.36	[24]
		Hepatocellular carcinoma diagnosis	DeepSeek-V3	DeepSeek-V3 is on par with senior radiologists in LR classifications	[25]
		Cognitive impairment detecting	GPT-4o	GPT-4o excels in CI classification and CDR scoring, with clinical potential	[26]
Multimodal diagnostic integration	Integrating multimodal data such as images and text to improve diagnostic accuracy and expand clinical applications	Dermatological diagnosis	SkinGPT-4	SkinGPT-4 provides reliable diagnoses	[27]
		Target volume contouring	LLMSeg	Superior to vision-only AI models, with strong generalization	[28]
		Diabetes care	DeepDR-LLM	Improved the quality of diabetes management and DR screening	[29]
		Generating dermatopathology reports	HistoGPT	Predicts tumor features, report quality similar to human	[30]
		General Pathology AI Assistant	PathChat	Performs excellently in diagnosis and pathology-related issues	[31]
		Choroid neoplasias	MMCBM	F1: 0.91	[32]
Comparative analysis and benchmarking	Focusing on model performance evaluation by comparing the performance of different LLMs in clinical diagnostic tasks	Oral lesions	ChatGPT-4o, DeepSeek-V3	DeepSeek-3 outperforms ChatGPT-4o in Top-3 accuracy	[33]
		Evaluate the ability to perform four different medical tasks.	DeepSeek-R1, ChatGPT-o1, Llama 3.1	DeepSeek-R1 outperforms other models in diagnostic reasoning	[34]
		Complex case diagnosis evaluation	DeepSeek-R1, GPT-4	DeepSeek-R1 matches GPT-4 in diagnostic accuracy but has lower differential diagnosis accuracy.	[35]
		Evaluate the model's performance on ophthalmology questions.	DeepSeek, ChatGPT	GPT o1 Pro achieved the highest accuracy on ophthalmology questions (83.4%)	[36]

Notes: AMIE: Articulate medical intelligence explorer model.

In clinical education and residency training, the application of LLMs has shown significant results. For example, comparative studies on the performance of DeepSeek-R1 and ChatGPT-4o in the Chinese National Medical Licensing Examination (CNMLE) showed that DeepSeek-R1 outperformed ChatGPT-4o in accuracy (92.0% vs. 87.2%) [37]. In addition, the application of ChatGPT-3.5 in residency training was also evaluated. The study found that ChatGPT-3.5 achieved a 45.1% correct rate in exams, with positive feedback from residents [38]. These studies suggest that LLMs can effectively support residency training and exam preparation. LLMs also show potential in medical specialty education. In fields like dermatology and radiology, LLMs have demonstrated strong generative and consultative capabilities. For example, a comparison between DeepSeek and ChatGPT in answering prostate cancer-related questions showed that DeepSeek performed better in Chinese environments, while the two models performed similarly in English [39]. Furthermore, the application of ChatGPT in cleft lip repair education was evaluated, achieving an average score of 2.9/4, with the highest scores in clarity and quality [40]. GPT-4 demonstrated high accuracy (4.45/5) and quality scores (4.28/5) in generating clinical cases for dermatology education [41]. In the field of radiology, the application of Retrieval-Augmented Generation (RAG) technology to enhance LLM performance showed that the RAG-enhanced Llama 3.2 11B model outperformed in accuracy, safety, and applicability [42].

Overall, the application of LLMs in medical education and training is driving the transformation of educational models. Automatic grading and assessment are an important application direction of LLMs in medical education. For example, the performance of GPT-4 and Gemini 1.0 Pro in automatic grading has been widely studied, with results showing that Gemini 1.0 Pro is highly consistent with human grading, while GPT-4's grading is lower [43]. In comparison with student scores, Gemini Pro performed the best, sparking discussions on the reliability and fairness of automatic grading [44]. Additionally, the application of GPT-3 in paper grading was explored, with its performance improving after incorporating language features [45]. These studies suggest that LLMs not only enhance the efficiency of medical education but also drive innovation in assessment and grading methods.

The application of LLMs in medical education and training demonstrates their enormous potential in improving teaching quality and assessment efficiency. From residency training to medical specialty education, and automatic grading and evaluation, LLMs are driving innovation in educational models. However, despite the promising performance of LLMs, the reliability and fairness of automatic grading remain issues that future research needs to address. Table 2 summarizes the applications of LLMs in medical education and training, showcasing their performance and results in different tasks.

Table 2. LLM's application in medical education and training.

Task	Characteristics	Object	Foundation Model	Result	Paper
LLMs integration in clinical education and training	Focusing on the application of LLMs in residency training and medical licensing exams to improve the quality and efficiency of clinical education	Compare DeepSeek-R1 and ChatGPT-4o on CNMLE	DeepSeek-R1, ChatGPT-4o	DeepSeek-R1 accuracy of 92.0%, ChatGPT-4o accuracy of 87.2%	[37]
		Evaluate ChatGPT's role in residency training	ChatGPT-3.5	45.1% correct on exams, positive feedback	[38]
LLMs in medical education and specialized fields	Focusing on the application of LLMs in medical specialty education, such as training and consultation in fields like dermatology and radiology	Compare DeepSeek and ChatGPT on prostate cancer questions	DeepSeek, ChatGPT	DeepSeek outperformed ChatGPT in Chinese. Similar in English	[39]
		Evaluate ChatGPT's role in cleft lip repair education	ChatGPT	Average rating of 2.9/4, highest in clarity and quality	[40]
		Explore GPT-4's role in generating clinical vignettes for dermatology education	GPT-4	High ratings in accuracy (4.45/5) and quality (4.28/5)	[41]
		Assess RAG's impact on local LLMs in radiology consultation	Llama 3.2, GPT-4o, Claude 3.5	RAG-enhanced Llama 3.2 11B outperformed others	[42]
LLMs in grading, assessment, and education transformation	Focusing on the role of LLMs in automatic grading, evaluation, and educational transformation, driving innovation and change in educational models	Evaluate GPT-4 and Gemini 1.0 Pro for automatic grading	GPT-4, Gemini 1.0 Pro	GPT-4 had lower grades, Gemini aligned well with human grading	[43]
		Compare LLM performance with student scores	Gemini Pro, Llama 3.1, Mistral-Large	Gemini Pro outperformed others, raising concerns about assessment	[44]
		Evaluate GPT-3 in essay scoring	GPT-3	GPT-3 scored moderately, improved with linguistic features	[45]

2.4. Clinical Documentation

With the continuous increase in data in the healthcare field, the processing and management of clinical documents have become particularly important. LLMs are gradually playing a significant role in simplifying, summarizing, and extracting information from clinical documents. By automating the generation, simplification, and structuring of clinical reports, LLMs can enhance document readability and improve healthcare work efficiency, while reducing the workload of doctors and medical staff. This section will discuss the application of LLMs in clinical documents, covering three main aspects: simplification and summarization of clinical documents, extraction and structuring of clinical information, and generation of clinical review reports.

In the simplification and summarization of clinical documents, LLMs have shown great potential. Studies have shown that ChatGPT has been used to simplify radiology reports. Although the reports are accurate, the

simplification process may lead to biases in the conclusions for patients [46]. Another study used GPT-4o to simplify discharge summaries and provide lifestyle recommendations. The results indicated an improvement in readability, though there was still a lack of personalization [47]. Furthermore, BrainGPT's performance in generating 3D brain CT reports suggests that LLMs can effectively improve the accuracy and relevance of reports [48].

In clinical information extraction and structuring, LLMs are widely used to extract key information from unstructured data. GatorTron has enhanced concept extraction and medical question-answering accuracy through optimized clinical natural language processing methods [49]. GPT-4o has made substantial advancements in pathological report information extraction, achieving 99% alignment with gold standards after multiple iterations of optimization [50]. Moreover, research indicates that fine-tuning smaller LLMs with synthetic data not only reduces computational costs but also maintains high accuracy [51]. In discharge summary generation, Russell-GPT 1.0 has shown time-saving and accuracy improvements, although doctor verification is still needed [52]. Lastly, LLMs have also demonstrated potential in generating clinical review reports. Studies evaluating various LLMs for generating clinical review reports acknowledge their accuracy, but there are still challenges in completeness and fluency [53].

Overall, LLMs demonstrate significant potential for application in clinical document processing, enhancing the efficiency and accuracy of medical documentation. However, challenges remain in areas such as personalization and fluency. Future research should focus on further optimizing LLMs to meet the high standards required for clinical document generation and information extraction. Table 3 summarizes the key applications and achievements of LLMs in simplifying and summarizing clinical documents, extracting and structuring clinical information, and generating clinical review reports.

Table 3. LLM's application in clinical documentation.

Task	Characteristics	Object	Foundation Model	Result	Paper
LLMs in simplifying and summarizing clinical documents	Focusing on the application of LLMs in simplifying and summarizing clinical documents to improve readability and patient comprehension	Evaluate ChatGPT in simplifying radiology reports	ChatGPT	Reports were accurate but could lead to incorrect conclusions	[46]
		Simplify discharge summaries and provide lifestyle advice	GPT-4o	Improved readability, but lacked personalization	[47]
		Evaluate BrainGPT in 3D brain CT report generation	BrainGPT	Outperformed baseline, with high accuracy and relevance	[48]
LLMs for clinical information extraction and structuring	Focusing on the application of LLMs in extracting and structuring information from clinical documents	Develop and evaluate GatorTron for clinical NLP tasks	GatorTron	Improved accuracy in tasks like concept extraction and Q&A	[49]
		Optimize LLM for clinical info extraction from pathology reports	GPT-4o	99% alignment with gold-standard after six refinements	[50]
		Fine-tune small LLMs using synthetic data for clinical info extraction	Llama-3.1-70B-Instruct (teacher), Llama-3.1-8B-Instruct (student)	Small models achieved high accuracy with reduced computational cost	[51]
		Propose PSOPL for automatic prompt design in medical info extraction	Alpaca-7B, GPT-J-6B, GLM-4	PSOPL improves efficiency in medical info extraction	[54]
		Evaluate Russell-GPT 1.0 for discharge summary generation	Russell-GPT 1.0 (Claude 2.1)	Saves time, improves accuracy, but requires physician validation	[52]
		Evaluate LLMs for structured data extraction from pathology reports	GPT-4, Llama2, Qwen2.5	Open-source LLMs perform similarly to GPT-4 in data extraction	[55]
		Evaluate LLMs for extracting structured data from unstructured MRI reports	Llama3.3, DeepSeek-R1, Phi4, Gemma-2	DeepSeek-R1-Llama3.3 achieved the highest accuracy in extraction	[56]
LLMs for clinical review generation	Exploring the application of LLMs in generating clinical reviews, reports, and other documents	Evaluate the effectiveness of LLMs in generating clinical reviews	A variety of large language models	Clinical reviews generated were accurate but lacked completeness and fluency	[53]

Notes: PSOPL: Particle swarm optimization-based prompt using a large language model. MRI: Magnetic Resonance Imaging.

2.5. Patient Interaction and Public Education

The application of LLMs in patient interaction and health education is rapidly expanding, encompassing various aspects, including online symptom assessment, disease information retrieval, emotional support, and the dissemination of health knowledge (see Table 4). In terms of patient interaction, LLMs can assist doctors in communicating with patients more efficiently and accurately, enhancing the completeness and empathy of the disease assessment. In public education, LLMs offer higher readability and coverage for the generation of medical science popularization and patient education materials. Existing research not only focuses on model accuracy and comprehensibility but also considers multiple evaluation dimensions such as language environment differences, patient trust, and adoption willingness, providing empirical support for the implementation of LLMs in healthcare services.

Table 4. LLM's application in patient interaction and public education.

Task	Characteristics	Object	Foundation Model	Result	Paper
LLMs in Patient Interaction	Focusing on the application of LLMs in patient interaction, improving communication between doctors and patients and symptom assessment	Online symptom assessment accuracy	A variety of large language models	Accuracy SSA: (11.5–90.0%); LLMs: (57.8–76.0%)	[57]
		Cardiovascular disease query performance	BARD, ChatGPT-3.5, ChatGPT-4.0, ERNIE	ChatGPT-4.0 is most accurate in English; ERNIE in Chinese	[58]
		Empathy in cancer-related responses	Claude V1, Claude V2, Claude V2 with CoT	AI models (Claude V2 with CoT) rated more empathetic than physicians	[59]
		Doctor-patient communication	ChatGPT 3.5, Gemini Pro, Co-Pilot	ChatGPT more accurate than Romanian Patient's Guide	[60]
		Evaluation of Clinical LLMs in Patient Interactions	GPT-4, GPT-3.5, Mistral, LLaMA-2-7b	Limited performance in diagnostic accuracy and history-taking	[61]
LLMs in Public Education	Focusing on the application of LLMs in health education, enhancing the generation of science popularization materials and improving the quality of patient education	Patient Education in Spinal Surgery	ChatGPT-4o, ChatGPT-o3 mini, DeepSeek-V3, DeepSeek-R1	DeepSeek-R1 most readable; overall quality fair	[62]
		Patient Education on Gonarthrosis & TKA	ChatGPT, DeepSeek	ChatGPT more accurate; DeepSeek more readable	[63]
		Trust and Adoption of DeepSeek for Health	DeepSeek	Greater trust and ease increased adoption; perceived risk reduced it	[64]
		Evaluation of AI Tools in ACL Surgery Education	ChatGPT-4o, DeepSeek-R1	GPT-4o more comprehensive; R1 clearer	[65]
		Effectiveness of LLMs vs. Search Engines in Health	LLMs (GPT-4, Llama3, MedLlama3), SEs (Google, Bing)	LLMs more accurate; RAG boosts smaller models	[66]
		Outpatient reception using nurse–LLM collaboration	Nurse-only vs Nurse+SSPEC (site-specific LLM chatbot)	Nurse+SSPEC better than nurse-only: higher satisfaction; fewer repeats; better empathy	[67]

Notes: TKA: Total knee arthroplasty. ACL: Anterior cruciate ligament.

Patient Interaction. The accuracy of online symptom self-assessment ranges from 57.8% to 76.0%, significantly higher than the lower limits of some traditional symptom assessment algorithms (11.5–90.0%) [57]. In the cardiovascular disease Q&A, ChatGPT-4.0 performed most accurately in English contexts, while ERNIE performed best in Chinese settings [58]. Empathy evaluations in cancer-related communication indicated that Claude V2, combined with CoT, was considered more empathetic [59]. In daily doctor-patient communication, ChatGPT outperformed Romanian patient guides [60]. However, in clinical interactions such as diagnostic accuracy and medical history collection, models like GPT-4/3.5, Mistral, and LLaMA-2-7b still exhibit certain limitations [61].

Public Education. The readability of patient education materials for spine surgery was best with DeepSeek-R1, but the overall quality was average [62]. In knee osteoarthritis and total knee arthroplasty (TKA) education, ChatGPT provided more accurate content, while DeepSeek was more readable [63]. In adoption research in the health field, higher trust levels corresponded to greater willingness to use DeepSeek, while perceived risks reduced adoption rates [64]. A comparison of anterior cruciate ligament (ACL) surgery education showed that GPT-4.0 had more comprehensive content coverage, while R1 was clearer in expression [65]. In general health information retrieval, LLMs have consistently outperformed traditional search engines, and the introduction of RAG technology has significantly enhanced the performance of smaller models [66]. In real-world outpatient settings, the "nurse + scenario-based LLM chatbot (SSPEC)" model significantly outperformed "nurse only", with improvements in patient satisfaction, repeated Q&A, empathy, and text completeness and readability [67].

LLMs in patient interaction and public education offer multiple advantages, including enhanced information accuracy, improved readability, emotional support, and an interactive experience. However, there are still shortcomings in clinical diagnostic interactions, cross-language consistency, and trust building. Future research should focus on deeper integration with real medical workflows, multilingual optimization, and localization adaptation, while also reducing hallucination risks through RAG, controllable generation, and fact-checking mechanisms.

2.6. Medical Research Assistance

The role of LLMs in accelerating and enhancing biomedical research is becoming increasingly prominent, providing comprehensive support throughout the data analysis, knowledge discovery, drug development, and public health research processes. Based on existing research (see Table 5), their applications can be primarily divided into three areas: biomedical data analysis and knowledge discovery, drug development and precision medicine, and support for clinical and public health research.

Table 5. LLM's application in medical research assistance.

Task	Characteristics	Object	Foundation Model	Result	Paper
Biomedical data analysis and knowledge discovery	Multi-omics and molecular data analysis, supporting basic and disease research	Genome-wide analysis of Drosophila behavior genes	GPT-3.5 + CoT	Identified 758 regulatory genes; validated novel factors (Mre11, NELF-B); low FP rate (7%)	[68]
		Discovering CRISPR-Cas system with self-processing pre-crRNA capability	CHOOSER (ESM-2)	Discovered 3477 systems; found 11 Cas λ	[69]
		Cancer proteomics analysis	DrBioRight 2.0	Supports natural language analysis and visualization; better multi-modal access	[70]
		Improve RNA structure prediction	RiNALMo	Achieved SOTA on multiple tasks, strong unseen-family generalization	[71]
		Evaluation of large language models for discovery of gene set function	GPT-4, Gemini Pro, Llama2	GPT-4 finds gene set functions accurately with low false positives	[72]
		Build simple single-cell embeddings from ChatGPT gene embeddings	GPT-3.5	Matches or outperforms large-scale pretrained models on classification tasks	[73]
Drug development and precision medicine	Multi-source information fusion, aiding drug prediction and personalized treatment	Boost efficiency and accuracy of patient–trial matching	CancerGPT	>90% recall with minimal retrieval; 87.3% accuracy; +43.8% ranking; 42.6% less screening time	[74]
		Unified framework for drug synergy prediction	BAITSAO	Best in 3/4 metrics; robust; identified key genes and targets	[75]
		Few/zero-shot drug pair synergy prediction in rare cancers	TrialGPT	Top accuracy in rare tissues without in-distribution data; strong zero/few-shot performance; valid reasoning	[76]
Support for clinical and public health research	Focusing on the extraction, analysis, and prediction of clinical and public health data, with an emphasis on the accuracy, interpretability, and timeliness of the results	Propose a teacher–GenAI collaboration framework to enhance learning skills	ChatGPT	Case studies from Australia, USA, and China show improved critical thinking, collaboration, and communication skills	[77]
		Identify Covid-19 transmission contexts from open-ended survey text	CamemBERT	Achieved 75% accuracy ; discovered new contexts beyond predefined options	[78]
		Automate case adjudication for phenotype algorithms	GPT-4, GPT-3.5, Llama-2, SDL2	LLMs matched human sensitivity/specificity, enabled large-scale PPV & sensitivity estimation, reduced manual workload	[79]
		Use LLMs for data extraction and risk-of-bias assessment in CAM RCTs	Claude-3.5-sonnet, Moonshot-v1-128k	LLM-only ≥95% accuracy; LLM-assisted ≥97%; 83–94% faster than conventional methods; high accuracy across domains	[80]
		Reformulate disease forecasting as text reasoning for multi-modal data integration	PandemicLLM	Outperformed existing models in 19-month COVID-19 forecasting across 50 U.S. states	[81]

Notes: CRISPR-Cas: Clustered regularly interspaced short palindromic repeats-CRISPR-associated. SOTA: State of the Art. CAM: Complementary and alternative medicine. RCT: Randomized controlled trials. PPV: Positive predictive value. BAITSAO: A scalable unified model for drug synergy prediction. CamemBERT: A RoBERTa-based French pre-trained language model.

Firstly, biomedical data analysis and knowledge discovery utilize LLMs to process and mine large-scale, multi-omics, and molecular-level data, thereby supporting basic and disease research. For example, GPT-3.5, combined with Chain-of-Thought (CoT) reasoning, successfully identified 758 regulatory genes in fruit fly behavioral genomics research, verifying new factors such as Mre11 and NELF-B, with a false positive rate of only 7% [68]. CHOOSER (ESM-2) discovered 3,477 new CRISPR-Cas systems, including 11 previously unknown Cas λ [69]. DrBioRight 2.0 excelled in natural language analysis and visualization of multi-modal cancer proteomics [70], while RiNALMo achieved SOTA performance in multiple RNA structure prediction tasks and demonstrated strong generalization abilities [71]. Additionally, GPT-4 demonstrated high accuracy and low false positive rates in gene set functional discovery [72] and single-cell embeddings built with GPT-3.5, which outperformed large-scale pre-trained models in classification tasks [73].

Secondly, drug development and precision medicine rely on the multi-source information integration capabilities of LLMs to enhance the accuracy and efficiency of drug prediction and personalized treatment. CancerGPT achieved a recall rate exceeding 90% and an accuracy rate of 87.3% in matching patients with clinical trials, significantly reducing screening time by 42.6% [74]. BAITSAO ranked first in three out of four drug synergy prediction metrics, simultaneously identifying key genes and targets [75]. TrialGPT achieved the best accuracy in rare cancer scenarios with few-shot and zero-shot conditions, demonstrating practical reasoning capabilities [76].

Lastly, clinical and public health research support focuses on the extraction, analysis, and prediction of clinical and public health data, highlighting the accuracy, interpretability, and timeliness of results. For example, ChatGPT improved critical thinking and communication collaboration skills in multi-country case studies within a teacher collaboration framework [77]. CamemBERT identified COVID-19 transmission scenarios with 75% accuracy

(91% after high-confidence filtering) and discovered scenario types beyond existing options [78]. Models like GPT-4, GPT-3.5, and Llama-2 matched human sensitivity and specificity in automatic phenotype determination, reducing manual workload [79]. Claude-3.5-sonnet and Moonshot-v1-128k achieved 95–97% accuracy in extracting complementary alternative medicine RCT data and bias risk assessment, with a speed improvement of 83–94% [80]. Additionally, PandemicLLM restructured infectious disease prediction into a text reasoning problem, outperforming existing models in predicting the COVID-19 pandemic across 50 U.S. states over 19 months [81].

Overall, LLMs are reshaping the model of research assistance in medicine, providing more efficient and intelligent research tools for fields such as biomedical science, drug development, and public health by connecting complex data sources and enhancing the depth and timeliness of analysis.

3. Discussions

Existing research indicates that LLMs have made substantial advancements in the medical field, encompassing various important areas, including clinical diagnosis and support, medical education and training, clinical document processing, patient interaction and public education, and medical research assistance. A consolidated overview of the major application categories, typical tasks, representative models is presented in Table 6, providing a structured summary of how LLMs are being applied across these domains. In clinical diagnosis and support, models are used for symptom interpretation, differential diagnosis, and triage, and are gradually incorporating multimodal information such as imaging to improve decision-making quality. In medical education and training, LLMs assist with exam preparation, clinical case generation, and personalized teaching, and are exploring automatic grading and objective evaluation. In the field of clinical document processing, LLMs demonstrate significant potential in simplifying and summarizing clinical documents, extracting information, and structuring data. In patient interaction and public education, LLMs help improve communication between doctors and patients, making it more efficient and accurate, and enhancing patient involvement in disease diagnosis and treatment decisions. In medical research assistance, LLMs provide important support in biomedical data analysis, drug development, and precision medicine. Overall, LLMs show great potential in diagnosis, medical education, clinical documentation, and patient education. However, they also reveal common challenges related to reliability, interpretability, privacy, and fairness, as well as engineering challenges in cross-institutional generalization and process integration.

Table 6. Overall summary of representative LLM applications in medicine across major task categories

Application Category	Subtasks/Scenarios	Representative Models	Data Sources
Clinical diagnosis and decision support	Symptom assessment, diagnostic reasoning, acute event prediction, decision support	GPT-4, Llama, DeepSeek, Med-PaLM 2	EHRs, clinical notes, simulated clinical cases
Medical education and training	Medical student training, specialty education, automated assessment	ChatGPT, DeepSeek, Llama, Claude	Examination datasets, OSCE scenarios, specialty training materials
Clinical documentation processing	Text simplification, summarization, information extraction, report generation	GPT-4, GatorTron, Llama	EHRs, radiology reports, discharge summaries
Patient interaction and public health education	Patient–LLM conversations, triage, health education	ChatGPT, PatientGPT, Llama, DeepSeek, Claude	Consumer health information materials, QA datasets
Biomedical and clinical research assistance	Knowledge discovery, multi-omics analysis, drug development, public health research	Llama, ChatGPT, Llama	Literature databases, multi-omics datasets

Notes: OSCE: Objective structured clinical examination.

3.1. Technical Roadmap and Foundational Research for Medical LLMs

From a technical trend perspective, research on medical LLMs is shifting from a focus on general conversational abilities to a combination of domain fine-tuning, retrieval augmentation, and multimodal integration, with accelerated progress in localization deployment and compliance [82–85]. Meanwhile, model evaluation methods are gradually moving away from dependence on offline question banks, shifting towards comprehensive evaluations that combine clinical contexts and physician involvement, which better reflect the real-world value and risks of the applications [34,86,87]. In this trend, the development framework for medical LLMs can be summarized as five core elements: data and knowledge construction, model training, multilingual and multimodal capabilities, security and fairness governance, and evaluation and clinical integration [88–91]. These elements support each other, providing a common technical foundation for various application scenarios and offering a clear developmental trajectory for the model's usability, reliability, and trustworthiness.

Data and knowledge construction form the basis of the model's capabilities. The medical domain requires high-quality data support, including clinical notes, radiology and pathology reports, discharge summaries, as well as textbooks, guidelines, and research literature. Recent studies show that through de-identification, standardization, and ontology

alignment, the accuracy of information extraction and question answering can be significantly improved [82,83,88]. Continuous updates and source traceability help mitigate the issue of outdated knowledge, ensuring that the model reflects the latest medical evidence.

The training methods for medical LLMs mainly include pre-training, fine-tuning, and prompt learning. In the pre-training method, the model is typically trained on large-scale medical corpora [82,83,88]. Through pre-training, the model learns the language structure, terminology, and relationships between concepts in the medical field, laying a solid foundation for subsequent tasks. The objectives of pre-training typically include tasks such as Masked Language Modeling and Next Token Prediction, which enable the model to understand medical terminology and concepts, thereby providing semantic understanding and reasoning capabilities for solving various medical tasks [92]. Next, fine-tuning adapts the general LLM to the specific needs of the medical field. Fine-tuning usually involves high-quality domain-specific datasets, such as doctor-patient dialogues, medical question-answering, and clinical literature [89,93,94]. Standard fine-tuning methods include Supervised Fine-Tuning (SFT), Instruction Fine-Tuning (IFT), and Parameter-Efficient Fine-Tuning (PEFT). Supervised Fine-Tuning optimizes the model's performance in understanding and generating medical texts by introducing domain-specific data; Instruction Fine-Tuning improves task performance by providing task-specific instructions to the model; Parameter-Efficient Fine-Tuning enhances model performance by optimizing only a small number of parameters, making it suitable for scenarios with limited computational resources [88]. These fine-tuning methods enable the model to be precisely adjusted according to the specific needs of medical tasks, allowing it to perform more effectively in practical applications. Prompt learning offers a more flexible and efficient approach to training medical LLMs. Through prompt learning, the model can reason about tasks based on the given prompt without requiring large-scale training or fine-tuning [95,96]. Standard prompt learning methods include In-context Learning (ICL), Chain-of-Thought Prompting (CoT), and RAG. In-context learning provides a few task examples, enabling the model to understand and complete the task quickly. Chain-of-Thought Prompting guides the model to generate reasoning steps, thereby improving transparency and accuracy when handling complex problems [97]. Retrieval-Augmented Generation combines external medical knowledge bases to expand the model's knowledge, prevent information loss, and improve the accuracy and timeliness of generated content [97,98]. The most significant advantage of prompt learning is that it does not require large amounts of data for fine-tuning; through well-designed prompt inputs, the model can quickly adapt to different medical task requirements and enhance its reasoning ability. Through these three methods—pre-training, fine-tuning, and prompt learning—medical LLMs demonstrate strong capabilities across various medical tasks, thus driving the application and development of medical artificial intelligence technologies. Multilingual and multimodal capabilities further extend the model's application boundaries.

In education and patient interactions, cross-lingual scenarios raise new demands for readability and terminological consistency [89]; in diagnosis and document generation, the integration of images and text enhances multimodal understanding and report generation capabilities [90]. The continuous progress in multilingual adaptation and multimodal fusion enables medical LLMs to cover a broader range of populations and tasks. Security and fairness governance are core safeguards for clinical applications. Studies have revealed differences in empathy and reliability during patient interactions, as well as biases across different populations [99–102]. Therefore, it is necessary to establish a full-link risk governance mechanism during training and inference, including data supply chain security, bias auditing, risk interception, and manual takeover, to ensure that the model's output meets ethical and clinical safety standards [103–105].

Evaluation and clinical integration determine whether the model can be practically implemented. Existing studies demonstrate significant performance differences between models in specialized tasks and complex cases [34,36,86], underscoring the need for context-specific benchmarks that are closer to clinical practice and mixed evaluations involving doctor participation [87]. Additionally, accumulating evidence through prospective pilots and quasi-randomized controlled trials, and positioning the model as a clinical "co-pilot" rather than an autonomous decision-maker, is a feasible path for integrating medical LLMs into medical workflows [106,107]. In summary, medical LLMs have gradually developed a systematic framework encompassing data construction, model training, multilingual/multimodal adaptation, security governance, and evaluation integration. This framework provides a unified technical foundation for different application scenarios and lays the groundwork for the usability and credibility of medical artificial intelligence.

3.2. Challenges

As the application of medical LLMs in healthcare continues to grow, they have demonstrated tremendous potential. However, they still face a series of challenges that affect their practical use in clinical settings.

Data limitations. The performance of medical large language models (LLMs) relies heavily on high-quality

training data. However, medical data is often constrained by multiple factors, including strict privacy protection, professional requirements, and limited accessibility. Such data is typically highly specialized and may exhibit strong regional, linguistic, and demographic variability. For example, due to the lack of public sharing, specialized datasets for certain diseases remain extremely scarce, resulting in insufficient training samples for model development.

In addition, some types of medical data such as doctor–patient dialogues, EHRs, radiology images, and corresponding reports—are unstructured and subject to stringent privacy regulations. This challenge is closely tied to the earlier discussion on clinical documentation applications (clinical documentation and clinical note processing), as these tasks require large amounts of real clinical text for LLMs to learn domain-specific medical language and expression patterns. However, due to privacy regulations (e.g., Health Insurance Portability and Accountability Act (HIPAA) and General Data Protection Regulation (GDPR)), hospitals rarely release raw clinical notes; even de-identified EHRs may still contain residual identifiable details that require further sanitization. Recent studies have shown that traditional de-identification or anonymization alone is insufficient to fully meet privacy and data-sharing requirements [108, 109].

Because high-quality, privacy-compliant clinical data is both scarce and imbalanced, models may become biased toward certain disease types or specific patient populations during training, ultimately reducing their generalizability in real clinical settings. Numerous studies emphasize that to advance the deployment of medical LLMs (and AI-driven medical models more broadly) in real-world clinical environments, it is essential to adopt privacy-preserving data-sharing and model-training mechanisms—such as federated learning, differential privacy, or synthetic data generation—to balance data security and model performance [110].

Hallucination. Medical LLMs may produce hallucinations when generating medical information, meaning they generate answers that seem reasonable but are actually incorrect. This phenomenon is particularly hazardous in high-risk tasks, such as medical diagnosis and treatment recommendations. The model's reasoning capabilities are based solely on statistical patterns and contextual information, lacking real clinical reasoning and evidence support. As a result, when faced with complex or uncommon clinical situations, it may provide incorrect advice or inaccurate predictions. Furthermore, medical tasks often require exact information, and any "plausible but incorrect" output can hurt patient health. Reducing hallucinations and improving the reasoning accuracy of LLMs is an urgent issue that needs to be addressed.

Lack of evaluation benchmarks and metrics. Currently, there is no unified evaluation standard for medical LLMs, particularly in assessing complex clinical tasks such as diagnostic decision-making and treatment recommendations. There is a lack of evaluation benchmarks that effectively reflect the performance of clinical tasks. Existing evaluations primarily focus on standardized test sets and benchmark tasks, overlooking the complexities of real-world clinical environments. The diversity and complexity of medical tasks require the evaluation of various model capabilities, including reasoning depth, knowledge coverage, and contextual adaptability. Although existing medical LLMs have achieved good results on specific standardized datasets, integrating standardized offline evaluation methods with clinical real-world needs to assess model performance in actual clinical settings remains an unresolved issue. Future evaluation standards should comprehensively cover multidimensional model performance, especially in specialized tasks, complex cases, and patient interactions. Establishing evaluation benchmarks that align more closely with clinical needs and ensuring they provide stable and practical support in real healthcare environments are necessary conditions for the successful implementation of medical LLMs.

Insufficient multimodal capabilities. As the demand for cross-modal data analysis in the medical field continues to grow, the multimodal fusion of medical LLMs has become an urgent technical challenge. Medical data encompasses not only traditional text data, such as clinical records and doctor-patient dialogues, but also various types of data, including medical images, genomic data, and laboratory test results. These different types of information are critical in clinical decision-making. Integrating heterogeneous data effectively so that the model can fully leverage all types of data when handling complex medical tasks is one of the key directions for current technological development. Multimodal fusion requires the model to understand and process different types of data, reasonably integrating them. For example, combining imaging data and text data helps the model make more accurate disease diagnoses and treatment recommendations. This fusion can not only improve diagnostic accuracy but also provide more comprehensive and intuitive support when generating medical reports. Additionally, multimodal models require cross-modal understanding and reasoning capabilities, enabling them to recognize and associate information from different data sources to make more precise and clinically valuable judgments. However, multimodal data in the medical field typically has varying formats, resolutions, and complexities. How to effectively integrate and represent these data is one of the main challenges currently faced by multimodal LLMs.

Regulatory challenges. As LLMs continue to develop in the medical field, ensuring that these models meet ethical standards and regulatory requirements has become an increasingly urgent issue. Medical LLMs, when

providing diagnostic advice, treatment plans, or drug recommendations to doctors, may have an impact on patient health. Therefore, ensuring the reliability and compliance of the models is crucial. The regulatory framework needs to cover all stages of model development, training, deployment, and application, particularly in areas such as data privacy, algorithm transparency, and output interpretability. Regulatory bodies must ensure that LLMs operate in accordance with existing laws and regulations and are traceable, so that accountability can be established in the event of adverse incidents. However, most existing medical regulations are designed for traditional methods, and the rapid development of AI technology may make it challenging for the current legal framework to cover all use cases of LLMs fully.

3.3. Research Directions

Establish comprehensive evaluation benchmarks and metrics. Currently, LLMs perform excellently in medical tasks but lack a unified evaluation benchmark and metric system, which limits their widespread application in various tasks and clinical environments. Future research should focus on establishing evaluation frameworks tailored for clinical environments, particularly evaluating the performance of models in real clinical scenarios, such as diagnostic accuracy in complex cases and the effectiveness of multidisciplinary decision support. Additionally, specific task-oriented evaluation standards should be developed to ensure comparability between different models under a unified standard, especially in high-risk areas such as clinical diagnosis and disease management.

Actively develop multimodal LLMs. Data in the medical field encompasses not only textual information but also multimodal data, including medical images, genomic data, and pathology slides. Future research should enhance the ability of medical LLMs to process multimodal data, integrating images, text, genomic data, and other types of information to provide more comprehensive support for disease diagnosis, treatment planning, and personalized healthcare. For example, combining images and text can enhance the application of medical LLMs in fields like imaging diagnosis and pathology analysis, improving the model's cross-modal reasoning and decision-making abilities. Moreover, researchers should explore how to address the complex relationships between multimodal data, enabling the model to accurately understand the interactions among various data types and provide more precise and multidimensional support for medical decision-making.

Strengthen the model's capabilities in personalized and precision medicine. With the ongoing development of precision medicine, personalized treatment and the processing of patient-specific data have become important research directions in medicine. Medical LLMs should be able to provide more personalized treatment plans by analyzing individual patient information, including medical history, genomic data, lifestyle, and family history. Future research needs to focus on how to integrate big data technology with personalized medical needs, using medical LLMs to integrate patient information from different sources and provide precise decision support to clinicians.

Enhance attention to medical ethics and regulations. In practical applications, medical LLMs must adhere to strict ethical standards and legal regulations. Future research should not only focus on the model's technical performance but also strengthen discussions on its ethical compliance, particularly in areas such as patient privacy protection, data security, and bias elimination.

4. Conclusions

In recent years, LLMs have made substantial advancements in the medical field, demonstrating their immense potential in clinical diagnosis, medical education, patient interaction, and research assistance. From enhancing clinical decision-making and providing accurate diagnostic advice to healthcare professionals, to efficiently managing clinical documentation and improving patient communication, LLMs have proven to be invaluable tools in healthcare. Models like ChatGPT and DeepSeek have already been applied in real-world healthcare settings, showing that LLMs can meet the specific needs of healthcare professionals and patients.

However, despite these advancements, the successful integration of LLMs in the medical field still faces numerous challenges. These challenges include data limitations, model hallucinations, a lack of evaluation benchmarks and metrics, and regulatory issues concerning medical AI systems. Additionally, the integration of multimodal data, the enhancement of personalized healthcare capabilities, and the establishment of mechanisms to ensure ethical and legal compliance remain key research areas that require attention.

Future research should focus on enhancing the interpretability, reliability, and adaptability of LLMs in clinical settings, including the development of standardized evaluation benchmarks that accurately reflect the complexity of real-world medical tasks. Exploring more advanced multimodal models, which combine text, images, and other forms of medical data, will be crucial to enhancing the capabilities of medical AI. In conclusion, while the potential of LLMs in medicine is vast, continuous development and refinement are needed to address existing challenges.

Author Contributions

Z.C., Y.L. (Yiwei Lu) and Y.Z.: conceptualization, methodology, data curation, writing—original draft preparation, investigation; D.T., Y.L. (Yun Li) and F.L.: conceptualization, supervision, validation, writing—reviewing and editing. All authors have read and agreed to the published version of the Manuscript.

Funding

This research received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

Not applicable.

Conflicts of Interest

The authors declare no conflict of interest.

Use of AI and AI-Assisted Technologies

No AI tools were utilized for this paper.

References

1. Chang, Y.; Wang, X.; Wang, J.; et al. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.* **2024**, *15*, 1–45.
2. Huang, J.; Xu, Y.; Wang, Q.; et al. Foundation models and intelligent decision-making: Progress, challenges, and perspectives. *Innovation* **2025**, *6*, 6100948.
3. Asif, S.; Wenhui, Y.; ur Rehman, S.; et al. Advancements and prospects of machine learning in medical diagnostics: Unveiling the future of diagnostic precision. *Arch. Comput. Methods Eng.* **2025**, *32*, 853–883.
4. Wang, D.; Zhang, S. Large language models in medical and healthcare fields: Applications, advances, and challenges. *Artif. Intell. Rev.* **2024**, *57*, 299.
5. Li, X.; Zhao, L.; Zhang, L.; et al. Artificial general intelligence for medical imaging analysis. *IEEE Rev. Biomed. Eng.* **2024**, *18*, 113–129.
6. Guo, D.; Yang, D.; Zhang, H.; et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv* **2025**, arXiv:2501.12948.
7. Gibney, E. China's cheap, open AI model DeepSeek thrills scientists. *Nature* **2025**, *638*, 13–14.
8. Shen, T.; Li, Y.; Cao, Y.; et al. Rapid deployment of large language model DeepSeek in Chinese hospitals demands a regulatory response. *Nat. Med.* **2025**, *31*, 3233–3238..
9. Liu, M.; Okuhara, T.; Chang, X.; et al. Performance of ChatGPT across different versions in medical licensing examinations worldwide: systematic review and meta-analysis. *J. Med. Internet Res.* **2024**, *26*, e60807.
10. Katz, U.; Cohen, E.; Shachar, E.; et al. GPT versus resident physicians—A benchmark based on official board scores. *Nejm Ai* **2024**, *1*, AIdbp2300192.
11. Kanjee, Z.; Crowe, B.; Rodman, A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. *Jama* **2023**, *330*, 78–80.
12. Achiam, J.; Adler, S.; Agarwal, S.; et al. Gpt-4 technical report. *arXiv* **2023**, arXiv:2303.08774.
13. Guo, D.; Yang, D.; Zhang, H.; et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature* **2025**, *645*, 633–638.
14. Touvron, H.; Lavril, T.; Izacard, G.; et al. Llama: Open and efficient foundation language models. *arXiv* **2023**, arXiv:2302.13971.
15. Bai, Y.; Kadavath, S.; Kundu, S.; et al. Constitutional ai: Harmlessness from ai feedback. *arXiv* **2022**, arXiv:2212.08073.
16. Liu, A.; Feng, B.; Xue, B.; et al. Deepseek-v3 technical report. *arXiv* **2024**, arXiv:2412.19437.
17. Sandmann, S.; Riepenhausen, S.; Plagwitz, L.; et al. Systematic analysis of ChatGPT, Google search and Llama 2 for clinical decision support tasks. *Nat. Commun.* **2024**, *15*, 2050.
18. Singhal, K.; Tu, T.; Gottweis, J.; et al. Toward expert-level medical question answering with large language models. *Nat. Med.* **2025**, *31*, 943–950.

19. McDuff, D.; Schaekermann, M.; Tu, T.; et al. Towards accurate differential diagnosis with large language models. *Nature* **2025**, *642*, 451–457.
20. Williams, C.Y.; Miao, B.Y.; Kornblith, A.E.; et al. Evaluating the use of large language models to provide clinical recommendations in the Emergency Department. *Nat. Commun.* **2024**, *15*, 8236.
21. Wu, X.; Huang, Y.; He, Q. A large language model improves clinicians' diagnostic performance in complex critical illness cases. *Crit. Care* **2025**, *29*, 230.
22. Perlis, R.H.; Goldberg, J.F.; Ostacher, M.J.; et al. Clinical decision support for bipolar depression using large language models. *Neuropsychopharmacology* **2024**, *49*, 1412–1416.
23. Shashikumar, S.P.; Mohammadi, S.; Krishnamoorthy, R.; et al. Development and prospective implementation of a large language model based system for early sepsis prediction. *NPJ Digit. Med.* **2025**, *8*, 290.
24. Kaygisiz, Ö.F.; Teke, M.T. Can deepseek and ChatGPT be used in the diagnosis of oral pathologies? *BMC Oral Health* **2025**, *25*, 638.
25. Zhang, J.; Liu, J.; Guo, M.; et al. DeepSeek-assisted LI-RADS classification: AI-driven precision in hepatocellular carcinoma diagnosis. *Int. J. Surg.* **2025**, *111*, 5970–5979.
26. Leng, Y.; He, Y.; Amini, S.; et al. A GPT-4o-powered framework for identifying cognitive impairment stages in electronic health records. *NPJ Digit. Med.* **2025**, *8*, 401.
27. Zhou, J.; He, X.; Sun, L.; et al. Pre-trained multimodal large language model enhances dermatological diagnosis using SkinGPT-4. *Nat. Commun.* **2024**, *15*, 5649.
28. Oh, Y.; Park, S.; Byun, H.K.; et al. LLM-driven multimodal target volume contouring in radiation oncology. *Nat. Commun.* **2024**, *15*, 9186.
29. Li, J.; Guan, Z.; Wang, J.; et al. Integrated image-based deep learning and language models for primary diabetes care. *Nat. Med.* **2024**, *30*, 2886–2896.
30. Tran, M.; Schmidle, P.; Guo, R.R.; et al. Generating dermatopathology reports from gigapixel whole slide images with HistoGPT. *Nat. Commun.* **2025**, *16*, 4886.
31. Lu, M.Y.; Chen, B.; Williamson, D.F.; et al. A multimodal generative AI copilot for human pathology. *Nature* **2024**, *634*, 466–473.
32. Wu, Y.; Liu, Y.; Yang, Y.; et al. A concept-based interpretable model for the diagnosis of choroid neoplasias using multimodal data. *Nat. Commun.* **2025**, *16*, 3504.
33. Hassanein, F.E.; El Barbary, A.; Hussein, R.R.; et al. Diagnostic Performance of ChatGPT-4o and DeepSeek-3 Differential Diagnosis of Complex Oral Lesions: A Multimodal Imaging and Case Difficulty Analysis. *Oral Dis.* **2025**. https://doi.org/10.1111/odi.70007.
34. Tordjman, M.; Liu, Z.; Yuce, M.; et al. Comparative benchmarking of the DeepSeek large language model on medical tasks and clinical reasoning. *Nat. Med.* **2025**, *31*, 2550–2555.
35. Chan, L.; Xu, X.; Lv, K. DeepSeek-R1 and GPT-4 are comparable in a complex diagnostic challenge: a historical control study. *Int. J. Surg.* **2025**, *111*, 4056–4059.
36. Shean, R.; Shah, T.; Pandiarajan, A.; et al. A comparative analysis of DeepSeek R1, DeepSeek-R1-Lite, OpenAi o1 Pro, and Grok 3 performance on ophthalmology board-style questions. *Sci. Rep.* **2025**, *15*, 23101.
37. Wu, J.; Wang, Z.; Qin, Y. Performance of DeepSeek-R1 and ChatGPT-4o on the Chinese national medical licensing examination: A vomparative atudy. *J. Med. Syst.* **2025**, *49*, 74.
38. Shang, L.; Li, R.; Xue, M.; et al. Evaluating the application of ChatGPT in China's residency training education: An exploratory study. *Med. Teach.* **2025**, *47*, 858–864.
39. Luo, P.W.; Liu, J.W.; Xie, X.; et al. DeepSeek vs ChatGPT: A comparison study of their performance in answering prostate cancer radiotherapy questions in multiple languages. *Am. J. Clin. Exp. Urol.* **2025**, *13*, 176.
40. Mahedia, M.; Rohrich, R.N.; Sadiq, K.O.; et al. Exploring the utility of chatgpt in cleft lip repair education. *J. Clin. Med.* **2025**, *14*, 993.
41. Rao, A.S.; Kim, J.; Mu, A.; et al. Synthetic medical education in dermatology leveraging generative artificial intelligence. *NPJ Digit. Med.* **2025**, *8*, 247.
42. Wada, A.; Tanaka, Y.; Nishizawa, M.; et al. Retrieval-augmented generation elevates local LLM quality in radiology contrast media consultation. *NPJ Digit. Med.* **2025**, *8*, 395.
43. Grévisse, C. LLM-based automatic short answer grading in undergraduate medical education. *BMC Med. Educ.* **2024**, *24*, 1060.
44. Hersh, W.; Fultz Hollis, K. Results and implications for generative AI in a large introductory biomedical and health informatics course. *NPJ Digit. Med.* **2024**, *7*, 247.
45. Mizumoto, A.; Eguchi, M. Exploring the potential of using an AI language model for automated essay scoring. *Res. Methods Appl. Linguist.* **2023**, *2*, 100050.
46. Jeblick, K.; Schachtner, B.; Dexl, J.; et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. *Eur. Radiol.* **2024**, *34*, 2817–2825.
47. Rust, P.; Frings, J.; Meister, S.; et al. Evaluation of a large language model to simplify discharge summaries and provide cardiological lifestyle recommendations. *Commun. Med.* **2025**, *5*, 208.

48. Li, C.Y.; Chang, K.J.; Yang, C.F.; et al. Towards a holistic framework for multimodal LLM in 3D brain CT radiology report generation. *Nat. Commun.* **2025**, *16*, 2258.
49. Yang, X.; Chen, A.; PourNejatian, N.; et al. A large language model for electronic health records. *NPJ Digit. Med.* **2022**, *5*, 194.
50. Hein, D.; Christie, A.; Holcomb, M.; et al. Iterative refinement and goal articulation to optimize large language models for clinical information extraction. *NPJ Digit. Med.* **2025**, *8*, 301.
51. Woo, E.G.; Burkhart, M.C.; Alsentzer, E.; et al. Synthetic data distillation enables the extraction of clinical information at scale. *NPJ Digit. Med.* **2025**, *8*, 267.
52. Chua, C.E.; Clara, N.L.Y.; Furqan, M.S.; et al. Integration of customised LLM for discharge summary generation in real-world clinical settings: a pilot study on RUSSELL GPT. *Lancet Reg. Health West. Pac.* **2024**, *51*, 101211.
53. Luo, Z.; Qiao, Y.; Xu, X.; et al. Cross sectional pilot study on clinical review generation using large language models. *NPJ Digit. Med.* **2025**, *8*, 170.
54. Zhang, T.; Ma, L.; Cheng, S.; et al. Automatic prompt design via particle swarm optimization driven LLM for efficient medical information extraction. *Swarm Evol. Comput.* **2025**, *95*, 101922.
55. Grothey, B.; Odenkirchen, J.; Brkic, A.; et al. Comprehensive testing of large language models for extraction of structured data in pathology. *Commun. Med.* **2025**, *5*, 96.
56. Di Palma, L.; Darvizeh, F.; Alì, M.; et al. Structured Transformation of Unstructured Prostate MRI Reports Using Large Language Models. *Tomography* **2025**, *11*, 69.
57. Kopka, M.; von Kalckreuth, N.; Feufel, M.A. Accuracy of online symptom assessment applications, large language models, and laypeople for self–triage decisions. *NPJ Digit. Med.* **2025**, *8*, 178.
58. Ji, H.; Wang, X.; Sia, C.H.; et al. Large language model comparisons between English and Chinese query performance for cardiovascular prevention. *Commun. Med.* **2025**, *5*, 177.
59. Chen, D.; Chauhan, K.; Parsa, R.; et al. Patient perceptions of empathy in physician and artificial intelligence chatbot responses to patient questions about cancer. *NPJ Digit. Med.* **2025**, *8*, 275.
60. Geantă, M.; Bădescu, D.; Chirca, N.; et al. The potential impact of large language models on doctor–patient communication: A case study in prostate cancer. *Healthcare* **2024**, *12*, 1548.
61. Johri, S.; Jeong, J.; Tran, B.A.; et al. An evaluation framework for clinical use of large language models in patient interaction tasks. *Nat. Med.* **2025**, *31*, 77–86.
62. Zhou, M.; Pan, Y.; Zhang, Y.; et al. Evaluating AI-generated patient education materials for spinal surgeries: Comparative analysis of readability and DISCERN quality across ChatGPT and deepseek models. *Int. J. Med. Inform.* **2025**, *198*, 105871.
63. Gurbuz, S.; Bahar, H.; Yavuz, U.; et al. Comparative Efficacy of ChatGPT and DeepSeek in Addressing Patient Queries on Gonarthrosis and Total Knee Arthroplasty. *Arthroplast. Today* **2025**, *33*, 101730.
64. Choudhury, A.; Shahsavar, Y.; Shamszare, H. User intent to use DeepSeek for health care purposes and their Ttrust in the large language model: Multinational survey study. *JMIR Hum. Factors* **2025**, *12*, e72867.
65. Gültekin, O.; Inoue, J.; Yilmaz, B.; et al. Evaluating deepResearch and deepThink in anterior cruciate ligament surgery patient education: ChatGPT-4o excels in comprehensiveness, DeepSeek R1 leads in clarity and readability of orthopaedic information. *Knee Surg. Sports Traumatol. Arthrosc.* **2025**, *33*, 3025–3031.
66. Fernández-Pichel, M.; Pichel, J.C.; Losada, D.E. Evaluating search engines and large language models for answering health questions. *NPJ Digit. Med.* **2025**, *8*, 153.
67. Wan, P.; Huang, Z.; Tang, W.; et al. Outpatient reception via collaboration between nurses and a large language model: a randomized controlled trial. *Nat. Med.* **2024**, *30*, 2878–2885.
68. Peng, D.; Zheng, L.; Liu, D.; et al. Large-language models facilitate discovery of the molecular signatures regulating sleep and activity. *Nat. Commun.* **2024**, *15*, 3685.
69. Li, W.; Jiang, X.; Wang, W.; et al. Discovering CRISPR-Cas system with self-processing pre-crRNA capability by foundation models. *Nat. Commun.* **2024**, *15*, 10024.
70. Liu, W.; Li, J.; Tang, Y.; et al. DrBioRight 2.0: an LLM-powered bioinformatics chatbot for large-scale cancer functional proteomics analysis. *Nat. Commun.* **2025**, *16*, 2256.
71. Penić, R.J.; Vlašić, T.; Huber, R.G.; et al. Rinalmo: General-purpose rna language models can generalize well on structure prediction tasks. *Nat. Commun.* **2025**, *16*, 5671.
72. Hu, M.; Alkhairy, S.; Lee, I.; et al. Evaluation of large language models for discovery of gene set function. *Nat. Methods* **2025**, *22*, 82–91.
73. Chen, Y.; Zou, J. Simple and effective embedding model for single-cell biology built from chatgpt. *Nat. Biomed. Eng.* **2025**, *9*, 483–493.
74. Li, T.; Shetty, S.; Kamath, A.; et al. CancerGPT for few shot drug pair synergy prediction using large pretrained language models. *NPJ Digit. Med.* **2024**, *7*, 40.
75. Liu, T.; Chu, T.; Luo, X.; et al. Building a unified model for drug synergy analysis powered by large language models. *Nat. Commun.* **2025**, *16*, 4537.
76. Jin, Q.; Wang, Z.; Floudas, C.S.; et al. Matching patients to clinical trials with large language models. *Nat. Commun.* **2024**, *15*, 9074.

77. Memon, T.D.; Kwan, P. A collaborative model for integrating Tteacher and genAI into future education. *TechTrends* **2025**, 1–15. https://doi.org/10.1007/s11528-025-01105-w.
78. Bizel-Bizellot, G.; Galmiche, S.; Lelandais, B.; et al. Extracting circumstances of Covid-19 transmission from free text with large language models. *Nat. Commun.* **2025**, *16*, 5836.
79. Schuemie, M.J.; Ostropolets, A.; Zhuk, A.; et al. Standardized patient profile review using large language models for case adjudication in observational research. *NPJ Digit. Med.* **2025**, *8*, 18.
80. Lai, H.; Liu, J.; Bai, C.; et al. Language models for data extraction and risk of bias assessment in complementary medicine. *NPJ Digit. Med.* **2025**, *8*, 74.
81. Du, H.; Zhao, Y.; Zhao, J.; et al. Advancing real-time infectious disease forecasting using large language models. *Nat. Comput. Sci.* **2025**, *5*, 467–480.
82. Wu, C.; Lin, W.; Zhang, X.; et al. PMC-LLaMA: toward building open-source language models for medicine. *J. Am. Med. Inform. Assoc.* **2024**, *31*, 1833–1843.
83. Peng, C.; Yang, X.; Chen, A.; et al. A study of generative large language model for medical research and healthcare. *NPJ Digit. Med.* **2023**, *6*, 210.
84. Zeng, D.; Qin, Y.; Sheng, B.; et al. DeepSeek's "Low-Cost" adoption across China's hospital systems: Too fast, too soon? *JAMA* **2025**, *333*, 1866–1869.
85. Deng, Z.; Ma, W.; Han, Q.L.; et al. Exploring DeepSeek: A Survey on advances, applications, challenges and future Directions. *IEEE/CAA J. Autom. Sin.* **2025**, *12*, 872–893.
86. Sandmann, S.; Hegselmann, S.; Fujarski, M.; et al. Benchmark evaluation of DeepSeek large language models in clinical decision-making. *Nat. Med.* **2025**, *31*, 2546–2549.
87. Tam, T.Y.C.; Sivarajkumar, S.; Kapoor, S.; et al. A framework for human evaluation of large language models in healthcare derived from literature review. *NPJ Digit. Med.* **2024**, *7*, 258.
88. Xie, Q.; Chen, Q.; Chen, A.; et al. Medical foundation large language models for comprehensive text analysis and beyond. *NPJ Digit. Med.* **2025**, *8*, 141.
89. Qiu, P.; Wu, C.; Zhang, X.; et al. Towards building multilingual language model for medicine. *Nat. Commun.* **2024**, *15*, 8384.
90. Perez-Garcia, F.; Sharma, H.; Bond-Taylor, S.; et al. Exploring scalable medical image encoders beyond text supervision. *Nat. Mach. Intell.* **2025**, *7*, 119–130.
91. Tran, M.; Balasooriya, C.; Jonnagaddala, J.; et al. Situating governance and regulatory concerns for generative artificial intelligence and large language models in medical education. *NPJ Digit. Med.* **2025**, *8*, 315.
92. Liu, F.; Zhou, H.; Gu, B.; et al. Application of large language models in medicine. *Nat. Rev. Bioeng.* **2025**, *3*, 445–464.
93. Griot, M.; Hemptinne, C.; Vanderdonckt, J.; et al. Large language models lack essential metacognition for reliable medical reasoning. *Nat. Commun.* **2025**, *16*, 642.
94. Kim, H.; Hwang, H.; Lee, J.; et al. Small language models learn enhanced reasoning skills from medical textbooks. *NPJ Digit. Med.* **2025**, *8*, 240.
95. Ferber, D.; Wiest, I.C.; Wölflein, G.; et al. GPT-4 for information retrieval and comparison of medical oncology guidelines. *Nejm Ai* **2024**, *1*, AIcs2300235.
96. Wang, L.; Chen, X.; Deng, X.; et al. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. *NPJ Digit. Med.* **2024**, *7*, 41.
97. Zhang, G.; Xu, Z.; Jin, Q.; et al. Leveraging long context in retrieval augmented language models for medical question answering. *NPJ Digit. Med.* **2025**, *8*, 239.
98. Wu, K.; Wu, E.; Wei, K.; et al. An automated framework for assessing how well LLMs cite relevant medical references. *Nat. Commun.* **2025**, *16*, 3615.
99. Bouguettaya, A.; Stuart, E.M.; Aboujaoude, E. Racial bias in AI-mediated psychiatric diagnosis and treatment: a qualitative comparison of four large language models. *NPJ Digit. Med.* **2025**, *8*, 332.
100. Omar, M.; Soffer, S.; Agbareia, R.; et al. Sociodemographic biases in medical decision making by large language models. *Nat. Med.* **2025**, *31*, 1873–1881.
101. Templin, T.; Fort, S.; Padmanabham, P.; et al. Framework for bias evaluation in large language models in healthcare settings. *NPJ Digit. Med.* **2025**, *8*, 414.
102. Chen, X.; Wang, T.; Zhou, J.; et al. Evaluating and mitigating bias in AI-based medical text generation. *Nat. Comput. Sci.* **2025**, *5*, 388–396.
103. Alber, D.A.; Yang, Z.; Alyakin, A.; et al. Medical large language models are vulnerable to data-poisoning attacks. *Nat. Med.* **2025**, *31*, 618–626.
104. Yang, J.; Xu, H.; Mirzoyan, S.; et al. Poisoning medical knowledge using large language models. *Nat. Mach. Intell.* **2024**, *6*, 1156–1168.
105. Pfohl, S.R.; Cole-Lewis, H.; Sayres, R.; et al. A toolbox for surfacing health equity harms and biases in large language models. *Nat. Med.* **2024**, *30*, 3590–3600.
106. Hager, P.; Jungmann, F.; Holland, R.; et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat. Med.* **2024**, *30*, 2613–2622.

107. Goh, E.; Bunning, B.; Khoong, E.C.; et al. Physician clinical decision modification and bias assessment in a randomized controlled trial of AI assistance. *Commun. Med.* **2025**, *5*, 59.
108. Fares, M.H.; Saad, A.M.S.E. Towards Privacy-Preserving Medical Imaging: Federated Learning with Differential Privacy and Secure Aggregation Using a Modified ResNet Architecture. *arXiv* **2024**, arXiv:2412.00687.
109. Zhao, H.; Sui, D.; Wang, Y.; et al. Privacy-Preserving Federated Learning Framework for Multi-Source Electronic Health Records Prognosis Prediction. *Sensors* **2025**, *25*, 2374.
110. Teo, Z.L.; Jin, L.; Liu, N.; et al. Federated machine learning in healthcare: A systematic review on clinical applications and technical architecture. *Cell Rep. Med.* **2024**, *5*, 101419.