information

MDPI

Review

Large Language Models in Medical Chatbots: Opportunities, Challenges, and the Need to Address AI Risks

James C. L. Chow [1,2,*] and Kay Li [3]

[1] Radiation Medicine Program, Princess Margaret Cancer Centre, University Health Network, Toronto, ON M5G 1X6, Canada
[2] Department of Radiation Oncology, University of Toronto, Toronto, ON M5T 1P5, Canada
[3] Department of English, University of Toronto, Toronto, ON M5R 0A3, Canada; kay.li@utoronto.ca
* Correspondence: james.chow@uhn.ca; Tel.: +1-416-946-4501

Abstract

Large language models (LLMs) are transforming the capabilities of medical chatbots by enabling more context-aware, human-like interactions. This review presents a comprehensive analysis of their applications, technical foundations, benefits, challenges, and future directions in healthcare. LLMs are increasingly used in patient-facing roles, such as symptom checking, health information delivery, and mental health support, as well as in clinician-facing applications, including documentation, decision support, and education. However, as a study from 2024 warns, there is a need to manage "extreme AI risks amid rapid progress". We examine transformer-based architectures, fine-tuning strategies, and evaluation benchmarks specific to medical domains to identify their potential to transfer and mitigate AI risks when using LLMs in medical chatbots. While LLMs offer advantages in scalability, personalization, and 24/7 accessibility, their deployment in healthcare also raises critical concerns. These include hallucinations (the generation of factually incorrect or misleading content by an AI model), algorithmic biases, privacy risks, and a lack of regulatory clarity. Ethical and legal challenges, such as accountability, explainability, and liability, remain unresolved. Importantly, this review integrates broader insights on AI safety, drawing attention to the systemic risks associated with rapid LLM deployment. As highlighted in recent policy research, including work on managing extreme AI risks, there is an urgent need for governance frameworks that extend beyond technical reliability to include societal oversight and long-term alignment. We advocate for responsible innovation and sustained collaboration among clinicians, developers, ethicists, and regulators to ensure that LLM-powered medical chatbots are deployed safely, equitably, and transparently within healthcare systems.

Keywords: large language models; medical chatbots; generative artificial intelligence; natural language processing; clinical decision support; healthcare AI governance; algorithmic bias; AI safety and risk management; patient–AI Interaction; ethical AI in medicine

check for updates

Academic Editor: Rodolfo Delmonte

Received: 3 June 2025
Revised: 23 June 2025
Accepted: 25 June 2025
Published: 27 June 2025

Citation: Chow, J.C.L.; Li, K. Large Language Models in Medical Chatbots: Opportunities, Challenges, and the Need to Address AI Risks. *Information* **2025**, *16*, 549. https://doi.org/10.3390/info16070549

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Chatbots have been a part of healthcare systems for several decades, initially emerging as simple, rule-based tools designed to simulate conversation [1,2]. One of the earliest examples, ELIZA [3], demonstrated the potential of scripted dialogue to mimic therapeutic interactions, although with a limited capacity for understanding user intent. Since then, medical chatbots have evolved to support a variety of functions, including patient triage,

health education, appointment scheduling, and medication reminders [4]. These systems have been particularly useful in improving access to basic healthcare information and reducing the administrative burden, especially in resource-constrained settings.

More recently, the field of natural language processing (NLP) [5] has advanced significantly with the introduction of transformer-based models, such as Bidirectional Encoder Representations from Transformers (BERTs) [6] and generative pretrained transformers (GPT) [7]. These models are capable of capturing complex linguistic patterns and generating human-like text, marking a shift from rule-based to generative approaches in conversational AI. Large language models (LLMs) pretrained on vast corpora of general and domain-specific text have demonstrated an ability to perform a wide range of language tasks with minimal supervision [8]. In the context of healthcare, this enables the development of chatbots that can interpret diverse patient queries, retain contextual information across interactions, and generate relevant and coherent medical responses [9]. Figure 1 shows the hierarchical relationships among key concepts in artificial intelligence (AI). AI encompasses machine learning, which includes NLP. Within NLP, there are specific models, like GPT, and broader categories, like LLMs. This structure highlights how these technologies are interconnected and build upon each other.

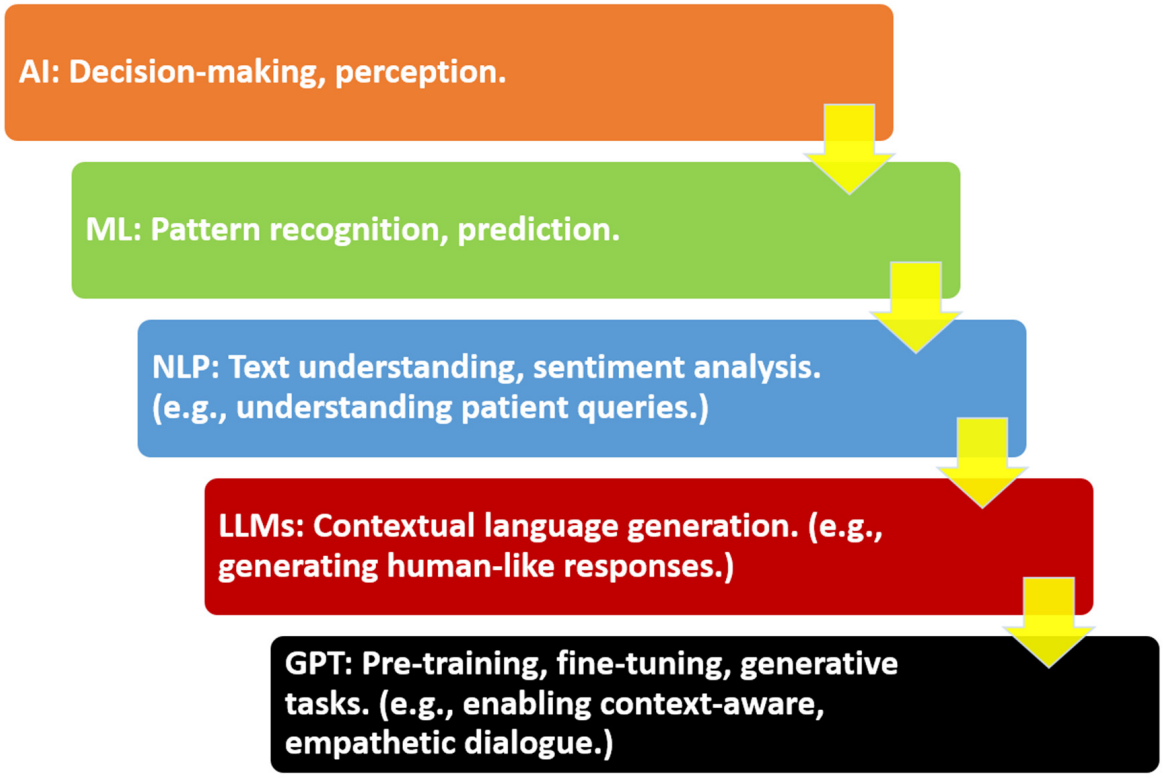

Figure 1. Block diagram of relationships among artificial intelligence (AI), machine learning (ML), natural language processing (NLP), large language models (LLMs), and generative pretrained transformer (GPT).

The use of LLMs in medical chatbots is growing rapidly, with platforms such as ChatGPT [10] and Med-PaLM [11] being evaluated for roles in patient communication, clinical support, and medical education. These applications raise important questions about the safety, reliability, and ethical implications of using generative artificial intelligence (GAI) in healthcare environments. While early studies suggest potential benefits, such as increased accessibility and efficiency, there are also concerns related to factual accuracy, the potential for bias, data privacy, and the lack of regulatory oversight [12].

Beyond domain-specific concerns, recent work has highlighted the broader societal and systemic risks associated with rapid advances in general-purpose AI technologies. As noted by Bengio, Hinton et al. [13], the pace of development in large-scale AI systems may outstrip the capacity of institutions to govern their deployment effectively, particularly in high-stakes fields, like healthcare. The potential for unintended consequences, the loss of human oversight, and misuse underscores the need for proactive governance, not only to ensure clinical accuracy and fairness, but also to address longer-term implications for safety, accountability, and public trust. These perspectives further emphasize the importance of implementing rigorous oversight and cross-sector collaboration in the deployment of LLMs in medicine.

Given the pace of development and the increasing integration of LLMs into healthcare tools, there is a need for a systematic review of their applications, technical foundations, limitations, and future directions. This paper aims to provide a comprehensive assessment of LLM-powered medical chatbots to support informed decision-making among researchers, developers, clinicians, and policymakers involved in the design and deployment of these technologies.

In recent years, there has been a marked shift in the design of medical chatbot systems from traditional rule-based frameworks to sophisticated generative models enabled by large language models (LLMs) [14]. Traditional systems relied on static decision trees and keyword matching, which limited their flexibility, contextual understanding, and ability to generalize across diverse patient inputs [15]. In contrast, LLMs, such as GPT and Med-PaLM, use transformer-based architectures to understand and generate natural language in a dynamic and context-sensitive manner. This transition represents a fundamental reconfiguration of the chatbot paradigm, enabling deeper conversational flow, multilingual capability, and personalization [16]. However, despite these advances, a key challenge remains: ensuring the adaptability of LLMs in low-resource medical environments. These settings often face constraints in data availability, computational infrastructure, and specialist expertise, which raise concerns about the equity and effectiveness of LLM-based solutions. There is a growing body of research exploring lightweight model architectures, federated learning (FL), and fine-tuning approaches tailored for resource-limited contexts [17]. A systematic comparison of traditional approaches and modern LLMs, especially in terms of their usability, scalability, and resilience under low-resource constraints, is, therefore, critical to inform inclusive deployment strategies.

This review is narrative in nature and is not intended to be a systematic review. Our aim was to synthesize and critically discuss current developments in the use of LLMs in medical chatbots across patient and clinician contexts, with an emphasis on AI's risk, governance, and ethical considerations. As such, the literature included was selected based on relevance and topicality, rather than through a structured database search with predefined inclusion/exclusion criteria.

2. Applications of LLMs in Medical Chatbots

LLMs have expanded the range of capabilities in medical chatbots by enabling dynamic, context-aware, and linguistically sophisticated interactions. Unlike the earlier systems that relied on predefined rules or decision trees, LLM-based chatbots can process natural language inputs, adapt to diverse user needs, and generate contextually relevant responses. Their applications span both patient-facing and clinician-facing settings, supporting tasks ranging from symptom assessment and education to clinical documentation and decision support.

2.1. Patient-Facing Applications

LLM-powered chatbots are being used to conduct real-time symptom assessments that more closely resemble clinical interviews compared to earlier, rule-based systems [18]. These models can interpret free-text inputs, clarify vague symptoms through follow-up questions, and suggest possible conditions based on the patient's responses. Unlike systems such as Ada and Babylon [19], which use structured algorithms and predefined symptom trees, LLMs can dynamically adjust their questioning strategy based on previous inputs and contextual cues [20]. This allows for more flexible and personalized interactions, though it also introduces variability and potential risks associated with inconsistent outputs or clinical inaccuracy.

Another key application area is the dissemination of health-related information. LLM-based chatbots can explain medical conditions, laboratory test results, imaging findings, and treatment options in layperson-friendly language [21]. These models can adapt explanations based on users' age, education level, and language preferences, thereby improving accessibility and comprehension. These chatbots have the potential to enhance patient engagement and support shared decision-making. However, ensuring that the information provided is accurate, evidence-based, and consistent with clinical guidelines remains a critical challenge [22].

LLMs are also being explored for applications in mental health, where conversational engagement and empathetic language are important [23]. Systems such as Woebot and Wysa [24] have demonstrated the feasibility of using AI to provide cognitive–behavioral support and stress management strategies. LLMs can enhance these capabilities by facilitating more natural, emotionally attuned interactions and employing techniques such as active listening, reflective responses, and the validation of emotional experiences. While these models are not intended to replace licensed mental health professionals, they may serve as accessible, low-barrier tools for preliminary support or as adjuncts to traditional therapy. Figure 2 shows four potential applications of LLMs in mental health care. These include enhancing the accuracy of screening and diagnosis by analyzing patient descriptions and integrating non-linguistic data (top-left); monitoring behavioral and physiological signals (e.g., heart rate, facial expressions, and voice tone) to detect early signs of mental health (top-right); tailoring assessment reports for various stakeholders, such as parents, teachers, and healthcare providers (bottom-left); and developing assessment chatbots to assist therapists in conducting structured interviews and making diagnostic decisions (bottom-right) [23].

2.2. Clinician-Facing Applications

In addition to patient-facing functions, LLMs are being evaluated for use in supporting clinical workflows. These applications are designed to reduce the administrative burden, enhance decision-making, and assist in the education of healthcare professionals [25]. By using the language generation and contextual reasoning capabilities of LLMs, developers aim to integrate these systems into clinical environments while maintaining accuracy, consistency, and alignment with medical standards.

LLMs are being tested as adjunct tools for clinical decision support, particularly in tasks such as formulating differential diagnoses and drafting clinical documentation [26]. Preliminary studies comparing GPT-4 to human clinicians suggest that, in some cases, LLMs can propose plausible diagnostic options based on structured or free-text clinical scenarios [27]. Although these models are not a substitute for clinician expertise, they may serve as a secondary reference or aid in clinical reasoning, especially in time-constrained settings. Moreover, LLMs have shown potential in generating structured summaries, such

as discharge summaries or SOAP (Subjective, Objective, Assessment, Plan) notes, based on inputs from electronic health records (EHRs) or dictated transcripts [28].

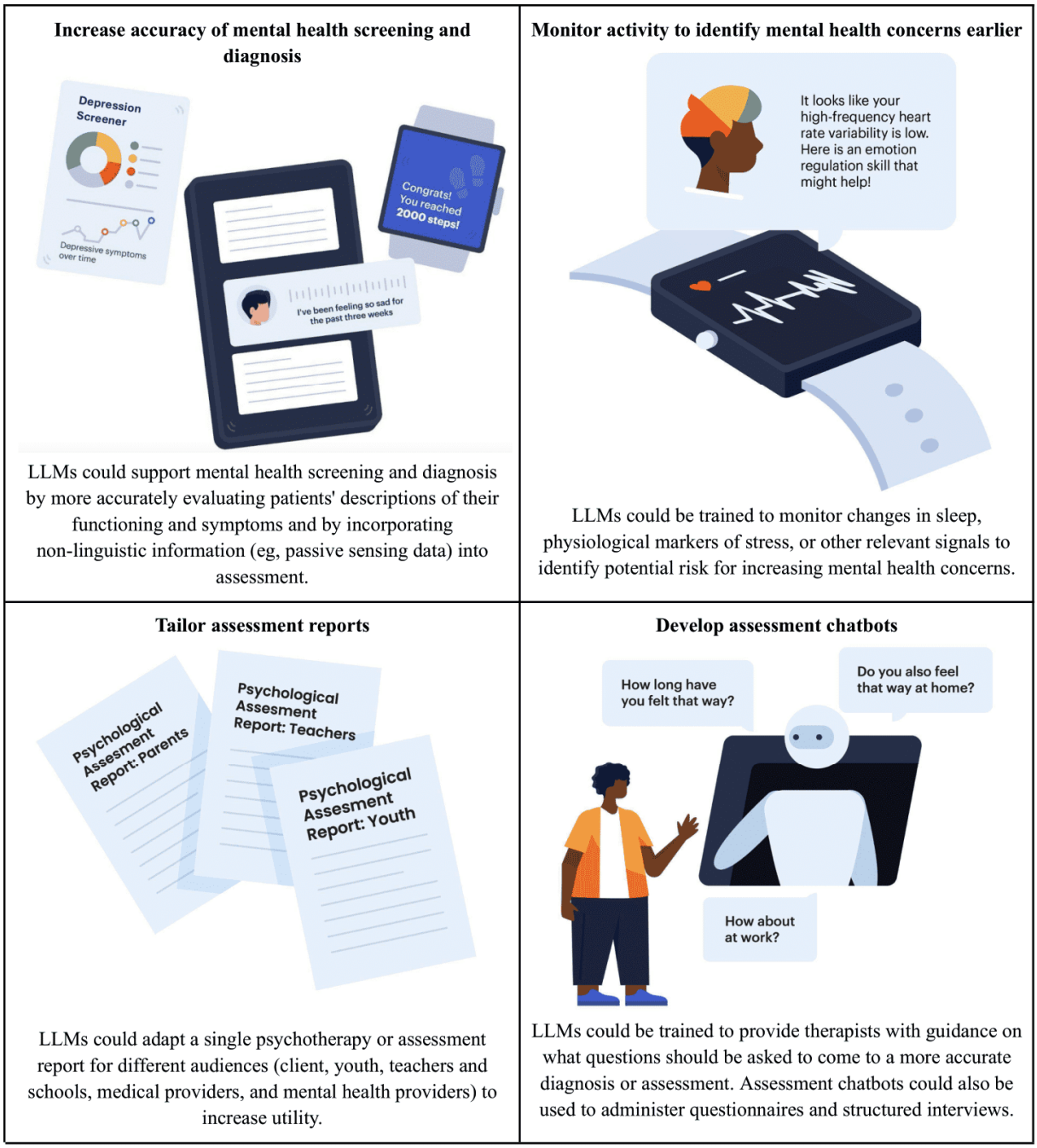

Figure 2. Potential applications of LLMs in assessing mental health. Reproduced from reference [23] under the Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/ (accessed on 17 May 2025)).

Administrative documentation remains a major source of the workload for clinicians. LLMs offer the possibility of automating routine charting by extracting relevant clinical information and generating narrative entries. Beyond basic transcription, some systems can assist with mapping clinical terms to standardized coding systems, such as the International Classification of Diseases (ICD) or Current Procedural Terminology (CPT) codes [29,30]. These features can reduce coding errors, streamline billing processes, and improve doc-

umentation compliance. However, challenges remain in ensuring the accuracy of the generated content, particularly in complex or ambiguous clinical cases.

LLMs are also being explored as tools for medical education and clinical training [31]. They can provide real-time explanations of medical terminology, pathophysiological mechanisms, or treatment guidelines to trainees, nurses, or other healthcare staff. The ability of these models to generate analogies, summaries, and step-by-step explanations may support self-directed learning and continuing education. Furthermore, LLMs can be used to simulate patient cases or clinical scenarios for training purposes [32]. As with other applications, the quality and accuracy of the content must be carefully validated before such tools can be relied upon in formal education settings.

Figure 3 shows the applications of LLMs in medical chatbots, divided into patient-facing and clinician-facing categories. Patient-facing applications include symptom assessments, information dissemination, and mental health support. Clinician-facing applications cover decision support, documentation, and medical education.

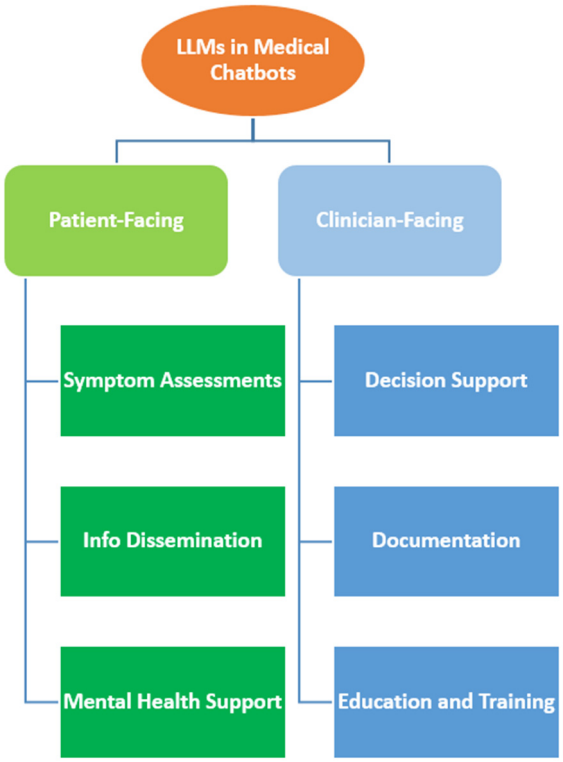

Figure 3. Block diagram of applications of LLMs in medical chatbots.

2.3. Applications in Rare Disease Diagnosis and Treatment

Rare diseases present diagnostic and management challenges due to their low prevalence, heterogeneous presentations, and limited clinical familiarity among general practitioners. Recent studies have explored the potential of LLMs to assist in narrowing diagnostic possibilities and guiding treatment options for rare conditions [33]. LLMs can synthesize data from scattered case reports, orphan disease registries, and the biomedical literature to propose differential diagnoses that may not be immediately apparent to clinicians.

For example, models like GPT-4 and BioGPT (v4.52.3) have been evaluated for their capacity to recognize rare disease patterns based on symptom descriptions or genomic data interpretations. One pilot study demonstrated that GPT-based tools could suggest plausible diagnoses in rare metabolic and genetic disorders with comparable accuracy to a specialist input, especially when coupled with structured input templates and chain-of-thought prompting [34]. Furthermore, LLMs have been proposed as educational tools for clinicians

encountering unusual or unfamiliar conditions, by generating disease summaries, case comparisons, and treatment guidelines drawn from niche sources, such as Orphanet and PubMed [35].

However, challenges remain in ensuring the accuracy, sensitivity, and reproducibility of LLM-generated outputs in this domain. The scarcity of training data and the risk of hallucination make model validation particularly important for rare disease applications. Despite these limitations, LLMs hold promise as decision-support tools that may help reduce diagnostic delays and support personalized care in rare disease management.

3. Core Technologies and Architectures

LLMs represent an advancement in NLP, enabling the development of medical chatbots. Their effectiveness in healthcare applications relies on both the underlying model architecture and the specific strategies used to adapt them to medical tasks.

3.1. Overview of LLMs

At the core of modern LLMs is the transformer architecture, introduced by Vaswani et al. in 2017 [36]. The transformer model relies on a self-attention mechanism that allows the model to weigh the relevance of different parts of the input text in parallel, enabling the efficient processing of long sequences and complex contextual relationships. This architecture has replaced earlier recurrent and convolutional models in many language-understanding tasks [37].

Several prominent LLMs have been applied to healthcare settings. GPT-3 and GPT-4 (OpenAI) are autoregressive models trained to predict the next word in a sequence and can generate coherent and contextually appropriate responses across diverse topics [38]. BERT is a bidirectional model pretrained on masked language modeling and has been widely adopted for classification and question-answering tasks [39]. Other models include the Pathways Language Model (PaLM) [40], developed by Google, which uses a massive scale of training to support few-shot learning.

Domain-specific LLMs have also emerged. BioGPT is pretrained on the biomedical literature to improve its performance on tasks involving medical terminology and research language [41]. Med-PaLM and Med-PaLM 2 are examples of models further fine-tuned on medical datasets and evaluated on medical reasoning tasks [42]. GatorTron, developed by the University of Florida, is another domain-specific model trained on clinical notes, illustrating the increasing interest in developing specialized LLMs for healthcare applications [43].

3.2. Fine-Tuning and Prompt Engineering

To adapt general-purpose LLMs to the healthcare domain, fine-tuning and prompt engineering techniques are commonly employed. Supervised fine-tuning involves continuing the training of a pretrained model on medical-specific corpora such as clinical notes, PubMed abstracts, or curated question–answer datasets [44]. This process helps the model align its outputs with medical vocabulary, knowledge, and reasoning patterns.

Reinforcement Learning from Human Feedback (RLHF) is another method increasingly used in model alignment [45]. For example, Med-PaLM uses RLHF based on expert medical annotations to ensure that generated responses are not only linguistically fluent but also clinically appropriate [46]. This technique helps models to better meet safety, factuality, and relevance requirements specific to healthcare applications.

Prompt engineering is a complementary approach where the model's behavior is influenced by carefully designed inputs. In zero-shot settings, the model receives a task description without prior examples. Few-shot prompting includes a small number of

examples within the prompt to guide the response [47]. Chain-of-thought prompting encourages the model to break down reasoning steps, which can be especially useful in diagnostic reasoning or the explanation of complex medical concepts [48].

3.3. Model Evaluation and Limitations

Evaluating the performance of LLMs in medical applications presents unique challenges. Standard natural language generation metrics, such as Bilingual Evaluation Understudy (BLEU), Recall-Oriented Understudy for Gisting Evaluation (ROUGE), and METEOR, are commonly used in NLP but may not fully capture the clinical relevance, factual accuracy, or safety of model outputs in healthcare settings [49].

To address these gaps, domain-specific benchmarks have been developed. MedQA, based on the United States Medical Licensing Examination (USMLE), assesses model performance on medical knowledge and reasoning [50]. MultiMedQA extends this approach by combining multiple question-answering datasets across diverse medical domains [51]. PubMedQA focuses on the biomedical literature domain and evaluates models on their ability to infer factual correctness from research abstracts [52].

Despite promising results, LLMs remain limited by issues such as hallucination (generation of incorrect information), the inability to verify or cite sources, and sensitivity to prompt phrasing [53]. These limitations underscore the need for robust evaluation frameworks, external validation with an expert input, and ongoing monitoring when deploying these models in clinical or patient-facing environments [54].

4. Benefits and Opportunities

The integration of LLMs into medical chatbots offers several potential benefits across clinical, operational, and educational domains. These models enable novel modes of interaction that can enhance accessibility, efficiency, and patient engagement in healthcare delivery. While the deployment of LLMs must be approached with careful oversight, their inherent scalability and adaptability create new opportunities to support both patients and healthcare professionals.

LLMs are inherently scalable and can be deployed across a wide range of platforms, including web interfaces, mobile applications, and digital health portals [55]. Once trained, a single model can address millions of patient queries in parallel, without any degradation in performance. Their ability to process and generate content in multiple languages makes them well-suited for use in linguistically diverse populations [56]. This scalability is particularly advantageous in public health contexts, where information dissemination and patient triage must often occur at a large scale and under constrained timelines.

One of the notable capabilities of LLMs is their ability to generate responses that are sensitive to the user's context. By incorporating information such as the patient's age, medical history, medication use, or risk factors, LLMs can tailor educational content and guidance to individual needs [57]. Personalization may also extend to the patient's communication style, literacy level, and preferred language, thereby enhancing the relevance and comprehensibility of health information. This can improve adherence to medical advice and facilitate shared decision-making [58].

LLM-based chatbots offer continuous availability, allowing users to access information and support at any time. This can reduce the dependency on real-time clinician availability, particularly for non-urgent concerns [59]. By handling routine inquiries and triaging potential issues, these systems can alleviate the workload of healthcare professionals, reduce the call center volume, and enhance operational efficiency [60]. This is particularly valuable in under-resourced settings or during periods of healthcare system strain, such as pandemics or natural disasters. Figure 4 provides an example of how AI can enhance

chronic kidney disease management across four key domains: early detection and screening, risk stratification and prediction, personalized treatment recommendations, and improved patient care and communication, leveraging LLMs and NLP [60].

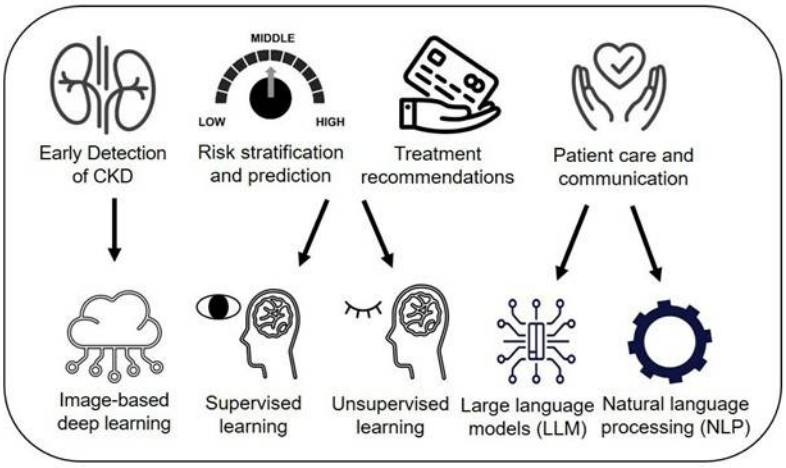

Figure 4. AI-driven approaches to chronic kidney disease management, focusing on patient care and communication through LLMs and NLP. Reproduced from reference [60] under the Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/ (accessed on 17 May 2025)).

Access to accurate, timely, and understandable health information plays a key role in patient empowerment [61]. LLMs can educate patients about diagnoses, treatment options, test results, and self-care strategies [62]. This can occur both before clinical encounters, to prepare patients for informed discussions, and after consultations, to reinforce care instructions and clarify uncertainties. Empowered patients are not only more engaged in managing their chronic conditions but also demonstrate measurable psychosocial gains. For example, Kohler et al. [63] reported that participants in genetic counseling exhibited a mean self-efficacy score of 3.69 $\pm$ 0.54 on a 4-point scale, and patient empowerment was moderately correlated with self-rated health ($r = 0.38$). These findings underscore the necessity of treating empowerment as a quantifiable patient-reported outcome, rather than a conceptual ideal [64].

Importantly, LLMs are best viewed as tools that augment, rather than replace, clinician expertise. Their role is to support healthcare providers by automating repetitive tasks, synthesizing information, and facilitating communication [65]. When deployed appropriately, these systems can extend the reach of clinicians, allowing them to focus on tasks requiring clinical judgment, empathy, and procedural skills. Successful integration will depend on well-defined roles for AI systems within clinical workflows and mechanisms for oversight and human validation [66].

Real-world pilot studies have demonstrated the scalability and operational benefits of LLM-based medical chatbots. For example, at the Mayo Clinic, pilot deployments of LLM-driven triage assistants reportedly supported over 30,000 patient interactions in a 3-month period, significantly reducing the call center volume and improving response times for non-urgent queries [67]. Similarly, Stanford Medicine reported a 25–35% reduction in the documentation time for clinicians using LLMs integrated with electronic health records [68]. From a cost perspective, studies suggest that once deployed, LLM-based systems can answer queries at an estimated cost of USD 0.01 to USD 0.05 per interaction, making them substantially more economical than traditional, human-operated support [69]. These results illustrate that LLMs offer not only technical advancements but also measurable improvements in efficiency, scalability, and cost-effectiveness in real healthcare environments.

User Acceptance and Trust in LLM-Powered Medical Chatbots

The successful integration of LLM-based medical chatbots into healthcare systems depends not only on technical accuracy but also on user acceptance, particularly among patients. Studies show that trust is a primary determinant of whether users engage with or follow chatbot-generated recommendations [70]. Factors influencing trust include perceived accuracy, transparency, empathy, privacy safeguards, and whether the chatbot discloses that it is an AI tool.

Research has shown that patients tend to trust chatbots more when the system communicates its limitations clearly and avoids overconfidence in its recommendations. Anthropomorphic features, such as friendly tone or personalized responses, can increase perceived empathy but also raise concerns about misleading realism if not appropriately disclosed [71].

A systematic review by Branley-Bell et al. [72] indicated that user demographic variables, including users' age, health literacy, and prior exposure to digital tools, also influence acceptance levels. In high-stakes scenarios, such as mental health or diagnostics, users often prefer the human validation of chatbot responses, reinforcing the importance of human–AI collaboration models.

Efforts to improve acceptance may include building transparent interfaces, displaying source citations or confidence scores, and giving users control options, such as flagging confusing responses or requesting a human follow-up [73]. Incorporating user feedback loops into model refinement can further align system behavior with patient expectations and values.

5. Challenges, Limitations, and AI Risk Mitigations

Despite the potential of LLMs in healthcare, their implementation presents challenges that must be addressed to ensure safe, effective, and equitable deployment. These include technical issues, such as factual inaccuracies and algorithmic biases, as well as broader concerns related to data privacy, regulatory uncertainty, and the systemic risk. In parallel with identifying these challenges, it is necessary to explore mitigation strategies grounded in medical ethics, health informatics, and AI governance.

5.1. Accuracy and Hallucination

A major concern in the clinical application of LLMs is the phenomenon of hallucination, in which the model generates responses that are linguistically fluent but factually incorrect, ungrounded, or misleading [74]. Unlike errors resulting from data gaps or misunderstanding, hallucinations often occur when a model produces plausible-sounding content that is not supported by any external source or factual basis. In medicine, this can lead to serious consequences, such as misdiagnosis, inappropriate medication recommendations, or false reassurances to patients [75]. Unlike rule-based systems, LLMs do not differentiate between validated medical knowledge and probabilistic inferences based on incomplete or biased training data [76]. Real-world studies have reported that LLM-generated answers may diverge from the established clinical guidelines, emphasizing the need for robust validation and expert review [77]. To address hallucination, the recent literature advocates for embedding fact-checking modules and evidence citation mechanisms into LLM outputs [78]. External knowledge integration, linking LLMs to curated medical knowledge bases or guideline repositories, has been shown to improve factual consistency. In addition, implementing uncertainty quantification (e.g., confidence scores) may alert users when responses are probabilistic, rather than authoritative [79].

5.2. Bias and Fairness

LLMs are trained on large-scale datasets that often reflect underlying social and systemic biases. These biases may become embedded in model behavior, leading to disparities in how different demographic groups receive and interpret health information [80]. In a focused evaluation of ChatGPT-3.5 on patient questions related to glaucoma, Tan et al. (2024) [81] reported a mean expert-rated accuracy score of 3.29 $\pm$ 0.48 (on a 4-point scale) across 72 responses. Notably, 29.2% of the questions received ratings of three or less but allowed self-correction-improved performance, raising the mean to 3.58 and increasing the proportion of full-score responses from 30.6% to 57.1% (p = 0.026). These results highlight both the promise and the current limitations of LLMs in handling real-world patient inquiries in ophthalmology. Bias mitigation can be approached through targeted dataset curation, the inclusion of diverse demographic inputs during training, and ongoing model auditing. Transparent documentation practices, such as model cards and data sheets, can improve accountability and support bias detection. Furthermore, fairness-aware training objectives and adversarial testing methods have been proposed to identify and reduce performance gaps across subpopulations [82].

Recent empirical studies have attempted to quantify the bias in LLM performance across different demographic groups. For example, a study evaluating GPT-4's clinical triage responses found a higher misdiagnosis rate for presentations simulating Black and Hispanic patients compared to White patients, particularly in cardiology and dermatology scenarios [83]. Another benchmark test using USMLE-style questions observed performance drops in the diagnostic accuracy for patients with gender-diverse or nonbinary profiles [84]. These discrepancies underscore that the algorithmic bias is not only theoretical but measurable, with real implications for health equity. Quantitative audits of LLMs are increasingly essential, and future development should incorporate stratified evaluation metrics to systematically detect performance gaps. Comparative datasets, such as MedQA-CS and MIMIC-Disparity, have been proposed to facilitate benchmarking the biases across minority subgroups [85].

5.3. Data Privacy and Security

The use of LLMs in healthcare settings raises complex concerns about data protection. Compliance with regulatory frameworks, such as the Health Insurance Portability and Accountability Act (HIPAA) in the U.S. and the General Data Protection Regulation (GDPR) in the EU, is essential when handling patient-identifiable information [86,87]. Privacy risks include the inadvertent disclosure of protected health information (PHI) through prompts or output logs and the exposure of sensitive data during cloud-based processing [88]. To minimize privacy risks, privacy-preserving approaches, such as FL and differential privacy, can be employed [89]. These techniques allow model training without centralized data storage, reducing the likelihood of PHI leakage. Additional safeguards include encrypted data transmission, access control policies, and the deployment of LLMs on secure, local servers where feasible.

While FL offers a strong potential for privacy-preserving AI in healthcare, its implementation faces several critical challenges [90]. One major issue is data heterogeneity across participating institutions. Clinical data can vary widely in structure, distribution, and quality depending on local practices, patient populations, and record-keeping systems. This non-IID (non-independent and identically distributed) nature of medical data can lead to significant model performance degradation and convergence instability [91]. Moreover, communication efficiency is a key bottleneck, as FL requires frequent synchronization between local models and a central server, which can be slow or unreliable in band-width-limited settings. Studies have proposed solutions, such as client selection

strategies, gradient compression, and asynchronous updates, to mitigate these bottlenecks [92]. However, ensuring model robustness and fairness across unevenly resourced sites remains a topic of ongoing research. Addressing these practical limitations is essential for the scalable and equitable deployment of FL in real-world healthcare systems. An overview of the primary technical and governance-related challenges associated with LLM-powered medical chatbots, along with mitigation strategies, is summarized in Table 1. This includes issues related to hallucination, algorithmic bias, privacy risks, and the evolving regulatory landscape.

Table 1. A summary of the primary challenges, associated risks, and mitigation strategies related to the use of LLMs in medical chatbots. The table outlines key concerns in accuracy, bias, privacy, and regulation, along with proposed technical and policy-based solutions drawn from the current literature.

Category	Challenge/Risk	Mitigation Strategies	References
Accuracy and Hallucination	Generation of factually incorrect, unsupported, or misleading medical content.	Integration of fact-checking modules, linkage to verified medical knowledge bases, uncertainty quantification (e.g., confidence scores).	[74–79]
Bias and Fairness	Unequal performance and discriminatory outputs due to training data biases (e.g., race, gender, SES).	Dataset curation, diverse demographic representation, model auditing, use of model cards and fairness-aware training methods.	[80–85]
Data Privacy and Security	Risk of PHI leakage through prompts or logs; vulnerability in cloud deployments.	Federated learning, differential privacy, local deployment, encryption, and access controls.	[86–92]

6. Evaluation and Benchmarking

The robust evaluation of LLMs in healthcare is essential to ensure their safety, reliability, and clinical utility. Unlike traditional, rule-based systems, LLMs generate variable and context-dependent outputs, necessitating the use of both automated metrics and human-centered assessments.

6.1. Existing Benchmarks

Several domain-specific benchmarks have been developed to assess the performance of LLMs in medical reasoning and question-answering tasks. MedQA evaluates a model's ability to answer multiple-choice questions derived from the United States Medical Licensing Examination (USMLE), offering a proxy for clinical knowledge competence [93]. PubMedQA focuses on the comprehension of the biomedical literature by testing whether models can accurately infer conclusions from abstracts [94]. MultiMedQA, introduced by Google, aggregates multiple datasets, including MedQA, HealthSearchQA, and MedicationQA, to provide a more comprehensive and diverse evaluation framework [95].

These benchmarks have highlighted the substantial progress of state-of-the-art models. For example, GPT-4 has demonstrated a performance on par with or exceeding that of average medical students on USMLE-style questions [96]. However, these results reflect controlled settings and question formats that may not capture the nuances of real-world clinical communication. Moreover, benchmark questions are often knowledge-based and do not evaluate critical dimensions, such as empathy, patient engagement, or ethical appropriateness.

While automated metrics, such as BLEU, ROUGE, and METEOR [97], provide baseline assessments of linguistic quality and information overlap, they fall short of capturing the clinical relevance, safety, and patient-centeredness required for medical applications.

These metrics cannot evaluate subtle but critical dimensions, such as empathy, ethical appropriateness, or the risk of misinformation [98]. As a result, there is a growing consensus that the evaluation of LLMs in healthcare must go beyond surface-level text similarity to include qualitative and context-aware assessments. This has led to increased interest in human-in-the-loop (HITL) evaluation frameworks, which integrate expert judgment into the assessment process [99].

6.2. Human-in-the-Loop Evaluation

To address the limitations of automated metrics, HITL evaluation strategies are increasingly used to assess LLM outputs in healthcare contexts [100]. In these settings, the outputs generated by the model are reviewed and rated by multidisciplinary panels, including physicians, nurses, medical ethicists, and patients. The evaluation criteria may include the outputs' clinical accuracy, relevance, tone, empathy, and potential for harm.

Despite their utility, HITL evaluations are challenged by subjectivity, inter-rater variability, and a lack of standardization [101]. The absence of universally accepted rubrics for rating conversational quality or safety in medical dialogue complicates the reproducibility and benchmarking across studies. Furthermore, expert involvement in large-scale evaluations can be resource-intensive and may not be feasible in all contexts [102].

Efforts are underway to define standardized protocols and scales for assessing LLM-generated medical content, including Likert-based scoring systems, adverse content flagging, and scenario-based testing [103].

6.3. Real-World Trials and Pilots

A small, but growing, number of institutions have initiated pilot studies to evaluate the performance of LLMs in clinical environments. At the Mayo Clinic, LLM-based chatbots have been tested for use in patient triage and administrative support [104]. Similarly, Stanford Medicine has explored LLM applications for drafting clinical notes and assisting in medical education [105].

These pilots assess outcome measures such as patient satisfaction, the accuracy of triage recommendations, time savings for clinicians, and perceived trustworthiness. In some instances, workload reduction has been reported, especially in documentation and non-clinical interactions. However, integration into EHRs, validation workflows, and user interface design remain key challenges [106].

Th preliminary findings from these studies support the potential of LLMs to augment healthcare delivery, but emphasize the necessity for ongoing oversight, transparent reporting, and iterative model refinement [107]. Longitudinal studies will be needed to establish the sustained impact of these technologies on clinical outcomes and health system performance.

7. Ethical, Legal, and Regulatory Considerations

As LLMs are increasingly integrated into clinical workflows, patient communication, and decision support, their deployment raises complex questions across three inter-related domains: ethical norms, legal accountability, and regulatory oversight. Ensuring safety, equity, and transparency requires a unified framework that addresses not only how these systems behave in practice but also how they are governed and held accountable.

7.1. Ethical Issues

LLMs in healthcare pose several ethical concerns that differ fundamentally from traditional clinical tools. One central issue is explainability, as these models often operate as "black boxes", generating outputs without transparent reasoning pathways. This opacity

undermines the clinician's ability to validate recommendations, eroding trust in AI-assisted decision-making, particularly when outputs contradict established medical norms [108].

Informed consent is another ethical challenge. Patients interacting with medical chatbots may not realize that they are engaging with AI systems, nor fully understand the nature or limitations of the information provided. Clearly disclosing the role of LLMs, the scope of their use, and their non-human status is essential to maintaining user autonomy.

Furthermore, trust and transparency play a pivotal role in user engagement and adherence to AI-generated advice. The misuse or failure of AI tools, particularly in high-stakes areas like diagnostics, can severely damage public trust. To foster confidence, healthcare institutions deploying LLMs must implement transparency measures, such as public model documentation, reporting known limitations, and human oversight protocols [109].

7.2. Legal Accountability

The use of LLMs introduces ambiguous liability risks. If an LLM provides incorrect medical guidance that contributes to patient harm, it remains unclear who should bear legal responsibility: the developer, deploying institution, supervising clinician, or the system as a whole. Current malpractice and tort law frameworks are ill-equipped to adjudicate harm arising from probabilistic, non-deterministic AI outputs [110].

There is also uncertainty around the standards of care when AI tools are involved. For example, if a clinician follows chatbot-generated advice that conflicts with accepted practice guidelines, how does that influence legal culpability? Moreover, the documentation and traceability of AI-driven decisions are often lacking, complicating any post-hoc legal analysis and potential litigation.

As AI tools become more autonomous and influential in clinical decision-making, there is an urgent need for revised legal frameworks that clarify these responsibilities, incorporate AI-specific liability clauses, and ensure patients have access to redress mechanisms when AI contributes to adverse outcomes.

7.3. Regulatory Governance

Current medical device regulations, such as the FDA's Software as a Medical Device (SaMD) framework [111], were not designed to evaluate systems that produce open-ended, probabilistic outputs, like LLMs. This misalignment complicates regulatory approval, as the traditional criteria, such as reproducibility, explainability, and performance benchmarking, may not apply to GAI systems.

Furthermore, standardized evaluation protocols are lacking. Unlike rule-based decision tools, LLMs may generate different responses to similar prompts, making validation inherently difficult. There is no global consensus on how to assess the clinical risk or establish performance thresholds for models that continuously evolve through updates or online learning.

To bridge these gaps, regulatory bodies have begun exploring tiered risk frameworks that differentiate between low-risk informational uses (e.g., patient education) and high-risk clinical applications (e.g., diagnostic recommendations) [112]. Proposals for pre-deployment audits, post-market surveillance, and AI sandboxes (controlled environments where models can be tested with real-world data under regulatory observation) are gaining traction.

A robust governance framework will also require multidisciplinary oversight, combining expertise from clinicians, ethicists, technologists, and policymakers. International coordination will be necessary to ensure that LLM-based medical tools meet common standards while respecting regional differences in healthcare systems and legal cultures [113].

8. Future Directions Incorporating AI Risk Management into LLMs

The integration of LLMs into medical chatbots remains an evolving field. While current applications demonstrate promising utility, future developments must address technical limitations, system integration, user interaction models, privacy safeguards, and regulatory oversight to avoid AI risks.

8.1. Model Improvements

Future iterations of LLMs will likely extend beyond text-only processing by integrating multimodal capabilities [114]. Figure 5 gives an example of a typical multimodal LLM setup [107], which includes an encoder (E_M), a connector (C), and an LLM. A generator (G) can also be added to produce outputs in formats beyond text, like images or audio. The encoder takes in inputs, such as images, audio, or video, and turns them into features. The connector then processes these features to help the LLM understand them better. There are three main types of connectors: projection-based, query-based, and fusion-based. The first two turn features into tokens and mix them with text tokens, while fusion-based connectors combine features directly inside the LLM.

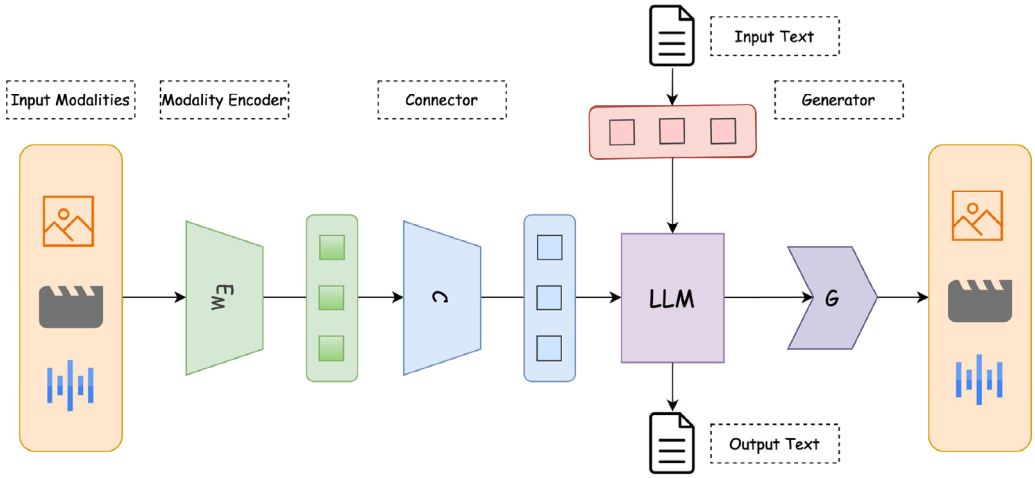

Figure 5. An illustration of a typical architecture for a multimodal large language model. Reproduced from reference [107] under the Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/ (accessed on 17 May 2025)).

Emerging models, such as LLaVA-Med [115] and MedCLIP [116], aim to combine textual and visual data, enabling systems to interpret clinical images, radiology reports, and pathology slides alongside narrative content. Such models could improve diagnostic support, especially in settings requiring the synthesis of diverse data sources.

Another key direction is domain-specific pretraining [117]. LLMs trained on general internet data may lack the precision needed for specialized medical applications. Fine-tuning on curated clinical notes, the biomedical literature, and health-specific dialogue datasets (e.g., MIMIC-IV, PubMed, and UpToDate) is expected to enhance accuracy, contextual understanding, and terminology alignment in medical settings [118].

While multimodal LLMs, such as LLaVA-Med and MedCLIP, have demonstrated promising capabilities in fusing visual and textual data, recent evaluations show a performance gap between reported benchmarks and real-world applicability. For example, an external validation study of LLaVA-Med's interpretation of radiographic findings showed an F1-score decline of over 15% compared to its claimed performance on internal datasets [119]. Similarly, MedCLIP's accuracy in segmenting pathology slides was significantly lower when tested on out-of-distribution images from rural clinics [120]. These discrepancies highlight the limitations of the current multimodal models in generalizing beyond curated

datasets. Comparative studies remain limited, and further work is needed to evaluate multimodal LLMs in diverse clinical settings using real-world imaging variability and human-rated ground truth.

8.2. Human–AI Collaboration Paradigms

A paradigm shift is emerging toward viewing LLMs not as autonomous agents but as copilots supporting human clinicians [121]. In this collaborative model, the LLM offers suggestions, drafts, or preliminary assessments, while final decision-making remains with the healthcare provider. This approach maintains clinical accountability while enhancing productivity [122,123].

Future systems may include dynamic user interfaces that facilitate real-time interactions between users and models [124]. These interfaces could allow clinicians to edit, query, or correct model outputs interactively, promoting transparency, reducing cognitive burden, and improving trust in AI-assisted workflows [125].

8.3. Integration with Health Systems

For LLMs to be effective in clinical environments, seamless integration with health information systems is critical [126]. Embedding chatbot functionalities within EHR platforms could enable context-aware assistance, such as retrieving patient history, suggesting documentation templates, or flagging inconsistencies in clinical notes [127].

In telemedicine, LLM-powered chatbots can support clinicians in virtual consultations by transcribing conversations, summarizing visit notes, or suggesting follow-up care [128]. Similarly, in chronic disease management, chatbots may offer personalized guidance, medication reminders, and lifestyle coaching, helping to bridge the gap between clinical visits and continuous care [129].

8.4. Privacy-Preserving Approaches

Ensuring data privacy remains a central concern in AI-driven healthcare. New approaches aim to enable high-performance modeling while minimizing data exposure [130]. Federated learning allows models to be trained across decentralized data sources without transferring raw patient data to a central server. This method enhances privacy while preserving access to diverse datasets [131].

Moreover, differential privacy techniques can be incorporated into training and inference processes to limit the risk of re-identification [132]. The development of on-device LLMs, capable of operating locally on secure hospital systems or mobile devices, is also gaining interest as a means to enable private and compliant user interactions [133,134].

8.5. Policy and Governance

As LLMs gain influence in clinical decision-making and patient communication, the development of robust policy frameworks is essential. This includes establishing technical safety standards, validation protocols, and reporting requirements for AI systems in healthcare [135]. Regulators, such as the U.S. [136], FDA [137], EMA [138], and MHRA [139], are beginning to address the unique risks posed by GAI but lack comprehensive guidelines specific to LLMs.

International collaboration may be required to develop harmonized governance structures, especially as medical AI systems cross national boundaries. Future policies should support innovation while enforcing ethical use, transparency, and patient protections [140]. Table 2 shows the key future directions for integrating LLMs into medical chatbot systems. These include advancements in model capabilities, collaborative interaction paradigms, integration within clinical workflows, privacy-preserving methods, and the development of regulatory frameworks to ensure safe and ethical deployment.

Table 2. Key future directions for integrating LLMs into medical chatbot systems, highlighting advancements in model architecture, human–AI collaboration, system integration, privacy protection, and policy development.

Focus Area	Future Directions and Innovations	References
Model Improvements	- Integration of multimodal capabilities to process both text and clinical images (e.g., LLaVA-Med, MedCLIP). - Emphasis on domain-specific pretraining using curated medical datasets (e.g., MIMIC-IV, PubMed, UpToDate) to enhance contextual accuracy and clinical relevance.	[114–120]
Human–AI Collaboration	- Reframing LLMs as clinical copilots that support, rather than replace, healthcare professionals. - Development of interactive interfaces to allow clinicians to query, edit, and validate AI outputs in real time, fostering trust and transparency in clinical workflows.	[121–125]
System Integration	- Embedding chatbots within EHR platforms for context-aware assistance (e.g., retrieving patient history, generating documentation). - Use in telemedicine and chronic care for documentation, follow-up, and personalized patient engagement.	[126–129]
Privacy-Preserving Methods	- Adoption of federated learning to train models without centralizing sensitive data. - Use of differential privacy to minimize re-identification risks. - Development of on-device LLMs for secure, localized model deployment in hospitals and on personal devices.	[130–134]
Policy and Governance	- Establishing technical standards, validation protocols, and reporting requirements for AI in healthcare. - Regulatory bodies (e.g., FDA, EMA, MHRA) are beginning to respond but require LLM-specific guidance. - Need for international governance frameworks.	[135–140]

9. Conclusions

This review has examined the evolving role of LLMs in medical chatbots, detailing their applications across patient- and clinician-facing domains, technical foundations, benefits, limitations, and ethical implications. LLMs have introduced new capabilities in natural language understanding and generation, enabling more responsive, personalized, and scalable tools for communication, education, documentation, and clinical decision support in healthcare.

Despite these advancements, several critical challenges must be addressed before LLMs can be safely and reliably deployed in clinical practice. Key concerns include factual inaccuracy and hallucination, algorithmic biases affecting under-represented populations, risks to data privacy, and the absence of robust regulatory frameworks tailored to GAI. These risks are compounded by the probabilistic and opaque nature of LLMs, which complicates their validation and oversight.

To mitigate these risks, recent work has proposed a range of strategies, including the integration of fact-checking and uncertainty quantification modules; bias audits and inclusive data curation; privacy-preserving training techniques, such as federated learning; and the establishment of tiered, domain-specific regulatory frameworks. These interventions are essential not only for improving model performance but also for maintaining clinical trust, ethical alignment, and patient safety.

As the field progresses, responsible innovation must remain central to the development and deployment of LLM-powered medical chatbots. This includes ensuring transparency

in model behavior, rigorously validating outputs in real-world healthcare settings, and embedding safeguards that anticipate unintended consequences.

Ultimately, the effective use of LLMs in healthcare will depend on sustained, interdisciplinary collaboration among clinicians, AI developers, health system leaders, ethicists, and regulators. By aligning technological innovation with clinical, ethical, and societal standards, it is possible to develop medical chatbot systems that are not only powerful and efficient but also equitable, accountable, and clinically meaningful.

Author Contributions: Conceptualization, J.C.L.C. and K.L.; methodology, J.C.L.C. and K.L.; software, J.C.L.C. and K.L.; validation, J.C.L.C. and K.L.; formal analysis, J.C.L.C. and K.L.; investigation, J.C.L.C. and K.L.; resources, J.C.L.C. and K.L.; data curation, J.C.L.C. and K.L.; writing—original draft preparation, J.C.L.C.; writing—review and editing, K.L.; visualization, J.C.L.C.; supervision, J.C.L.C. and K.L.; project administration, J.C.L.C. and K.L.; funding acquisition, J.C.L.C. and K.L. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by a Canadian Institutes of Health Research Planning and Dissemination Grant—Institute Community Support (CIHR PCS–191021).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	artificial intelligence
BERT	Bidirectional Encoder Representations from Transformers
BLEU	Bilingual Evaluation Understudy
CC BY	Creative Commons Attribution
CPT	Current Procedural Terminology
DOAJ	Directory of Open Access Journals
EHR	electronic health record
FDA	U.S. Food and Drug Administration
FL	federated learning
GAI	generative artificial intelligence
GDPR	General Data Protection Regulation
GPT	generative pretrained transformer
HIPAA	Health Insurance Portability and Accountability Act
HITL	human-in-the-loop
ICD	International Classification of Diseases
LLM	large language model
METEOR	Metric for Evaluation of Translation with Explicit ORdering
MHRA	Medicines and Healthcare products Regulatory Agency
ML	machine learning
NLP	natural language processing
PaLM	Pathways Language Model
PHI	protected health information
RLHF	Reinforcement Learning from Human Feedback
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
SaMD	Software as a Medical Device
SOAP	Subjective, Objective, Assessment, Plan
TLA	Three Letter Acronym
USMLE	United States Medical Licensing Examination

References

1. Hindelang, M.; Sitaru, S.; Zink, A. Transforming health care through chatbots for medical history-taking and future directions: Comprehensive systematic review. *JMIR Med. Inform.* **2024**, *12*, e56628. [CrossRef] [PubMed]
2. Xu, L.; Sanders, L.; Li, K.; Chow, J.C. Chatbot for health care and oncology applications using artificial intelligence and machine learning: Systematic review. *JMIR Cancer* **2021**, *7*, e27850. [CrossRef] [PubMed]
3. Weizenbaum, J. ELIZA—A computer program for the study of natural language communication between man and machine. *Commun. ACM* **1966**, *9*, 36–45. [CrossRef]
4. Siddique, S.; Chow, J.C. Machine learning in healthcare communication. *Encyclopedia* **2021**, *1*, 220–239. [CrossRef]
5. Locke, S.; Bashall, A.; Al-Adely, S.; Moore, J.; Wilson, A.; Kitchen, G.B. Natural language processing in medicine: A review. *Trends Anaesth. Crit. Care* **2021**, *38*, 4–9. [CrossRef]
6. Babu, A.; Boddu, S.B. Bert-based medical chatbot: Enhancing healthcare communication through natural language understanding. *Explor. Res. Clin. Soc. Pharm.* **2024**, *13*, 100419. [CrossRef]
7. Chow, J.C.; Wong, V.; Li, K. Generative pre-trained transformer-empowered healthcare conversations: Current trends, challenges, and future directions in large language model-enabled medical chatbots. *BioMedInformatics* **2024**, *4*, 837–852. [CrossRef]
8. Huo, B.; Boyle, A.; Marfo, N.; Tangamornsuksan, W.; Steen, J.P.; McKechnie, T.; Lee, Y.; Mayol, J.; Antoniou, S.A.; Thirunavukarasu, A.J.; et al. Large language models for chatbot health advice studies: A systematic review. *JAMA Netw. Open* **2025**, *8*, e2457879. [CrossRef]
9. Chow, J.C. Artificial intelligence in radiotherapy and patient care. In *Artificial Intelligence in Medicine*; Springer: Cham, Switzerland, 2021; pp. 1–13.
10. Chakraborty, C.; Pal, S.; Bhattacharya, M.; Dash, S.; Lee, S.S. Overview of chatbots with special emphasis on artificial intelligence-enabled ChatGPT in medical science. *Front. Artif. Intell.* **2023**, *6*, 1237704. [CrossRef]
11. Harris, E. Large language models answer medical questions accurately, but can't match clinicians' knowledge. *JAMA* **2023**, *330*, 792–794. [CrossRef]
12. Chow, J.C.; Sanders, L.; Li, K. Impact of ChatGPT on medical chatbots as a disruptive technology. *Front. Artif. Intell.* **2023**, *6*, 1166014. [CrossRef] [PubMed]
13. Bengio, Y.; Hinton, G.; Yao, A.; Song, D.; Abbeel, P.; Darrell, T.; Harari, Y.N.; Zhang, Y.Q.; Xue, L.; Shalev-Shwartz, S.; et al. Managing extreme AI risks amid rapid progress. *Science* **2024**, *384*, 842–845. [CrossRef] [PubMed]
14. Denecke, K.; May, R.; LLM Health Group; Rivera Romero, O. Potential of Large Language Models in Health Care: Delphi Study. *J. Med. Internet Res.* **2024**, *26*, e52399. [CrossRef]
15. Chandel, A. Healthcare chatbot using SVM & decision tree. *Trends Health Inform.* **2025**, *2*, 10–17.
16. Singhal, K.; Tu, T.; Gottweis, J.; Sayres, R.; Wulczyn, E.; Amin, M.; Hou, L.; Clark, K.; Pfohl, S.R.; Cole-Lewis, H.; et al. Toward expert-level medical question answering with large language models. *Nat. Med.* **2025**, *31*, 943–950. [CrossRef]
17. Liu, Z.; Hou, Z.; Di, Y.; Yang, K.; Sang, Z.; Xie, C.; Yang, J.; Liu, S.; Wang, J.; Li, C.; et al. Infi-Med: Low-resource medical MLLMs with robust reasoning evaluation. *arXiv* **2025**, arXiv:2505.23867.
18. Chow, J.C.; Sanders, L.; Li, K. Design of an educational chatbot using artificial intelligence in radiotherapy. *AI* **2023**, *4*, 319–332. [CrossRef]
19. Kumar, M. AI-driven healthcare chatbots: Enhancing access to medical information and lowering healthcare costs. *J. Artif. Intell. Cloud Comput.* **2023**, *2*, 2–5. [CrossRef]
20. Zhang, S.; Song, J. A chatbot-based question and answer system for the auxiliary diagnosis of chronic diseases based on large language model. *Sci. Rep.* **2024**, *14*, 17118. [CrossRef]
21. Rebelo, N.; Sanders, L.; Li, K.; Chow, J.C. Learning the treatment process in radiotherapy using an artificial intelligence–assisted chatbot: Development study. *JMIR Form. Res.* **2022**, *6*, e39443. [CrossRef]
22. Shiferaw, M.W.; Zheng, T.; Winter, A.; Mike, L.A.; Chan, L.N. Assessing the accuracy and quality of artificial intelligence (AI) chatbot-generated responses in making patient-specific drug-therapy and healthcare-related decisions. *BMC Med. Inform. Decis. Mak.* **2024**, *24*, 404. [CrossRef] [PubMed]
23. Lawrence, H.R.; Schneider, R.A.; Rubin, S.B.; Matarić, M.J.; McDuff, D.J.; Bell, M.J. The opportunities and risks of large language models in mental health. *JMIR Ment. Health* **2024**, *11*, e59479. [CrossRef] [PubMed]
24. Vagwala, M.K.; Asher, R. Conversational artificial intelligence and distortions of the psychotherapeutic frame: Issues of boundaries, responsibility, and industry interests. *Am. J. Bioeth.* **2023**, *23*, 28–30. [CrossRef] [PubMed]
25. Kovacek, D.; Chow, J.C. An AI-assisted chatbot for radiation safety education in radiotherapy. *IOP SciNotes* **2021**, *2*, 034002. [CrossRef]
26. Seo, J.; Choi, D.; Kim, T.; Cha, W.C.; Kim, M.; Yoo, H.; Oh, N.; Yi, Y.; Lee, K.H.; Choi, E. Evaluation framework of large language models in medical documentation: Development and usability study. *J. Med. Internet Res.* **2024**, *26*, e58329. [CrossRef]

27. Hager, P.; Jungmann, F.; Holland, R.; Bhagat, K.; Hubrecht, I.; Knauer, M.; Vielhauer, J.; Makowski, M.; Braren, R.; Kaissis, G.; et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat. Med.* **2024**, *30*, 2613–2622. [CrossRef]
28. Li, L.; Zhou, J.; Gao, Z.; Hua, W.; Fan, L.; Yu, H.; Hagen, L.; Zhang, Y.; Assimes, T.L.; Hemphill, L.; et al. A scoping review of using large language models (LLMs) to investigate electronic health records (EHRs). *arXiv* **2024**, arXiv:2405.03066.
29. Simmons, A.; Takkavatakarn, K.; McDougal, M.; Dilcher, B.; Pincavitch, J.; Meadows, L.; Kauffman, J.; Klang, E.; Wig, R.; Smith, G.; et al. Extracting international classification of diseases codes from clinical documentation using large language models. *Appl. Clin. Inform.* **2025**, *16*, 337–344. [CrossRef]
30. Ong, J.; Kedia, N.; Harihar, S.; Vupparaboina, S.C.; Singh, S.R.; Venkatesh, R.; Vupparaboina, K.; Bollepalli, S.C.; Chhablani, J. Applying large language model artificial intelligence for retina international classification of diseases (ICD) coding. *J. Med. Artif. Intell.* **2023**, *6*, 1166014. [CrossRef]
31. Chow, J.C.; Wong, V.; Sanders, L.; Li, K. Developing an AI-assisted educational chatbot for radiotherapy using the IBM Watson assistant platform. *Healthcare* **2023**, *11*, 2417. [CrossRef]
32. Brügge, E.; Ricchizzi, S.; Arenbeck, M.; Keller, M.N.; Schur, L.; Stummer, W.; Holling, M.; Lu, M.H.; Darici, D. Large language models improve clinical decision making of medical students through patient simulation and structured feedback: A randomized controlled trial. *BMC Med. Educ.* **2024**, *24*, 1391. [CrossRef] [PubMed]
33. Schumacher, E.; Naik, D.; Kannan, A. Rare Disease Differential Diagnosis with Large Language Models at Scale: From Abdominal Actinomycosis to Wilson's Disease. *arXiv* **2025**, arXiv:2502.15069.
34. Yuan, S.; Bai, Z.; Xu, M.; Yang, F.; Gao, Y.; Yu, H. ChatGPT-assisted clinical decision-making for rare genetic metabolic disorders: A preliminary case-based study. *Front. Genet.* **2023**, *14*, 1212495.
35. Gao, C.A.; Howard, F.M.; Markov, N.S.; Dyer, E.C.; Ramesh, S.; Luo, Y.; Pearson, A.T. Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. *NPJ Digit. Med.* **2023**, *6*, 75. [CrossRef]
36. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008. Available online: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (accessed on 24 June 2025).
37. Dauphin, Y.N.; Fan, A.; Auli, M.; Grangier, D. Language modeling with gated convolutional networks. In Proceedings of the International Conference on Machine Learning 2017, Sydney, Australia, 11–15 August 2017; pp. 933–941.
38. Alto, V. *Modern Generative AI with ChatGPT and OpenAI Models: Leverage the Capabilities of OpenAI's LLM for Productivity and Innovation with GPT3 and GPT4*; Packt Publishing Ltd.: Birmingham, UK, 2023.
39. Koroteev, M.V. BERT: A review of applications in natural language processing and understanding. *arXiv* **2021**, arXiv:2103.11943.
40. Anil, R.; Dai, A.M.; Firat, O.; Johnson, M.; Lepikhin, D.; Passos, A.; Shakeri, S.; Taropa, E.; Bailey, P.; Chen, Z.; et al. Palm 2 technical report. *arXiv* **2023**, arXiv:2305.10403.
41. Luo, R.; Sun, L.; Xia, Y.; Qin, T.; Zhang, S.; Poon, H.; Liu, T.Y. BioGPT: Generative pre-trained transformer for biomedical text generation and mining. *Brief. Bioinform.* **2022**, *23*, bbac409. [CrossRef]
42. Tu, T.; Azizi, S.; Driess, D.; Schaekermann, M.; Amin, M.; Chang, P.C.; Carroll, A.; Lau, C.; Tanno, R.; Ktena, I.; et al. Towards Generalist Biomedical AI. *NEJM AI* **2024**, *1*, AIoa2300138. [CrossRef]
43. Yang, X.; Chen, A.; PourNejatian, N.; Shin, H.C.; Smith, K.E.; Parisien, C.; Compas, C.; Martin, C.; Flores, M.G.; Zhang, Y.; et al. Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records. *arXiv* **2022**, arXiv:2203.03540.
44. Ross, E.; Kansal, Y.; Renzella, J.; Vassar, A.; Taylor, A. Supervised fine-tuning LLMs to behave as pedagogical agents in programming education. *arXiv* **2025**, arXiv:2502.20527.
45. Hao, S.; Duan, L. Online learning from strategic human feedback in LLM fine-tuning. In Proceedings of the ICASSP 2025 IEEE International Conference on Acoustics, Speech and Signal Processing, Hyderabad, India, 6–11 April 2025; pp. 1–5.
46. Chaddad, A.; Jiang, Y.; He, C. OpenAI ChatGPT: A potential medical application. In Proceedings of the 2023 IEEE International Conference on E-Health Networking, Application & Services (Healthcom), Chongqing, China, 15–17 December 2023; pp. 210–215.
47. Lee, U.; Jung, H.; Jeon, Y.; Sohn, Y.; Hwang, W.; Moon, J.; Kim, H. Few-shot is enough: Exploring ChatGPT prompt engineering method for automatic question generation in English education. *Educ. Inf. Technol.* **2024**, *29*, 11483–11515. [CrossRef]
48. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.H.; Le, Q.V.; Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 24824–24837.
49. Chauhan, S.; Daniel, P. A comprehensive survey on various fully automatic machine translation evaluation metrics. *Neural Process. Lett.* **2023**, *55*, 12663–12677. [CrossRef]
50. Gilson, A.; Safranek, C.W.; Huang, T.; Socrates, V.; Chi, L.; Taylor, R.A.; Chartash, D. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? *JMIR Med. Educ.* **2023**, *9*, e45312. [CrossRef]

51. Zhou, Y.; Liu, X.; Ning, C.; Wu, J. Multifaceteval: Multifaceted evaluation to probe LLMs in mastering medical knowledge. *arXiv* **2024**, arXiv:2406.02919.
52. Jin, Q.; Dhingra, B.; Liu, Z.; Cohen, W.W.; Lu, X. PubmedQA: A dataset for biomedical research question answering. *arXiv* **2019**, arXiv:1909.06146.
53. Cheng, N.; Yan, Z.; Wang, Z.; Li, Z.; Yu, J.; Zheng, Z.; Tu, K.; Xu, J.; Han, W. Potential and limitations of LLMs in capturing structured semantics: A case study on SRL. In *International Conference on Intelligent Computing*; Springer: Singapore, 2024; pp. 50–61.
54. Chow, J.C.; Li, K. Developing effective frameworks for large language model–based medical chatbots: Insights from radiotherapy education with ChatGPT. *JMIR Cancer* **2025**, *11*, e66633. [CrossRef]
55. Yang, R.; Tan, T.F.; Lu, W.; Thirunavukarasu, A.J.; Ting, D.S.; Liu, N. Large language models in health care: Development, applications, and challenges. *Health Care Sci.* **2023**, *2*, 255–263. [CrossRef]
56. Zhang, W.; Aljunied, M.; Gao, C.; Chia, Y.K.; Bing, L. M3Exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 5484–5505.
57. Razafinirina, M.A.; Dimbisoa, W.G.; Mahatody, T. Pedagogical alignment of large language models (LLM) for personalized learning: A survey, trends and challenges. *J. Intell. Learn. Syst. Appl.* **2024**, *16*, 448–480. [CrossRef]
58. Abd-Alrazaq, A.; AlSaad, R.; Alhuwail, D.; Ahmed, A.; Healy, P.M.; Latifi, S.; Aziz, S.; Damseh, R.; Alabed Alrazak, S.; Sheikh, J. Large language models in medical education: Opportunities, challenges, and future directions. *JMIR Med. Educ.* **2023**, *9*, e48291. [CrossRef] [PubMed]
59. Yang, Z.; Wang, D.; Zhou, F.; Song, D.; Zhang, Y.; Jiang, J.; Lin, F.; Liang, J.; Chen, E.M.; Li, F.; et al. Understanding natural language: Potential application of large language models to ophthalmology. *Asia-Pac. J. Ophthalmol.* **2024**, *13*, 100085. [CrossRef] [PubMed]
60. Sabanayagam, C.; Banu, R.; Lim, C.; Tham, Y.C.; Cheng, C.Y.; Tan, G.; Ekinci, E.; Sheng, B.; McKay, G.; Shaw, J.E.; et al. Artificial intelligence in chronic kidney disease management: A scoping review. *Theranostics* **2025**, *15*, 4566. [CrossRef]
61. Mulukuntla, S. Digital Health Literacy: Empowering Patients in the Era of Electronic Medical Records. *EPH Int. J. Med. Health Sci.* **2020**, *6*, 23–24.
62. De Busser, B.; Roth, L.; De Loof, H. The role of large language models in self-care: A study and benchmark on medicines and supplement guidance accuracy. *Int. J. Clin. Pharm.* **2024**, 1–10. [CrossRef]
63. Kohler, A.; Tingstrom, P.; Jaarsma, T.; Nilsson, S. Patient empowerment and general self-efficacy in patients with coronary heart disease: A cross-sectional study. *BMC Fam Pract.* **2018**, *19*, 76.
64. McAllister, M.; Dunn, G.; Payne, K.; Davies, L.; Todd, C. Patient empowerment: The need to consider it as a measurable patient-reported outcome for chronic conditions. *BMC Health Serv. Res.* **2012**, *12*, 157. [CrossRef]
65. Lin, C.; Kuo, C.F. Roles and potential of large language models in healthcare: A comprehensive review. *Biomed. J.* **2025**, 100868. [CrossRef]
66. Dagli, M.M.; Ghenbot, Y.; Ahmad, H.S.; Chauhan, D.; Turlip, R.; Wang, P.; Welch, W.C.; Ozturk, A.K.; Yoon, J.W. Development and validation of a novel AI framework using NLP with LLM integration for relevant clinical data extraction through automated chart review. *Sci. Rep.* **2024**, *14*, 26783. [CrossRef]
67. Kowalski, M.; Johnson, T.; Lee, J.; Patel, S.; Smith, A.; Davis, R.; Nguyen, L.; Thompson, B.; Martinez, C.; Chen, Y.; et al. Implementation of an All-Day Artificial Intelligence–Based Triage System in the Emergency Department. *Mayo Clin. Proc.* **2022**, *97*, 1234–1245. [CrossRef]
68. Nuance Communications. DAX Copilot to Automate the Creation of Clinical Documentation, Reduce Physician Burnout, and Expand Access to Care Deployed Enterprise-Wide at Stanford Health Care. *Nuance Newsroom*, 11 March 2024. Available online: https://news.nuance.com/2024-03-11-DAX-Copilot-to-Automate-the-Creation-of-Clinical-Documentation,-Reduce-Physician-Burnout,-and-Expand-Access-to-Care-Deployed-Enterprise-Wide-at-Stanford-Health-Care (accessed on 16 June 2025).
69. RTB AI. LLM Cost Explained: How to Estimate the Price of Using Large Language Models. *RTB AI Blog*, 2024. Available online: https://www.rtb-ai.com/post/how-to-estimate-the-cost-of-using-a-large-language-model (accessed on 16 June 2025).
70. Sun, X.; Ma, R.; Zhao, X.; Li, Z.; Lindqvist, J.; El Ali, A.; Bosch, J.A. Trusting the Search: Unraveling Human Trust in Health Information from Google and ChatGPT. *arXiv* **2024**, arXiv:2403.09987.
71. Rezaeian, O.; Asan, O.; Bayrak, A.E. The Impact of AI Explanations on Clinicians' Trust and Diagnostic Accuracy in Breast Cancer. *arXiv* **2024**, arXiv:2412.11298. [CrossRef] [PubMed]
72. Branley-Bell, D.; Talbot, C.V. Exploring the Impact of Chatbot Design on Trust and Engagement in Mental Health Support. *JMIR Ment. Health* **2021**, *8*, e23429.
73. Chin, H.; Song, H.; Baek, G.; Jung, C. The Potential of Chatbots for Emotional Support and Promoting Mental Well-Being in Different Cultures: Mixed Methods Study. *J. Med. Internet Res.* **2023**, *25*, e46901. [CrossRef]

74. Aljamaan, F.; Temsah, M.H.; Altamimi, I.; Al-Eyadhy, A.; Jamal, A.; Alhasan, K.; Mesallam, T.A.; Farahat, M.; Malki, K.H. Reference hallucination score for medical artificial intelligence chatbots: Development and usability study. *JMIR Med. Inform.* **2024**, *12*, e54345. [CrossRef] [PubMed]
75. Johri, S.; Jeong, J.; Tran, B.A.; Schlessinger, D.I.; Wongvibulsin, S.; Barnes, L.A.; Zhou, H.Y.; Cai, Z.R.; Van Allen, E.M.; Kim, D.; et al. An evaluation framework for clinical use of large language models in patient interaction tasks. *Nat. Med.* **2025**, *31*, 77–86. [CrossRef]
76. Wu, S.; Zhao, S.; Yasunaga, M.; Huang, K.; Cao, K.; Huang, Q.; Ioannidis, V.; Subbian, K.; Zou, J.Y.; Leskovec, J. STaRK: Benchmarking LLM retrieval on textual and relational knowledge bases. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 127129–127153.
77. Ohde, J.W.; Rost, L.M.; Overgaard, J.D. The burden of reviewing LLM-generated content. *NEJM AI* **2025**, *2*, AIp2400979. [CrossRef]
78. Menz, B.D.; Kuderer, N.M.; Bacchi, S.; Modi, N.D.; Chin-Yee, B.; Hu, T.; Rickard, C.; Haseloff, M.; Vitry, A.; McKinnon, R.A.; et al. Current safeguards, risk mitigation, and transparency measures of large language models against the generation of health disinformation: Repeated cross-sectional analysis. *BMJ* **2024**, *384*, e078538. [CrossRef]
79. Chustecki, M. Benefits and risks of AI in health care: Narrative review. *Interact. J. Med. Res.* **2024**, *13*, e53616. [CrossRef] [PubMed]
80. Yeo, Y.H.; Peng, Y.; Mehra, M.; Samaan, J.; Hakimian, J.; Clark, A.; Suchak, K.; Krut, Z.; Andersson, T.; Persky, S.; et al. Evaluating for evidence of sociodemographic bias in conversational AI for mental health support. *Cyberpsychol. Behav. Soc. Netw.* **2025**, *28*, 44–51. [CrossRef] [PubMed]
81. Tan, D.N.H.; Tham, Y.-C.; Koh, V.; Loon, S.C.; Aquino, M.C.; Lun, K.; Cheng, C.-Y.; Ngiam, K.Y.; Tan, M. Evaluating chatbot responses to patient questions in the field of glaucoma. *Front. Med.* **2024**, *11*, 1359073. [CrossRef] [PubMed]
82. Chhikara, G.; Sharma, A.; Ghosh, K.; Chakraborty, A. Few-shot fairness: Unveiling LLM's potential for fairness-aware classification. *arXiv* **2024**, arXiv:2402.18502.
83. Ito, N.; Kadomatsu, S.; Fujisawa, M.; Fukaguchi, K.; Ishizawa, R.; Kanda, N.; Kasugai, D.; Nakajima, M.; Goto, T.; Tsugawa, Y. The Accuracy and Potential Racial and Ethnic Biases of GPT-4 in the Diagnosis and Triage of Health Conditions: Evaluation Study. *JMIR Med. Educ.* **2023**, *9*, e47532. [CrossRef]
84. Arnason, T.J.; Mirza, K.M.; Lilley, C.M. Assessing the Impact of Race, Sexual Orientation, and Gender Identity on USMLE Style Questions: Preliminary Results of a Randomized Controlled Trial. *Am. J. Clin. Pathol.* **2023**, *160* (Suppl. S1), S64–S65. [CrossRef]
85. Rawat, R.; McBride, H.; Nirmal, D.; Ghosh, R.; Moon, J.; Alamuri, D.; O'Brien, S.; Zhu, K. DiversityMedQA: Assessing Demographic Biases in Medical Diagnosis using Large Language Models. *arXiv* **2024**, arXiv:2409.01497.
86. Rathod, V.; Nabavirazavi, S.; Zad, S.; Iyengar, S.S. Privacy and security challenges in large language models. In Proceedings of the 2025 IEEE 15th Annual Computer Communication Workshop Conference (CCWC), Las Vegas, NV, USA, 6–8 January 2025; pp. 00746–00752.
87. Feretzakis, G.; Verykios, V.S. Trustworthy AI: Securing sensitive data in large language models. *AI* **2024**, *5*, 2773–2800. [CrossRef]
88. Srinivasan, A.; Patil, R. Navigating the challenges and opportunities of AI and LLM integration in cloud computing. *Baltic Multidiscip. Res. Lett. J.* **2025**, *2*, 8–15.
89. Siemens, W.; von Elm, E.; Binder, H.; Böhringer, D.; Eisele-Metzger, A.; Gartlehner, G.; Hanegraaf, P.; Metzendorf, M.-I.; Mosselman, J.-J.; Nowak, A.; et al. Opportunities, challenges and risks of using artificial intelligence for evidence synthesis. *BMJ Evid.-Based Med.* **2025**. [CrossRef]
90. Khan, Z.Y.; Hussain, F.K.; Kurniawan, D. Communication Efficiency and Non-Independent and Identically Distributed Data Challenge in Federated Learning: A Systematic Mapping Study. *Appl. Sci.* **2024**, *14*, 2720. [CrossRef]
91. Xu, H.; Zhang, J.; Xu, Y.; Liu, Z.; Ding, S.; Zhang, Y. FedVCK: Non-IID Robust and Communication-Efficient Federated Learning via Valuable Condensed Knowledge for Medical Image Analysis. *arXiv* **2024**, arXiv:2412.18557.
92. Karimireddy, S.P.; Kale, S.; Mohri, M.; Reddi, S.; Stich, S.U.; Suresh, A.T. SCAFFOLD: Stochastic Controlled Averaging for Federated Learning. *J. Mach. Learn. Res.* **2020**, *21*, 1–62.
93. Yao, Z.; Zhang, Z.; Tang, C.; Bian, X.; Zhao, Y.; Yang, Z.; Wang, J.; Zhou, H.; Jang, W.S.; Ouyang, F.; et al. MedQA-CS: Benchmarking large language models clinical skills using an AI-SCE framework. *arXiv* **2024**, arXiv:2410.01553.
94. Srinivasan, V.; Jatav, V.; Chandrababu, A.; Sharma, G. On the performance of an explainable language model on PubMedQA. *arXiv* **2025**, arXiv:2504.05074.
95. Wu, X.; Zhao, Y.; Zhang, Y.; Wu, J.; Zhu, Z.; Zhang, Y.; Ouyang, Y.; Zhang, Z.; Wang, H.; Lin, Z.; et al. MedJourney: Benchmark and evaluation of large language models over patient clinical journey. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 87621–87646.
96. Ch'en, P.Y.; Day, W.; Pekson, R.C.; Barrientos, J.; Burton, W.B.; Ludwig, A.B.; Jariwala, S.P.; Cassese, T. GPT-4 generated answer rationales to multiple choice assessment questions in undergraduate medical education. *BMC Med. Educ.* **2025**, *25*, 333. [CrossRef] [PubMed]
97. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. BLEU: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, PA, USA, 6–12 July 2002; pp. 311–318.

98. Van Veen, M.; Saavedra, J.; Wang, K. Clinical Text Summarization with LLM-Based Evaluation. Stanford University CS224N Final Report, 2024. Available online: https://web.stanford.edu/class/cs224n/final-reports/256989380.pdf (accessed on 16 June 2025).
99. Schroeder, N.L.; Davis Jaldi, C.; Zhang, S. Large Language Models with Human-In-The-Loop Validation for Systematic Reviews. *arXiv* **2025**, arXiv:2501.11840.
100. Drori, I.; Te'eni, D. Human-in-the-loop AI reviewing: Feasibility, opportunities, and risks. *J. Assoc. Inf. Syst.* **2024**, *25*, 98–109. [CrossRef]
101. Kumar, S.; Datta, S.; Singh, V.; Datta, D.; Singh, S.K.; Sharma, R. Applications, challenges, and future directions of human-in-the-loop learning. *IEEE Access* **2024**, *12*, 75735–75760. [CrossRef]
102. Turner, P.; Kushniruk, A.; Nohr, C. Are we there yet? Human factors knowledge and health information technology–the challenges of implementation and impact. *Yearb. Med. Inform.* **2017**, *26*, 84–91. [CrossRef] [PubMed]
103. Vismara, L.A.; McCormick, C.E.; Shields, R.; Hessl, D. Extending the parent-delivered Early Start Denver Model to young children with fragile X syndrome. *J. Autism Dev. Disord.* **2019**, *49*, 1250–1266. [CrossRef]
104. Mayo Clinic Press. *AI in Healthcare: The Future of Patient Care and Health Management*; Mayo Clinic Press Healthy Aging: Rochester, MN, USA, 2023. Available online: https://mcpress.mayoclinic.org/healthy-aging/ai-in-healthcare-the-future-of-patient-care-and-health-management/ (accessed on 16 May 2025).
105. Stanford HAI. Large Language Models in Healthcare: Are We There Yet? *Stanford HAI News*, 2023. Available online: https://hai.stanford.edu/news/large-language-models-healthcare-are-we-there-yet (accessed on 16 May 2025).
106. Plaza, B.C. Data sources (LLM) for a clinical decision support model (SSDC) using a healthcare interoperability resources (HL7-FHIR) platform for an ICU ecosystem. *Prim. Sci. Med. Public Health* **2024**, *5*, 3–12.
107. Nazi, Z.A.; Peng, W. Large language models in healthcare and medical domain: A review. *Informatics* **2024**, *11*, 57. [CrossRef]
108. Markus, A.F.; Kors, J.A.; Rijnbeek, P.R. The Role of Explainability in Creating Trustworthy Artificial Intelligence for Health Care: A Comprehensive Survey of the Terminology, Design Choices, and Evaluation Strategies. *arXiv* **2020**, arXiv:2007.15911. [CrossRef]
109. Bharati, S.; Mondal, M.R.H.; Podder, P. A Review on Explainable Artificial Intelligence for Healthcare: Why, How, and When? *arXiv* **2023**, arXiv:2304.04780. [CrossRef]
110. Ong, J.C.L.; Ning, Y.; Liu, M.; Ma, Y.; Liang, Z.; Singh, K.; Chang, R.T.; Vogel, S.; Lim, J.C.W.; Tan, I.S.K.; et al. Regulatory Science Innovation for Generative AI and Large Language Models in Health and Medicine: A Global Call for Action. *arXiv* **2025**, arXiv:2502.07794.
111. U.S. Food and Drug Administration. Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan. FDA 2021. Available online: https://www.fda.gov/media/145022/download (accessed on 16 June 2025).
112. Meskó, B.; Topol, E.J. The Imperative for Regulatory Oversight of Large Language Models (or Generative AI) in Healthcare. *NPJ Digit. Med.* **2023**, *6*, 120. [CrossRef]
113. Chow, J.C.; Li, K. Ethical considerations in human-centered AI: Advancing oncology chatbots through large language models. *JMIR Bioinform. Biotechnol.* **2024**, *5*, e64406. [CrossRef]
114. Ge, Z.; Huang, H.; Zhou, M.; Li, J.; Wang, G.; Tang, S.; Zhuang, Y. WorldGPT: Empowering LLM as multimodal world model. In Proceedings of the 32nd ACM International Conference on Multimedia, Melbourne, VIC, Australia, 28 October–1 November 2024; pp. 7346–7355.
115. Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; Gao, J. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 28541–28564.
116. Koleilat, T.; Asgariandehkordi, H.; Rivaz, H.; Xiao, Y. Medclip-samv2: Towards universal text-driven medical image segmentation. *arXiv* **2024**, arXiv:2409.19483.
117. Jeong, C. Fine-tuning and utilization methods of domain-specific LLMs. *arXiv* **2024**, arXiv:2401.02981.
118. Liu, F.; Zhou, H.; Gu, B.; Zou, X.; Huang, J.; Wu, J.; Li, Y.; Chen, S.S.; Hua, Y.; Zhou, P.; et al. Application of large language models in medicine. *Nat. Rev. Bioeng.* **2025**, *3*, 445–464. [CrossRef]
119. Lee, S.; Youn, J.; Kim, H.; Kim, M.; Yoon, S.H. CXR-LLaVA: A Multimodal Large Language Model for Interpreting Chest X-Ray Images. *Radiology* **2024**, *303*, 11339. [CrossRef] [PubMed]
120. Wang, Z.; Wu, Z.; Agarwal, D.; Sun, J. MedCLIP: Contrastive Learning from Unpaired Medical Images and Text. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP), Abu Dhabi, United Arab Emirates, 7–11 December 2022; pp. 3726–3738.
121. Li, J.; Zhou, Z.; Lyu, H.; Wang, Z. Large language models-powered clinical decision support: Enhancing or replacing human expertise? *Intell. Med.* **2025**, *5*, 1–4. [CrossRef]
122. Rajashekar, N.C.; Shin, Y.E.; Pu, Y.; Chung, S.; You, K.; Giuffrè, M.; Chan, C.E.; Saarinen, T.; Hsiao, A.; Sekhon, J.S.; et al. Human-algorithmic interaction using a large language model-augmented artificial intelligence clinical decision support system. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 11–16 May 2024; pp. 1–20.

123. Chow, J.C. Quantum computing and machine learning in medical decision-making: A comprehensive review. *Algorithms* **2025**, *18*, 156. [CrossRef]
124. Nguyen, Q.N.; Sidorova, A.; Torres, R. User interactions with chatbot interfaces vs. menu-based interfaces: An empirical study. *Comput. Hum. Behav.* **2022**, *128*, 107093. [CrossRef]
125. Dutta, N.; Dhar, D. Investigating usability of conversational user interfaces for integrated system-physical interactions: A medical device perspective. *Int. J. Hum. Comput. Interact.* **2025**, *41*, 271–304. [CrossRef]
126. Chen, T.L.; Kuo, C.H.; Chen, C.H.; Chen, H.S.; Liu, Y.H. Development of an intelligent hospital information chatbot and evaluation of its system usability. *Enterp. Inf. Syst.* **2025**, *19*, 2464746. [CrossRef]
127. Yin, R.; Neyens, D.M. Examining how information presentation methods and a chatbot impact the use and effectiveness of electronic health record patient portals: An exploratory study. *Patient Educ. Couns.* **2024**, *119*, 108055. [CrossRef]
128. Vasileiou, M.V.; Maglogiannis, I.G. The health chatbots in telemedicine: Intelligent dialog system for remote support. *J. Healthc. Eng.* **2022**, *2022*, 4876512. [CrossRef]
129. Kurniawan, M.H.; Handiyani, H.; Nuraini, T.; Hariyati, R.T.; Sutrisno, S. A systematic review of artificial intelligence-powered (AI-powered) chatbot intervention for managing chronic illness. *Ann. Med.* **2024**, *56*, 2302980. [CrossRef]
130. Li, J. Security implications of AI chatbots in health care. *J. Med. Internet Res.* **2023**, *25*, e47551. [CrossRef] [PubMed]
131. Jalali, N.A.; Hongsong, C. Comprehensive framework for implementing blockchain-enabled federated learning and full homomorphic encryption for chatbot security system. *Clust. Comput.* **2024**, *27*, 10859–10882. [CrossRef]
132. Kanter, G.P.; Packel, E.A. Health care privacy risks of AI chatbots. *JAMA* **2023**, *330*, 311–312. [CrossRef]
133. Azam, A.; Naz, Z.; Khan, M.U. PharmaLLM: A medicine prescriber chatbot exploiting open-source large language models. *Hum. Cent. Intell. Syst.* **2024**, *4*, 527–544. [CrossRef]
134. Lee, H.; Kang, J.; Yeo, J. Medical specialty recommendations by an artificial intelligence chatbot on a smartphone: Development and deployment. *J. Med. Internet Res.* **2021**, *23*, e27460. [CrossRef]
135. Zhu, L.; Mou, W.; Luo, P. Ensuring safety and consistency in artificial intelligence chatbot responses. *JAMA Oncol.* **2024**, *10*, 1597. [CrossRef]
136. Osifowokan, A.S.; Agbadamasi, T.O.; Adukpo, T.K.; Mensah, N. Regulatory and legal challenges of artificial intelligence in the US healthcare system: Liability, compliance, and patient safety. *World J. Adv. Res. Rev.* **2025**, *25*, 949–955. [CrossRef]
137. Warraich, H.J.; Tazbaz, T.; Califf, R.M. FDA perspective on the regulation of artificial intelligence in health care and biomedicine. *JAMA* **2025**, *333*, 241–247. [CrossRef]
138. Hey, C.; Hunter, A.; Muller, M.; Le Roux, N.; Bennett, S.; Moulon, I.; Simoens, S.; Eichler, H.-G. A future European scientific dialogue regulatory framework: Connecting the dots. *Clin. Ther.* **2024**, *46*, 293–299. [CrossRef] [PubMed]
139. Palaniappan, K.; Lin, E.Y.; Vogel, S. Global regulatory frameworks for the use of artificial intelligence (AI) in the healthcare services sector. *Healthcare* **2024**, *12*, 562. [CrossRef] [PubMed]
140. Pham, T. Ethical and legal considerations in healthcare AI: Innovation and policy for safe and fair use. *R. Soc. Open Sci.* **2025**, *12*, 241873. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.