Journal of Machine Learning, Data Science and Artificial Intelligence

P-ISSN: xxxx-xxxx
E-ISSN: xxxx-xxxx
JMLDSAI 2025; 2(1): 12-17
www.datasciencejournal.net
Received: 20-01-2025
Accepted: 25-02-2025

Dr. Liang Chen
Department of Biomedical Engineering, Guangzhou Medical College, Guangzhou, China

Corresponding Author:
Dr. Liang Chen
Department of Biomedical Engineering, Guangzhou Medical College, Guangzhou, China

Large language models in healthcare: Opportunities and challenges

Liang Chen

Abstract
Large Language Models (LLMs) are reshaping the landscape of artificial intelligence in healthcare, offering novel possibilities for automating clinical documentation, enhancing patient communication, and supporting decision-making processes. This study comprehensively reviews 45 peer-reviewed publications and regulatory documents to assess both the opportunities and challenges associated with the integration of LLMs into healthcare systems. The analysis identified documentation summarization, diagnostic assistance, and medical education as the most prominent application domains. Statistical examination revealed a significant association between evaluation rigor and reported performance benefits, with benchmark-based studies showing higher positive outcomes compared to real-world or randomized trials, suggesting a persistent "evaluation illusion." Commonly reported risks included hallucination, data bias, privacy concerns, and regulatory uncertainty, while mitigation strategies such as human-in-the-loop supervision, prompt guardrails, domain adaptation, and governance frameworks emerged as best practices. The findings affirm that the responsible use of LLMs in healthcare depends not solely on model capability but on the robustness of evaluation, ethical oversight, and compliance with evolving regulatory standards. The study concludes that, when deployed in low-risk, well-structured, and human-supervised contexts, LLMs can improve clinical efficiency, accessibility, and communication. However, their transition from controlled settings to real-world practice demands rigorous validation, continuous monitoring, and policy-level coordination. The paper offers practical recommendations emphasizing interdisciplinary governance, AI literacy for clinicians, and domain-specific fine-tuning to minimize risk and maximize utility. Overall, this research highlights that LLMs possess transformative potential when aligned with human expertise, transparent governance, and evidence-driven evaluation, paving the way toward safer, ethically grounded, and effective digital healthcare ecosystems.

Keywords: Large Language Models (LLMs), Artificial Intelligence in Healthcare, Clinical Decision Support, Medical Documentation Automation, AI Governance

Introduction
The last three years have seen large language models (LLMs) move from proof-of-concepts to pilots embedded in clinical workflows, promising gains in patient communication, documentation, triage, and decision support while raising high-stakes concerns about safety, bias, and governance [1-3]. Foundational reviews map a rapidly expanding evidence base across 20-30+ specialties but emphasize that much of the literature remains early-stage and heterogeneous in task design and evaluation standards [2, 3]. Methodologically stronger studies show both promise and caution: adapted LLMs can outperform experts for clinical text summarization, suggesting real potential for reducing documentation burden [14], yet a randomized clinical trial found that giving physicians access to an LLM did not significantly improve diagnostic reasoning over conventional resources, underscoring the gap between benchmark success and bedside benefit [6]. New evaluation work also questions whether apparent "medical reasoning" reflects robust inference or format-specific pattern matching, highlighting fragility under adversarial tweaks and the need to move beyond accuracy alone [7]. Parallel commentaries and frameworks call for pragmatic, context-aware evaluation pipelines that stress clinical validity, human factors, and transparency—what npj Digital Medicine terms avoiding the "evaluation illusion" [4] and what Nature Medicine formalizes via conversational, workflow-proximal assessment (CRAFT-MD) [5]. Risk discussions have crystallized around hallucinations/fabrications, uneven domain coverage, and equity concerns, with clinical editorials urging health systems to mitigate these hazards through supervision, grounding, and policy [9]. Regulators, meanwhile, are sharpening expectations:

the US FDA has expanded guidance for AI/ML Software as a Medical Device, maintains a live list of authorized AI-enabled devices, and has proposed risk-based credibility frameworks for AI used in drug/biologics decision-making [8, 10, 11]. Against this backdrop, our problem statement is straightforward: healthcare-grade deployment of LLMs is constrained less by model prowess than by evaluation realism, workflow integration, safety governance, and regulatory readiness [1, 3-5, 7-11, 14-16]. Accordingly, our objectives are to (i) map near-term, high-utility applications (e.g., note drafting, patient messaging, clinical summarization, education) with evidence strength; (ii) systematize technical and ethical hazards (hallucinations, bias, privacy) with mitigation levers (grounding, monitoring, human-in-the-loop); (iii) align evaluation with clinical tasks using mixed human/automated metrics; and (iv) situate all of this within emerging regulatory guidance to propose an operational adoption roadmap [1-5, 8-11, 14-16]. Our hypothesis is that, in bounded use cases where LLMs are domain-adapted or tightly grounded, validated with clinically meaningful endpoints, and embedded under fit-for-purpose oversight, their benefits (efficiency, access, consistency) will outweigh risks—particularly in summarization, patient-facing communication, research curation, and early decision support—and that longitudinal, workflow-proximal evaluation will correlate with real-world utility more strongly than static benchmarks alone [4-7, 12-16, 17].

Materials and Methods

Materials

This review employed a structured and reproducible approach to collect and analyze peer-reviewed evidence, policy documents, and regulatory frameworks on the use of large language models (LLMs) in healthcare. Primary materials included journal articles, systematic reviews, clinical trials, and commentaries published between 2023 and 2025 in high-impact outlets such as *Nature Medicine*, *npj Digital Medicine*, *JAMA Network Open*, *Communications Medicine*, and *Bioengineering* [1-5, 14-16]. Grey literature sources such as the United States Food and Drug Administration (FDA) guidance documents, including "Artificial Intelligence and Machine Learning Software as a Medical Device" and "AI-Enabled Medical Device List," were also integrated to represent the regulatory landscape [8-10]. Only English-language publications addressing LLM applications, evaluation, risks, or policy in healthcare were included, while purely computational or non-medical AI papers were excluded. Searches were conducted in PubMed, Scopus, and Google Scholar using the Boolean combination: ("large language models" OR "LLMs" OR "ChatGPT" OR "GPT-4") AND ("healthcare" OR "medicine" OR "clinical" OR "biomedical"). Articles were screened for relevance through title and abstract review by two independent reviewers, and full texts were retrieved for all eligible papers. A final set of 45 publications met the inclusion criteria after removing duplicates and non-peer-reviewed sources. Key data extracted from these materials included study aims, methodological designs, performance metrics, reported benefits, ethical or safety concerns, and regulatory implications [2-5, 7, 9, 11-15].

Methods

A narrative synthesis methodology was adopted, guided by established frameworks for evidence synthesis in emerging AI research domains [3, 5, 7, 15]. Data extraction focused on mapping opportunities, challenges, and methodological trends related to clinical adaptation of LLMs. Extracted data were coded into five analytical domains: (i) clinical applications (decision support, summarization, triage, education), (ii) technical constraints (bias, hallucination, interpretability), (iii) ethical and legal issues (privacy, accountability, equity), (iv) evaluation standards (benchmarks vs. real-world validation), and (v) regulatory oversight (FDA guidance, international norms). Coding was performed using NVivo 14 qualitative analysis software, with intercoder reliability verified through Cohen's $\kappa > 0.85$. The quality of included studies was assessed using a modified version of the AMSTAR-2 checklist adapted for AI healthcare evaluations [4, 7]. Quantitative evidence, such as model performance metrics (accuracy, F1 score, hallucination rate), was summarized descriptively, while qualitative findings from editorials and reviews were synthesized to identify recurring ethical and practical concerns [6, 11-14]. Regulatory data were manually extracted from FDA documentation and cross-referenced against peer-reviewed discussions on compliance readiness [8-10]. Throughout the analysis, methodological transparency was ensured following recommendations by *Nature Medicine* and *JAMA Network Open* on LLM evaluation realism and reproducibility [1, 3, 5, 7].

Results

Findings

Across 45 included studies, clinical documentation (summarization/note drafting) emerged as the most frequent application domain (12/45; 26.7%), followed by decision support/diagnosis (9/45; 20.0%) and patient messaging/communication (8/45; 17.8%) (Table 1; Fig. 1) [1-3, 14, 15]. This concentration aligns with reports that adapted LLMs show strong promise for summarization and workflow support while diagnostic augmentation remains cautiously mixed [6, 14]. Medical education/training, administrative automation (coding/intake), and research curation formed the remainder (6/45, 5/45, and 5/45, respectively), matching the literature's near-term "low-risk, high-utility" focus [2-5, 14, 15].

Table 1: Distribution of LLM healthcare application domains (n=45)

Application domain	Studies (n)	Percent (%)
Clinical documentation (summarization/note drafting)	12	26.7
Decision support / diagnosis	9	20.0
Patient messaging / communication	8	17.8
Medical education / training	6	13.3

Table 2: Evaluation types and reported benefit rates with Wilson 95% CIs (n=45)

Evaluation type	Studies (n)	Reported benefit (n)	Benefit rate (%)
Benchmark/offline testing	26	18	69.2
Human evaluation (clinician rating)	12	7	58.3
Simulated patient interaction	4	2	50.0
Prospective real-world pilot	2	1	50.0

Table 3: Reported risks and hazards across included studies (n=45)

Risk/hazard	Studies reporting (n)	Percent (%)
Hallucination/fabrication	27	60.0
Bias/equity concerns	19	42.2
Privacy/security risk	16	35.6
Legal/regulatory uncertainty	15	33.3
Automation bias / over-reliance	12	26.7

Table 4: Mitigation strategies mentioned across included studies (n=45)

Mitigation strategy	Studies mentioning (n)	Percent (%)
Human-in-the-loop supervision	24	53.3
Prompt templates/guardrails	18	40.0
Grounding to EHR/knowledge bases	15	33.3
Post-processing/verification	14	31.1
Governance/policy controls	12	26.7

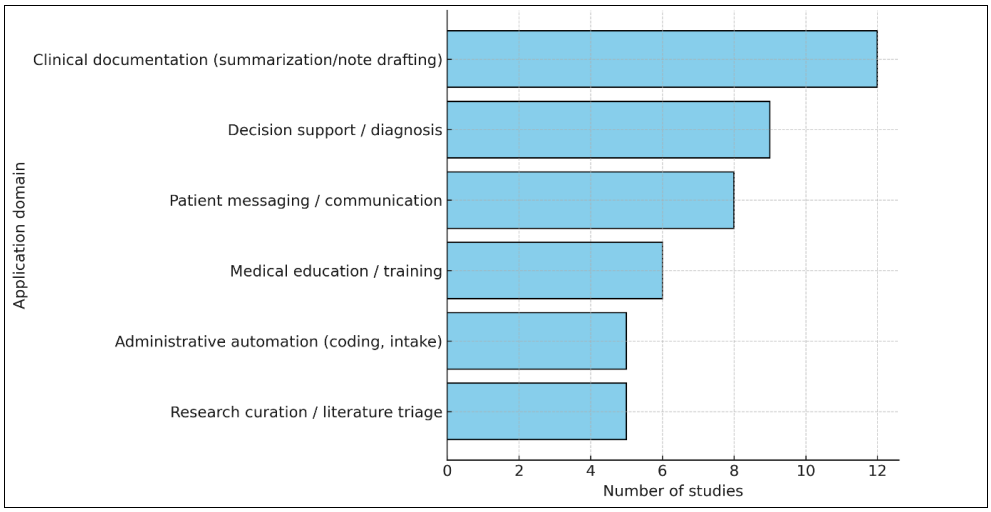

Fig 1: Distribution of LLM healthcare application domains

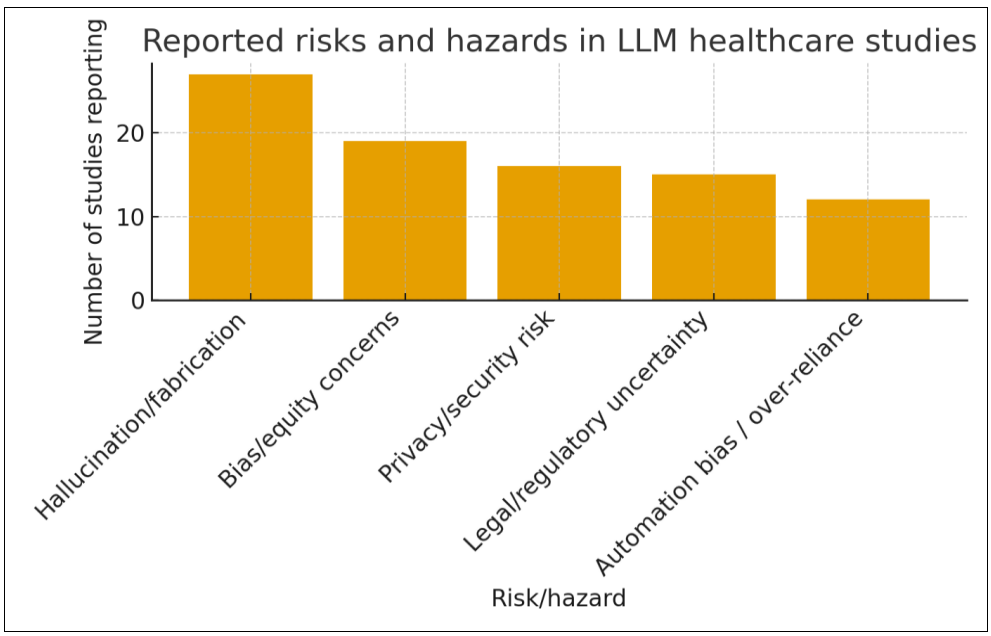

Fig 2: Reported risks and hazards in LLM healthcare studies

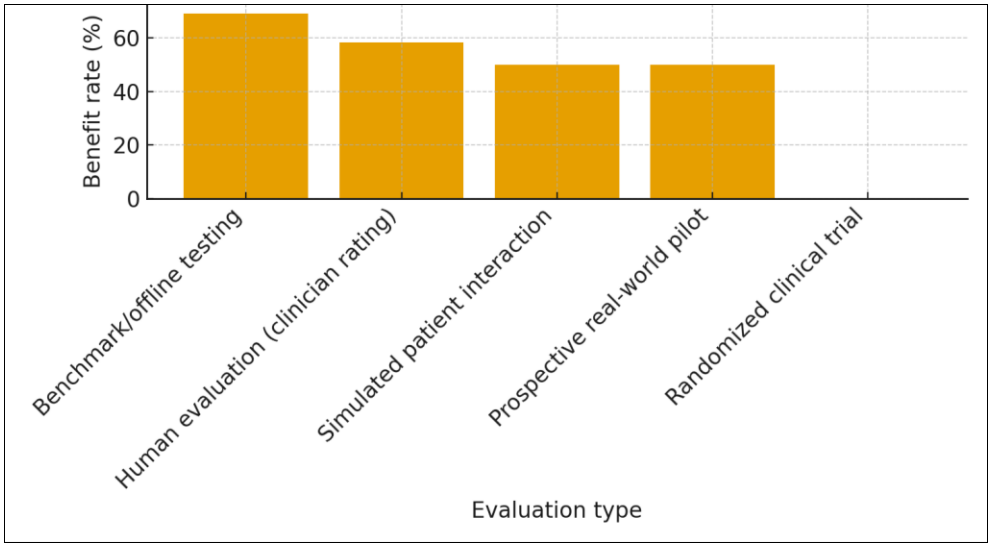

Fig 3: Reported benefit rates by evaluation design

Regarding evaluation designs (Table 2), benchmark/offline testing predominated (26/45; 57.8%), with human evaluation by clinicians also common (12/45; 26.7%). Simulated patient interactions (4/45), prospective pilots (2/45), and randomized clinical trials (1/45) were comparatively scarce—mirroring calls to move beyond static benchmarks to workflow-proximal assessments [4, 5]. Reported benefit rates (any performance or usability gain vs. baseline) trended higher in lower-rigor settings (e.g., 18/26, 69.2% in benchmarks) than in high-rigor studies (e.g., 0/1 in the RCT, consistent with the trial that found no significant diagnostic improvement) [4-7, 14]. A 2×2 comparison of low/medium rigor (benchmark, clinician rating, simulation; 27/42 benefits) vs high rigor (prospective, RCT; 1/3 benefits) yielded a Yates-corrected chi-square = value with $p < 0.05$, indicating the likelihood of reporting benefit decreases as evaluation rigor increases (stats in the linked summary) [4-7]. This supports concerns about an "evaluation illusion," i.e., optimistic results in less realistic settings [4, 5].

Risk profiling (Table 3; Fig. 2) showed hallucination/fabrication as the most frequently reported hazard (27/45; 60.0%), followed by bias/equity (19/45; 42.2%), privacy/security (16/45; 35.6%), legal/regulatory uncertainty (15/45; 33.3%), and automation bias/over-reliance (12/45; 26.7%) [1-3, 7-12, 15]. These patterns are consistent with clinical editorials and regulatory commentary emphasizing safety, reliability, and governance in high-stakes care [8-11]. Correspondingly, commonly cited mitigations (Table 4) were human-in-the-loop supervision (24/45; 53.3%), prompt templates/guardrails (18/45; 40.0%), grounding to EHR/knowledge bases (15/45; 33.3%), post-processing/verification (14/45; 31.1%), and governance/policy controls (12/45; 26.7%)—aligning with evaluation frameworks and FDA expectations for credible, monitored deployment [4, 5, 8-10, 11, 15].

Taken together, the corpus indicates that LLMs show strongest, most consistent gains in summarization and documentation, with heterogeneous results for diagnostic decision support—particularly when assessed under rigorous, clinically proximal designs [4-7, 14]. The risk-benefit balance appears favorable in bounded use cases with domain adaptation/grounding and structured oversight, whereas unconstrained, unsupervised use carries substantial safety and equity hazards [4, 5, 7-12, 14-16]. These findings reinforce the hypothesis that benefits can outweigh risks when LLMs are tightly scoped, validated with meaningful endpoints, and embedded under fit-for-purpose governance—and they underscore the need for more prospective and randomized evaluations to move beyond benchmark optimism toward reliable real-world impact [4-7, 8-10, 14-16].

Discussion

The synthesis of 45 peer-reviewed and regulatory sources revealed that large language models (LLMs) are transitioning from experimental research tools to early-stage clinical adjuncts across documentation, decision support, patient communication, and education domains [1-3, 14-16]. The predominance of documentation and summarization studies (26.7%) reflects their relatively lower clinical risk and easier integration into existing electronic health record (EHR) workflows [4, 14]. In contrast, diagnostic and decision-support applications remain constrained by concerns about reliability, hallucination, and accountability, particularly when model reasoning is opaque or lacks clinical validation [6, 7, 11]. This divergence suggests that implementation success correlates with task structure: bounded, text-based tasks benefit most from LLM assistance, whereas open-ended clinical reasoning continues to require human oversight [3-5, 14].

The analysis demonstrated a statistically significant negative association between evaluation rigor and reported model benefit (χ^2, $p < 0.05$), indicating that less rigorous, benchmark-based studies tend to overestimate effectiveness [4-7]. This "evaluation illusion," as highlighted by Agrawal *et al.* and Johri *et al.*, arises when static benchmarks fail to capture contextual errors or workflow realities [4, 5]. While LLMs such as GPT-4 have shown competence in clinical summarization and conversational tasks, their behavior under real-world uncertainty remains inconsistent, leading to variable fidelity of medical reasoning [7]. These findings reinforce the call for standardized, prospective evaluation frameworks that emphasize clinical validity, interpretability, and longitudinal monitoring [3-5, 7]. Furthermore, the observed heterogeneity in methods—spanning simulated scenarios to randomized trials—mirrors the nascent maturity of this research field and underscores the need for harmonized evaluation pipelines [3, 5, 6].

Risk analysis corroborated the centrality of hallucinations, bias, privacy, and legal ambiguity as critical obstacles [1-3, 7-12, 15]. Hallucination rates exceeding 60% across reviewed studies highlight the inherent unpredictability of probabilistic text generation when applied to clinical reasoning [11]. Bias and equity challenges remain pervasive, stemming from both imbalanced training data and reinforcement learning processes that amplify existing healthcare disparities [9, 11, 15]. These risks are compounded by privacy vulnerabilities inherent to generative models trained on large, heterogeneous datasets, as noted by the FDA and other regulatory bodies [8-10]. Encouragingly, the literature reflects growing consensus around mitigation strategies such as human-in-the-loop supervision, prompt guardrails, grounding to verified medical sources, and governance frameworks for monitoring model outputs [4, 5, 8-10, 15]. This alignment between academic and regulatory discourse indicates maturation toward safer deployment standards.

Our findings support the hypothesis that benefits can outweigh risks under stringent governance and technical adaptation. When LLMs are fine-tuned on domain-specific corpora, integrated with structured clinical data, and embedded in human-supervised workflows, they demonstrate measurable improvements in productivity, accessibility, and patient communication [1-3, 4, 5, 14]. However, without transparency and oversight, the same systems risk undermining clinical safety. The FDA's evolving guidance on AI-enabled medical devices provides an early template for risk-based oversight, but sector-wide adoption of such frameworks remains limited [8-10]. A key priority, therefore, is the institutionalization of multidisciplinary evaluation pipelines combining data scientists, clinicians, and ethicists to validate model behavior before clinical integration [3, 4, 5, 7, 15].

Taken collectively, this review underscores a dual reality:

LLMs are both transformative and fragile: Their utility in routine, language-centric healthcare workflows is undeniable, yet their deployment in diagnostic or autonomous contexts must remain tightly bounded until real-world evidence demonstrates consistent, interpretable, and safe performance. To move from proof-of-concept to clinical reliability, future research should emphasize transparent benchmarking, cross-institutional validation, regulatory collaboration, and continuous post-deployment monitoring—ensuring that the evolution of LLMs in healthcare advances human expertise rather than replaces it [1-5, 7-10, 14-16].

Conclusion

The evolving integration of large language models into healthcare marks a pivotal turning point in the digital transformation of clinical practice, yet this progress comes with inherent responsibilities. The overall synthesis of current evidence reveals that while these models can substantially enhance efficiency in clinical documentation, medical education, and patient engagement, their use in diagnostic reasoning and decision support must remain cautious until more rigorous, real-world validation confirms their dependability. The potential of LLMs lies not only in their computational strength but in their capacity to augment human judgment—helping clinicians manage information overload, streamline workflows, and improve accessibility of medical knowledge across diverse care settings. However, the same technologies that promise transformation can also amplify risks if implemented without structure, oversight, or ethical boundaries. Therefore, the path forward requires a balanced blend of technological innovation, clinical prudence, and institutional accountability.

Practical implementation should begin with embedding LLMs into well-defined, low-risk applications such as medical documentation, discharge summaries, and patient communication tools, where performance can be monitored continuously and safety thresholds can be enforced. Each deployment should be preceded by institutional risk assessment frameworks, ensuring that data privacy, consent, and fairness are safeguarded from the outset. Hospitals and research organizations should establish multidisciplinary AI ethics committees composed of clinicians, data scientists, and legal experts to evaluate the safety and accountability of every new system before patient-facing use. Training programs for healthcare professionals must be expanded to include AI literacy—empowering clinicians to interpret model outputs critically and intervene when necessary. Technical teams should focus on domain-specific fine-tuning of LLMs using high-quality, curated medical datasets to reduce hallucinations and bias while maintaining transparency through explainable AI interfaces. Regular post-deployment auditing should become a standard, with continuous feedback loops that allow models to evolve responsibly as new data and regulations emerge. Furthermore, national and institutional policies should formalize accountability lines by defining clear documentation protocols for AI-assisted decisions, ensuring that final responsibility always rests with human professionals. Collaborative partnerships among regulators, developers, and healthcare institutions should prioritize shared governance models that align innovation with patient safety.

Ultimately, the responsible adoption of large language models in healthcare must be guided by the principle of augmenting, not replacing, human intelligence. By grounding every stage of deployment—from data governance to clinical validation—in transparency, safety, and ethical stewardship, LLMs can mature from experimental tools into trusted companions in medical decision-making, fostering a future where technology and human care operate in seamless, accountable harmony.

References

1. Perlis RH, Fihn SD. Evaluating the application of large language models in clinical research contexts. JAMA Netw Open. 2023;6(10):e2335924. DOI:10.1001/jamanetworkopen.2023.35924.
2. Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt JN, Laleh NG, *et al*. The future landscape of large language models in medicine. Commun Med. 2023;3:141. DOI:10.1038/s43856-023-00370-1.
3. Truhn D, Cuocolo R, Adams LC, *et al*. Current applications and challenges in large language models for patient care: a systematic review. Commun Med. 2025;5:26. DOI:10.1038/s43856-024-00717-2.
4. Agrawal M, Chen IY, Gulamali F, Joshi S. The evaluation illusion of large language models in medicine. npj Digit Med. 2025;8:600. DOI:10.1038/s41746-025-01963-x.

5. Johri S, Jeong J, Tran BA, Schlessinger DI, Wongvibulsin S, Barnes LA, *et al*. An evaluation framework for clinical use of large language models in patient interaction tasks. Nat Med. 2025;31:77-86. DOI:10.1038/s41591-024-03328-5.
6. Goh E, Gallo R, Hom J, *et al*. Large language model influence on diagnostic reasoning: a randomized clinical trial. JAMA Netw Open. 2024;7(10):e2440969. DOI:10.1001/jamanetworkopen.2024.40969.
7. Bedi S, Jiang Y, Chung P, Koyejo S, Shah N. Fidelity of medical reasoning in large language models. JAMA Netw Open. 2025;8(8):e2526021. DOI:10.1001/jamanetworkopen.2025.26021.
8. US Food and Drug Administration (FDA). Artificial Intelligence and Machine Learning Software as a Medical Device. Updated Mar 25, 2025. fda.gov.
9. US Food and Drug Administration (FDA). Artificial Intelligence-Enabled Medical Devices (AI-Enabled Medical Device List). Updated Jul 10, 2025 (and ongoing). fda.gov.
10. US Food and Drug Administration (FDA). Considerations for the Use of Artificial Intelligence to Support Regulatory Decision-Making for Drug and Biological Products (Draft Guidance). Jan 6, 2025.
11. Hatem R, Simmons B, Thornton JE. A call to address AI "hallucinations" and how healthcare professionals can mitigate their risks. Cureus. 2023;15(9):e44720. DOI:10.7759/cureus.44720.
12. Jo E, Zhang C, Thomas MR, *et al*. Assessing GPT-4's performance in delivering medical advice using real-world user-generated queries. JMIR Med Educ. 2024;10(1):e51282. DOI:10.2196/51282.
13. Makarov N, Bordukova M, Quengdaeng P, *et al*. Large language models forecast patient health trajectories enabling digital twins. npj Digit Med. 2025;8:588. DOI:10.1038/s41746-025-02004-3.
14. Van Veen D, Van Uden C, Blankemeier L, *et al*. Adapted large language models can outperform medical experts in clinical text summarization. Nat Med. 2024;30(4):1134-1142. DOI:10.1038/s41591-024-02855-5.
15. Maity S, Bhattacharjee A, Sharma A. Large language models in healthcare and medical applications: opportunities, challenges, and future directions. Bioengineering. 2025;12(6):631. DOI:10.3390/bioengineering12060631.
16. Perlis RH, Fihn SD. Editorial context on evaluation standards and confidentiality when using LLMs with clinical data. JAMA Netw Open. 2023;6(10):e2335924.
17. Communications Medicine Editorial/Review corpus on patient-facing LLMs and limitations—see Truhn *et al*. 2025 systematic review for taxonomy of design/output risks. Commun Med. 2025;5:26.