REVIEW **Open access**

Large Language Models in Clinical Medicine from 2017 to 2025: A Systematic Review of Performance on Medical Licensing Examinations, Clinical Documentation, Decision Support, and Safety Concerns

Gabriel Costa[1*], Rafael Mendes[1], Bruno Teixeira[2], Lucas Ribeiro[1]

Abstract

Large language models (LLMs) have rapidly advanced since the transformer architecture was introduced in 2017, with systems such as GPT-3, GPT-4, Med-PaLM, and Claude increasingly explored for applications in medical education, clinical documentation, decision support, and patient communication, raising both optimism and concerns regarding safety and reliability. This systematic review synthesizes evidence across studies retrieved from PubMed, arXiv, ACL Anthology, IEEE Xplore, and Google Scholar that empirically evaluated LLMs in clinical settings using quantitative performance metrics, with risk of bias assessed using an adapted PROBAST framework for machine learning research. Findings show that LLMs achieve 60–90% accuracy on USMLE-style examinations, with leading models such as GPT-4 and Med-PaLM 2 reaching or surpassing passing thresholds, while in clinical documentation tasks they can reduce physician workload by approximately 30–50% in generating outputs such as discharge summaries, though human review remains consistently required. Performance in clinical decision support is more variable and specialty-dependent, and hallucination rates ranging from 5–30% have been reported, alongside persistent issues of bias and overconfidence in incorrect outputs. Overall, while LLMs demonstrate strong capabilities in structured medical knowledge tasks and documentation support, current limitations including hallucinations, bias, and lack of prospective clinical validation prevent safe autonomous deployment, making clinician oversight and robust safety safeguards essential for any clinical use.

Keywords Decision support, Large language models, Clinical documentation, Clinical medicine, USMLE, Hallucination

*Correspondence:
Gabriel Costa
gabriel.costa@gmail.com
[1] Department of Healthcare AI Engineering, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
[2] Department of Clinical Intelligence Systems, University of Campinas, Campinas, Brazil

Introduction

Large language models have evolved rapidly since the transformer architecture was introduced in 2017, with GPT-3 (2020), GPT-4 (2023), and Med-PaLM (2023) demonstrating capabilities that extend into clinical medicine and generating substantial interest in healthcare applications [1]. Domain-specific models such as Med-PaLM 2 have been fine-tuned on medical corpora, achieving performance approaching clinician levels on medical question-answering benchmarks and

prompting exploration of diverse clinical use cases [2]. The pace of model development and commercialization has, however, outstripped systematic safety evaluation, creating an evidence gap that this review addresses.

LLM applications in clinical medicine span four principal domains corresponding to distinct stages of clinical workflow and training: medical education and knowledge assessment, where models have been extensively evaluated on licensing examinations including the USMLE [3, 4]; clinical documentation, where LLMs generate discharge summaries, progress notes, and referral letters with reported time savings but persistent accuracy concerns [5, 6]; clinical decision support, encompassing differential diagnosis and treatment recommendations with variable reliability [7, 8]; and patient interaction, which raises additional challenges of empathy and safety.

Safety concerns documented across multiple studies include hallucination—factually incorrect outputs occurring at rates of 5–30% depending on domain—as well as bias arising from unrepresentative training data that may exacerbate healthcare disparities [9-11]. Additional concerns include overconfidence in incorrect outputs, vulnerability to adversarial prompts, privacy risks from training data memorization, and absent liability frameworks for AI-assisted clinical decisions [12, 13]. A systematic characterization of these concerns across the LLM evaluation literature is needed to inform clinical deployment and regulatory policy.

This review aims to synthesize evidence on LLM performance across medical licensing examinations, clinical documentation, decision support, and safety, and to characterize the nature and prevalence of documented safety concerns. Specific objectives are: first, to quantify LLM performance on standardized clinical knowledge assessments; second, to evaluate LLM effectiveness in documentation and decision support; and third, to systematically categorize safety concerns including hallucination, bias, and overconfidence [14].

Figure 1 illustrates the hierarchical evolution of large language models and their integration across key clinical domains, culminating in cross-domain safety constraints and the need for clinician oversight.

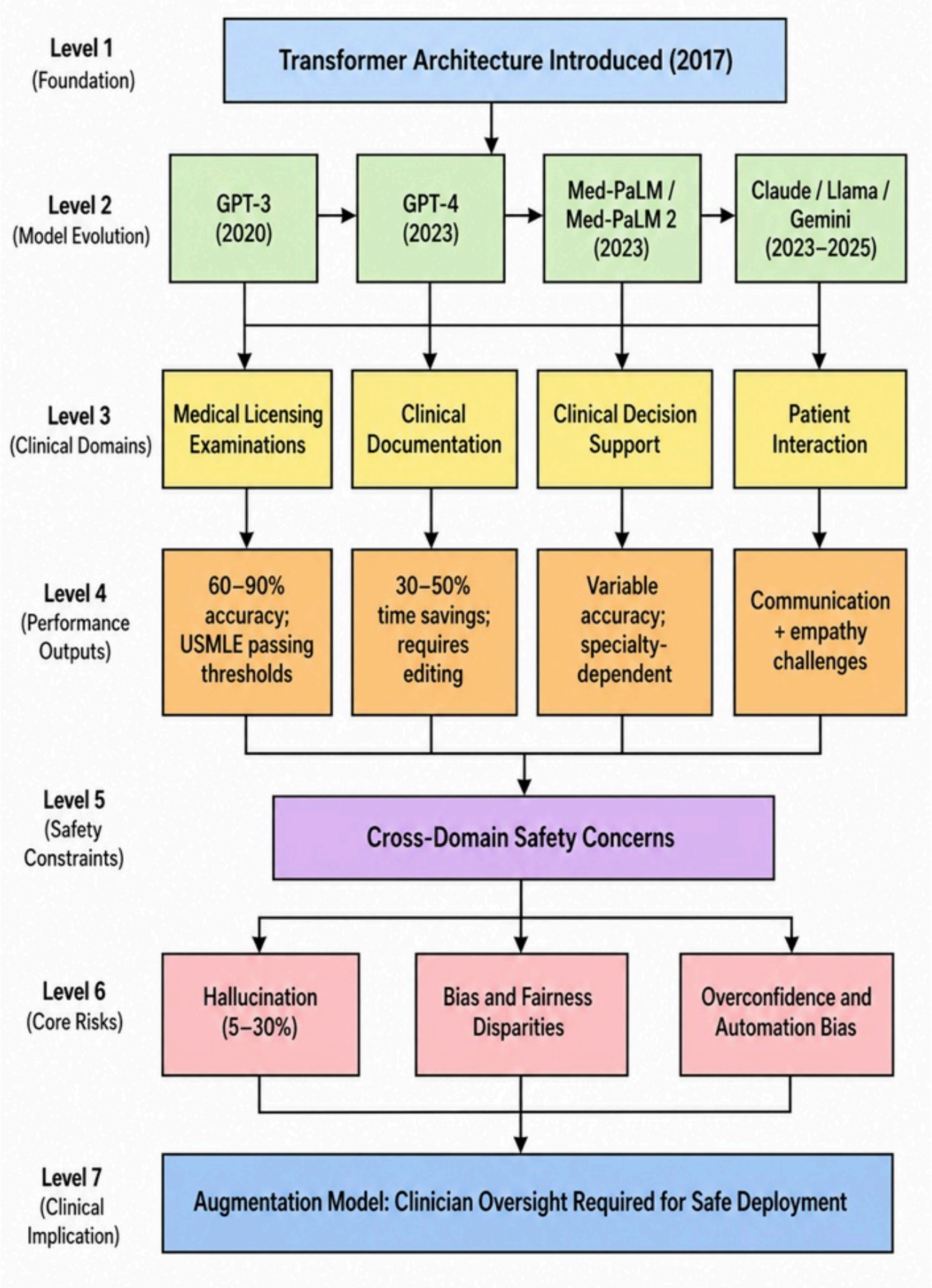

Level 1
(Foundation)
Transformer Architecture Introduced (2017)
Level 2
(Model Evolution)
GPT-3
(2020)
GPT-4
(2023)
Med-PaLM /
Med-PaLM 2
(2023)
Claude / Llama /
Gemini
(2023–2025)
Level 3
(Clinical Domains)
Medical Licensing
Examinations
Clinical
Documentation
Clinical Decision
Support
Patient
Interaction
Level 4
(Performance
Outputs)
60–90%
accuracy;
USMLE passing
thresholds
30–50%
time savings;
requires
editing
Variable
accuracy;
specialty-
dependent
Communication
+ empathy
challenges
Level 5
(Safety
Constraints)
Cross-Domain Safety Concerns
Level 6
(Core Risks)
Hallucination
(5–30%)
Bias and Fairness
Disparities
Overconfidence and
Automation Bias
Level 7
(Clinical
Implication)
Augmentation Model: Clinician Oversight Required for Safe Deployment

Figure 1. Evolution and Clinical Integration of Large Language Models in Medicine (2017–2025)

Materials and Methods

Search strategy

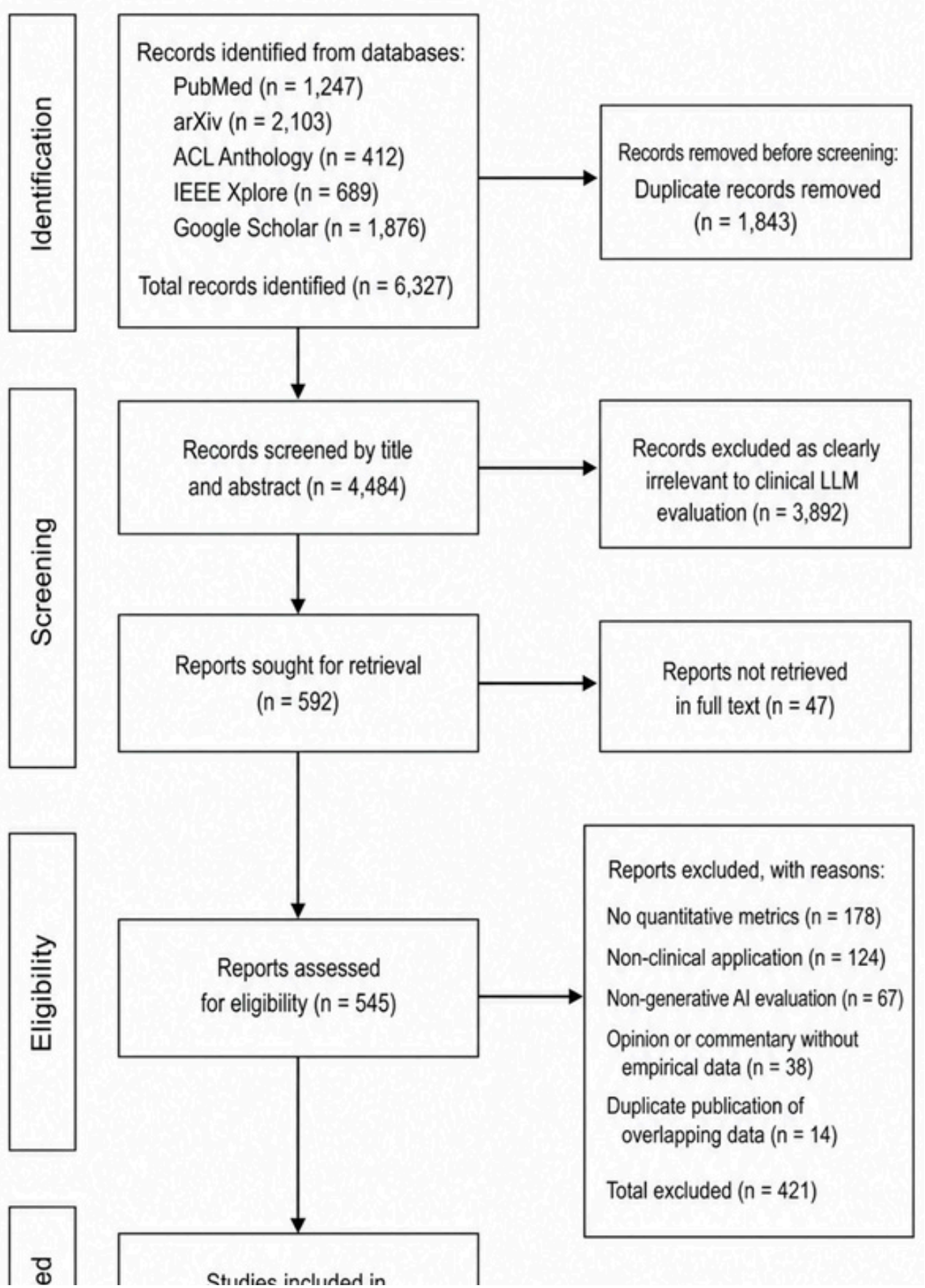
A systematic search was conducted across PubMed, arXiv, ACL Anthology, IEEE Xplore, and Google Scholar for studies published between January 2017 and December 2025. Search terms combined large language model identifiers (GPT-3, GPT-4, ChatGPT, Med-PaLM, Claude, Llama, Bard, Gemini) with clinical application terms (USMLE, medical licensing examination, clinical documentation, discharge summary, clinical decision support, differential diagnosis, hallucination, bias, safety). Reference lists of included studies and relevant reviews were hand-searched, and preprint servers were included to capture rapidly evolving research.

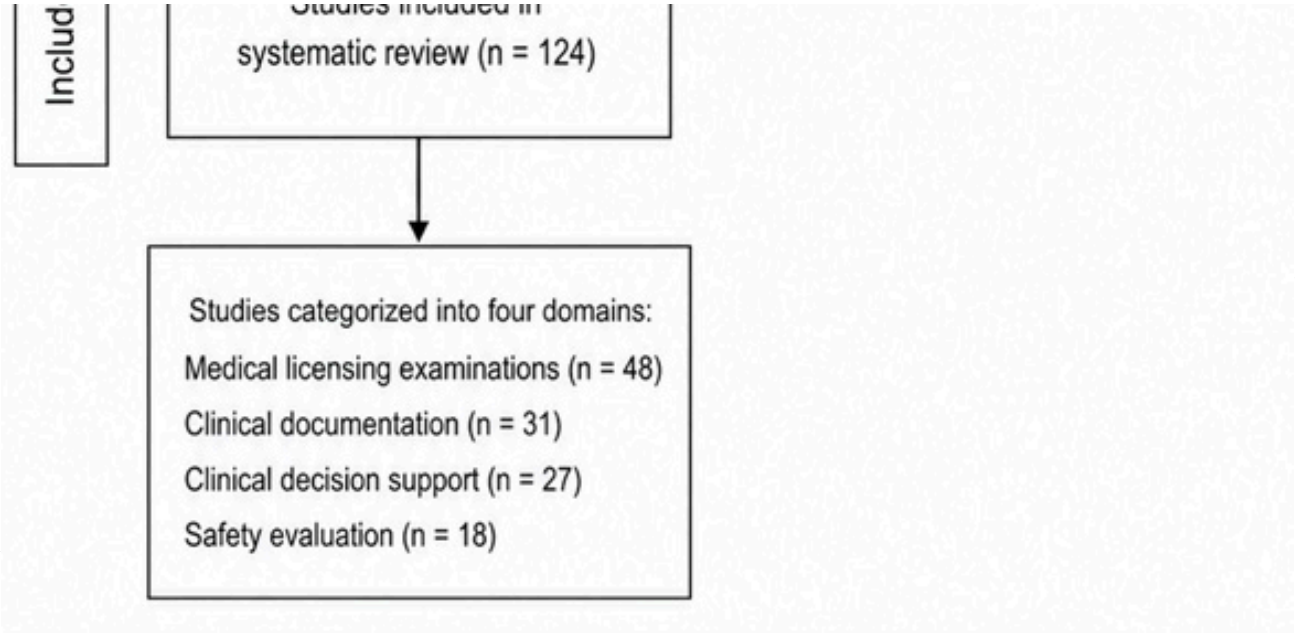
Inclusion and exclusion criteria

Studies were included if they provided empirical evaluation of at least one LLM on a clinical medicine task with quantitative performance metrics, were published in English, and had full text available. Eligible LLMs included GPT-3 through GPT-4, Med-PaLM series, Claude series, and Llama series models. Eligible clinical domains comprised medical licensing examinations, clinical documentation, decision support, and safety evaluation. Studies evaluating non-generative AI only, lacking clinical application, or without extractable quantitative results were excluded, as were opinion pieces without empirical data.

PRISMA flow diagram

The study selection process is summarized in **Figure 2** using a PRISMA 2020 flow diagram detailing identification, screening, eligibility, and inclusion stages.


Identification
Records identified from databases:
PubMed (n = 1,247)
arXiv (n = 2,103)
ACL Anthology (n = 412)
IEEE Xplore (n = 689)
Google Scholar (n = 1,876)
Total records identified (n = 6,327)
Records removed before screening:
Duplicate records removed
(n = 1,843)
Screening
Records screened by title
and abstract (n = 4,484)
Records excluded as clearly
irrelevant to clinical LLM
evaluation (n = 3,892)
Reports sought for retrieval
(n = 592)
Reports not retrieved
in full text (n = 47)
Eligibility
Reports assessed
for eligibility (n = 545)
Reports excluded, with reasons:
No quantitative metrics (n = 178)
Non-clinical application (n = 124)
Non-generative AI evaluation (n = 67)
Opinion or commentary without
empirical data (n = 38)
Duplicate publication of
overlapping data (n = 14)
Total excluded (n = 421)



ratings, and editing requirements were recorded. One reviewer performed initial extraction with verification by a second reviewer for all data points to ensure accuracy and completeness.

Risk of bias assessment

Risk of bias was assessed using an adapted PROBAST framework evaluating participant selection, predictor specification, outcome validity, and analysis adequacy. Particular attention was paid to benchmark contamination risk—the possibility that LLM training data included evaluation dataset questions—which would inflate apparent performance relative to genuine clinical reasoning capability. Each domain was rated as low, high, or unclear risk of bias, and ratings were incorporated into narrative synthesis interpretation [17].

Synthesis methods

Narrative synthesis was employed due to substantial heterogeneity in study designs, evaluation tasks, and performance metrics across the literature. Studies were grouped by clinical domain for primary synthesis and further stratified by model family, model generation, and evaluation methodology. Quantitative performance ranges are reported to characterize central tendencies, but formal meta-analysis was not conducted due to methodological heterogeneity and risk of overlapping evaluation datasets. Synthesis focused on identifying consistent performance patterns and documented safety concerns [18].

Table 1 provides a structured comparison of LLM performance characteristics across clinical domains, highlighting functional strengths and operational limitations

Table 1. Cross-Domain Performance and Functional Capabilities of Large Language Models in Clinical Medicine

Clinical Domain	Primary Tasks	Performance Range	Strengths	Limitations	Clinical Readiness
Medical Licensing Examinations	USMLE, MedQA, national exams	60–90% accuracy	Strong factual recall; benchmark consistency	Weak multi-step reasoning; possible dataset contamination	Moderate (education use)

Clinical Documentation	Discharge summaries, notes, referrals	30–50% time reduction	Efficiency gains; structured summarization	Requires full clinician editing; omission errors	High (assistive use only)
Clinical Decision Support	Diagnosis, triage, prescribing	Variable (50–80%)	Useful for common conditions; rapid synthesis	Poor rare disease handling; contextual errors	Low (requires supervision)
Patient Interaction	Explanation, communication	Qualitative (high fluency)	Clear language; scalable communication	Risk of misinformation; empathy inconsistency	Low (risk-sensitive use)

Results and Discussion

Medical licensing examinations

LLM performance on medical licensing examinations is the most extensively studied clinical application domain. Singhal *et al*. demonstrated that Med-PaLM achieved 67.6% accuracy on the MedQA dataset, with Med-PaLM 2 reaching 86.5% and approaching clinician-level performance [1, 2]. GPT-4 has achieved approximately 75–90% accuracy across USMLE Steps 1, 2, and 3, consistently exceeding the passing threshold [3, 4]. Systematic reviews confirm GPT-4 outperforms GPT-3.5 by 15–25 percentage points, while Claude and open-source models show lower but potentially useful performance [9, 10]. International examinations have also been evaluated, with GPT-4 achieving passing-level performance on the Japanese National Medical Licensing Examination [19].

Performance varies across question types, with LLMs performing better on factual recall than multi-step clinical reasoning. Wu *et al*. found that GPT-4 and Claude 2 showed accuracy variation across nephrology subtopics, performing better on electrolyte disorders than glomerular diseases [20]. Strong *et al.* compared LLM and medical student performance on free-response clinical reasoning examinations, finding that LLMs generated plausible responses but exhibited characteristic failures including difficulty with probabilistic reasoning and overreliance on presented information [14]. These findings suggest that multiple-choice examination accuracy may overestimate genuine clinical reasoning capability in unstructured environments.

Clinical documentation

LLM-assisted clinical documentation, particularly discharge summary generation, demonstrates substantial time savings with consistent need for clinician editing. Studies report that physicians rate AI-generated drafts favorably at approximately 4 out of 5 for usefulness, with documentation time reduced by 30–50% compared to manual composition [5, 6]. Song *et al.* evaluated an LLM assistant for emergency department discharge documentation and found reduced editing time relative to de novo composition, though physician editing remained universally necessary [11]. Williams *et al.* directly compared physician-generated and LLM-generated summaries, finding comparable overall quality but noting that LLM drafts were more comprehensive in capturing documented findings while occasionally including extraneous or incorrectly emphasized information [16].

Evaluation of other document types including progress notes and referral letters confirms similar patterns. Peng *et al.* and Yang *et al.* demonstrated that LLMs can generate coherent clinical notes from structured electronic health record data with acceptable accuracy for routine findings but higher error rates for abnormal results and medication details [21, 22]. Croxford *et al*. found that LLM-generated summaries captured key clinical concepts but exhibited systematic errors including omission of pertinent negatives and hallucination of undocumented findings [17]. Asgari *et al*. developed a structured safety framework identifying specific error categories—factual inaccuracies, omissions, and contradictions—that pose patient safety risks if undetected [23].

Clinical decision support

LLM decision support performance varies substantially by clinical context and specialty. Rutledge *et al.* found GPT-4 diagnostic accuracy above 80% for common presentations but substantially lower for rare diseases and atypical presentations [24]. Gaber *et al.* demonstrated that structured prompting improved triage and referral performance, though clinically significant errors persisted in complex cases requiring nuanced judgment [7]. Ong *et al.* evaluated prescribing error identification and found moderate sensitivity for common errors such as drug-drug interactions but poor performance on context-dependent errors requiring patient-specific factor integration [25].

Comparative studies reveal performance relative to clinicians. Katz *et al*. found GPT-4 outperformed residents on fact-based board questions but underperformed on management questions requiring contextual integration [26]. Shan *et al*. systematically reviewed diagnostic accuracy comparisons, finding LLMs approached non-specialist clinicians on structured tasks but consistently underperformed specialists, particularly for physical examination interpretation and visual diagnosis [27]. Jin *et al.* identified specific multimodal GPT-4 failure modes including medical image misinterpretation and plausible-sounding but incorrect interpretations of ambiguous visual data [28].

Hallucination rates

Hallucination—factually incorrect statements presented with confidence—is consistently documented across clinical LLM applications. Weissman *et al.* demonstrated that unregulated LLMs produce medical device-like decision support containing clinically significant errors in authoritative language [8]. Asgari *et al.* found error rates of 5–15% in LLM-generated medical notes, with higher rates for medication details, laboratory values, and temporal relationships [23]. Gaber *et al.* showed that retrieval-augmented generation reduced but did not eliminate hallucinated content, suggesting architectural limitations beyond current mitigation strategies [7].

The clinical significance of hallucination varies by use case, with different risk profiles for documentation versus diagnostic recommendations. Workum *et al*. found that LLMs frequently generated plausible but incorrect explanations for wrong answers, potentially misleading clinicians lacking domain expertise [29]. Lee *et al*. identified hallucination as a primary safety concern, noting that users may not distinguish accurate from hallucinated content due to authoritative linguistic style [5]. Peng *et al*. noted that clinician reviewers themselves may be susceptible to automation bias, potentially overlooking errors in well-structured AI-generated content [21].

Bias and fairness

Performance disparities across demographic groups raise concerns about LLM-exacerbated healthcare inequities. Singhal *et al*. identified accuracy variations related to patient race, ethnicity, and socioeconomic status indicators reflecting training data biases [2]. Peng *et al.* and Yang *et al.* discussed how overrepresentation of certain populations in medical literature may lead LLMs to generate less appropriate recommendations for underrepresented groups [21, 22]. Mbakwe *et al*. argued that strong USMLE performance partly reflects the examination's emphasis on pattern recognition over diverse-patient clinical reasoning, potentially masking bias issues [27].

Bias mitigation remains incompletely developed. Siam *et al*. found that performance disparities across question domains persisted despite prompt engineering interventions [30]. Wang *et al.* showed that prompts reducing one error type sometimes increased others, complicating universal mitigation strategies [31]. Chen *et al*. identified systematic error pattern differences between ChatGPT and Bard reflecting underlying training data and architecture differences [32]. Current evidence supports that clinical LLM bias is multidimensional, requiring attention to training data, model architecture, prompting, and deployment context.

Other safety concerns

Privacy risks, overconfidence, and adversarial vulnerabilities compound the safety concerns identified above. Weissman *et al*. and Lee *et al*. discussed potential training data memorization including protected health information inadvertently included in web-scraped corpora [5, 8]. Overconfidence—high confidence in incorrect outputs—makes errors harder to detect without independent verification [12]. Goh *et al*. demonstrated in a randomized trial that clinicians exposed to LLM diagnostic suggestions showed susceptibility to anchoring bias, being less likely to revise diagnoses despite contradictory evidence [6].

Adversarial vulnerabilities further complicate clinical deployment. Jin *et al*. identified susceptibility to misleading visual features that could degrade diagnostic performance [28]. Katz *et al*. noted that adversarial prompting could override safety guardrails to generate harmful medical advice [26]. Rust *et al*. found that LLM simplification of discharge summaries occasionally introduced clinically significant meaning distortions when simplifying complex medical language for patient-facing applications [13]. These diverse concerns underscore the need for comprehensive pre-deployment evaluation extending beyond accuracy to encompass robustness, privacy, and clinician-AI interaction effects.

Summary of principal findings

This review synthesizes evidence that LLMs have achieved substantial clinical capabilities while exhibiting safety limitations precluding autonomous deployment. On medical licensing examinations, GPT-4 and Med-PaLM 2 achieve 75–90% accuracy across USMLE components, exceeding passing thresholds [1, 2, 4]. In documentation, LLM assistance reduces time by 30–50% with favorable physician ratings, though clinician editing remains universally necessary [5, 6]. Decision support performance is variable and context-dependent, with moderate accuracy on structured questions but significant degradation on complex cases [7, 8]. Safety concerns including hallucination, bias, and overconfidence are consistently documented across domains and model versions, reflecting fundamental limitations rather than transient challenges.

The automation vs augmentation debate

Evidence strongly supports an augmentation paradigm wherein LLMs serve as assistive tools rather than autonomous decision-makers. Goh *et al.* demonstrated that LLM exposure influenced clinician reasoning in ways that could both help and hinder decision quality, with clinicians showing anchoring on LLM-suggested diagnoses [6]. Lee *et al*. argued GPT-4 is best conceptualized as an extender of clinician capability—processing information and drafting documentation under supervision [5]. Strong *et al*. found LLM performance characterized by distinct failure modes differing from human error patterns [14]. This complementary capability profile supports a collaborative model pairing LLM information synthesis with clinician contextual judgment rather than attempting full clinical decision-making replication.

Hallucination as a fundamental limitation

Consistent hallucination documentation across models, domains, and methodologies indicates this phenomenon represents a fundamental architectural limitation. Asgari *et al*. categorized hallucination types including fabrications, omissions, and logical contradictions that resist elimination through prompting or fine-tuning [23]. Gaber *et al*. found retrieval-augmented generation reduced but did not eliminate clinically significant errors, indicating hallucination arises from the language modeling objective of predicting plausible text rather than verifying factual accuracy [7]. Croxford *et al*. found error rates higher for clinically significant content than background information, while Rust *et al*. identified meaning-altering errors in patient-facing simplification [13, 17]. Peng *et al*. and Yang *et al*. noted automation bias may compound direct hallucination risk by reducing clinician vigilance [21, 22]. These findings argue for a precautionary approach assuming hallucination will occur and implementing robust detection workflows.

Table 2 synthesizes the multidimensional safety risks associated with clinical LLM deployment, providing a structured taxonomy of failure modes and their clinical implications

Table 2. Taxonomy of Safety Risks and Failure Modes in Clinical Large Language Model Applications

Safety Domain	Failure Mode	Description	Reported Range	Clinical Impact	Mitigation Status
Hallucination	Fabrication	Generation of false facts	5–30%	Misdiagnosis, incorrect documentation	Partial (RAG reduces but not eliminates)
Hallucination	Omission	Missing critical clinical details	Variable	Incomplete care decisions	Limited
Bias	Demographic disparity	Performance variation across populations	Not consistently quantified	Healthcare inequity risk	Early-stage
Overconfidence	False certainty	Incorrect outputs presented confidently	Common	Reduced clinician vigilance	Minimal
Automation Bias	Cognitive anchoring	Clinician over-reliance on AI	Demonstrated experimentally	Diagnostic error propagation	Unresolved
Privacy Risk	Data memorization	Leakage of sensitive training data	Unclear prevalence	Legal and ethical violations	Poorly addressed
Adversarial Vulnerability	Prompt manipulation	Safety guardrail bypass	Demonstrated	Harmful recommendations	Limited
Multimodal Error	Image misinterpretation	Incorrect visual analysis	Specialty-dependent	Diagnostic inaccuracies	Early-stage

Regulatory and liability implications

Current regulatory frameworks are not designed for generative AI systems producing variable outputs across clinical tasks. Weissman *et al.* argued unregulated LLMs effectively function as medical devices when generating diagnostic recommendations without pre-market safety evaluation [8]. Lee *et al*. discussed the need for frameworks accommodating iterative LLM improvement while maintaining safety oversight [10]. Liability models assuming a responsible human decision-maker are strained when LLM recommendations influence clinician judgment in difficult-to-audit ways [8]. Goh *et al*. experimentally demonstrated LLM influence on diagnostic reasoning, raising unresolved questions about liability apportionment when AI-influenced decisions cause harm [6]. Evidence supports that clinicians must retain ultimate responsibility, LLM outputs should be clearly labeled as AI-generated, and accountability frameworks must be clarified.

Limitations

Review limitations

Several methodological limitations affect this review. Publication bias is significant in the rapidly evolving LLM literature, as positive results are more likely to be published and indexed than negative findings [15]. Rapid model development creates temporal validity challenges, as benchmarks for specific model versions may be outdated upon publication [16]. Benchmark contamination—where training data inadvertently includes evaluation questions—represents a pervasive concern that may inflate apparent performance relative to genuine clinical reasoning capability, though studies cannot reliably verify its absence [27]. These limitations collectively suggest the published literature may overestimate LLM clinical capabilities relative to independent prospective evaluations.

Evidence base limitations

The evidence base itself constrains conclusions. Most studies were conducted in silico using benchmark datasets rather than in real clinical workflows with actual patients, limiting ecological validity [17]. Clinician-in-the-loop studies evaluating workflow impact and patient outcomes are rare, with few randomized trials identified [6]. Long-term safety outcomes following sustained clinical deployment have not been evaluated in any study [18]. Heterogeneity in evaluation methodologies, metrics, and reporting standards complicates synthesis and prevents formal meta-analysis for most outcomes. These limitations reflect the early stage of clinical LLM research and highlight the need for rigorous, clinically embedded evaluation with standardized reporting.

Comparison with prior reviews

Prior systematic reviews have focused on narrower domains, most commonly USMLE performance, without comprehensive safety integration. Brin *et al*. reviewed GPT model performance on the USMLE and confirmed GPT-4 consistently achieved passing scores, aligning with this review's findings [9]. Gilson et al. reviewed ChatGPT's USMLE performance and concluded LLMs showed potential for medical education while acknowledging clinical reasoning limitations [4]. Woo *et al*. systematically reviewed LLM use in clinical documentation, identifying consistent time savings and acceptable quality within nursing contexts but not extending to physician documentation or decision support [18]. These reviews provide domain-specific confirmation while highlighting the need for cross-domain synthesis integrating safety.

The present review extends prior work by providing integrated performance and safety synthesis essential for clinical readiness assessment. Whereas prior reviews reported examination performance without detailed characterization of hallucination, bias, or clinician-AI interaction effects, this review demonstrates these dimensions are inseparably linked: models achieving passing USMLE scores simultaneously generate hallucinations at 5–30% and exhibit demographic performance disparities [7, 23, 27]. Williams *et al*. and Croxford *et al*. identified documentation quality concerns complementing safety findings here but did not address broader bias, privacy, and interaction dimensions [16,17]. By synthesizing evidence across examination, documentation, decision support, and safety domains, this review provides a more comprehensive evidence base for deployment decisions than previously available.

Recommendations

For researchers

Researchers should adopt standardized safety evaluation frameworks assessing hallucination, bias, and overconfidence alongside accuracy metrics, reporting safety outcomes with equal prominence to performance results. Kresevic *et al*. demonstrated that optimizing clinical guideline interpretation required systematic attention to both accuracy and reliability [33]. Wang *et al*. showed prompt engineering could improve consistency for evidence-based questions, but optimization for one dimension sometimes degraded others, highlighting the need for comprehensive evaluation [31]. Researchers should report complete model specifications including version, fine-tuning strategy, and prompt templates, and prospectively register evaluation protocols to reduce selective reporting. Current practice of reporting only favorable metrics creates an incomplete evidence base that may encourage premature clinical deployment.

For journal editors and reviewers

Editors and reviewers should require systematic safety evaluation as a publication condition and apply heightened scrutiny to studies reporting only accuracy without addressing hallucination, bias, or robustness. Weissman *et al.* argued unregulated LLMs produce medical device-like decision support, underscoring publishing ecosystem responsibility to accurately characterize limitations [8]. Mbakwe *et al*. demonstrated that USMLE passing scores can create misleading impressions of readiness when safety is not simultaneously evaluated [27]. Publishing only favorable performance without safety assessment contributes to hype cycles pressuring clinical institutions toward premature adoption. Reviewers should require explicit disclosure of model limitations, performance reporting across clinically relevant subgroups, and discussion of error rates' clinical significance in intended use contexts.

For clinical institutions

Clinical institutions should implement staged deployment beginning with low-stakes applications under robust human oversight and establish clear AI governance structures. Goh *et al*. demonstrated LLM exposure influences clinician reasoning in both beneficial and potentially harmful ways, indicating deployment must consider clinician-AI interaction effects not predictable from bench evaluations [6]. Hains *et al*. evaluated LLM discharge summary preparation using real documentation and found that while time savings were achievable, clinician editing remained essential, suggesting universal human review rather than selective verification [15]. Institutions should train clinicians on known LLM failure modes—hallucination patterns, overconfidence, bias—as awareness is essential for effective oversight. Clear policies should govern documentation of LLM use and disclosure to patients, as transparency is both an ethical imperative and liability necessity.

For regulatory bodies

Regulatory bodies should develop specific evaluation frameworks for clinical LLMs addressing generative AI's unique characteristics, including adaptability across clinical contexts and variable outputs across populations. Ong *et al*. found meaningful accuracy variation across prescribing error types, illustrating the challenge of characterizing safety across clinical contexts [25]. Weissman *et al*. argued LLMs producing decision support function as de facto medical devices requiring oversight commensurate with potential harm [8]. Frameworks should require pre-market evaluation of hallucination rates, bias across demographics, and robustness before deployment above defined risk thresholds. Post-market surveillance should detect emerging safety signals during real-world use. Clear labeling should mandate AI-generated content be identifiable to enable verification, and liability frameworks should establish that clinicians remain accountable for decisions made with AI assistance

Research gaps

Prospective clinical trials

The most significant evidence gap is near-complete absence of prospective trials evaluating LLM impact on patient outcomes rather than in silico benchmarks that dominate the literature. Goh *et al*. conducted one of few randomized trials, evaluating diagnostic reasoning influence experimentally rather than in clinical practice with actual patient outcomes [6]. Gaber *et al*. evaluated decision support workflows retrospectively without prospective deployment [7]. Future research must prioritize randomized trials comparing LLM-assisted workflows to standard care with patient-relevant outcomes including diagnostic accuracy, time to treatment, medication errors, and morbidity endpoints. Trials should be powered for clinically meaningful differences and evaluate outcomes across diverse populations to detect differential effects that might exacerbate healthcare disparities.

Hallucination mitigation

Hallucination research has produced detection frameworks but not solutions reducing error rates to clinically acceptable levels for autonomous use. Asgari *et al*. developed structured safety assessment frameworks providing characterization tools but not yet resolution strategies [23]. Croxford *et al*. identified error patterns that could inform targeted mitigation but did not test interventions in clinical contexts [17]. Retrieval-augmented generation shows promise but, as Gaber *et al*. demonstrated, reduces rather than eliminates hallucinations, with clinically significant errors persisting despite access to verified knowledge [7]. Research is urgently needed on multimodal verification cross-referencing LLM outputs against structured data, medical fact-checking modules, confidence calibration enabling uncertainty recognition, and refusal mechanisms preventing speculative content when evidence is insufficient.

Real-world bias evaluation

Bias evaluation has relied on aggregate metrics masking clinically significant disparities, without prospective monitoring in real deployments. Singhal *et al.* evaluated Med-PaLM 2 across demographic subgroups on benchmark datasets, identifying variation warranting investigation but not characterizing real clinical manifestations [2]. Researchers

benchmarked multiple LLMs and identified performance variations potentially reflecting biases, but the examination format limited characterization of how disparities manifest in individual patient encounters [30]. Research is needed prospectively evaluating performance across populations varying by race, ethnicity, gender, language, socioeconomic status, and disease presentation. Mitigation strategies including training data curation and algorithmic fairness interventions must be validated specifically for clinical applications, with evaluation assessing whether clinically meaningful outcome disparities are reduced, developed in collaboration with historically underserved communities.

Implications

For research practice

The research landscape overemphasizes benchmark performance, particularly USMLE accuracy, with insufficient attention to safety, interaction effects, and workflow integration. Workum *et al*. and Siam *et al*. demonstrated performance varies meaningfully across evaluation methodologies, underscoring the need for standardized protocols capturing multiple capability dimensions [29, 30]. The community should shift from benchmark chasing toward clinically meaningful evaluation including safety outcomes, prospective studies, and long-term monitoring. Funding agencies should prioritize research addressing identified gaps rather than incremental benchmark improvements. Implementation science frameworks should be applied to understand how LLMs are actually used, barriers limiting deployment, and unintended consequences during sustained use.

For clinical practice

Evidence supports cautious LLM adoption as assistive tools for low-stakes applications while strongly cautioning against autonomous deployment for direct patient care. Song *et al*. demonstrated emergency department discharge documentation assistance reduced burden, and Ganzinger *et al*. showed structured data-based discharge summary generation was feasible with acceptable quality, but both required human review [11, 12]. Current LLMs are best deployed for documentation drafting, information summarization, and education where clinician verification integrates efficiently into workflows. For higher-stakes diagnostic or therapeutic applications, evidence does not support deployment outside controlled research settings. Clinicians should maintain awareness of documented failure modes and approach AI content with critical scrutiny appropriate for any clinical information source.

For policy and regulation

Weissman *et al.* provided compelling evidence that unregulated clinical LLMs constitute a regulatory gap with patient harm potential; this review confirms safety concerns are sufficiently prevalent to warrant oversight [8]. The FDA should classify LLMs for clinical decision support as medical devices requiring pre-market evaluation proportional to risk: those generating unsupervised diagnostic recommendations as high-risk requiring rigorous evidence, and documentation assistance with mandated human review subject to less intensive oversight. Post-market surveillance should detect safety signals during real-world use across clinical contexts and populations. International regulatory coordination is needed given global LLM deployment. Liability policies should clarify that clinicians remain accountable for AI-assisted decisions, institutions bear responsibility for ensuring deployed systems meet standards, and patients have recourse when AI-influenced care causes harm.

Conclusion

This systematic review synthesized evidence on large language models in clinical medicine from 2017 to 2025, encompassing medical licensing examination performance, clinical documentation, decision support, and safety concerns. The evidence demonstrates substantial capabilities: GPT-4 and Med-PaLM 2 pass the USMLE with accuracy exceeding 80%, LLM-assisted documentation reduces clinician time by 30–50%, and decision support approaches clinician levels for structured tasks. These achievements suggest meaningful potential for LLMs to reduce burden and enhance healthcare delivery when appropriately deployed.

The gap between benchmark performance and clinical readiness remains substantial and is defined primarily by unresolved safety concerns. Hallucination rates of 5–30%, documented demographic bias, overconfidence in incorrect outputs, and demonstrated clinician susceptibility to AI-influenced diagnostic error represent barriers that current technology has not overcome. The persistence of these concerns across model versions and absence of effective mitigation indicate fundamental architectural limitations rather than transient challenges. The near-complete absence of prospective clinical trials evaluating patient outcomes further underscores the prematurity of autonomous deployment.

Coordinated action across research, clinical, and regulatory domains is required. Standardized safety evaluation frameworks must be developed and required. Prospective trials with patient-centered outcomes must be prioritized over continued benchmark evaluation. Regulatory frameworks must accommodate generative AI characteristics while maintaining safety oversight. Clinicians must be trained on LLM limitations and retain ultimate authority and accountability for clinical decisions.

The vision emerging from this evidence is of LLMs as clinician assistants rather than replacements—tools reducing documentation burden, suggesting diagnostic possibilities, and summarizing information under human supervision. This augmentation paradigm preserves demonstrated benefits while protecting patients from unresolved autonomous decision-making risks. Achieving this vision requires sustained investment in safety research, regulatory development, and clinician education commensurate with the potential benefits and risks large language models present for clinical medicine.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 29 May 2025 Revised: 20 Jul 2025 Accepted: 22 Aug 2025
Published online: 20 January 2026

References

Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nat. 2023;620(7972):172-180.

Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Amin M, et al. Toward expert-level medical question answering with large language models. Nat Med. 2025;31(3):943-950.

Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198.

Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ. 2023;9:e45312.

Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med. 2023;388(13):1233-9.

Goh E, Gallo R, Hom J, Strong E, Weng Y, Kerman H, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. JAMA Netw Open. 2024;7(10):e2440969.

Gaber F, Shaik M, Allega F, Bilecz AJ, Busch F, Goon K, et al. Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis. NPJ Digit Med. 2025;8(1):263.

Weissman GE, Mankowitz T, Kanter GP. Unregulated large language models produce medical device-like output. NPJ Digit Med. 2025;8(1):148.

Brin D, Sorin V, Konen E, Nadkarni G, Glicksberg BS, Klang E. How GPT models perform on the United States medical licensing examination: a systematic review. Discov Appl Sci. 2024;6(10):500.

Mihalache A, Huang RS, Popovic MM, Muni RH. ChatGPT-4: an assessment of an upgraded artificial intelligence chatbot in the United States Medical Licensing Examination. Med Teach. 2024;46(3):366-72.

Song JW, Park J, Kim JH, You SC. Large language model assistant for emergency department discharge documentation. JAMA Netw Open. 2025;8(10):e2538427.

Ganzinger M, Kunz N, Fuchs P, Lyu CK, Loos M, Dugas M, et al. Automated generation of discharge summaries: leveraging large language models with clinical data. Sci Rep. 2025;15(1):16466.

Rust P, Frings J, Meister S, Fehring L. Evaluation of a large language model to simplify discharge summaries and provide cardiological lifestyle recommendations. Commun Med. 2025;5(1):208.

Strong E, DiGiammarino A, Weng Y, Kumar A, Hosamani P, Hom J, et al. Chatbot vs medical student performance on free-response clinical reasoning examinations. JAMA Intern Med. 2023;183(9):1028-30.

Hains L, Kleinig O, Murugappa A, Gluck S, Marks J, Gilbert T, et al. Large language model discharge summary preparation using real world electronic medical record data shows promise. Intern Med J. 2025;55(7):1188-92.

Williams CY, Subramanian CR, Ali SS, Apolinario M, Askin E, Barish P, et al. Physician- and large language model–generated hospital discharge summaries. JAMA Intern Med. 2025;185(7):818-25.

Croxford E, Gao Y, First E, Pellegrino N, Schnier M, Caskey J, et al. Evaluating clinical AI summaries with large language models as judges. NPJ Digit Med. 2025;8(1):640.

Woo BF, Cato K, Cho H, You SB, Song J. The use of large language models in clinical documentation: a scoping review. Int J Nurs Stud. 2025:105322.

Mbakwe AB, Lourentzou I, Celi LA, Mechanic OJ, Dagan A. ChatGPT passing USMLE shines a spotlight on the flaws of medical education. PLOS Digit Health. 2023;2(2):e0000205.

Wu S, Koo M, Blum L, Black A, Kao L, Fei Z, et al. Benchmarking open-source large language models, GPT-4 and Claude 2 on multiple-choice questions in nephrology. NEJM AI. 2024;1(2):AIdbp2300092.

Peng C, Yang X, Chen A, Smith KE, PourNejatian N, Costa AB, et al. A study of generative large language model for medical research and healthcare. NPJ Digit Med. 2023;6(1):210.

Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, et al. A large language model for electronic health records. NPJ Digit Med. 2022;5(1):194.

Asgari E, Montaña-Brown N, Dubois M, Khalil S, Balloch J, Yeung JA, et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. NPJ Digit Med. 2025;8(1):274.

Rutledge GW. Diagnostic accuracy of GPT-4 on common clinical scenarios and challenging cases. Learn Health Syst. 2024;8(3):e10438.

Ong JC, Jin L, Elangovan K, San Lim GY, Lim DY, Sng GG, et al. Large language model as clinical decision support system augments medication safety in 16 clinical specialties. Cell Rep Med. 2025;6(10):101869.

Katz U, Cohen E, Shachar E, Somer J, Fink A, Morse E, et al. GPT versus resident physicians—a benchmark based on official board scores. NEJM AI. 2024;1(5):AIdbp2300192.

Shan G, Chen X, Wang C, Liu L, Gu Y, Jiang H, et al. Comparing diagnostic accuracy of clinical professionals and large language models: systematic review and meta-analysis. JMIR Med Inform. 2025;13(1):e64963.

Jin Q, Chen F, Zhou Y, Xu Z, Cheung JM, Chen R, et al. Hidden flaws behind expert-level accuracy of multimodal GPT-4 vision in medicine. NPJ Digit Med. 2024;7(1):190.

Siam MK, Varela A, Faruk MJ, Cheng JQ, Gu H, Maruf AA, et al. Benchmarking large language models on the United States medical licensing examination for clinical reasoning and medical licensing scenarios. Sci Rep. 2025;15(1):38421.

Chen Y, Huang X, Yang F, Lin H, Lin H, Zheng Z, et al. Performance of ChatGPT and Bard on medical licensing examinations varies across different cultures: comparison study. BMC Med Educ. 2024;24(1):1372.

Wang L, Chen X, Deng X, Wen H, You M, Liu W, et al. Prompt engineering in consistency and reliability with evidence-based guideline for LLMs. NPJ Digit Med. 2024;7(1):41.

Tanaka Y, Nakata T, Aiga K, Etani T, Muramatsu R, Katagiri S, et al. Performance of generative pretrained transformer on the national medical licensing examination in Japan. PLOS Digit Health. 2024;3(1):e0000433.

Kresevic S, Giuffrè M, Ajcevic M, Accardo A, Crocè LS, Shung DL. Optimization of hepatological clinical guidelines interpretation by large language models: a retrieval augmented generation-based framework. NPJ Digit Med. 2024;7(1):102