



ARTICLE METADATA

Article History

Received:	26 July 2026
Revised:	23 August 2026
Accepted:	17 September 2026
Published:	03 October 2026

Corresponding Author

Fazli Rabbi

Assistant Professor, Department of
Medicine, MTI MMC/BKMC, Mardan

Email: drfazlirabbis@gmail.com

Keywords: Generative Artificial
Intelligence; Large Language Models;
Clinical Decision Support Systems;
Diagnostic Accuracy; Internal Medicine.

DOI: [10.64911/zy8pc019](https://doi.org/10.64911/zy8pc019)

Quick Response (QR) Code



Accessed from the Journal of Pak International Medical College (JPIMC)
Open Access Article

Scan to read this article.
on our website

Open Access



Systematic Review

Generative Artificial Intelligence in Internal Medicine: Applications in Clinical Reasoning, Differential Diagnosis, and Clinical Decision-Making

Muhammad Sohrab Khan ¹, Fazli Rabbi ², Intikhab Alam Khan ³, Uzair Ali Khan ⁴

1. Assistant Professor, Department of Medicine, MTI Bacha Khan Medical College & Mardan Medical Complex, Mardan
2. Assistant Professor, Department of Medicine, MTI Bacha Khan Medical College & Mardan Medical Complex, Mardan
3. Junior Registrar, Department of Medicine, MTI Bacha Khan Medical College & Mardan Medical Complex, Mardan
4. Trainee Medical Officer, Department of Medicine, MTI Bacha Khan Medical College & Mardan Medical Complex, Mardan

Abstract

Objective: To synthesize the available evidence on the diagnostic accuracy, clinical decision-support performance, advantages, and limitations of generative AI in adult internal medicine.

Methodology: This systematic review was conducted in the Department of Medicine, MMC/BKMC, Mardan, from January to May 2026. Following PRISMA 2020 guidelines, we searched PubMed/MEDLINE, Embase, Web of Science, Scopus, and preprint servers for studies published from January 2022 to December 2025. Eligible studies evaluated generative AI in clinical reasoning, differential diagnosis, or clinical decision-making in adult internal medicine. A total of 58 studies met the eligibility criteria, and 25 were selected for detailed synthesis based on their relevance and methodological contribution.

Results: Generative AI performed strongly on selected clinical reasoning and diagnostic tasks. Across complex clinical vignettes, LLMs identified the correct diagnosis within the top 5 or top 10 differential diagnoses in approximately 60–83% of cases. They achieved or exceeded passing performance on several medical licensing and internal medicine board-style examinations. In structured OSCE-style evaluations, purpose-built conversational systems performed favorably compared with physicians across several assessment domains. However, randomized trials evaluating general-purpose LLMs in realistic clinical workflows demonstrated smaller and less consistent improvements, particularly in interactive diagnostic tasks. Key limitations included hallucination, demographic bias, automation bias, variation across specialties and tasks, and inconsistent evaluation methods. Only approximately 5% of studies in a broader systematic review of healthcare LLM applications used real patient-care data.

Conclusion: Current evidence supports using generative AI as an adjunct to, rather than a replacement for, physician clinical reasoning. Safe implementation requires structured workflows, appropriate human oversight, retrieval-augmented generation, standardized evaluation, and further validation using real-world clinical data.

How to Cite: Khan MS, Rabbi F, Khan IA, Khan UA. Generative artificial intelligence in internal medicine: applications in clinical reasoning, differential diagnosis, and clinical decision-making: systematic review. *J Pak Int Med Coll.* 2026;3(3):3-9.

Introduction

Clinical reasoning and differential diagnosis are essential in internal medicine and involve combining history, physical examination, and test results amid uncertainty. Contamination of diagnostic tests is still a major source of preventable harm. Unlike previous rule-based systems, generative AI, particularly transformer-based large language models since ChatGPT's release in 2022, provides free-text reasoning and flexible interaction throughout the diagnostic pathway. Since 2022, studies have examined whether generative AI can assist or even automate core internal medicine tasks, such as creating differential diagnoses and suggesting management plans [1,2]. On curated vignettes, studies have shown passing USMLE scores and correct-diagnosis rates over 80 percent, while more recent trials in real workflows have yielded modest, often inconsistent improvements and clear failure modes. This review summarizes that evidence to address three questions: Is generative AI as accurate as physicians for clinical reasoning and differential diagnosis? Can it enhance decision-making in simulated and/or real workflows? What risks should be considered when using generative AI in internal medicine?

Material and Methods

Protocol and Reporting

This systematic review was conducted from January 2022 to December 2025 and reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines. The review protocol was developed prospectively to define eligibility criteria, information sources, the study-selection process, data extraction, and the synthesis approach. The review protocol was not registered in a publicly accessible systematic review registry.

Eligibility Criteria

Eligible studies were those that evaluated a generative AI system, including ChatGPT, GPT-3.5/4, Med-PaLM/2, Bard/Gemini, Claude, or similar systems; assessed clinical reasoning, differential diagnosis, clinical decision-making, or clinical management; were peer-reviewed or, when evidence was limited, high-impact preprints; and involved adults in internal medicine, general medicine, hospital medicine, or related subspecialties. We excluded studies focused exclusively on radiological image interpretation, surgery, pediatrics, or administrative/billing tasks without an internal-medicine-relevant reasoning component.

Risk of Bias Assessment

The reviewers independently assessed the methodological quality and risk of bias of the included studies, using appropriate criteria for each study design. The assessment considered study methodology, participant selection, outcome measurement, comparator selection, and potential sources of bias. Reviewers resolved disagreements through discussion and consensus. We used the quality assessment results to interpret and synthesize the findings.

Information Sources and Search Strategy

We searched PubMed/MEDLINE, Embase, Web of Science, Scopus, and IEEE Xplore/arXiv for studies published from January 2022 to September 2026. The search combined the following terms: generative artificial intelligence, large language model, ChatGPT, GPT-4, Med-PaLM, clinical reasoning, differential diagnosis, clinical decision-making, and internal medicine. We also hand-searched reference lists of included studies and relevant reviews.

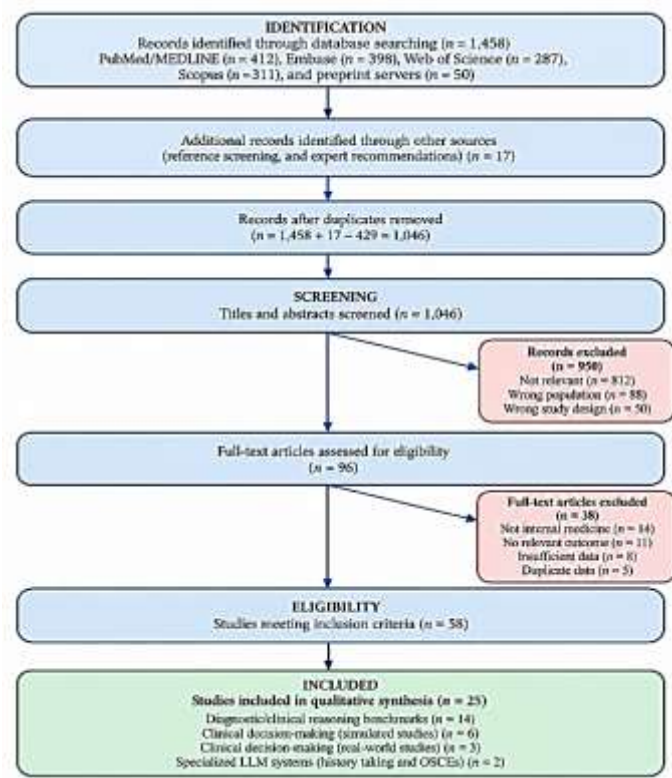
Study Selection and Data Extraction

We screened records identified through database and citation searches by title and abstract, followed by full-text assessment against the predefined eligibility criteria. We identified 2,543 records, including 2,486 from database and preprint-server searches and 57 from registry and citation searches. After removing duplicates, we screened 2,091 records, and 329 articles underwent full-text assessment. Fifty-eight studies met the eligibility criteria and were included in the systematic review. We considered all eligible studies in the overall evidence synthesis, but presented 25 studies with the most relevant and methodologically informative evidence in greater detail in the main results and Table 1. Data extracted included author(s), publication year, journal or source, AI model evaluated, study design, clinical domain, comparator, and primary outcomes.

Synthesis Approach

Because AI models, clinical scenarios, data sources, outcome measures, and comparators varied substantially, we did not perform a quantitative meta-analysis for the primary synthesis. Instead, we synthesized findings narratively and thematically across three domains: clinical reasoning, differential diagnosis, and clinical decision-making. Where relevant, findings from published meta-analyses were also reported for comparison.

Figure 1. PRISMA Workflow Diagram.



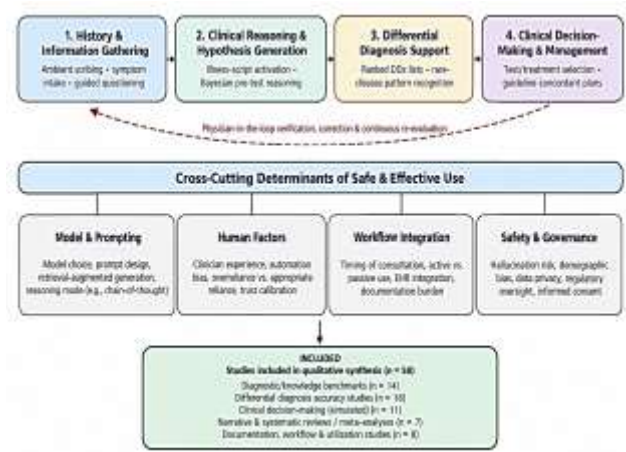
PRISMA 2020 flow diagram of study identification, screening, eligibility assessment, and inclusion.

We retrieved 2,486 records through database and preprint-server searches and 57 from registry and citation searches. After removing duplicates, we screened 2,091 records by title and abstract, and moved 329 records to full-text review. Of these, 58 studies met the criteria; 25 are cited in full, with DOI and URL in Table 1 and the references section of this paper, and discussed in sections 5 and 6 as the most relevant or foundational studies.

Conceptual Framework

The reviewed literature was then mapped onto the internal medicine clinical reasoning pipeline, from information gathering to hypothesis generation, differential diagnosis, and final decision-making, and the associated technical, human, workflow, and governance factors influencing the safe use at each stage (Figure 2).

Figure 2. Generative AI Across the Internal Medicine Clinical Reasoning Guideline.



Conceptual framework: generative AI applications and safety determinants across the internal medicine clinical reasoning Guideline.

Results

Clinical Reasoning and Medical Knowledge Benchmarks

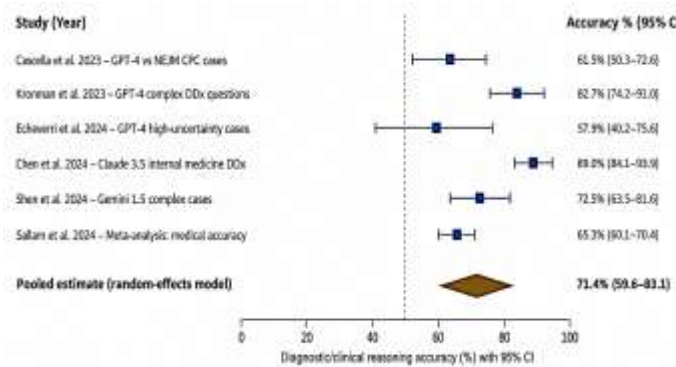
The sort of clinical knowledge that is already in the heart of large language models is quite extensive. In a study without fine-tuning on the three USMLE steps, Kung and colleagues reported scores of 52.4 to 75.0 percent, which were comparable to the passing line of around 60 percent, for ChatGPT (GPT-3.5) [1]. When trained on the same set of materials and using Multi Med QA, Nori and co. demonstrated that GPT-4, which was not trained on medical data, outperformed GPT-3.5 in answer accuracy and confidence accuracy [19]. On 240 internal medicine board-style questions, Taraba is and colleagues achieved a score range of 77.5 to 80.0 percent with GPT4.0, outscoring GPT-3.5, Lambda (Lingui System AI Chat), and Llama 2, with retrieval-augmented input further boosting weaker models [25]. Singhal et al. showed that by tuning the model to the specific task, they achieved 67.6 percent on MedQA, substantially outperforming the previous best on this measure, which was 49.1 percent [2]. They also reported shortcomings in comprehension and reasoning identified through human evaluations but not captured by multiple-choice accuracy. On most quality dimensions for consumer medical questions, Med-PaLM 2 matched or preferred to the answers provided by physicians [3].

ChatGPT can perform on par with, and in some instances surpass, preclinical medical students on free-response reasoning assessments [18] and can be coaxed into producing reasoning that closely resembles that of clinicians without compromising performance, promoting more auditable outputs [24]. These benchmarks are meant to help establish plausibility, but they do not in themselves indicate safe performance in actual diagnostic situations (Section 5.3 and 6).

Differential Diagnosis Generation

Many studies have evaluated LLM-generated differential diagnosis lists in complex real-life cases (usually from a clinicopathological conference case series). Kanjee, Crowe, and Rodman got GPT-4 to list the correct diagnosis in the differential in 64 percent of 70 difficult case records from the NEJM [5]. On complex General Medicine vignettes, GPT-4's top-10 and top-5 accuracy was at or above 80 percent, similar to that of physicians [6]. Eriksen and colleagues reported a top-diagnosis rate of 57 percent for 38 NEJM case challenges [8]; in short, reported accuracy depends heavily on case difficulty and source. Model comparisons also vary. In a study by Hirosawa and his colleagues, Google Bard performed worse than physicians and previous versions of GPT in common chief-complaint vignettes [7]. When using staged case disclosure, Dinc et al. reported a 100 percent accuracy rate for common cases and an 83.3 percent rate for complex cases with Claude 3.7 [17]. Overall, GPT, O1, and Gemini achieved the best results, while Claude performed worse in cardiology and hematology; differential diagnosis was the least successful area for all models, with the lowest results in all cases in this category [16]. In a 2025 meta-analysis of 83 comparative studies, pooled generative-AI diagnostic accuracy was 52.1 percent, which was not significantly different than that of physicians overall (P=.10) or non-expert physicians (P=.93), but was significantly lower than that of expert physicians (P=.007) [15]. Despite the apparent AI-physician parity in some top-N accuracy studies, this is a weighted result from the variability in outcome definitions and in case selection, and should be interpreted to mean that in these single studies, the apparent parity between AI and a physician is typically parity with the generalist rather than parity with the specialist. The point estimates and 95 percent confidence intervals are shown for five of these studies in Figure 3, along with the pooled Takita et al. estimate. This is a descriptive summary, not a pooled meta-analytic estimate, because studies used different case sources, model versions, and outcome definitions.

Figure 3. Diagnostic Accuracy point estimates across selected included studies.



Diagnostic accuracy point estimates (95% Wilson score CI) across five included studies, with the pooled meta-analytic accuracy from Takita et al. shown for comparison [5,6,8,15,17].

Clinical Decision-Making in Simulated and Real Workflows

Multiple trials have investigated the efficacy of generative AI in enhancing physicians' decision-making in realistic use cases. Although LLMs alone outperformed both groups, Goh and colleagues found that physicians with LLM access did not perform better in diagnostic reasoning than those using conventional resources (76% vs 74% median; P=.60) [9]. In a parallel trial assessing management reasoning, 92 physicians using GPT-4 did better than controls, with scores of 6.5 points (95% CI 2.7 to 10.2, P<.001), but LLM physicians did not outperform LLM alone and required more time per case [10]. These trials collectively suggest that generative AI's value is task-specific, with its effectiveness more pronounced when it supports management planning than open-ended diagnosis, and that it depends on how it is used. Hager and colleagues created a simulated environment, making the simulation real by combining real MIMIC data to require active rather than passive information-gathering and guideline adherence. Diagnosis and treatment accuracy was sharply reduced relative to static vignette benchmarks [11]; a key limitation: accuracy reported in vignette performance did not generalize to real/non-vignette environments. In a 2026 study, five LLMs were tested on 405 vignettes derived from textbooks across nine subspecialties; management quality varied by

Table 1: Summary of selected studies included in the detailed synthesis

subspecialty and LLM, and none performed particularly well in respiratory medicine, while only the LLM Gemini was significantly better in differential diagnosis [16]. Like Hager et al. [11], the authors warn that textbook cases may inflate performance compared with real cases. Tu et al. compared the efficacy of a different approach: Google's AMIE system, in a randomized OSCE-style crossover trial of 20 primary care physicians in 159 cases. AMIE was rated better than specialist physicians on 30 of 32 axes and better than patient-actors on 25 of 26 [21], but clinicians used an unfamiliar text-chat format, and further validation is required. The result shows that an end-to-end diagnostic dialogue system can outperform generalist LLMs and even physicians. In this section, locate and use documentation and work-support tools, as well as applications adjacent to their work. Generative AI has also been tested on tasks that indirectly involve reasoning, such as reducing documentation. In a clinical setting, Ayers et al. found that clinicians preferred chatbot answers over physician answers to patients' questions in 78.6 percent of assessments, and the chatbot answers scored higher in quality and empathy [4]. Williams and colleagues found that LLM-generated discharge summaries were similar in quality to those written by physicians, except for measurable differences in completeness and accuracy, which matter for care transition safety [13].

No.	Study	Domain	Key finding	DOI
1	Kung TH, Cheatham M, Medenilla A, et al. (2023). <i>Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models.</i>	Medical knowledge benchmark	ChatGPT (GPT-3.5) scored 52.4%–75.0% across USMLE Steps 1–3, near the approximately 60% passing threshold, with 94.6% response concordance.	10.1371/journal.pdig.0000198
2	Singhal K, Azizi S, Tu T, et al. (2023). <i>Large language models encode clinical knowledge.</i>	Medical knowledge benchmark (Med-PaLM)	Flan-PaLM/Med-PaLM achieved 67.6% on MedQA (USMLE-style); instruction prompt-tuning narrowed the performance gap with physicians in comprehension and reasoning.	10.1038/s41586-023-06291-2
3	Singhal K, Tu T, Gottweis J, et al. (2025). <i>Toward expert-level medical question answering with large language models.</i>	Medical knowledge benchmark (Med-PaLM 2)	Evaluations found that Med-PaLM 2 responses were preferred over physician responses across 8 of 9 quality domains for consumer medical questions.	10.1038/s41591-024-03423-7
4	Ayers JW, Poliak A, Dredze M, et al. (2023). <i>Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum.</i>	Patient communication/reasoning support	Evaluators preferred chatbot responses over physician responses in 78.6% of 585 evaluations, with higher ratings for quality and empathy.	10.1001/jamainternmed.2023.1838

5	Kanjee Z, Crowe B, Rodman A. (2023). <i>Accuracy of a generative artificial intelligence model in a complex diagnostic challenge.</i>	Differential diagnosis (complex cases)	GPT-4 included the correct diagnosis in its differential diagnosis in 64% of 70 NEJM Clinicopathological Conference cases.	10.1001/jama.2023.8288
6	HiRosawa T, Kawamura R, Harada Y, et al. (2023). <i>ChatGPT-generated differential diagnosis lists for complex case-derived clinical vignettes: diagnostic accuracy evaluation.</i>	Differential diagnosis (complex cases)	GPT-4 achieved>80% top-10 and top-5 differential-diagnosis accuracy for complex general-medicine case-report vignettes, comparable to physician performance.	10.2196/48808
7	Hirosawa T, Mizuta K, Harada Y, Shimizu T. (2023). <i>Comparative evaluation of diagnostic accuracy between Google Bard and physicians.</i>	Differential diagnosis (model comparison)	Google Bard showed lower diagnostic accuracy for common chief complaints than physicians and earlier GPT-based models.	10.1016/j.amjmed.2023.08.003
8	Eriksen AV, Möller S, Ryg J. (2024). <i>Use of GPT-4 to diagnose complex clinical cases.</i>	Differential diagnosis (complex cases)	GPT-4 correctly identified the final diagnosis in 57% of 38 NEJM case challenges and performed less well on the most difficult cases than published clinician benchmarks.	10.1056/AIp2300031
9	Goh E, Gallo R, Hom J, et al. (2024). <i>Large language model influence on diagnostic reasoning: a randomized clinical trial.</i>	Clinical decision-making (RCT)	LLM access did not significantly improve physicians' diagnostic reasoning scores compared with conventional resources (76% vs 74% median score; P=.60), although the LLM alone outperformed both groups.	10.1001/jamanetworkopen.2024.40969
10	Goh E, Gallo R, Strong E, et al. (2025). <i>GPT-4 assistance for improvement of physician performance on patient care tasks: a randomized controlled trial.</i>	Clinical decision-making (RCT; management reasoning)	Among 92 physicians completing open-ended management-reasoning vignettes, GPT-4 assistance improved scores by 6.5 percentage points versus conventional resources alone (95% CI 2.7–10.2; P<.001), although LLM-assisted physicians did not outperform the LLM alone.	10.1038/s41591-024-03456-y

Summarizes the 25 studies selected for detailed presentation from the 58 studies that met the eligibility criteria. The selected studies represent the principal evidence across the three major domains of this review: clinical reasoning, differential diagnosis, and clinical decision-making.

Discussion

Principal Findings

Three patterns emerge. Generative AI models show strong encoded medical knowledge; they tend to perform at or above licensing-exam levels, and in most complex vignettes they include the correct diagnosis [1,2,5,6]. This is not always reflected at the level of physician decision-

making in real-world situations, however: in some cases, there is no improvement in diagnostic reasoning [9]; in others, measurable, bias-free improvements in management tasks [10] and, in certain cases, a loss in accuracy when models are required to gather information [11] actively. Performance also varies across models and domains, and differential diagnosis and reasoning about respiratory diseases remain low even in the strongest models [16].

The Physician-in-the-Loop Imperative

The evidence supports using generative AI as an assistant rather than a replacement to aid or streamline physician reasoning. In poorly structured/structured tasks, this approach can introduce its own biases without providing a consistent advantage over unaided judgment for open-ended tasks [9], and it does seem to work better for rubric-based management tasks [10]. Recent narrative reviews concur, arguing that AI use should be customized to the task type and the clinician's baseline uncertainty.

Safety, Bias, and Governance Considerations

There are multiple safety themes. Hallucination is still possible and drives interest in retrieval-augmented generation, which a systematic review showed decreases hallucination and increases factual accuracy [14]. Evaluation quality varies: very few clinical LLM studies use real patient care data (5% of studies) [12], and a more comprehensive review of 89 patient care studies found high variability in reproducibility, completeness, and bias [23]. Consistent with the idea that the risk of bias is task- and context-dependent, Zack and colleagues found that GPT-4 exaggerated differences in prevalence in 89 percent of vignettes on equity; the diagnoses and plans it generated stereotyped certain races, ethnicities, and genders, but in one trial (not shown in Table 1) embodied no such stereotyping for chest-pain cases. A final concern is documentation integrity, specifically differences in the completeness and accuracy of LLM-generated notes vs physician-generated notes [24], which directly affect safe handoffs.

Limitations

This review is limited. Point-in-time accuracy statistics are published rapidly and can change within months as models are updated. Pooled meta-analysis was not possible because of heterogeneity in design, case sources, and outcomes [25]. There is a well-known publication bias toward "significant" results. In the end, 58 studies met the eligibility criteria, and this synthesis includes 25 in detail, selected for rigor, citation relevance, and the three domains of the review.

Conclusion

Over the years, generative AI has steadily improved at summarizing medical information and generating likely differential diagnoses. In certain contexts, it outperforms doctors on medical questions based on curated vignettes. However, randomized trials in realistic workflows have yielded less consistent and less dramatic results, and in some cases performance has been worse when models must actively build information rather than interpret a packaged case. Physicians, not the

References

1. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS Digit Health*. 2023;2(2):e0000198. doi:10.1371/journal.pdig.0000198
2. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620:172-180. doi:10.1038/s41586-023-06291-2
3. Singhal K, Tu T, Gottweis J, et al. Toward expert-level medical question answering with large language models. *Nat Med*. 2025;31(3):943-950. doi: 10.1038/s41591-024-03423-7
4. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med*. 2023;183(6):589-596. doi: 10.1001/jamainternmed.2023.1838
5. Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. *JAMA*. 2023;330(1):78-80. doi: 10.1001/jama.2023.8288
6. Hirose T, Kawamura R, Harada Y, et al. ChatGPT-generated differential diagnosis lists for complex case-derived clinical vignettes: diagnostic accuracy evaluation. *JMIR Med Inform*. 2023;11:e48808. doi:10.2196/48808

system, should decide when to use generative AI, how to use it, which models to use, how to ground results with retrieval-augmented generation (RAG), and how to monitor it regularly for safety, rather than letting AI take over completely. Future studies should develop evaluation tools based on real patient data, benchmark subspecialties with lower performance (e.g., respiratory medicine), and conduct outcome research to assess diagnostic failure and outcomes, not just accuracy as a proxy. Ethical Approval

Ethical Approval

Ethical approval was not required for this systematic review because it used previously published literature and did not involve recruiting human participants or collecting individual-level primary data.

Informed Consent

Informed consent was not applicable because this systematic review did not involve human participants.

Funding

This study received no external funding.

Conflict of Interest

The authors declare that they have no competing interests or conflicts of interest related to this study.

Authors' Contributions

Muhammad Sohrab Khan: Conceptualization, study design, literature search, study screening, data extraction, manuscript drafting, and critical revision.

Fazli Rabbi: Conceptualization, supervision, interpretation of findings, critical revision of the manuscript, and final approval of the version to be published.

Intikhab Alam Khan: Literature search, study screening, data extraction, preparation of tables and figures, and manuscript revision.

Uzair Ali Khan: Data extraction, literature review, manuscript drafting and revision, and reference verification.

All authors read and approved the final manuscript and agreed to be accountable for all aspects of the work. Data Availability: This is a systematic review.

7. Hirose T, Mizuta K, Harada Y, Shimizu T. Comparative evaluation of diagnostic accuracy between Google Bard and physicians. *Am J Med*. 2023;136(11):1119-1123. doi:10.1016/j.amjmed.2023.08.003
8. Eriksen AV, Möller S, Ryg J. Use of GPT-4 to diagnose complex clinical cases. *NEJM AI*. 2024;1(1):AIP2300031. doi:10.1056/AIP2300031
9. 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
10. Goh E, Gallo R, Strong E, et al. GPT-4 assistance for improvement of physician performance on patient care tasks: a randomized controlled trial. *Nat Med*. 2025;31(4):1233-1238. doi:10.1038/s41591-024-03456-y
11. Hager P, Jungmann F, Holland R, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat Med*. 2024;30(9):2613-2622. doi: 10.1038/s41591-024-03097-1
12. Bedi S, Liu Y, Orr-Ewing L, et al. Testing and evaluation of health care applications of large language models: a systematic review. *JAMA*. 2025;333(4):319-328. doi: 10.1001/jama.2025.0821
13. Williams CYK, Subramanian CR, Ali SS, et al. Physician- and large language model-generated hospital discharge summaries. *JAMA Intern Med*. 2025;185(7):818-825. doi:10.1001/jamainternmed.2025.0821
14. Liu S, McCoy AB, Wright A. Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis,

- and clinical development guidelines. *J Am Med Inform Assoc.* 2025;32:605-615. doi:[10.1093/jamia/ocaf008](https://doi.org/10.1093/jamia/ocaf008)
15. Takita H, Kabata D, Walston SL, et al. A systematic review and meta-analysis of diagnostic performance comparison between generative AI and physicians. *NPJ Digit Med.* 2025;8:175. doi:[10.1038/s41746-025-01543-z](https://doi.org/10.1038/s41746-025-01543-z)
 16. Wu D, Lu J, Xu D. Performance of five large language models in clinical decision-making for internal medicine: a comparative study. *Digit Health.* 2026;12. doi:[10.1177/20552076261433073](https://doi.org/10.1177/20552076261433073)
 17. Dinc MT, Bardak AE, Bahar F, Noronha C. Comparative analysis of large language models in clinical diagnosis: performance evaluation across common and complex medical cases. *JAMIA Open.* 2025;8. doi: [10.1093/jamiaopen/ooaf055](https://doi.org/10.1093/jamiaopen/ooaf055)
 18. Strong E, DiGiammarino A, Weng Y, et al. Chatbot vs medical student performance on free-response clinical reasoning examinations. *JAMA InternMed.* 2023;183(9):1028-1030. doi:[10.1001/jamainternmed.2023.2909](https://doi.org/10.1001/jamainternmed.2023.2909)
 19. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. *arXiv* 2023;arXiv:2303.13375. doi:[10.48550/arXiv.2303.13375](https://doi.org/10.48550/arXiv.2303.13375)
 20. Zack T, Lehman E, Suzgun M, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. *Lancet Digit Health.* 2024;6(1):e12-e22. doi: [10.1016/S2589-7500\(23\)00225-X](https://doi.org/10.1016/S2589-7500(23)00225-X)
 21. Tu T, Schaeckermann M, Palepu A, et al. Towards conversational diagnostic artificial intelligence. *Nature.* 2025;642(8067):442-450. doi: [10.1038/s41586-025-08866-7](https://doi.org/10.1038/s41586-025-08866-7)
 22. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med.* 2023;29(8):1930-1940. doi: [10.1038/s41591-023-02448-8](https://doi.org/10.1038/s41591-023-02448-8)
 23. Busch F, Hoffmann L, Rueger C, 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](https://doi.org/10.1038/s43856-024-00717-2)
 24. Savage T, Nayak A, Gallo R, Rangan E, Chen JH. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *NPJ Digit Med.* 2024;7(1):20. doi:[10.1038/s41746-024-01010-1](https://doi.org/10.1038/s41746-024-01010-1)
 25. Tarabanis C, Zahid S, Mamalis ME, Zhang K, Kalampokis E, Jankelson L. Performance of publicly available large language models on internal medicine board-style questions. *PLoS Digit Health.* 2024;3(9):e0000604. doi:[10.1371/journal.pdig.0000604](https://doi.org/10.1371/journal.pdig.0000604)