NEXT JOURNAL OF AI IN HEALTHCARE · Vol. 1 · No. 1 · 2025 · NGHC Publications

https://doi.org/10.00000/njaih.2025.0001 • Open Access • CC BY 4.0

ORIGINAL RESEARCH

Large language models in clinical decision support: a systematic evaluation of diagnostic accuracy and safety considerations across five medical specialties

Benchmarking GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro on standardised clinical vignettes with human physician comparison

Jahed Hossain,[1,2] Amara Diallo,[3] Lena Schreiber,[4] Mohammed Al-Rashid[5]

[1] Health Protection and Communicable Diseases, Ministry of Public Health, Doha, Qatar [2] Department of Medicine, Faculty of Medicine and Dentistry, University of Alberta, Edmonton, Canada [3] School of Public Health, University of Ghana, Accra, Ghana [4] Institute of Medical Informatics, Charité – Universitätsmedizin Berlin, Germany [5] Department of Internal Medicine, Hamad Medical Corporation, Doha, Qatar

Received: 14 January 2025 · Accepted: 22 March 2025 · Published: 1 May 2025

Correspondence: j.hossain@nghc.ca

ABSTRACT

Background: Large language models (LLMs) are increasingly proposed as adjuncts to clinical decision support systems; however, rigorous comparative evaluations of their diagnostic accuracy against practising physicians remain scarce, particularly across diverse specialty contexts.

Methods: We constructed a benchmark dataset of 450 standardised clinical vignettes drawn equally from five specialties: internal medicine, emergency medicine, paediatrics, rheumatology, and infectious disease. Three frontier LLMs — GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro — were evaluated against a panel of 30 board-certified physicians (6 per specialty). Primary outcomes included diagnostic accuracy (top-1 and top-3), safety classification of diagnostic errors, and consistency across three independent model invocations.

Results: Physician panels achieved a mean top-1 diagnostic accuracy of 84.2% (95% CI: 81.1–87.3%). GPT-4o achieved 79.6%, Claude 3.5 Sonnet 81.4%, and Gemini 1.5 Pro 74.9%. LLM performance was highest in infectious disease (87.3% for Claude 3.5 Sonnet) and lowest in emergency medicine for all models. Critical safety errors occurred at rates of 3.1%, 2.4%, and 5.8% for GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro respectively, compared with 1.1% for physician panels.

Conclusion: Contemporary LLMs approach but do not yet match board-certified physician diagnostic accuracy across mixed specialty clinical vignettes. The disproportionate safety error rate in time-sensitive scenarios represents the principal barrier to unsupervised clinical deployment and warrants targeted evaluation frameworks.

Keywords: large language models · clinical decision support · diagnostic accuracy · artificial intelligence · patient safety · GPT-4o · generative AI

1. Introduction

The integration of artificial intelligence into clinical medicine has accelerated markedly since the introduction of transformer-based large language models capable of natural-language reasoning

across complex, multi-step problems.[1,2] Unlike earlier, narrowly-scoped clinical AI tools — which were typically trained on structured data for discrete prediction tasks — contemporary LLMs demonstrate emergent capabilities in differential diagnosis generation, clinical documentation synthesis, and patient communication that have attracted considerable attention from health systems worldwide.[3,4]

Several evaluations have demonstrated that frontier LLMs can achieve passing scores on standardised medical licensing examinations, including the United States Medical Licensing Examination (USMLE), suggesting a breadth of biomedical knowledge sufficient for basic clinical reasoning tasks.[5,6] However, examination performance does not necessarily predict behaviour in clinically realistic, contextually complex scenarios where incomplete histories, atypical presentations, and time-critical decision-making are routine features of practice.[7]

Furthermore, existing comparative studies have been limited in scope, frequently restricted to single specialties, single models, or outcome measures that do not adequately capture the safety-relevant dimensions of diagnostic error.[8] The distinction between an error of omission in a low-acuity outpatient setting and a missed time-sensitive diagnosis in an emergency context carries profoundly different implications for patient safety, yet this distinction has rarely been operationalised within LLM evaluation frameworks.[9]

The present study addresses these gaps by conducting a systematic evaluation of three frontier LLMs — GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro — against panels of board-certified physicians across five distinct medical specialties, with explicit classification of errors according to their potential for clinical harm.

2. Methods

2.1 Vignette construction and specialty selection

Clinical vignettes were developed by a multidisciplinary panel comprising two senior clinicians per specialty, one clinical informaticist, and one patient safety specialist. Each vignette was structured to replicate a realistic clinical encounter, incorporating presenting complaint, relevant history, examination findings, and available investigation results. Vignettes were designed to admit a single defensible primary diagnosis with two plausible alternative diagnoses, enabling meaningful differentiation between top-1 and top-3 accuracy. Specialties were selected to represent a range of cognitive demands, acuity levels, and knowledge domains: internal medicine, emergency medicine, paediatrics, rheumatology, and infectious disease.

2.2 Safety error classification

A novel four-tier safety classification framework was applied to all diagnostic errors. Tier 1 errors comprised failure to identify immediately life-threatening conditions; Tier 2 comprised failure to identify conditions where delayed diagnosis results in significant preventable morbidity; Tier 3 comprised clinically significant but non-urgent diagnostic errors; and Tier 4 comprised minor errors without material clinical consequence. The present analysis focuses on Tier 1 errors (hereafter described as “critical safety errors”) as the primary safety outcome.

2.3 Model evaluation protocol

Each vignette was presented to each model on three separate occasions with identical prompting to assess output consistency. The prompt instructed the model to act as a consulting clinician, to provide a primary diagnosis and two differential diagnoses in ranked order, and to indicate the urgency of the clinical presentation. No system-level instructions instructing refusal of clinical

reasoning were applied. Outputs were evaluated by blinded clinical adjudicators who were not involved in vignette construction.

3. Results

3.1 Overall diagnostic accuracy

Across the full 450-vignette benchmark, physician panels achieved a mean top-1 diagnostic accuracy of 84.2% (95% CI: 81.1–87.3%). Among the three LLMs evaluated, Claude 3.5 Sonnet demonstrated the highest overall accuracy at 81.4%, followed by GPT-4o at 79.6% and Gemini 1.5 Pro at 74.9%. The difference between physician panels and Claude 3.5 Sonnet was 2.8 percentage points ($p = 0.042$); the difference between physician panels and Gemini 1.5 Pro was 9.3 percentage points ($p < 0.001$). Top-3 accuracy substantially attenuated observed differences, with all three LLMs achieving greater than 93% top-3 accuracy compared with 96.8% for physician panels.

Evaluator	Top-1 Accuracy (%)	Top-3 Accuracy (%)	Critical Safety Errors (%)	Consistency (κ)
Physician panels	84.2 (81.1–87.3)	96.8 (95.2–98.4)	1.1	—
GPT-4o	79.6 (76.2–83.0)	93.4 (91.3–95.5)	3.1	0.84
Claude 3.5 Sonnet	81.4 (78.1–84.7)	94.6 (92.7–96.5)	2.4	0.89
Gemini 1.5 Pro	74.9 (71.2–78.6)	91.2 (88.8–93.6)	5.8	0.76

Table 1. *Summary of primary outcomes across all specialties. Values in parentheses represent 95% confidence intervals. Consistency (κ) denotes Fleiss' kappa across three independent model invocations per vignette. Critical safety errors defined as Tier 1 errors per the study safety classification framework.*

3.2 Safety error analysis

Critical safety errors — failures to identify time-sensitive diagnoses — occurred at rates of 3.1%, 2.4%, and 5.8% for GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro respectively, representing 2.8-fold, 2.2-fold, and 5.3-fold elevation over physician-panel rates. In emergency medicine specifically, critical safety error rates for LLMs ranged from 7.3% (Claude 3.5 Sonnet) to 12.1% (Gemini 1.5 Pro). Qualitative analysis of Tier 1 errors revealed two principal failure modes: overweighting of common diagnoses at the expense of rare but immediately life-threatening alternatives, and anchoring bias following presentation of a plausible but incorrect working diagnosis in the vignette narrative.

4. Discussion

The principal finding of this study is that contemporary frontier LLMs approach, but have not yet achieved, board-certified physician-level diagnostic accuracy across a multi-specialty clinical vignette benchmark. The magnitude of this performance gap — approximately 2.8 percentage points for the best-performing model — is modest in absolute terms; however, its clinical significance is amplified by the disproportionate concentration of LLM errors in high-acuity, time-sensitive scenarios.

The finding that Claude 3.5 Sonnet achieved near-equivalent accuracy to physician panels in infectious disease (87.3% versus 87.4%) is noteworthy and may reflect the relative prevalence of

infectious disease content in the pre-training corpora of contemporary LLMs, given the volume of published literature generated during and after the COVID-19 pandemic.[10] By contrast, the substantially lower performance across all models in emergency medicine is consistent with prior observations that LLMs under-perform in scenarios requiring contextual synthesis under time pressure or with incomplete information.[11]

The output consistency metric warrants particular attention from a deployment perspective. Claude 3.5 Sonnet demonstrated the highest inter-invocation consistency ($\kappa = 0.89$), suggesting that its diagnostic outputs are more reproducible across independent queries — a property of practical significance for clinical governance frameworks that require predictable, auditable AI behaviour.[12]

4.1 Implications for clinical deployment

The present findings do not support the unsupervised deployment of any of the evaluated LLMs as autonomous diagnostic agents. Rather, the performance profile identified — high top-3 accuracy, modest top-1 accuracy, and disproportionate critical safety error rates — is consistent with a role as augmentative second-opinion tools operating under structured clinical oversight. Implementation frameworks should include mandatory escalation pathways for high-acuity presentations, real-time safety filtering for Tier 1 diagnoses, and regular prospective performance auditing against updated benchmarks as model versions evolve.

5. Conclusions

This study provides evidence that frontier LLMs demonstrate clinically meaningful, though not yet equivalent, diagnostic accuracy relative to board-certified physicians across five medical specialties. The critical safety error differential — which is most pronounced in emergency medicine — represents the primary challenge to responsible clinical deployment and should form the central focus of future evaluation and mitigation research. Development of standardised, specialty-stratified, safety-weighted benchmarking frameworks is urgently needed to support evidence-based governance of LLM integration into clinical decision support systems.

Declarations

Competing interests: The authors declare no competing interests. **Funding:** No external funding was received for this study. **Ethics:** This study involved no human participants; ethics review was not required. **Data availability:** The benchmark vignette dataset and anonymised model outputs are available on the NJAIH Open Data Repository at https://data.njaih.org/2025/0001. **Author contributions:** JJH conceived the study. JJH, AD, and LS developed the vignette framework. MA-R led clinical adjudication. All authors contributed to interpretation and approved the final manuscript.

References

1. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172–180.
2. Nori H, King N, McKinney SM, et al. Capabilities of GPT-4 on medical challenge problems. arXiv. 2023; arXiv:2303.13375.
3. Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. Nat Med. 2023;29(8):1930–1940.
4. Moor M, Banerjee O, Abad ZSH, et al. Foundation models for generalist medical artificial intelligence. Nature. 2023;616(7956):259–265.

5. 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.
6. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination? JMIR Med Educ. 2023;9:e45312.
7. Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. JAMA. 2023;330(1):78–80.
8. Brodeur PG, Buckley TA, Ahmad FS, et al. Diagnostic accuracy of GPT-4 in an unfiltered cohort of patients presenting to an internal medicine resident clinic. medRxiv. 2024; preprint.
9. Rodman A, Buckley TA, Ahmad FS, et al. Evaluating large language model accuracy in clinical encounters. NPJ Digit Med. 2024;7:48.
10. Hossain JJ. Artificial intelligence in infectious disease surveillance: opportunities, limitations and the role of immunisation registry data. J Infect Public Health. 2025;18(3):102–109.
11. Rao A, Kim J, Kamineni M, et al. Evaluating GPT as an adjunct for radiologic decision making. J Am Coll Radiol. 2023;20(8):801–808.
12. Shah NH, Entwistle D, Pfeffer MA. Creation and adoption of large language models in medicine. JAMA. 2023;330(9):866–869.

