Communication

Beyond the Leaderboard: The Limitations of LLM Benchmarks and the Case for Real-World Clinical Evaluation

Sandeep Reddy *

Posted Date: 20 November 2025

doi: 10.20944/preprints202511.1572.v1

Keywords: artificial intelligence; LLMs; benchmarks; medical AI; real-world evaluation

 doi:10.20944/preprints202511.1572.v1

Communication

Beyond the Leaderboard: The Limitations of LLM Benchmarks and the Case for Real-World Clinical Evaluation

Sandeep Reddy

Queensland University of Technology, Australia; sandeep.reddy@qut.edu.au

Abstract

This article critically examines the limitations of current large language model (LLM) benchmarks, particularly in healthcare and clinical evaluation. While standardised leaderboards and benchmarks have driven rapid technical progress and shaped industry perceptions, they are increasingly undermined by issues like benchmark data contamination and narrow assessment criteria. The article explains that benchmark leakage can overstate results, and traditional evaluation using multiple-choice questions does not reflect the complexity of clinical practice. Specialised medical benchmarks, though more targeted, still overlook essential attributes such as reliability, calibration, and safety, and often lack representation of diverse healthcare contexts and languages. A shift toward real-world evaluation frameworks that emphasise scenario-based simulations, multisite validation, and comprehensive translational assessment is required. The Translational Evaluation of Healthcare AI (TEHAI) framework is presented as a robust alternative that integrates technical, utility, and adoption criteria and explicitly addresses ethical and contextual factors. Genuine clinical benefit and patient safety can be ensured only through continuous, context-specific evaluation that transcends traditional benchmarking.

Keywords: artificial intelligence; LLMs; benchmarks; medical AI; real-world evaluation

Introduction

The landscape of large language model (LLM) development in 2025 is characterised by intense competition, with vendors racing to top performance leaderboards through standardised benchmarks. Platforms such as Vellum AI, Artificial Analysis, and LLM-Stats provide real-time rankings across diverse metrics, including reasoning capabilities, coding proficiency, speed, and cost-effectiveness. [1] Leading models like Gemini 3 Pro, GPT 5.1, and Claude Sonnet 4.5 consistently dominate benchmarks such as GPQA Diamond, AIME 2025, and SWE Bench, showcasing remarkable technical achievements in artificial intelligence. [1,2] These benchmark successes drive substantial marketing narratives and shape investment decisions, positioning high-performing models as the gold standard for deployment across industries, including healthcare.

However, mounting evidence reveals fundamental limitations in relying solely on benchmark performance to evaluate LLM capabilities. Benchmark data contamination occurs when evaluation data is accidentally or intentionally included in training datasets, posing a serious challenge that can compromise the accuracy of performance measurements. [3,4] Studies demonstrate that LLMs can achieve inflated scores through exposure to test questions during training rather than genuine capability development. [5] Research employing Kernel Divergence Score and other detection methodologies has revealed that many leading models exhibit significant contamination, with benchmark leakage occurring across popular datasets including MMLU, HellaSwag, and TruthfulQA.[6,7] Furthermore, even when contamination is detected, the complexity of modern training pipelines and the proprietary nature of commercial models make systematic decontamination challenging.

 doi:10.20944/preprints202511.1572.v1

Beyond contamination concerns, traditional benchmarks fail to capture the nuanced requirements of real-world applications. Generic LLM benchmarks prioritise pattern recognition and knowledge retrieval through multiple-choice questions and constrained response formats that poorly reflect authentic use cases. [8] Studies evaluating LLM performance in game-based scenarios reveal significant discrepancies between benchmark scores and practical strategic reasoning capabilities, with models frequently exhibiting rigid strategies and struggling to adapt to unfamiliar scenarios.[9,10] These limitations become particularly pronounced in high-stakes domains such as healthcare, where the consequences of model failures extend far beyond academic interest.

Benchmarking Medical LLMs

In the medical domain, specialised benchmarks including MedQA, MMLU Medicine, and BioASQ have emerged to evaluate clinical knowledge and reasoning.[11] Recent frameworks such as MedCheck introduce lifecycle-oriented assessment covering 46 criteria across design integrity, clinical fidelity, and governance. [12] Despite these advances, medical LLM benchmarks share fundamental limitations with their generic counterparts. A systematic review of 761 studies evaluating LLMs in clinical settings found that 93.55% focused on general-domain models such as ChatGPT and GPT-4, with accuracy as the predominant evaluation metric (21.78% of studies). [13] This narrow focus on matching-based metrics fails to assess critical attributes, including reliability, trustworthiness, calibration, and safety parameters essential for clinical deployment. [14,15]

Moreover, medical benchmarks predominantly utilise exam-style multiple-choice questions that diverge substantially from real clinical practice. [16] Research demonstrates that LLMs achieving near-saturation scores on medical examination benchmarks (80-90% accuracy) exhibit concerning overconfidence regardless of answer correctness, with only marginal calibration improvements in higher-tier models.[17] This overconfidence poses significant clinical risks, potentially leading to misinformed decisions and eroded trust. Furthermore, benchmark datasets often over-represent high-income disease profiles while underrepresenting regionally prevalent conditions and practical safety considerations, limiting their applicability across diverse healthcare systems.[18] Studies specifically evaluating LLMs in non-English medical contexts, such as the Swedish Medical LLM Benchmark, confirm that performance varies substantially across languages and regional medical practices, with models demonstrating significantly reduced capability in low-resource languages. [19]

Real-World Evaluation

The messiness of real-world clinical situations demands evaluation approaches that transcend traditional benchmarks. Clinical decision-making involves managing incomplete information, addressing contextual ambiguity, navigating evolving patient presentations, and integrating multidisciplinary perspectives—dimensions that static test questions inadequately capture.[20] Deployment pilots incorporating LLMs into actual clinical workflows have exposed critical limitations, including unpredictable outputs, demographic biases, and workflow misalignment that remain invisible in benchmark evaluations.[21,22] Case-based simulations using longitudinal electronic health records reveal that while LLMs may excel on structured tests, they significantly lag behind physicians in adaptive decision-making with evolving clinical information.[23] Randomised controlled trials evaluating physician-LLM collaboration in real cases have identified concerning patterns of over-reliance, failure to improve diagnostic accuracy, and narrowed differential diagnoses—insights absent from benchmark assessments.[24]

Alternative evaluation mechanisms that incorporate real-world validation and comprehensive translational assessment offer more robust pathways for evaluating healthcare AI. Multisite retrospective and prospective validation studies utilising live hospital data across institutions enable quantification of generalizability and operational impact beyond single-site deployments.[21,22] Scenario simulators creating clinically realistic patient journeys expose LLMs to variable timing,

 doi:10.20944/preprints202511.1572.v1

incomplete data, and iterative action-observation cycles that better approximate authentic clinical reasoning.[25] These approaches consistently demonstrate that benchmark improvements do not necessarily translate to real-world clinical benefits, highlighting the critical importance of context-specific validation.

Figure 1. Alternative real-world evaluation measures for LLMs.

TEHAI

The Translational Evaluation of Healthcare AI (TEHAI) framework is a comprehensive, validated approach that addresses these evaluation gaps.[26] Developed through international expert consensus and grounded in translational research principles, TEHAI encompasses three core components: capability, utility, and adoption. The capability component assesses intrinsic technical performance by evaluating objectives, dataset integrity, internal and external validity, performance metrics, and use-case justification. The utility component evaluates real-world applicability, including generalizability, safety and quality considerations, transparency, privacy protections, and non-maleficence. The adoption component examines translational value through assessment of healthcare setting integration, technical implementation, multi-site deployment, and domain alignment.(26) Unlike traditional benchmarks, TEHAI explicitly incorporates ethical dimensions and can be applied iteratively across development, deployment, and discernment phases of AI system lifecycles.

TEHAI distinguishes itself by emphasising translational and ethical features that are often overlooked in conventional evaluation frameworks. The framework recognises that comprehensive evaluation must extend beyond technical performance to encompass contextual relevance, workflow integration, stakeholder utility, and sustained real-world effectiveness. [26] By providing structured assessment across 15 subcomponents with weighted scoring reflecting relative importance, TEHAI offers both flexibility for resource-constrained settings and rigour for regulatory applications. The framework recognises the fundamental reality that many AI systems that demonstrate promise in controlled environments or single-site deployments fail when scaled to diverse healthcare settings, underscoring the need for continuous, comprehensive evaluation.

 doi:10.20944/preprints202511.1572.v1

Conclusions

In summary, benchmark leaderboards help monitor technical progress but have drawbacks like contamination risk, limited assessment scope, and low real-world relevance. More thorough evaluation methods are needed. In healthcare, particularly, the stakes of deployment demand evaluation frameworks that assess not only technical capability but also safety, ethical alignment, contextual appropriateness, and translational value. Frameworks such as TEHAI, combined with real-world validation studies and iterative deployment pilots, offer more robust pathways to ensure AI systems deliver genuine clinical benefit while maintaining patient safety and trust. The future of medical AI evaluation lies not in gaming benchmarks but in demonstrating sustained value within the complex, messy reality of clinical practice.

References

1. Vellum AI. LLM Leaderboard. Available from: https://www.vellum.ai/llm-leaderboard [Accessed 20 Nov 2025].
2. Artificial Analysis. Models Leaderboard. Available from: https://artificialanalysis.ai/models [Accessed 20 Nov 2025].
3. Balloccu S, Schmidtová P, Lango M, et al. Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source LLMs. In: *Proceedings of the 19th Conference of the European Chapter of the ACL (EACL 2024)*. 2024. Available from: https://leak-llm.github.io/ [Accessed 20 Nov 2025].
4. Yang Z, Li Y, Wang L, et al. Benchmark Data Contamination of Large Language Models: A Survey. arXiv. 2024. arXiv:2406.04244.
5. Bordt S, Singh A, Gokaslan A, et al. How Much Can We Forget about Data Contamination? In: *International Conference on Learning Representations (ICLR 2025)*. 2025. Available from: https://openreview.net/forum?id=8ivK2TngIW.
6. Zhang Y, Chen X, Wang M, et al. How Contaminated Is Your Benchmark? Measuring Dataset Leakage in Large Language Models with Kernel Divergence. arXiv. 2025. arXiv:2502.00678.
7. Ashfri NS, Wijaya R, Chen WJ. Simulating Training Data Leakage in Multiple-Choice Benchmarks for LLM Evaluation. arXiv. 2025. arXiv:2505.24263.
8. Zhou Z, Zhang J, Wang Y, et al. Benchmarking Benchmark Leakage in Large Language Models. arXiv. 2024. arXiv:2404.18824.
9. Topsakal O, Edell CJ, Harper JB. Evaluating Large Language Models with Grid-Based Game Competitions: An Extensible LLM Benchmark and Leaderboard. arXiv. 2024. arXiv:2407.07796.
10. Topsakal O, Harper JB. Benchmarking Large Language Model (LLM) Performance for Game Playing via Tic-Tac-Toe. *Electronics*. 2024;13(8):1532.
11. Intuition Labs. Large Language Model Benchmarks – Life Sciences Overview. Available from: https://intuitionlabs.ai/articles/large-language-model-benchmarks-life-sciences-overview [Accessed 20 Nov 2025].
12. Emergent Mind. Medical LLM Benchmarks. Available from: https://www.emergentmind.com/topics/medical-llm-benchmarks [Accessed 20 Nov 2025].
13. Shool S, Adimi S, Amleshi RS, et al. A systematic review of large language model (LLM) evaluations in clinical medicine. *BMC Med Inform Decis Mak*. 2025;25(1):117.
14. Budler LC, Chen H, Chen A, Topaz M, Tam W, Bian J, Stiglic G. A brief review on benchmarking for large language models evaluation in healthcare. WIREs Data Mining Knowl Discov. 2025;15(2):e70010. doi:10.1002/widm.70010.
15. Wang H, Liu J, Zhang Y, et al. Large Language Models in Healthcare: A Comprehensive Benchmark. medRxiv. 2024. doi:10.1101/2024.04.24.24306315.
16. Zhang M, Chen L, Wang Q, et al. LLMEval-Med: A Real-world Clinical Benchmark for Medical LLMs with Physician Validation. arXiv. 2025. arXiv:2506.04078.
17. Omar M, Agbareia R, Glicksberg BS, et al. Benchmarking the Confidence of Large Language Models in Answering Clinical Questions: Cross-sectional Evaluation Study. *JMIR Med Inform*. 2025;13:e66917.

 doi:10.20944/preprints202511.1572.v1

18. Mutisya J, Kamau P, Ochieng D, et al. Mind the Gap: Evaluating the Representativeness of Quantitative Medical Language Reasoning LLM Benchmarks for African Disease Burdens. arXiv. 2025. arXiv:2507.16322.
19. Moëll B, Hertzberg L, Aronsson E, et al. Swedish Medical LLM Benchmark: development and evaluation of a framework for assessing large language models in the Swedish medical domain. *Front Artif Intell*. 2025;10:1557920.
20. Li Y, Zhang H, Wang C, et al. Benchmarking Large Language Models on Answering and Explaining Challenging Medical Questions. arXiv. 2024. arXiv:2402.18060.
21. Qiu P, Wu C, Liu S, et al. Quantifying the reasoning abilities of LLMs on clinical cases. *Nat Commun*. 2025;16:9799. doi:10.1038/s41467-025-64769-1.
22. 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.
23. Chen S, Fiscella K, Rucci A, et al. The effect of using a large language model to respond to patient messages in primary care. *Lancet Digit Health*. 2024. doi:10.1016/S2589-7500(24)00060-8.
24. Laverde N, et al. Integrating large language model-based agents into a virtual patient chatbot for clinical anamnesis training. 2025. Available from: https://www.sciencedirect.com/science/article/pii/S2001037025001850.
25. Gao C, Li N, Li M, et al. Large language models empowered agent-based modeling and simulation. *Humanit Soc Sci Commun*. 2024. doi:10.1057/s41599-024-03611-3.
26. Reddy S, Rogers W, Makinen V-P, et al. Evaluation framework to guide implementation of AI systems into healthcare settings. *BMJ Health Care Inform*. 2021;28(1):e100444.