npj | digital medicine

Comment

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01963-x

The evaluation illusion of large language models in medicine

Monica Agrawal, Irene Y. Chen, Freya Gulamali & Shalmali Joshi

While large language models (LLMs) hold promise for transforming clinical healthcare, current comparisons and benchmark evaluations of large language models in medicine often fail to capture real-world efficacy. Specifically, we highlight how key discrepancies arising from choices of data, tasks, and metrics can limit meaningful assessment of translational impact and cause misleading conclusions. Therefore, we advocate for rigorous, context-aware evaluations and experimental transparency across both research and deployment.

Large language models (LLMs) hold transformative potential for re-imagining complex workflows in medicine including clinical decision support, health record documentation and retrieval, and patient communication[1–3]. However, the open-ended generation capabilities facilitated by foundation models can be nuanced to evaluate, particularly in medicine. Many current experimental designs fall short of adequately comparing and validating the real-world utility, safety, and contextual effectiveness of both models and end-to-end tools. The breakneck speed of LLM research, translation, and dissemination has often outpaced rigorous evaluation and review, which can lead to a disconnect between underlying technological capabilities and public perception. As the community of authors, reviewers, consumers, and purveyors of clinical NLP research expands exponentially, it is imperative that this community be able to precisely communicate and reason about claims on LLM performance in clinical settings. For example, the lack of rigorous empirical evaluations can led to misleading conclusions about whether bespoke and clinically fine-tuned LLMs outperform general-purpose LLMs[4,5]. Achieving this goal requires being upfront about the research questions a specific experiment design is equipped to address, factoring in the common gaps between evaluation in research and what's required for robust workflow integration. In this work, we describe key discrepancies arising in current evaluation of LLMs in medicine and the real-world desiderata for robust evaluation (Fig. 1). These observations are tailored towards autoregressive LLMs, both general-purpose and medically fine-tuned[6].

Discrepancy in data

A strong indicator of the potential for the success of a clinical LLM tool, irrespective of the specific clinical task, is whether the evaluation data reflects the real-world data on which the tool will be deployed[7]. A recent survey across a broad range of clinical tasks found that only 5% of recent evaluations of LLMs in medicine have been conducted on actual electronic health record (EHR) data[8]. The rest were evaluated on benchmarks consisting of data like clinical vignettes, subject matter expert-generated questions, or multiple choice medical question-answering exams. However, real EHR data comes with the added complexity of clinical jargon, fragmentation in data collection, heterogeneity across healthcare institutions, and harder-to-process formats (e.g., scanned faxes). Moreover, evaluation datasets themselves introduce bias through patient population selection, institutional practices, and data collection protocols that may not represent the diversity of real-world deployment contexts[9]. These dataset biases can systematically favor certain model architectures or training approaches while masking performance disparities across demographic groups[10]. For most tasks in medicine, having all relevant clinical information synthesized into a short paragraph would be highly unrealistic and time-consuming to produce.

Discrepancy in tasks

Using EHR data alone is insufficient for construct validity, i.e., measuring the underlying target of evaluation[11]. For example, the popular MedNLI benchmark aims to measure how well models can interpret EHR text; given a pair of clinical sentences (a premise and a hypothesis about a patient), the goal is to see whether the hypothesis can be inferred, refuted, or neither by the premise[12]. However, using only the hypothesis and no premise (which should be equivalent to random), one can achieve approximately double the F1 score of a random baseline, indicating significant artifacts in the dataset[13]. The most popular category of benchmarks are multiple-choice exams, whose questions by design must include all relevant information[14]. However, this doesn't mirror clinical practice, in which additional measurements must be continuously elicited by the clinician[15]; unfortunately, language models have been shown to have significantly poorer performance in partial information scenarios in which they need to interactively uncover more evidence for diagnosis[16,17]. Further exacerbating this discrepancy is the fact that high performance on multiple-choice benchmarks can be driven by flawed rationales[18].

Discrepancy in automated metrics

Researchers often continue to rely on automated metrics as a source of evaluation for open-ended generation tasks. While such metrics from the model development process may serve as a valuable first step–for example to filter out tools that may not meet the most preliminary utility threshold—reliance on crude criterion is misguided. There is strong evidence to believe that accuracy measured using automated metrics lifted from the natural language processing (NLP) literature, such as ROUGE, BLEU, BERTScore, and its variants[19–21] do not reflect accuracy when measured via human evaluations across many clinical text generation tasks. To quantify this notion, we conducted a scoped literature search that measure how well automated metrics correlate with human perceptions for clinical summarization tasks. Papers were initially filtered from arXiv, the most popular

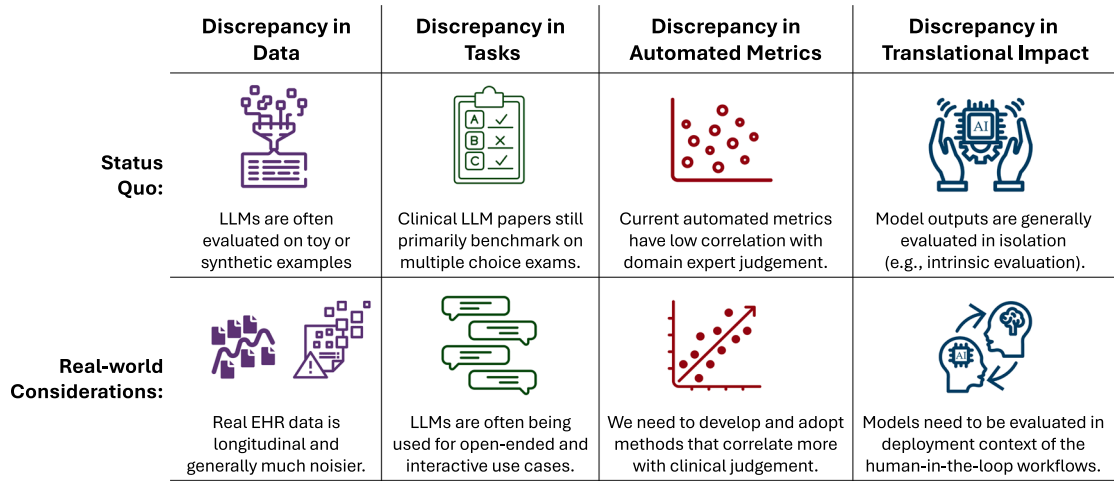

Fig. 1 | Overview of the four major discrepancies in current evaluation of clinical large language models. Discrepancies in data, tasks, automated metrics, and translational impact between the current status quo and real-world deployment lead to insufficient or misleading evaluations of large language models in medical contexts.

Table 1 | Literature search on concordance of common automated metrics for clinical summarization

Metric Name	Dimension	Pearson Correlation	Spearman Correlation
BERTScore	Completeness	0.28[32], 0.44[33]	0.15–0.2[34], 0.645[35]
	Correctness	0.23[32], 0.52[33], 0.022[36]	0.15–0.2[34], 0.530[35], −0.77[36]
BLEU	Completeness	0.22[32]	0.2–0.25[34], 0.685[35]
	Correctness	0.13[32]	0.1–0.15[34], 0.447[35]
ROUGE	Completeness	0.30[32], 0.42[33], 0.479[37]	0.326[24], −0.389[24], 0.2–0.25[34], 0.610[35]
	Correctness	0.16[32], 0.53[33], −0.01[36], 0.416[37]	0.15–0.2[34], 0.479[35], −0.66[36]

Correlation of completeness and correctness measures between human evaluators and three common automated metrics (BERTScore, BLEU, and ROUGE), as measure on clinical summarization tasks.

computer science archive, and Google Scholar, based on relevant terms in the title and abstract. For full literature search details, see Supplementary Note 1 and Tables 1, 2. We report the correlations for three commonly used metrics across the dimensions of completeness and correctness in Table 1; further dimensions and metrics can be found in the supplemental table. Unfortunately, many evaluations found that automated metrics only loosely reflected human preference, with correlations sometimes even being negative. Similarly, automated readability scores (e.g., based on syllables per word) have correlations of 0.1–0.3 with layperson perception[22]. This suggests that even preliminary evaluations warrant small-scale human-evaluations[23]. While the clinical NLP community has explored improved measures and metrics (e.g., DocLens), we present a call to the broader NLP community to test out these new measures of accuracy across use cases and develop better metrics for tasks like readability[24].

Several benchmarking efforts, such as HealthBench[25], MedArena[26], have recently focused on grounded rubric-based or preference-based human-in-the-loop evaluations as an alternative. However, challenges persist. For example, HealthBench consists largely of synthetically expert-generated examples; evaluation requires custom rubrics per example, which is difficult to scale more broadly, particularly considering changing guidelines and complex patient context.

Discrepancy in translational impact

Finally, there is a disconnect between the dimensions emphasized in current evaluations and the potential for utility of LLM deployments in healthcare. For example, for in-basket messaging applications, preliminary results have shown benefits not in saved time for drafting responses, the primary motivation for the pilot studies, but in the perceived cognitive burden of the task[27–29], for which tangible and standardized metrics have not been established. Current human evaluations fall short of rigorous experiment design to ensure that studies can surface relevant insights (tangibly measured with statistical significance) on all evaluation targets, such as utility, edge cases, across clinical domains and specialties of interest, patient representation, etc. There is a broad swath of classical NLP and informatics literature emphasizing the need for extrinsic evaluations, both in-situ and in-vivo evaluations[30,31]. While an *intrinsic* evaluation might measure if a summary is accurate, an *extrinsic* evaluation could test whether real-world clinicians' workflows are improved in time and accuracy with access to the summary. For question-answering and broadly close-ended applications of clinical LLMs, we emphasize that the bar for evidence must be higher and require a discussion of failure modes, and rare edge cases to help the broader community concretely operationalize potential safety features, enabling the healthcare AI research community to triage algorithmic improvements. While this may entail increased cognitive burden and deliberation for staff responsible for evaluations, it can increase the chance of success of a tool once deployed[7].

Bringing clinical and human context into evaluation

Clinical LLMs are unlikely to be deployed as standalone tools in the foreseeable future highlighting the need for evaluations that account for the

complexity of human-AI collaborations. Here rigorous experiment design must include standard practices such as on-boarding and training of clinicians to the use of the AI tool, and transparently reporting these steps as part of the study design due to their significant impact on the downstream (perceived and actual) utility of the tool, ability of researchers and the broader healthcare community to assess a piece of evidence, and in general raising the bar for the quality of reported evidence of clinical LLM tools.

Meaningful integration of LLMs in healthcare requires a multidisciplinary approach to rigorous evaluation, with vigilance from authors, reviewers, and readers. We must incentivize researchers to be forthcoming about the set of discrepancies present in their work, not just for other researchers, but so non-computational audiences don't have to read between the lines to infer the nuances in a piece of evidence. While every work cannot comprehensively evaluate tools end-to-end, the community has a strong mandate to be upfront and precise about the scope and limitations of experimental evidence provided. This shift requires strategic investments in interdisciplinary expertise—combining implementation scientists, clinical NLP researchers, data scientists, and healthcare practitioners—to create sophisticated, context-aware evaluation methodologies. By establishing standards that systematically validate the utility, limitations, and potential impacts of AI technologies, we can responsibly harness the transformative potential of LLMs and AI more broadly while providing grounded evaluations for impactful systems.

Data availability

No datasets were generated or analyzed during the current study.

Monica Agrawal[1,2], Irene Y. Chen[3,4,5] ✉, Freya Gulamali[2] & Shalmali Joshi[6]

[1]Department of Biostatistics and Biomedical Informatics, Duke University, Durham, NC, USA. [2]Department of Computer Science, Duke University, Durham, NC, USA. [3]Computational Precision Health, University of California, Berkeley, Berkeley, CA, USA. [4]University of California, San Francisco, San Francisco, CA, USA. [5]Department of Electrical Engineering and Computer Science, University of California, Berkeley, Berkeley, CA, USA. [6]Department of Biomedical Informatics, Columbia University, New York, NY, USA. ✉e-mail: iychen@berkeley.edu

Received: 19 April 2025; Accepted: 17 August 2025;
Published online: 07 October 2025

References

1. Shah, S. J. et al. Ambient artificial intelligence scribes: physician burnout and perspectives on usability and documentation burden. *J. Am. Med. Inform. Assoc.* **32**, 375–380 (2025).
2. Rajashekar, N. C. et al. Human-algorithmic interaction using a large language model-augmented artificial intelligence clinical decision support system. In *Proc. CHI Conf. Hum. Factors Comput. Syst.* Vol. 442, 1–20 (Association for Computing Machinery, 2024).
3. Liu, S. et al. Using large language model to guide patients to create efficient and comprehensive clinical care message. *J. Am. Med. Inform. Assoc.* **31**, 1665–1670 (2024).
4. Jeong, D. P., Garg, S., Lipton, Z. C. & Oberst, M. Medical adaptation of large language and vision-language models: Are we making progress? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (eds Al-Onaizan, Y., Bansal, M. & Chen, Y.-N.) 12143–12170. https://aclanthology.org/2024.emnlp-main.677/ (Association for Computational Linguistics, 2024).
5. Li, Y., Harrigian, K., Zirikly, A. & Dredze, M. Are clinical T5 models better for clinical text? In *Proc. 4th Machine Learning for Health Symposium*, vol. 259 of *Proceedings of Machine Learning Research* (eds Hegselmann, S. et al.) 636–667. https://proceedings.mlr.press/v259/li25a.html (PMLR, 2025).
6. Ling, C. et al. Domain specialization as the key to make large language models disruptive: a comprehensive survey. *ACM Comput. Surv.* https://doi.org/10.1145/3764579 (Association for Computing Machinery, 2025).
7. Joshi, S. et al. AI as an intervention: improving clinical outcomes relies on a causal approach to AI development and validation. *J. Am. Med. Inform. Assoc.* **32**, 589–594 (2025).
8. Bedi, S. et al. Testing and evaluation of health care applications of large language models: a systematic review. *JAMA* **333**, 319–328, https://doi.org/10.1001/jama.2024.21700 (2025).
9. Chen, I. Y. et al. Ethical machine learning in healthcare. *Annu. Rev. Biomed. Data Sci.* **4**, 123–144 (2021).
10. Zack, T. et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. *The Lancet Digit. Health* **6**, e12–e22 (2024).
11. Alaa, A. et al. Medical large language model benchmarks should prioritize construct validity. *arXiv preprint* https://doi.org/10.48550/arXiv.2503.10694 (2025).
12. Romanov, A. & Shivade, C. Lessons from natural language inference in the clinical domain. In *Proc. Conf. Empir. Methods Nat. Lang. Process* 1586–1596 https://doi.org/10.18653/v1/D18-1187 (Association for Computational Linguistics, 2018).
13. Herlihy, C. & Rudinger, R. Mednli is not immune: natural language inference artifacts in the clinical domain. In *Proc. Annu. Meet. Assoc.Comput. Linguist. Int. Jt. Conf. Nat. Lang. Process*. Vol. 2, 1020–1027 https://doi.org/10.18653/v1/2021.acl-short.129 (Association for Computational Linguistics, 2021).
14. Raji, ID., Daneshjou, R. & Alsentzer, E. It's time to bench the medical exam benchmark. *NEJM AI* **2**, Ale2401235, https://doi.org/10.1056/Ale2401235 (2025).
15. Ball, J. R., Miller, B. T. & Balogh, E. P. *Improving Diagnosis in Health Care* (National Academies Press, 2015).
16. Li, S. et al. Mediq: question-asking LLMs and a benchmark for reliable interactive clinical reasoning. *Adv. Neural Inf. Process. Syst.* **37**, 28858–28888 (2024).
17. Johri, S. et al. An evaluation framework for clinical use of large language models in patient interaction tasks. *Nat. Med.* **31**, 77–86 (2025).
18. Jin, Q. et al. Hidden flaws behind expert-level accuracy of multimodal GPT-4 vision in medicine. *npj Digit. Med.* **7**, 190 (2024).
19. Lin, C.-Y. Rouge: A package for automatic evaluation of summaries. *Proc. Workshop Text Summarization Branches Out,* 74–81 (Association for Computational Linguistics, 2004).
20. Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proc. 40th Annu. Meet.Assoc. Comput. Linguist* 311–318 (Association for Computational Linguistics, 2002).
21. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. & Artzi, Y. Bertscore: evaluating text generation with Bert. *Proc. Int. Conf. Learn. Rep.* https://openreview.net/forum?id=SkeHuCVFDr (2019).
22. Zheng, J. & Yu, H. Readability formulas and user perceptions of electronic health records difficulty: a corpus study. *J. Med. Internet Res.* **19**, e59 (2017).
23. Tam, T. Y. C. et al. A framework for human evaluation of large language models in healthcare derived from literature review. *NPJ Digit. Med.* **7**, 258 (2024).
24. Xie, Y. et al. Doclens: multi-aspect fine-grained evaluation for medical text generation. In *Proc. 62nd Annu. Meet. Assoc. Comput. Linguist* Vol. 1, 649–679 https://doi.org/10.18653/v1/2024.acl-long.39 (Association for Computational Linguistics, 2024).
25. Arora, R. K. et al. Healthbench: evaluating large language models towards improved human health. *arXiv preprint arXiv*:2505.08775 https://arxiv.org/abs/2505.08775 (2025).
26. Wu, E., Wu, K. & Zou, J. MedArena: Comparing LLMs for medicine in the wild. *Stanford HAI News* https://hai.stanford.edu/news/medarena-comparing-llms-for-medicine-in-the-wild (2025).
27. Tai-Seale, M. et al. AI-generated draft replies integrated into health records and physicians' electronic communication. *JAMA Netw. Open* **7**, e246565–e246565 (2024).
28. Baxter, S. L., Longhurst, C. A., Millen, M., Sitapati, A. M. & Tai-Seale, M. Generative artificial intelligence responses to patient messages in the electronic health record: early lessons learned. *JAMIA open* **7**, ooae028 (2024).
29. Garcia, P. et al. Artificial intelligence–generated draft replies to patient inbox messages. *JAMA Netw. Open* **7**, e243201–e243201 (2024).
30. Pivovarov, R. & Elhadad, N. Automated methods for the summarization of electronic health records. *J. Am. Med. Inform. Assoc.* **22**, 938–947 (2015).
31. Demner-Fushman, D. & Elhadad, N. Aspiring to unintended consequences of natural language processing: a review of recent developments in clinical and consumer-generated text processing. *Yearb. Med. Inform.* **25**, 224–233 (2016).
32. Chao, C.-J. et al. Evaluating large language models in echocardiography reporting: opportunities and challenges. *Eur. heart j. Digit. health* **6**, 326–339 (2025).
33. Abacha, A. B. et al. An investigation of evaluation methods in automatic medical note generation. In *Findings of the Association for Computational Linguistics: ACL* (2023).
34. Van Veen, D. et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nat. Med.* **30**, 1134–1142 (2024).
35. Moramarco, F. et al. Human evaluation and correlation with automatic metrics in consultation note generation. In *Proc. 60th Annual Meeting of the Association for Computational Linguistics* Vol. 1, 5739–5754 (Association for Computational Linguistics, 2022).
36. Wang, L. L. et al. Automated metrics for medical multi-document summarization disagree with human evaluations. In *Proc. conference. Association for Computational Linguistics. Meeting* Vol. 1, 9871–9889 (2023).
37. Abacha, A. B., Yim, W.-w., Fan, Y. & Lin, T. An empirical study of clinical note generation from doctor-patient encounters. In *Proc. 17th Conference of the European Chapter of the Association for Computational Linguistics* 2291–2302 (2023).

Acknowledgements

M.A. is grateful for funding from a Duke Whitehead Scholar Award.

Author contributions

M.A., I.Y.C., F.G. and S.J. all contributed to the writing of the work. F.G. prepared Table 1.

Competing interests
M.A. is a co-founder of Layer Health.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-025-01963-x.

Correspondence and requests for materials should be addressed to Irene Y. Chen.