Letter to the Editor

Author's Reply: Critical Limitations in Systematic Reviews of Large Language Models in Health Care

Andre Python[1,2,3], PhD; HongYi Li[1,4], BS; Jun-Fen Fu[5,6,7], MD, PhD

[1]Center for Data Science, Zhejiang University, Hangzhou, China
[2]School of Medicine, Zhejiang University, Hangzhou, China
[3]Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom
[4]School of Mathematical Sciences, Zhejiang University, Hangzhou, China
[5]School of Medicine, Children's Hospital of Zhejiang University, Hangzhou, China
[6]National Clinical Research Center for Child Health, Hangzhou, China
[7]National Regional Center for Children's Health, Hangzhou, China

Corresponding Author:

Andre Python, PhD
Centre for Human Genetics, Nuffield Department of Medicine
University of Oxford
Roosevelt Drive
Oxford OX3 7BN
United Kingdom
Phone: 44 01865 287500
Email: andre.python@well.ox.ac.uk

Related Articles:
Comment on: https://www.jmir.org/2025/1/e81769
Comment on: https://www.jmir.org/2025/1/e71916

J Med Internet Res 2025;27:e82729; doi: 10.2196/82729

Keywords: large language model; LLM; clinical; artificial intelligence; AI; digital health; LLM review; review; letter

Introduction

We thank the correspondent for engaging with our original work [1] and raising constructive points in their Letter [2].

Citation Threshold Bias

We acknowledge that the citation criteria applied to select journals may exclude relevant studies from emerging or specialized venues. Our criteria were not only desirable but necessary to balance comprehensiveness with methodological quality considering the rapidly expanding literature. To mitigate the risk of omission of innovative research, we (1) screened and incorporated all relevant articles from main database platforms as well as e-prints and (2) made available an interactive online guideline offering an up-to-date guide to clinicians.

Definition of "Best Performance"

We acknowledge the concerns associated with the performance comparison of models across heterogeneous contexts. To avoid ambiguity and misinterpretation, we stated and discussed in detail that, in our study, the term "best performance" is solely associated with the findings from the reviewed studies. Our analysis helps identify models successfully applied in clinical studies, without aiming at or implying comparison across domains. We direct readers to the excellent recent work by Liu et al [3] for a comparison of lightweight large language models (LLMs) for medical tasks.

Quality Assessment of the Included Studies

We carried out a thorough quality assessment following PRISMA guidelines [4]. This might have escaped the correspondent's attention, as the details are provided in Multimedia Appendix 2 of our work [1].

Clinical Workflow

The suggested 5-stage workflow does not ignore nor intend to capture the complexity of clinical practice. Rather, it serves as

a framework to associate the reported use of LLMs with tasks and processes familiar to clinicians, in line with a previous study [5]. Our workflow offers a practical assessment of the role and extent of LLMs applied in clinically relevant sectors of activities and tasks.

Clinical Validation Gap

We acknowledge and discuss the challenges in assessing the practicality of their deployment in clinical applications. Complementary to benchmarking LLMs on research datasets, our review covers studies using LLMs in both research and clinical settings. While we identified key challenges of LLMs in real-world applications, a comprehensive assessment of discrepancies between research and clinical settings is clearly beyond the scope.

Safety and Risk Analyses

While our review discusses key concerns of the use of LLMs in clinical settings including hallucination risks and ethical considerations, a comprehensive risk assessment is beyond scope. Future research dedicated to tackle this key topic would require substantial efforts.

Economic Evaluation

Our review assesses the associated costs of the graphics processing unit memory and its cooling requirements by process and clinical tasks. Our interactive online guideline will regularly incorporate future changes in the requirements and costs, as exemplified by the recent rise of lightweight LLMs that may offer excellent performance on consumer-grade hardware. However, a comprehensive cost-effectiveness or return-on-investment analysis is beyond the study scope.

Conclusion

These observations are a timely reminder that our current understanding of the application of LLMs in clinical settings remains provisional and that we need continual reassessment of their current and future roles in health care practice.

Acknowledgments

We declare that no part of this submission has been generated by AI.

Conflicts of Interest

None declared.

References

1. Li H, Fu JF, Python A. Implementing large language models in health care: clinician-focused review with interactive guideline. J Med Internet Res. Jul 11, 2025;27:e71916. [doi: 10.2196/71916] [Medline: 40644686]
2. Weizman Z. Critical limitations in systematic reviews of large language models in health care. J Med Internet Res. 2025;27:e81769. [doi: 10.2196/81769]
3. Liu F, Zhou H, Gu B, et al. Application of large language models in medicine. Nat Rev Bioeng. 2025;3(6):445-464. [doi: 10.1038/s44222-025-00279-5]
4. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. Mar 29, 2021;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
5. Betzler BK, Chen H, Cheng CY, et al. Large language models and their impact in ophthalmology. Lancet Digit Health. Dec 2023;5(12):e917-e924. [doi: 10.1016/S2589-7500(23)00201-7] [Medline: 38000875]

Abbreviations

LLM: large language model

Edited by Tiffany Leung; This is a non–peer-reviewed article; submitted 20.08.2025; final revised version received 26.08.2025; accepted 29.08.2025; published 24.09.2025

Please cite as:
Python A, Li H, Fu JF
Author's Reply: Critical Limitations in Systematic Reviews of Large Language Models in Health Care
J Med Internet Res 2025;27:e82729
URL: https://www.jmir.org/2025/1/e82729
doi: 10.2196/82729