Peer-Review Report

Peer Review of "The Performance of DeepSeek R1 and Gemini 3 in Complex Medical Scenarios: Comparative Study"

Mayank Kejriwal

University of Southern California, Los Angeles, CA, United States

Related Articles:
Preprint (medRxiv): https://www.medrxiv.org/content/10.1101/2025.04.29.25326666v1
Authors' Response to Peer-Review Reports: https://med.jmirx.org/2026/1/e96220
Published Article: https://med.jmirx.org/2026/1/e76822

JMIRx Med 2026;7:e96227; doi: 10.2196/96227

Keywords: large reasoning model; LRM; large language model; LLM; accuracy; medical scenario; DeepSeek R1; Gemini 3

This is a peer-review report for "The Performance of DeepSeek R1 and Gemini 3 in Complex Medical Scenarios: Comparative Study."

Round 1 Review

General Comments

This paper [1] reports on an experimental study to analyze the Massive Multitask Language Understanding Pro (MMLU-Pro) Q&A dataset. The authors find that DeepSeek R1 had an accuracy rate of 95.1% in 162 medical scenarios after reconciliation with subject matter experts on 23 questions. The findings contribute to the growing body of knowledge on large language model applications in health care and provide insights into the strengths and limitations of DeepSeek R1 in this domain.

Specific Comments

Major Comments

1. The results are not appropriately qualified with results on statistical significance, and/or are lacking comparisons with other language models. Even if we know how other models perform overall, it would still be good to have more details, such as a comparison of where one model is right and another is wrong. Those kinds of deep insights are lacking in this paper. All we really know is that DeepSeek performs at a level roughly equivalent to the other leading models (nothing surprising there) and that it sometimes has incomplete or inexplicable behavior. I feel the paper needs to have more results and analysis to be a good fit for this journal.

2. Maybe you could add a workflow diagram/figure to better illustrate the methods?

3. I would like Table 1 to be augmented. Perhaps you can add an example question with answer choices? Right now, it looks very trivial. The alternative is to create a simple bar graph instead of a table, but the former would be more useful.

Conflicts of Interest

None declared.

References

1. Bajwa M, Hoyt R, Knight D, Haider M. The performance of DeepSeek R1 and Gemini 3 in complex medical scenarios: comparative study. JMIRx Med. 2026;7:e76822. [doi: 10.2196/76822]

Abbreviations

MMLU-Pro: Massive Multitask Language Understanding Pro

Edited by Amy Schwartz; This is a non–peer-reviewed article; submitted 26.Mar.2026; final revised version received 26.Mar.2026; accepted 26.Mar.2026; published 27.Apr.2026

Please cite as:
Kejriwal M
Peer Review of "The Performance of DeepSeek R1 and Gemini 3 in Complex Medical Scenarios: Comparative Study"
JMIRx Med 2026;7:e96227
URL: https://med.jmirx.org/2026/1/e96227
doi: 10.2196/96227