Open-source LLM DeepSeek on a par with proprietary models in clinical decision making

By systematic analysis of patient cases, we evaluated the clinical utility of open-source large language models (LLMs), such as the DeepSeek models, for implementation in medical applications. Their performance on clinical decision-making tasks was comparable to and partly better than proprietary models GPT-4o and Gemini-2.0 Flash Thinking Experimental, respectively.

This is a summary of:
Sandmann, S. et al. Benchmark evaluation of DeepSeek large language models in clinical decision-making. *Nat. Med.* https://doi.org/10.1038/s41591-025-03727-2 (2025).

Published online: 18 July 2025

The project

Large language models (LLMs) have the potential to improve patient care by enhancing clinical decision-making and automating administrative tasks[1]. However, compliance with data privacy and medical device regulations pose a challenge for implementation of proprietary LLMs such as GPT-4o[2,3]. Currently, these models do not align with strict medical regulations. Furthermore, a previous study of ours, evaluating GPT-4, GPT-3.5 and Google search for clinical decision-making, identified clear shortcomings of the models, especially in the diagnosis of rare diseases[4]. Open-source LLMs represent a potential alternative for clinical application. Although benchmarking has typically demonstrated an inferior performance of open-source LLMs compared to proprietary state-of-the-art models, open-source LLMs are starting to catch up[5]. DeepSeek was released in January 2025 and news of its superior performance was soon omnipresent; thus, we decided to conduct a systematic benchmarking.

The observation

To assess the performance of open-source LLMs (DeepSeek-V3 and DeepSeek-R1) in comparison to proprietary models GPT-4o and Gemini-2.0 Flash Thinking Experimental (Gem2FTE), we analyzed a set of $n = 125$ patient cases, evaluating clinical decision tasks on differential diagnosis and treatment recommendation. Power calculation was performed to ensure sufficient statistical power for systematic pairwise model testing, adjusting for multiple testing. To enable direct comparison of the results with our previous study that analyzed GPT-3.5, GPT-4 and Google search, an overlapping set of $n = 110$ patient cases was considered. Patient cases covered a balanced set of multiple specialties (internal medicine, neurology, surgery, gynecology and pediatrics) as well as disease frequencies (frequent, less frequent and rare).

With respect to diagnosis, we observed clearly superior performance of all models compared to previously evaluated GPT-4, GPT-3.5 and Google search (Fig. 1). DeepSeek-R1 and GPT-4o performed significantly better than Gem2FTE ($P < 0.001$). However, the open-source LLM DeepSeek-R1 was on a par with the proprietary model GPT-4o. Considering the effect of reasoning empowerment (DeepSeek-R1 versus DeepSeek-V3), no improvement was detected. The reasoning module, which is included in DeepSeek-R1, solely led to longer, but less concise, responses, without any substantial improvement in clinical accuracy. Subgroup analysis showed no difference in the diagnosis of rare diseases compared to frequent diseases for GPT-4o and DeepSeek. Solely in the case of Gem2FTE, we saw an inferior performance for rare diseases. Similar results were observed for evaluating treatment recommendation. Both GPT-4o and DeepSeek-R1 showed superior performance compared to GPT-4 and GPT-3.5. Furthermore, GPT-4o and DeepSeek-R1 significantly outperformed Gem2FTE ($P = 0.0016$ and $P = 0.0235$, respectively). Again, no significant differences could be detected between open-source model DeepSeek-R1 and its proprietary counterpart GPT-4o.

The implications

Our benchmarking study indicates that open-source models, such as DeepSeek, might be a viable solution for the implementation of LLMs in real-world medical applications. These models would allow for secure model training in accordance with data privacy and medical device regulations. Individual, institution-specific training could be realized, providing a secure and cost-effective approach for the implementation of LLMs in clinical settings.

However, inaccuracies in model predictions were sill observed, highlighting the need for expert oversight. Furthermore, it should be noted that our study only covers a portion of potential clinical use cases. Although we considered two highly relevant tasks in clinical decision-making (namely diagnosis and treatment recommendations), further applications for the potential use of LLMs in clinical settings remain to be evaluated; this could range from assisting in the process of clinical documentation, for example, extracting relevant information from doctor–patient interviews, to complex decision support tasks at interdisciplinary meetings such as molecular tumor boards.

The next step will be the implementation of an LLM into real-world clinical use, which will require a prospective clinical observational trial for assessing the performance and benefit of LLMs in routine clinical applications.

Sarah Sandmann[1] & Roland Eils[2]
[1]University of Münster, Münster, Germany.
[2]Fudan University, Shanghai, China.

EXPERT OPINION

"This paper evaluated open-source LLM DeepSeek with proprietary models such as GPT and Google's LLM on clinical decision-making tasks. There were two tasks to be evaluated: diagnosis and treatment recommendations. This is a timely study given that the DeepSeek models have attracted a lot of attention and have shown great potential in many tasks." **Jie Yang, Brigham and Women's Hospital & Harvard Medical School, Boston, MA, USA.**

FIGURE

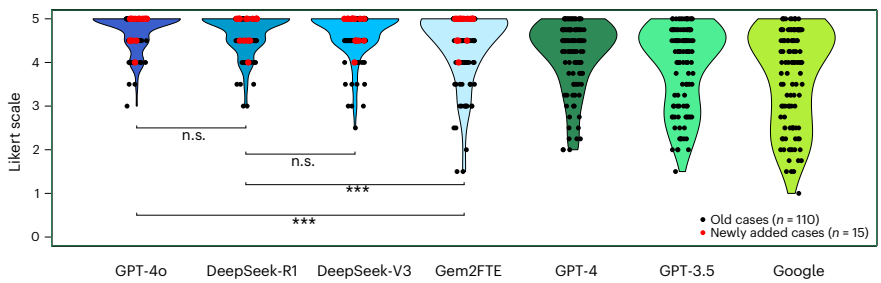

Fig. 1 | Performance of the LLMs for diagnosis tasks. Violin plots show the performance of GPT-4o, DeepSeek-R1, DeepSeek-V3 and Gem2FTE in comparison with GPT-4, GPT-3.5 and Google, evaluated in our previous study[4]. Significant results (***$P < 0.001$) can be observed for GPT-4o versus Gem2FTE and DeepSeek-R1 versus Gem2FTE. Comparison between GPT-4o and open-source model DeepSeek-R1 does not show significant differences, nor does a comparison between DeepSeek-R1 and reasoning model DeepSeek-V3 (n.s., not significant). Black dots show performance on $n = 110$ old cases, already analyzed in our previous study, red dots highlight $n = 15$ new cases.

BEHIND THE PAPER

The release of DeepSeek was fascinating news. The day we read about it, we were keen to test it. Is it really as good as they say? How does it compare with the latest release of OpenAI's GPT models? And how does it perform in comparison to Google's Gemini-2.0 Flash Thinking Experimental, at that moment leading the general nonmedical benchmark on lmarena.ai?

Our previous evaluation of GPT-3.5, GPT-4 and Google search seemed to be the optimum test case to assess the power of DeepSeek's latest LLMs. What followed was an intensive couple of days in which we all worked together very closely. When evaluating individual cases, we immediately realized that the open-source models generated remarkable answers. The final evaluation brought certainty: the open-source models really were on a par with the best currently available proprietary approaches. Now, we hope to take the next step soon, and test a fine-tuned open-source model in routine clinical care. **S.S. & R.E.**

REFERENCES

1. Quer, G. & Topol, E. J. The potential for large language models to transform cardiovascular medicine. *Lancet Digit. Health* **6**, e767–e771 (2024).
   A review article that presents opportunities and limitations of artificial intelligence models in the field of cardiovascular medicine.
2. de Hond, A. et al. From text to treatment: the crucial role of validation for generative large language models in health care. *Lancet Digit. Health* **6**, e441–e443 (2024).
   A comment on the challenge of validating LLMs in healthcare, suggesting general, task-specific and clinical validation.
3. Ong, J. C. L. et al. Medical ethics of large language models in medicine. *NEJM AI* https://doi.org/10.1056/Aira2400038 (2024).
   A review article that presents bioethical principles to promote the responsible use of LLMs, enabling their use ethically, equitably and effectively in medicine.
4. Sandmann, S. et al. Systematic analysis of ChatGPT, google search and Llama 2 for clinical decision support tasks. *Nat Commun.* **15**, 2050 (2024).
   A benchmarking article showing the potential and shortcomings of commercial LLMs for clinical decisions.
5. Hou, G. & Lian, Q. Benchmarking of commercial large language models: ChatGPT, Mistral, and Llama. Preprint at *Research Square* https://doi.org/10.21203/rs.3.rs-4376810/v1 (2024).
   A benchmarking article presenting a critical look at LLMs, showing the need for ongoing evaluations and the potential of hybrid models (that is, combining LLMs and existing systems).

FROM THE EDITOR

"As large language models become increasingly a commonplace tool, it becomes paramount to assess their ability to perform relevant tasks in a comparative manner. The authors here present an early evaluation of DeepSeek, a family of models setting out to challenge the current panel of available tools. The authors use real clinical cases and test the models in clinically relevant tasks, posing the basis for wider evaluations." **Editorial Team, *Nature Medicine*.**