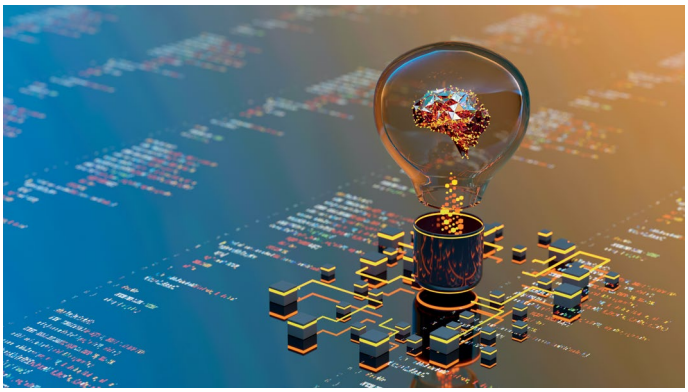
Machine learning

https://doi.org/10.1038/s41591-025-03606-w

Making large language models into reliable physician assistants

Sílvia Mamede & Henk G. Schmidt

A randomized study highlights the potential of large language models to assist clinicians in patient care; future research should harness our understanding of the cognitive processes underlying clinical reasoning to further optimize their utility.



Large language models (LLMs), artificial intelligence applications that can generate humanlike language, have caused a flurry of excitement in the healthcare field. ChatGPT (OpenAI), for example, surprised the medical community by obtaining passing scores on the United States Medical Licensing Examination[1] and performing as well as physicians on board certification exams[2]. Nevertheless, many stakeholders remain (rightly) cautious about the readiness of LLMs for clinical use. An answer to the question of whether superior performance in answering medical questions implies similar performance on clinical tasks is pending. A recent systematic review revealed that the bulk of research on LLMs focuses on their ability to answer medical questions[3]. Performance on diagnosis or treatment accounted for, respectively, less than 20% and 10% of the studies. In this issue of *Nature Medicine*, Goh et al.[4] report on a randomized controlled trial that assessed whether LLM assistance improves physician performance on management reasoning tasks compared with the use of conventional resources alone.

In the study from Goh et al.[4], physicians answered open-ended management reasoning questions for five clinical vignettes in a simulated setting, by using either conventional resources alone (the control condition) or GPT-4 plus conventional resources. Physicians using the LLM scored significantly higher on expert-developed scoring rubrics than the control condition, but did not significantly differ from GPT-4 alone. The authors conclude that these findings seem to support the use of LLMs to assist physicians in management decisions and even as standalone tools in certain scenarios – but they would of course need to be validated in clinical practice.

The study is a welcome effort to explore the performance of LLMs in making case management recommendations, an area that has hitherto received little attention and yielded mixed results. For example, whereas ChatGPT management recommendations for a pool of real patient cases of colorectal cancer achieved high concordance with decisions made by a multidisciplinary team[5], ChatGPT responses to urology cases were insufficiently appropriate and of poor quality[6]. Indeed, recent research evaluating the use of LLMs throughout the clinical workflow with actual patient data recommend caution when considering their application in clinical practice. One study[7], which used a curated database of 2,400 real patient cases of 4 common abdominal pathologies, requested several LLMs to gather and synthesize information to arrive at a diagnostic and treatment plan; the LLMs performed significantly worse than physicians on diagnostic accuracy and followed neither diagnostic nor treatment guidelines.

Whereas these somewhat mixed findings urge caution, they should not obscure the potential of LLMs to contribute to improved physician performance. On the contrary, existing pitfalls of LLMs and challenges for their use should be treated as opportunities and targeted for further study and optimization, to bring us closer to realizing their potential.

First, concerns have been raised over the risk of different types of bias while using LLMs in healthcare. LLMs unavoidably inherit biased information contained in the datasets used in their training[8]. Advances have been made in reducing bias in the responses of the models, but further adjustments are much needed to ensure that their use help reduce real-world disparities in healthcare rather than amplify them.

Second, biases may also come from the interaction between the physician and the model. Physicians tend to rely on the responses of the LLM, especially when they have limited expertise and trust the model. If the response is wrong, overreliance will lead to error, a phenomenon called automation bias[8]. Interventions designed to counteract automation bias in other fields, such as providing aircrews with training on automation bias or displaying prompts for verification, have largely failed.[9] The challenge of fostering verification of the outputs of the LLM may be even harder. Many models cannot explain how they 'reasoned' to arrive at their outputs. It has been advocated that proper prompting can lead the model to provide a rationale for its responses, which would compel physicians to verify their validity[8]. Whether this will actually happen, however, remains to be seen.

Third, errors produced by an LLM may also be caused by misleading information in the disease history of a patient, for instance irrelevant references to a previous illness. ChatGPT has been demonstrated to be equally sensitive to such biasing information as practicing physicians[10].

These potential biases have been recognized and receive attention in research on the use of LLMs to improve clinicians' performance. What remains largely untouched is the following: the most problematic

aspect for the use of LLMs is possibly that they cannot observe a patient independently, at least in the fields of medicine that directly rely on the patient–physician encounter. Presently, LLMs cannot see, hear, smell or perform a physical examination. An LLM relies for its input on the physician's interpretive observations of the patient's clinical features, which are not necessarily objective nor relevant to the patient's actual problem[11]. This input reflects the physician's impressions of the patient's problem, and the LLM cannot do better than the information provided by the physician allows it to do.

To realize their potential, LLMs should not be used in the same (largely ineffective) way as the existing decision support systems have done so far[11], but rather align with what is known from a large body of research on the cognitive processes underlying clinical reasoning. It is well established that physicians, early in a clinical encounter, generate one or a few diagnostic hypotheses based on a limited set of clinical findings – a largely intuitive process of matching the problem at hand with 'scripts' of diseases and clinical scenarios stored in memory. Subsequent data gathering is driven by these early diagnostic hypotheses. The appropriateness of such hypotheses largely determines the ultimate diagnostic accuracy[12]. On the other hand, their early presence may blind the physician to possible alternatives, opening the door to error. An appropriate LLM should therefore challenge the physician to analytically compare actual findings of their patient with findings to be expected if their early diagnostic hypotheses were correct.

Although research on clinical reasoning has not yet been brought to bear on the task of conceiving the best use of LLMs, it has provided evidence on sources of errors in physicians' diagnostic reasoning and strategies to counteract them. If LLMs can be used in such way to help physicians apply these approaches in clinical reasoning, they can probably be even more successful helpers. Of note, these approaches come from research focused on diagnostic reasoning, whereas Goh et al.[4] studied decisions in case management. Based on the performance of the participants (and the LLM) in their study, there may be value in exploring such cognitive reasoning approaches to further optimize the use of LLMs as an adjunct to case management decision-making.

Sílvia Mamede[1,2] ✉ & Henk G. Schmidt[3]

[1]Wenckebach Institute (WIOO), Lifelong Learning, Education and Assessment Research Network (LEARN), University Medical Center Groningen, Groningen, The Netherlands. [2]Institute of Medical Education Research Rotterdam, Erasmus Medical Center, Rotterdam, The Netherlands. [3]Department of Psychology, Education and Child Studies, Erasmus University Rotterdam, Rotterdam, The Netherlands.
✉e-mail: silviamamede@gmail.com

Published online: 24 March 2025

References

1. Kung, T. H. et al. *PLOS Digit. Health* **2**, e0000198 (2023).
2. Jarou, Z. J., Dakka, A., McGuire, D. & Bunting, L. *Ann. Emerg. Med.* **83**, 87–88 (2024).
3. Bedi, S. et al. *JAMA* **333**, 319–328 (2025).
4. Goh, E. et al. *Nat. Med.* https://doi.org/10.1038/s41591-024-03456-y (2025).
5. Horesh, N. et al. *Dis. Colon Rectum* **68**, 41–47 (2025).
6. Cocci, A. et al. *Prostate Cancer Prostatic Dis.* **27**, 103–108 (2024).
7. Hager, P. et al. *Nat. Med.* **30**, 2613–2622 (2024).
8. Zhang, P., Shi, J. & Kamel Boulos, M. N. *Future Internet* **16**, 462 (2024).
9. Lyell, D. & Coiera, E. *J. Am. Med. Inform. Assoc.* **24**, 423–431 (2017).
10. Schmidt, H. G., Rotgans, J. I. & Mamede, S. *J. Gen. Intern. Med.* https://doi.org/10.1007/s11606-024-09177-9 (2024).
11. Schmidt, H. G. & Mamede, S. *Diagnosis* **10**, 38–42 (2023).
12. Norman, G. Research in clinical reasoning: past history and current trends. *Med. Educ.* **39**, 418–427 (2005).

Competing interests

The authors declare no competing interests.