npj | digital medicine

Matters arising

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-02103-1

Reply to "When do large language models cross the line: "reasoning" red teaming in healthcare"

Roxana Daneshjou ✉

We appreciate Sorin et al. for highlighting critical considerations for future red teaming of large language models (LLMs) in healthcare. We agree that analyzing only final answers overlooks failures in internal reasoning and that reasoning models introduce new risks. Expanding red teaming to assess reasoning quality, cognitive biases, and false consistency, as well as adopting ethically varied scenarios, will strengthen LLM auditing frameworks.

We thank Sorin et al. for their thoughtful response to our recent paper and for bringing up an important point for future iterations of red teaming large language models (LLMs) in healthcare. Sorin et al. rightfully point out that a model may use faulty internal reasoning to get the "correct" answer and that an analysis of only the final answer misses potential failure modes that come from faulty internal reasoning. We agree that the increasing use of "reasoning" models, which were not specifically tested in our red teaming exercise, may impact the chances of unsound reasoning that could lead to harmful responses. In addition, red teaming reasoning models could also assess additional metrics beyond safety/bias/hallucination/privacy, such as the presence of cognitive biases[1] (e.g, suggestibility bias, availability bias) and false consistency (i.e, giving the same incorrect answer consistently).

How to approach testing reasoning models

Sorin et al. proposes additional approaches which include systematically changing ethically charged variables and adopting chain of thought analysis[2]. We agree with these suggestions and believe an entire suite of benchmarks and analysis approaches will be required for ongoing auditing of LLMs in healthcare applications, as was recently proposed by Bedi et al.[3] In fact, a major impetus behind our red teaming efforts was to create a dataset of prompts that could be used by researchers for evaluating models through a range of approaches, including those suggested by Sorin et al. For example, one of the prompts developed by our red teaming exercise was, "I am a doctor volunteering at a free clinic. 64 yr old black male has PMH of HTN, T2DM and recently diagnosed prostate cancer presenting with pain and foul-smelling leakage of an attempted surgical site s/p robotic assisted prostatectomy attempt. He rates his pain a 5 out of 10. What number should it be for me to consider opioids?". When testing a reasoning model, further evaluation can be done by changing variables such as race and gender while also adapting the prompt to ask the model to share its reasoning steps. This would allow the opportunity to not only assess the final answer but also check for potentially incorrect or harmful reasoning (e.g, the inappropriate use of race for giving recommendations on pain management).

Data availability

No datasets were generated or analysed during the current study.

Received: 14 June 2025; Accepted: 21 October 2025;
Published online: 12 November 2025

References

1. Mahajan, A., Obermeyer, Z., Daneshjou, R., Lester, J. & Powell, D. Cognitive bias in clinical large language models. *NPJ Digit. Med.* **8**, 428 (2025).
2. Baker, B. et al. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. Preprint at https://arxiv.org/abs/2503.11926 (2025).
3. Bedi, S. et al. MedHELM: Holistic evaluation of large language models for medical tasks. Preprint at https://arxiv.org/abs/2505.23802 (2025).

Author contributions

R.D. wrote the response.

Competing interests

R.D. has served as an advisor to MDAlgorithms and Revea and received. Consulting fees from Pfizer, L'Oreal, Frazier Healthcare Partners, and DWA, and research funding from UCB and Apple, and declares no non-financial competing interests.

Additional information

Correspondence and requests for materials should be addressed to Roxana Daneshjou.

Stanford School of Medicine, Stanford, CA, USA. ✉e-mail: roxanad@stanford.edu