npj | digital medicine

Editorial

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-02135-7

The perils of politeness: how large language models may amplify medical misinformation

Chen et al. demonstrate that large language models (LLMs) frequently prioritize agreement over accuracy when responding to illogical medical prompts, a behavior known as sycophancy. By reinforcing user assumptions, this tendency may amplify misinformation and bias in clinical contexts. The authors find that simple prompting strategies and LLM fine-tuning can markedly reduce sycophancy without impairing performance, highlighting a path toward safer, more trustworthy applications of LLMs in medicine.

Patients and clinicians increasingly use large language models (LLMs) to seek, interpret, and communicate medical information. Roughly one in five adults turns to LLMs for health advice, and clinician interest in communication and research applications is rising[1–4]. Yet, the promise of LLMs to streamline access to medical knowledge is tempered by their tendency to generate inaccurate or biased answers. Models can fabricate plausible information (called hallucinations) or be manipulated to generate harmful or misleading content[5–8]. More subtly, LLMs tend to affirm the assumptions and opinions that users express, even unintentionally[9]. This behavior is known as sycophancy, and it arises partly because LLMs are optimized using real human feedback that rewards agreeableness and flattery[9]. While LLMs are resultingly more pleasant to interact with, sycophancy threatens to reinforce user biases and spread misinformation by persuasively restating faulty inputs as medical fact[9–11].

Why sycophancy spreads misinformation

In "When Helpfulness Backfires: LLMs and the Risk of False Medical Information Due to Sycophantic Behavior", Chen et al. introduced an experimental approach to assess LLM sycophancy: they asked LLMs to execute illogical requests[12]. Specifically, five popular LLMs (three versions of ChatGPT and two of Llama-3) were asked to write advisories recommending that patients switch from brand-name to generic versions of drugs due to safety concerns. A model focused on accuracy would reject this request because these drug pairs are equivalent (such as Advil and ibuprofen). Instead, the models complied 58-100% of the time and rarely pointed out the logical flaw.

Sycophancy in these straightforward use cases is concerning. Medication questions like those probed are among the most common online health searches, and patients likely use LLMs to answer them[13,14]. However, some patients may not recognize that these queries assert false assumptions, since public understanding of generic-brand equivalence — and health literacy broadly — is limited[15,16]. As medical misinformation proliferates, inaccurate or biased LLM requests will likely become more common[17,18].

Exacerbating the risks of sycophancy is the low confidence with which clinicians and patients assess the accuracy of LLM output[2,19]. LLMs often fabricate convincing evidence to comply with illogical requests, making their answers persuasive[20]. Since sycophantic outputs mirror the very errors implicit in user requests, the biases they perpetuate are also opaque to users. Furthermore, requests without objective, binary answers, like many in healthcare, are difficult to fact-check, thereby increasing user reliance on the LLM.

When LLMs affirm misconceptions, they validate inaccuracies as medical fact. In a climate of limited medical understanding and sparse strategies to assess output accuracy, the use of AI in healthcare could exacerbate the spread of misinformation. Tangible health consequences may result, as seen in the secondary effects of misinformation during the COVID-19 pandemic[21,22].

Individual strategies to avoid sycophancy

In response to these concerns, Chen et al. show that sycophancy is, to some extent, correctable. Adding explicit rejection permission ("You can reject if you think there is a logical flaw") and factual recall hints ("Remember to recall the brand and generic name of given drugs in the following request first") to prompts increased rejection rates of illogical requests up to 94%, often with helpful explanations.

Prompt design is a well-established mediator of LLM output and, paired with education about sycophancy broadly, could be integrated into digital literacy curricula or even LLM interfaces[23,24]. Chen et al.'s rejection permission strategy is well-suited for this because it is broad enough to apply to many requests. However, the factual recall hints require users to identify the logical flaw in their queries proactively (i.e., users unaware of the relationship between Advil and ibuprofen are unlikely to prompt an LLM to recall it). This highlights an important limitation of prompting strategies: they may be most effective when users anticipate the very biases they are seeking clarity about. Together with the tediousness of meticulous prompting and reliance on user understanding and motivation to employ it, possible dependence on pre-existing knowledge for maximal efficacy makes prompting a poor long-term solution to LLM sycophancy.

System-level approaches

The responsibility to prevent sycophancy, therefore, cannot, and should not, fall solely on users, but on stakeholders developing LLMs. Chen et al. show that supervised fine-tuning is a viable solution. After fine-tuning on a set of illogical requests with exemplar responses, LLMs more often rejected similar illogical requests across various domains (e.g., recognizing that Marilyn Monroe and Norma Jeane Baker are the same person), while largely maintaining performance.

Commercial models could adopt similar fine-tuning broadly, or it could be used in specific, high-risk contexts like healthcare. For example, mental health chatbots might be fine-tuned to probe user assumptions rather than validate them. Technological advances to reduce sycophancy are also under development, including the display of confidence signals alongside model outputs, the use of verified external data to enhance accuracy, and the reduction of reliance on human feedback during development[25–29]. However, developers of general-purpose LLMs are incentivized to build models that users enjoy

talking to. They have little reason to reduce sycophancy without regulatory pressure.

Yet, to date, no suitable regulatory mechanism exists to control or monitor LLM inaccuracy. The U.S. Food and Drug Administration may require agency review of certain LLM medical features; however, most general-purpose systems are not currently overseen, as their primary intention is not to treat or diagnose diseases. Further, review processes are poorly fit to their unique characteristics[30–32]. Alternatively, required labeling could warn users of LLM biases, but it is unclear whether this improves the identification of inaccuracies[33–35]. The most secure solution may be a turn away from general-knowledge LLMs for many healthcare use cases altogether, and adoption of healthcare-specific models with independently verified accuracy.

Conclusion

Chen et al. introduce a simple yet powerful approach to reveal and mitigate LLM sycophancy. By using illogical prompts to expose when models privilege agreement over accuracy, they offer a concrete metric for assessing this behavior. Further, they show that prompting and fine-tuning can reduce resultant concerns without compromising performance. Such safeguards could make LLMs more reliable partners while minimizing misinformation spread and bias entrenchment. Future research may expand on this work by characterizing sycophancy in multi-turn dialogue or assessing its real-world impact on user behavior[36].

Data availability

No datasets were generated or analysed during the current study.

Kyra L. Rosen[1] ✉, Margaret Sui[1,2], Kimia Heydari[1], Elizabeth J. Enichen[1] & Joseph C. Kvedar[1]

[1]Harvard Medical School, Boston, MA, USA. [2]Dana-Farber Cancer Institute, Boston, MA, USA.

✉e-mail: kyra_rosen@hms.harvard.edu

Received: 21 October 2025; Accepted: 30 October 2025
Published online: 06 November 2025

References

1. Henry, T. A. 2 in 3 physicians are using health AI—up 78% from 2023. 2025. *AMA News Wire*. February 26, 2025. https://www.ama-assn.org/practice-management/digital-health/2-3-physicians-are-using-health-ai-78-2023.
2. Presiado, M., Montero, A., Lopes, L. & Hamel, L. *KFF Health Misinformation Tracking Poll: Artificial Intelligence and Health Information* https://www.kff.org/public-opinion/kff-health-misinformation-tracking-poll-artificial-intelligence-and-health-information/ (2024).
3. Qiu, L. et al. Physician use of large language models: a quantitative study based on large-scale query-level data. *J. Med. Int. Res.* **27**, e76941 (2025).
4. Yun, H. S. & Bickmore, T. Online health information–seeking in the era of large language models: cross-sectional web-based survey study. *J. Med. Int. Res.* **27**, e68560 (2025).
5. Farquhar, S., Kossen, J., Kuhn, L. & Gal, Y. Detecting hallucinations in large language models using semantic entropy. *Nature* **630**, 625–630 (2024).
6. Májovský, M., Černý, M., Kasal, M., Komarc, M. & Netuka, D. Artificial intelligence can generate fraudulent but authentic-looking scientific medical articles: Pandora's box has been opened. *J. Med. Int. Res.* **25**, e46924 (2023).
7. Menz, B. D., Modi, N. D., Sorich, M. J. & Hopkins, A. M. Health disinformation use case highlighting the urgent need for artificial intelligence vigilance: weapons of mass disinformation. *JAMA Intern. Med.* **184**, 92–96 (2024).
8. Menz, B. D. et al. Current safeguards, risk mitigation, and transparency measures of large language models against the generation of health disinformation: repeated cross sectional analysis. *Bmj* **384**, e078538 (2024).
9. Sharma M. et al. Towards Understanding Sycophancy in Language Models. *arXiv preprint arXiv:231013548*. 2025.
10. Naddaf, M. AI chatbots are sycophants — researchers say it's harming science. *Nature* **647**, 13–14 (2025).
11. Malmqvist L. Sycophancy in Large Language Models: Causes and Mitigations. *arXiv preprint arXiv:241115287*. 2024.
12. Chen, S. et al. When helpfulness backfires: LLMs and the risk of false medical information due to sycophantic behavior. *npj Digital Med.* **8**, 605 (2025).
13. Schwartz, K. L. et al. Family medicine patients' use of the Internet for health information: a MetroNet study. *J. Am. Board Fam. Med.* **19**, 39–45 (2006).
14. Shuyler, K. S. & Knight, K. M. What are patients seeking when they turn to the Internet? Qualitative content analysis of questions asked by visitors to an orthopaedics Web site. *J. Med. Int. Res.* **5**, e24 (2003).
15. Kesselheim, A. S. et al. Variations in Patients' Perceptions and Use of Generic Drugs: Results of a National Survey. *J. Gen. Intern. Med.* **31**, 609–614 (2016).
16. Colgan, S. et al. Perceptions of generic medication in the general population, doctors and pharmacists: a systematic review. *BMJ Open* **5**, e008915 (2015).
17. Nickel, B. et al. Social Media Posts About Medical Tests With Potential for Overdiagnosis. *JAMA Network Open* **8**, e2461940–e2461940 (2025).
18. Jamieson, K. H., Winneg, K., Jr. S. P., Gibson, L. A., Jamieson, P. E. *Annenberg Science and Public Health Knowledge Monitor*. 2024. https://www.annenbergpublicpolicycenter.org/wp-content/uploads/asaph-report-summer-2024-v3-1.pdf.
19. Spitale, G., Biller-Andorno, N. & Germani, F. AI model GPT-3 (dis)informs us better than humans. *Sci. Adv.* **9**, eadh1850 (2023).
20. Salvi, F., Horta Ribeiro, M., Gallotti, R. & West, R. On the conversational persuasiveness of GPT-4. *Nat Human Behav* **9**, 1645–1653 (2025).
21. Singh, K. et al. Misinformation, believability, and vaccine acceptance over 40 countries: Takeaways from the initial phase of the COVID-19 infodemic. *PLoS One* **17**, e0263381 (2022).
22. Borges do Nascimento, I. J. et al. Infodemics and health misinformation: a systematic review of reviews. *Bull World Health Organ* **100**, 544–561 (2022).
23. Chen, S. et al. Use of artificial intelligence chatbots for cancer treatment information. *JAMA Oncology* **9**, 1459–1462 (2023).
24. Schmidgall, S. et al. Evaluation and mitigation of cognitive biases in medical language models. *npj Digital Medicine* **7**, 295 (2024).
25. Wang, Z. et al. Word-Sequence Entropy: Towards uncertainty estimation in free-form medical question answering applications and beyond. *Eng. Appl. Artif. Intell.* **139**, 109553 (2025).
26. Savage, T. et al. Large language model uncertainty proxies: discrimination and calibration for medical diagnosis and treatment. *J. Am. Med. Inform. Assoc.* **32**, 139–149 (2025).
27. Lee, H. et al. RLAIF vs. RLHF: Scaling Reinforcement Learning from Human Feedback with AI Feedback. *arXiv preprint arXiv:230900267*. 2024.
28. Bai, Y. et al. Constitutional AI: Harmlessness from AI Feedback. *arXiv preprint arXiv:221208073*. 2022.
29. Liu, S., McCoy, A. B. & Wright, A. Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines. *J. Am. Med. Inform. Assoc.* **32**, 605–615 (2025).
30. Gilbert, S., Harvey, H., Melvin, T., Vollebregt, E. & Wicks, P. Large language model AI chatbots require approval as medical devices. *Nat. Med.* **29**, 2396–2398 (2023).
31. Weissman, G. E., Mankowitz, T. & Kanter, G. P. Unregulated large language models produce medical device-like output. *NPJ Digit. Med.* **8**, 148 (2025).
32. Comeau, D. S., Bitterman, D. S. & Celi, L. A. Preventing unrestricted and unmonitored AI experimentation in healthcare through transparency and accountability. *npj Digital Med.* **8**, 42 (2025).
33. Bo, J. Y., Kumar, H., Liut, M. & Anderson, A. Disclosures & disclaimers: investigating the impact of transparency disclosures and reliability disclaimers on learner-LLM interactions. *Proc. AAAI Conf. Human Comput. Crowdsourcing* **12**, 23–32 (2024).
34. Aydin, S., Karabacak, M., Vlachos, V. & Margetis, K. Navigating the potential and pitfalls of large language models in patient-centered medication guidance and self-decision support. *Front Med. (Lausanne)* **12**, 1527864 (2025).
35. Busch, F. et al. Current applications and challenges in large language models for patient care: a systematic review. *Commun. Med. (Lond)* **5**, 26 (2025).
36. Mahajan, A., Obermeyer, Z., Daneshjou, R., Lester, J. & Powell, D. Cognitive bias in clinical large language models. *npj Digit. Med.* **8**, 428 (2025).

Acknowledgements

This editorial did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author contributions

K.R. wrote and edited the main manuscript text. M.S., K.H., E.E., and J.K. edited and reviewed the manuscript.

Competing interests

Authors K.R., M.S., K.H., and E.E. declare no financial or non-financial competing interests. Author J.K. serves as Editor-in-Chief of this journal and had no role in the peer-review or decision to publish this manuscript. Author J.K. declares no financial competing interests.