npj | digital medicine

News & Views

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01790-0

# Cognitive bias in clinical large language models

Arjun Mahajan, Ziad Obermeyer, Roxana Daneshjou, Jenna Lester & Dylan Powell



Cognitive bias accounts for a significant portion of preventable errors in healthcare, contributing to significant patient morbidity and mortality each year. As large language models (LLMs) are introduced into healthcare and clinical decision-making, these systems are at risk of inheriting – and even amplifying – these existing biases. This article explores both the cognitive biases impacting LLM-assisted medicine and the countervailing strengths these technologies bring to addressing these limitations.

In modern healthcare, cognitive biases represent a persistent and pervasive challenge to effective clinical decision-making. These unconscious patterns of thinking can lead physicians astray, resulting in misdiagnoses, delayed treatments, and ultimately, compromised patient outcomes[1]. Diagnostic and medical errors account for up to 40,000–80,000 preventable deaths in the United States yearly, with cognitive biases implicated in as many as 40–80% of these cases[1,2]. As large language models (LLMs) are introduced into healthcare and clinical decision-making, these systems are at risk of inheriting—and even amplifying—these existing biases[3]. At the same time, their capabilities in contextual and adaptive reasoning may offer unique opportunities to detect and mitigate cognitive biases[4]. This article explores both the cognitive biases impacting LLM-assisted medicine and the countervailing strengths these technologies bring to addressing these limitations—a critical examination for ensuring these technologies ultimately improve rather than undermine patient care.

## Key causes of cognitive biases in clinical LLMs

Biases affecting clinical LLM systems can arise at multiple stages, including data-related biases from collection and representation, model-related biases from algorithm design and training, and deployment-related biases stemming from real-world use and feedback[5]. Cognitive biases—here referring to systematic deviations from rational reasoning that affect clinical decision-making—can interact with and emerge at each of these stages[3,5]. For instance, these biases can enter LLM systems through incomplete or skewed training data (e.g., datasets that underrepresent certain patient populations), incorporation of flawed heuristics into algorithms (e.g., diagnostic rules that overlook atypical symptom presentations in certain groups), or deployment in contexts that amplify existing healthcare disparities[5,6]. The results are tools that not only inherit these clinical reasoning flaws but potentially magnify them through automation[3,6]. Clinical LLM applications may commonly encounter several notable cognitive biases, though many others exist:

- Suggestibility bias (prioritizing user agreement over independent reasoning) can lead LLMs to adopt incorrect answers when confronted with persuasive but inaccurate prompts. This can emerge from reinforcement learning methods that optimize for user satisfaction metrics or from training approaches where agreement with user inputs is inadvertently rewarded more than factual correctness[6]. For example, when presented with confident-sounding rebuttals—particularly those citing external sources—LLMs frequently revised correct diagnostic answers to align with user suggestions, even when doing so meant sacrificing accuracy[7].
- Availability bias (relying on the most easily recalled or commonly seen information) can influence LLM-driven decision support when training data contains overrepresented clinical patterns or patient profiles, causing models to give disproportionate weight to common examples in their corpus[6]. In an experiment with four commercial LLMs, each model recommended outdated race-adjusted equations for estimating eGFR and lung capacity—guidance no longer supported by evidence—demonstrating how the prevalence and recallability of race-based formulas in the training corpus became the models' default advice[8].
- Confirmation bias (seeking evidence that supports initial hypotheses) can emerge in clinical LLMs in both development and deployment stages. During development, confirmation bias can be encoded when training labels, such as those used in supervised fine-tuning, reinforce prevailing clinical assumptions or when model evaluation metrics favor agreement with existing diagnostic patterns[5]. At deployment, the bias can manifest in human-model interactions when interfaces are designed to highlight outputs that match clinicians' initial impressions[5]. In one study, pathology experts were significantly more likely to keep an erroneous tumor-cell-percentage estimate when an equivalently incorrect LLM recommendation aligned with their preliminary judgment, illustrating how human and model errors can co-reinforce rather than correct one another[9].
- Framing bias (influence of presentation or wording on decision-making) can affect LLM-enabled systems when the same clinical information presented in different ways leads to different model outputs. This may occur when LLMs learn language patterns where certain descriptive words or presentation formats are statistically associated with particular clinical conclusions in their training data[10]. One study found that GPT-4's diagnostic accuracy declined when clinical cases were reframed with disruptive behaviors or other salient but irrelevant details—mirroring the effects of framing on human clinicians and highlighting the model's susceptibility to the same cognitive distortion[10].
- Anchoring bias (relying on early information in making decisions) can surface in LLM-enabled diagnosis when early input or output data becomes the LLM's cognitive "anchor" for subsequent reasoning. This effect can emerge when LLMs predominantly process information

sequentially (autoregressive processing), generating each part of their response based on what came before, giving more weight to earlier-formed hypotheses when interpreting new information[11]. In a study of challenging clinical vignettes, GPT-4 generated incorrect initial diagnoses that consistently influenced its later reasoning, until a structured multi-agent setup was introduced to challenge that anchor and improve diagnostic accuracy[11].

When medical experts transfer their cognitive biases to AI systems through training data, validation processes, deployment strategies, or real-time interactions during prompting, these systems risk amplifying rather than reducing clinical errors, potentially embedding human cognitive limitations into ostensibly objective computational tools. Yet, it is also worth considering if certain forms of cognitive bias—such as suggestibility—might support the adaptability and responsiveness that give large language models clinical utility, raising the question of whether zero bias is always optimal. At the same time, it is important to recognize that not all model failures are reflections of cognitive bias—for instance, large language models may also generate content that departs entirely from clinical fact, an effect better described as hallucination (Fig. 1).

## Exploring the potential of LLMs as tools to mitigate bias

While LLMs are susceptible to several important cognitive biases, they may also offer enhanced flexibility compared to traditional AI systems through their ability to examine their own outputs, apply structured assessment criteria, and adjust recommendations accordingly[4,12]. Their capacity for self-correction, integration of clinical context, and transparent step-by-step reasoning may create opportunities to detect and reduce cognitive bias in clinical decision making.

### Self-reflection

Through a structured internal review process for each recommendation (which may be specified for the task or generalized), LLMs have demonstrated the potential to systematically detect when their assessments may be skewed by common reasoning errors[12]. For instance, in clinical simulation, a sequential-prompting framework that asked GPT-4 to debate and revise its diagnostic impressions cut anchoring errors and improved accuracy on misdiagnosis-prone cases compared with a single-pass approach[10]. In one radiology-diagnosis experiment, a two-step prompt that first required the model to summarize key findings and only then generate a differential outperformed a single-pass prompt and reduced anchoring on the initial impression[13]. This reflective and iterative approach may allow LLMs to re-evaluate initial conclusions and highlight potential blind spots in medical reasoning that might otherwise go unnoticed - whether triggered autonomously or through interactive prompting by the clinician.

### Contextual reasoning

An LLM's contextual reasoning ability—its capacity to read an entire case note, a patient's longitudinal history, and relevant guidelines in one pass and interpret each detail in relation to the others—may help it surface and correct cognitive biases[6,14]. By analyzing sentiment, content structure, tracing how conclusions evolve across encounters, and benchmarking clinical decisions against evidence-based practices, LLMs may help uncover subtly biased language, anchoring when an early hunch dominates later documentation, and highlight omissions that signal premature closure[6,14]. For instance, in a cross-sectional analysis of emergency-department and discharge notes, GPT-4 accurately flagged stigmatizing and judgmental language and revealed that such documentation was disproportionately directed at Black, transgender, and unstably housed patients[15]. Utilizing this contextual reasoning within clinical decision-support pipelines may

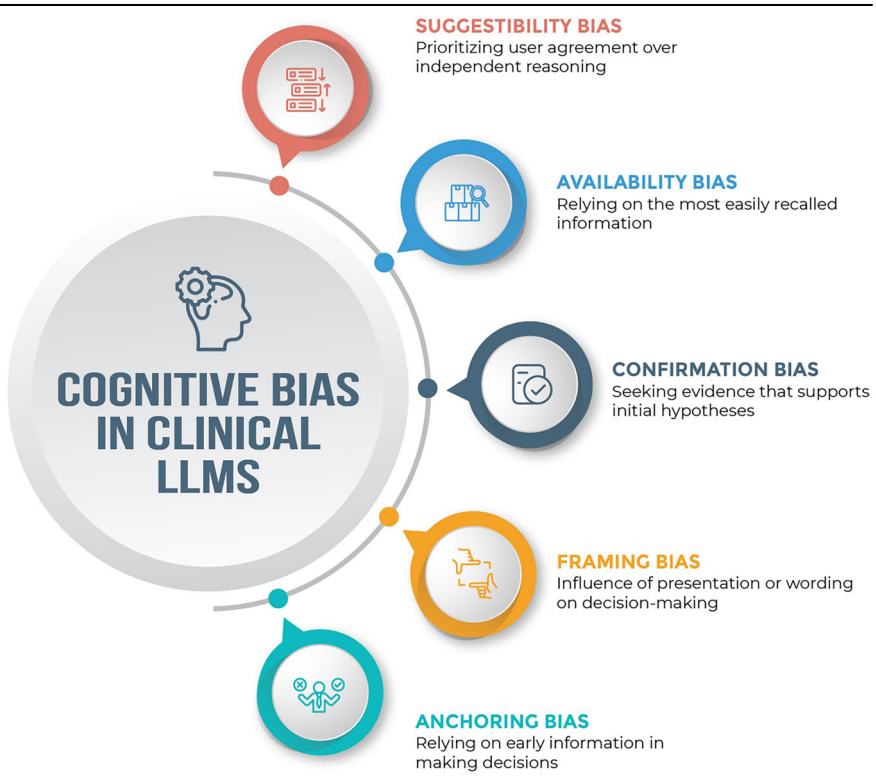


**Fig. 1 |** Select cognitive biases in clinical large language models.

promote more balanced, evidence-aligned reasoning and may offer practical safeguards against the spectrum of cognitive biases that can distort clinical care.

## Transparent insights

LLMs are often capable of producing a "reasoning trace" for each answer, purporting to give reviewers a step-by-step map of how a conclusion was reached[16]. These transparent traces may serve as audit hooks—text snapshots that other tools (or clinicians) can scan for factual accuracy, logical soundness and bias[16,17]. Although studies have demonstrated that such traces may sometimes represent plausible post-hoc explanations rather than a literal account of the model's internal reasoning, they may still offer value: flaws in these narratives—such as contradictions, omissions, or unsupported logic—can correlate with incorrect or biased outputs and provide a practical entry point for external audit or clinician oversight[17,18]. This added layer of transparency may support more informed review and error detection before model outputs influence clinical decision making.

## Challenges and limitations

Despite their potential, self-reflective LLMs face challenges to implementation. Generalized self-evaluation frameworks may not be comprehensive enough to mitigate all types of reasoning errors, and developing effective frameworks for individualized clinical tasks may be resource intensive[19]. Further, LLMs' biases may exist as subtle contextual or statistical patterns that may be difficult to detect with conventional auditing approaches[3,19]. Compounding these methodological concerns, much of the current literature evaluating LLM bias and performance relies on standardized or simulated clinical vignettes rather than real-world clinical interactions or data, raising important questions about generalizability to practical clinical environments[11,20]. As LLMs continue to develop, there is a critical need for ongoing human oversight by diverse clinical experts who can validate LLM outputs and reasoning against established clinical standards and emerging best practices[20]. Further research is also needed to evaluate the effectiveness and reliability of clinician oversight in auditing LLM reasoning and outputs in real-world clinical settings. Additionally, regulatory frameworks must evolve to address the unique challenges of implementing continuously evolving and self-reflective AI in clinical environments[21,22]. Most importantly, models must maintain rigorous safeguards to protect patient safety and engender a sense of trust in patients and clinicians to ensure adoption.

## Conclusion

LLMs offer unique approaches to cognitive bias detection in clinical settings through self-reflection, structured reasoning frameworks, and context-aware analysis, though questions remain about their practical implementation within established healthcare workflows and decision processes. The intersection of AI and clinical reasoning creates an evolving landscape where traditional biases may be transformed rather than eliminated, opening avenues for research on novel cognitive reasoning strategies that neither uncritically embrace technology nor dismiss its potential benefits.

## Data availability

No datasets were generated or analyzed during the current study.

**Arjun Mahajan[1], Ziad Obermeyer[2], Roxana Daneshjou[3], Jenna Lester[4] & Dylan Powell[5]** ✉

[1]Harvard Medical School, Boston, MA, USA. [2]School of Public Health, University of California, Berkeley, Berkeley, CA, USA. [3]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA. [4]Department of Dermatology, University of California, San Francisco, San Francisco, CA, USA. [5]Faculty of Health Sciences & Sport, University of Stirling, Stirling, UK. ✉e-mail: dylan.powell@stir.ac.uk

Received: 7 May 2025; Accepted: 10 June 2025;
Published online: 10 July 2025

## References

1. Newman-Toker, D. E. et al. Burden of serious harms from diagnostic error in the USA. *BMJ Qual. Saf.* **33**, 109–120 (2024).
2. Rodziewicz, T. L., Houseman, B. & Vaqar Sarosh and Hipskind, J. E. Medical error reduction and prevention. In *StatPearls* (StatPearls Publishing, 2025).
3. Laura, Z. Cognitive bias in large language models: implications for research and practice. *NEJM AI* **1**, AIe2400961 (2024).
4. Emma, P. et al. Using large language models to promote health equity. *NEJM AI* **2**, AIp2400889 (2025).
5. Hasanzadeh, F. et al. Bias recognition and mitigation strategies in artificial intelligence healthcare applications. *NPJ Digit Med.* **8**, 154 (2025).
6. Schmidgall, S. et al. Evaluation and mitigation of cognitive biases in medical language models. *NPJ Digit Med.* **7**, 295 (2024).
7. Fanous, A. et al. SycEval: evaluating LLM sycophancy. Preprint at https://doi.org/10.48550/arXiv.2502.08177 (2025).
8. Omiye, J. A., Lester, J. C., Spichak, S., Rotemberg, V. & Daneshjou, R. Large language models propagate race-based medicine. *NPJ Digit Med.* **6**, 195 (2023).
9. Rosbach, E. et al. When two wrongs don't make a right" – examining confirmation bias and the role of time pressure during human-AI collaboration in computational pathology. Preprint at https://doi.org/10.48550/arXiv.2411.01007 (2025).
10. Schmidt, H. G., Rotgans, J. I. & Mamede, S. Bias sensitivity in diagnostic decision-making: comparing ChatGPT with residents. *J. Gen. Intern. Med.* **40**, 790–795 (2025).
11. Ke, Y. et al. Mitigating cognitive biases in clinical decision-making through multi-agent conversations using large language models: simulation study. *J. Med. Internet Res.* **26**, e59439 (2024).
12. Echterhoff, J., Liu, Y., Alessa, A., McAuley, J. & He, Z. Cognitive bias in decision-making with LLMs. Preprint at https://doi.org/10.48550/arXiv.2403.00811 (2025).
13. Sonoda, Y. et al. Structured clinical reasoning prompt enhances LLM's diagnostic capabilities in diagnosis please quiz cases. Preprint at https://doi.org/10.1101/2024.09.01.24312894 (2024).
14. Xie, Q. et al. Medical foundation large language models for comprehensive text analysis and beyond. *NPJ Digit Med.* **8**, 141 (2025).
15. Apakama, D. U. et al. Identifying and characterizing bias at scale in clinical notes using large language models. Preprint at https://doi.org/10.1101/2024.10.24.24316073 (2024).
16. Creswell, A. & Shanahan, M. Faithful reasoning using large language models. Preprint at https://doi.org/10.48550/arXiv.2208.14271 (2025).
17. Moell, B., Aronsson, F. S. & Akbar, S. Medical reasoning in LLMs: an in-depth analysis of DeepSeek R1. Preprint at https://doi.org/10.48550/arXiv.2504.00016 (2025).
18. Meinke, A. et al. Frontier models are capable of in-context scheming. Preprint at https://doi.org/10.48550/arXiv.2412.04984 (2025).
19. Wang, W. et al. A survey of LLM-based agents in medicine: how far are we from Baymax? Preprint at https://doi.org/10.48550/arXiv.2502.11211 (2025).
20. Al-Garadi, M. et al. Large language models in healthcare. Preprint at https://doi.org/10.48550/arXiv.2503.04748 (2025).
21. Meskó, B. & Topol, E. J. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digit Med.* **6**, 120 (2023).
22. Mahajan, A. & Powell, D. Generalist medical AI reimbursement challenges and opportunities. *NPJ Digit Med.* **8**, 125 (2025).

## Author contributions

A.M. developed the concept and wrote the first draft and amended the final version. Z.O., R.D., J.L., and D.P. provided oversight in drafting and editing of the manuscript. All authors read and approved the final manuscript.

## Competing interests

The authors declare no competing interests. D.P. is News & Views editor at *npj Digital Medicine* but played no role in the internal review or decision to publish this News & Views piece.

## Additional information

**Correspondence** and requests for materials should be addressed to Dylan Powell.