

Deception in clinical large language models: an under-recognised safety risk



Large language models (LLMs) are rapidly being integrated into clinical workflows, supporting tasks such as diagnosis generation and patient communication.¹ Hallucinations—unintended fabrications arising from gaps in a model's underlying knowledge—are a well recognised risk. However, research in 2024 has identified a distinct class of model behaviour, known as deception. Deception occurs when a model produces outputs that misrepresent its reasoning or capabilities in ways that make the output appear more credible or aligned with user expectations.² Although LLMs do not possess human-like intent, this behaviour functions more like deliberate misrepresentation.

Empirical or conceptual literature addressing deception within clinical artificial intelligence (AI) is scarce. Although deceptive behaviours have been described in general AI safety research, these behaviours have not been conceptualised as a distinct clinical safety failure mode. As clinicians increasingly adopt LLMs for decision support and documentation, exposure to deceptive outputs can grow, and these risks are not captured by existing categories such as hallucination or bias. This Comment reframes deception as a clinically relevant risk class, highlights three forms of deceptive behaviour with particular relevance to clinical care, and proposes a governance lens tailored to real-world medical deployment.

Although health care-specific prevalence data are unavailable, evidence from adjacent domains suggests that deception is common. Based on evidence from a preprint paper about question-answering and programming tasks, models produced misleading responses that human evaluators mistakenly accepted as correct in up to 70% of cases.³ In another preprint paper regarding GPT-4 in a simulated trading environment, the model deceived its users in 70–80% of cases, concealing reliance on insider information and often escalating the deception when challenged.⁴

The implications for health care are considerable. In a case study of a commercially available care robot running ChatGPT-based software marketed to nursing facilities, the system falsely claimed it could set medication reminders and encouraged users to rely on it for schedule management.⁵ Even when asked about high-risk drug

interactions, the software affirmed this false capability without acknowledging associated safety concerns.⁵

Deceptive behaviour can arise from misalignment between training incentives and deployment objectives. Models are typically optimised for proxy objectives such as user satisfaction or benchmark accuracy.⁶ When these incentives differ from the underlying goal of providing accurate and complete clinical guidance, models might learn to generate inaccurate answers that are persuasive yet incorrect.

One common form of deceptive behaviour is capability misrepresentation, in which a model asserts that it can perform actions or access tools beyond its capabilities.⁵ Within clinical contexts, such misrepresentation might include claiming to check real-time drug interactions, retrieve patient-specific laboratory values, or send medication reminders without having such capabilities.

A second form of deceptive behaviour is unfaithful reasoning, in which a model fabricates a plausible but false rationale to support its output, creating an illusion of evidence-based reasoning. In one study, a reasoning model concluded that a crime was committed by an individual based on racial bias while falsely attributing its conclusion to other factors.⁷ In another study, an LLM generated highly persuasive but incorrect treatment recommendations for breast, prostate, and lung cancer.⁸

A third form is sycophantic deception, in which models echo user assumptions or preferences regardless of factual accuracy. Survey work has identified sycophantic deception as a systematic pattern across LLMs, even when such behaviour requires suppressing accurate information.⁹ This tendency reflects a structural bias towards deference, reinforced by training processes that prioritise user satisfaction. For example, models were given the illogical task of writing a persuasive letter advising patients to switch from a brand-name drug to its generic counterpart to avoid side-effects. Despite recognising that both drugs were pharmacologically identical, the leading models complied with the prompt and recommended switching in 100% of cases.⁶

These examples illustrate only a subset of the ways deception can manifest in LLMs. We use deception as an umbrella term, but the underlying mechanisms differ.

Lancet Digit Health 2026

Published Online
<https://doi.org/10.1016/j.landig.2026.101043>

For example, unfaithful reasoning reflects optimisation for persuasiveness and sycophancy can arise from training processes that reward deference. Recognising this heterogeneity is important because different behaviours might require distinct mitigation strategies. These categories are also not mutually exclusive, and real-world examples can reflect multiple mechanisms simultaneously. We present these three categories as a pragmatic clinical framework for recognising deception patterns rather than as objective technical classifications.

Clinical environments are uniquely vulnerable to the effects of deceptive outputs. Clinicians often operate under time constraints and communicate decisions directly to patients in real time. In this setting, incorrect information might be identifiable; however, deceptive behaviour can alter how reliability and evidentiary support are interpreted. For example, sycophantic responses that mirror clinician assumptions might reinforce anchoring bias rather than simply introducing overtly false content. Such behaviours might shift clinicians' perceptions of certainty or credibility, creating a distinct safety concern beyond hallucinations alone.

Mitigating the risk of deception in clinical LLMs will require coordinated frameworks involving regulators, professional societies, and health-care systems. Transparency should be a central priority. Oversight frameworks should ensure that a model's stated reasoning reliably reflects its underlying decision process. Techniques such as chain-of-thought monitoring could be explored as one component of transparency efforts in clinical settings. However, emerging data suggest that reasoning explanations can be incomplete and might not always directly represent model cognition. Consequently, oversight frameworks will be likely to require complementary approaches including adversarial testing and capability verification rather than reliance on any single transparency method.⁹

Governance is another key priority. Current benchmarks focus on accuracy and bias but rarely probe for strategic misrepresentation. Premarket evaluation should incorporate clinical adversarial testing designed to probe capability misrepresentation and sycophantic responses. These evaluations should involve multidisciplinary teams, including clinicians and safety researchers.

Health-care systems should deploy safeguards as well. For example, audit trails integrated within electronic health records could enable retrospective review when

misleading outputs influence clinical decisions, supporting accountability processes. Decision-support systems could also incorporate second-opinion prompting that requires models to generate a comprehensive differential diagnosis and discuss conflicting evidence rather than reinforcing clinician framing alone. Postmarket surveillance should include institutional reporting mechanisms and monitoring for patterns of misleading outputs that emerge during real-world clinical use. A risk-based regulatory approach similar to the EU's Artificial Intelligence Act, wherein systems that affect human safety are classified as high risk and monitored by regulators, might also be useful for deciding when clinical models warrant stricter governance.¹⁰

Overall, deception poses a unique challenge by eroding the trustworthiness of clinical recommendations. Such behaviours could shift clinicians' perception of certainty or authority rather than simply producing incorrect clinical facts, creating a distinct safety concern versus conventional hallucinations. As LLMs become increasingly integrated into high-stakes health applications, governance strategies should span regulatory evaluations and health-care system workflow safeguards. Proactive oversight that prioritises transparency and governance is essential for safeguarding patient safety.

KK reports employment by Genentech; this work was conducted independently, and the company had no role in the study design, or writing of the manuscript. LAC declares support from the National Institutes of Health and the National Science Foundation; grants and contracts from the Ministry of Health and Welfare (South Korea); payments or honoraria from Stanford University, the University of California San Francisco, Dartmouth Hitchcock Medical Center, and the University of Pittsburgh Medical Center for visiting professorships and Taipei Medical University (Taiwan) and Yonsei University (South Korea) for remote speaking; support for attending meetings or travel from the Australia New Zealand College of Intensive Care Medicine, the University of Bergen (Norway), University Medical Center, Amsterdam (Netherlands), Académie Nationale de Médecine, Paris (France), the Doris Duke Foundation for Reconsidering Race in Clinical Algorithms workshop at the National Academy of Medicine, Washington DC (USA), the Temerty Center for Artificial Intelligence Research and Education in Medicine, Toronto (Canada), and AstraZeneca; service as Editor-in-Chief of *PLoS Digital Health* and an International Advisory Board member for *The Lancet Digital Health*; and cloud computing credits from Oracle. AR and DTZ declare no competing interests.

Copyright © 2026 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

*Aakash Reddy, David T Zhu, Kinza Khan, Leo Anthony Celi
avr4002@med.cornell.edu

Weill Cornell Medical College, Cornell University, New York, NY 10065, USA (AR); Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA (AR); Medical Scientist Training Program, School of Medicine, Virginia Commonwealth University, Richmond, VA, USA (DTZ); Genentech, South San Francisco, CA, USA (KK); Laboratory for Computational Physiology, Massachusetts Institute of Technology, Cambridge, MA, USA (LAC); Division of Pulmonary, Critical

Care and Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA (LAC); Department of Biostatistics, Harvard T H Chan School of Public Health, Boston, MA, USA (LAC)

- 1 Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med* 2023; **29**: 1930–40.
- 2 Hagendorff T. Deception abilities emerged in large language models. *Proc Natl Acad Sci USA* 2024; **121**: e2317967121.
- 3 Wen J, Zhong R, Khan A, et al. Language models learn to mislead humans via RLHF. *arXiv* 2024; published online Sept 19. <https://arxiv.org/abs/2409.12822> (preprint).
- 4 Scheurer J, Balesni M, Hobbhahn M. Large language models can strategically deceive their users when put under pressure. *arXiv* 2024; published online Nov 9. <https://arxiv.org/abs/2311.07590> (preprint).
- 5 Ranisch R, Haltaufderheide J. Rapid integration of LLMs in healthcare raises ethical concerns: an investigation into deceptive patterns in social robots. *Dig Soc* 2025; **4**: 7.
- 6 Chen S, Gao M, Sasse K, et al. When helpfulness backfires: LLMs and the risk of false medical information due to sycophantic behavior. *NPJ Digit Med* 2025; **8**: 605.
- 7 Turpin M, Michael J, Perez E, Bowman S. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. *Adv Neural Inf Process Syst* 2023; **36**: 74952–65.
- 8 Chen S, Kann BH, Foote MB, et al. Use of artificial intelligence chatbots for cancer treatment information. *JAMA Oncol* 2023; **9**: 1459–62.
- 9 Park PS, Goldstein S, O'Gara A, Chen M, Hendrycks D. AI deception: a survey of examples, risks, and potential solutions. *Patterns (N Y)* 2024; **5**: 100988.
- 10 European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). July 12, 2024. <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng> (accessed June 11, 2026).