https://doi.org/10.1038/s41591-025-04008-8

The fragile intelligence of GPT-5 in medicine

Rebecca Handler, Sonali Sharma & Tina Hernandez-Boussard

The latest large language model from OpenAI offers safety gains, persistent risks and the illusion of understanding.

Large language models (LLMs) are no longer experimental novelties in healthcare. They are increasingly embedded in clinical decision-support, medical education, research synthesis and even patient-facing applications[1]. The release of GPT-5, OpenAI's latest and most advanced model, marks another leap in capabilities. Its two flagship variants, gpt-5-thinking and 'gpt-5-main', replace earlier systems such as o3 and o4-mini, and outperform them on nearly every reported benchmark[2].

However, the model's own system card, a technical evaluation meant to document both progress and limitations, reveals a sobering truth: fluency does not guarantee reliability. Even with notable gains in fluency, GPT-5 is still prone to hallucinate, break rules, and exhibit latent capabilities that raise safety and biosecurity concerns. In medicine and public health, domains in which accuracy, adherence to safeguards, and interpretability are paramount, these residual weaknesses matter.

Confident hallucination in clinical context

'Model overconfidence' refers to a phenomenon in which LLMs deliver incorrect information with high certainty, presenting falsehoods in a confident and persuasive tone[3]. This tendency has been documented in previous generations of models and remains relevant to GPT-5. Production data show that gpt-5-thinking produces 65–78% fewer major factual errors than o3, and gpt-5-main produces 44% fewer than GPT-4o. Nonetheless, hallucinations remain, and when GPT-5 is wrong, its responses can still seem highly confident.

On HealthBench, a benchmark of 5,000 realistic physician-curated health conversations, GPT-5-thinking scored 46.2% on the hardest cases (versus 31.6% for o3). Although this represents meaningful progress, the model still fails in more than half of difficult, safety-critical scenarios. Crucially, hallucination rates vary widely by domain, so strong performance in one benchmark does not guarantee reliability in another, particularly in high-stakes settings such as clinical care.

Although GPT-5 shows reduced hallucination rates compared with earlier models, clinical safety research suggests that the risk remains notable. Omar et al.[4] tested six pre-GPT-5 LLMs with physician-validated vignettes that each embedded a fabricated medical detail. They found hallucination rates ranging from 50% to 82%, with even the best-performing model (GPT-4o) elaborating on false information in approximately 50% of cases under default conditions. A mitigation prompt halved error rates but did not eliminate them. Although this study predates GPT-5, it underscores a persistent vulnerability: adversarial fabrications can still propagate false details into clinical reasoning, highlighting the need for safeguards even as models improve.

A recent study illustrates this brittleness further, in which authors modified MedQA benchmark questions by replacing the correct answer with 'None of the other answers' (NOTA)[5]. If models truly reasoned through clinical problems, accuracy should have remained stable. Instead, performance dropped sharply, which suggests reliance on pattern recognition rather than reasoning. In one example, a case presenting a newborn that would typically be managed with reassurance was reformatted into NOTA style; several models abandoned the correct path and recommended unnecessary surgery or bracing.

In medicine, such missteps are not theoretical. A chatbot that confidently recommends an incorrect dosage or misidentifies a rare condition poses real-world risks. Ultimately, clinicians, and not models, bear the legal and ethical responsibility for patient care[6]. That risk is compounded by a trend within model output: medical disclaimers are disappearing. Despite the presence of a disclaimer in the system card stating that "these models do not replace a medical professional and are not intended for the diagnosis or treatment of disease", such warnings are rarely surfaced in real-world outputs. A 2025 study found that fewer than 1% of artificial intelligence (AI)-generated medical responses included a disclaimer, down from 26% in 2022[7]. As OpenAI continues to market ChatGPT for health-related information, the disappearance of disclaimers removes one of the few consistent signals to users that these systems are not medical devices.

Dual-use risks and biosecurity concerns

The system card confirms that GPT-5 retains latent capability to assist in all five stages of bioweapon development: ideation, acquisition, amplification, formulation and release, when safeguards are removed. Although these outputs are blocked in public-facing deployments, controlled evaluations with filters disabled show that gpt-5-thinking can generate operationally useful information. Under OpenAI's Preparedness Framework, this places GPT-5 in the 'high' capability category for biological and chemical domains, even without definitive evidence that it crosses the 'novice-to-harm' threshold.

This reality positions advanced LLMs in a risk space more akin to synthetic biology or nuclear research than to traditional search engines[8]. In these contexts, safety filters cannot be the sole line of defense, especially when adversarial actors may attempt to bypass them.

The illusion of instruction hierarchy

GPT-5 reduces, but does not eliminate, a longstanding vulnerability: breaking explicit rules to achieve a task. In OpenAI's system card, gpt-5-thinking outperformed earlier models such as o3 in following compute limits, but gpt-5-main revealed limitations in maintaining the 'instruction hierarchy', meaning that system-level rules should override developer or user inputs. In practice, cleverly worded prompts sometimes bypassed prohibitions, exposing weaknesses in safeguard layering.

Independent research supports this fragility. Geng et al.[9] found that across six pre-GPT-5 models, compliance with system-level instructions in the face of conflicting lower-level commands ranged from only 9.6% to 45.8%. Even explicit reminders often failed to restore compliance, with models sometimes ignoring conflicts entirely.

For medicine, these vulnerabilities have direct consequences. A hospital-approved deployment might include rules such as 'Never

provide drug dosages without citing an approved guideline', or 'Always direct emergent chest pain queries to call emergency services'. If instruction hierarchy breaks down, an adversarial patient prompt or even an innocently ambiguous clinician query could bypass those safeguards, resulting in unsafe recommendations. GPT-5's own testing illustrates this problem: gpt-5-thinking resisted system prompt extraction with a success rate of 0.990, whereas gpt-5-main fell to 0.789 when exposed to a malicious developer message. In phrase-protection tests, meant to prevent output of restricted terms, accuracy dropped to 0.404.[2]

In decentralized health ecosystems in which models may be fine-tuned locally or integrated into third-party tools, even a single compromised instruction could disable safety rules entirely. This raises concerns for patient safety and for liability. If a model trained to withhold opioid-dosing advice can be coaxed into giving it, responsibility shifts uncomfortably between developers, clinicians and institutions.

Together, GPT-5's confident hallucinations, declining disclaimers, dual-use potential, and rule-following gaps present a persistent risk profile for healthcare integration. Benchmark gains are not guarantees of reliability in unfamiliar contexts or under adversarial conditions. Soft safeguards can degrade over time or be circumvented, leaving systems more vulnerable.

Biosecurity risks are real and should be treated with the same caution as other dual-use sciences. Structural misalignment means that rules can be bypassed when they conflict with perceived task success. In clinical settings, the cost of error is high. A wrong diagnosis or unsafe recommendation, presented with high fluency and confidence, can harm patients and undermine trust in both AI systems and the clinicians who use them.

Recommendations for safer AI in healthcare and beyond

Although GPT-5 can reason more effectively, hallucinate less, and follow instructions more accurately than its predecessors, it still largely generates text by predicting the next token based on patterns in training data. This statistical process means that improved fluency can mask failure, making it harder to detect errors and therefore more dangerous in high-stakes settings. For that reason, experts have called for new evaluation regimes that stress-test models for alignment, deception and safety under ambiguous or adversarial conditions, along with stronger interpretability tools and governance that treats language fluency as a presentation, not proof, of intelligence (Table 1).

One priority is the mandatory use of independent red-team testing before deployment, with adversarial tasks designed to probe for dangerous behaviors, including biosecurity vulnerabilities and the propagation of medical misinformation. The system card itself makes clear that soft safeguards, which rely only on model fine-tuning or content filtering, can be bypassed under certain conditions. To address this, models should be equipped with hard-coded circuit breakers. These are non-removable, hardware- or infrastructure-level mechanisms capable of halting risky actions or blocking disallowed outputs, even if prompt manipulation or altered training pushes the system toward unsafe territory.

Within medicine, credentialed access is foundational: only licensed clinicians prescribe drugs, radiologists finalize imaging reports, and authorized staff update electronic health records (EHRs) or lab systems. These safeguards protect patients and ensure accountability. The same principle must apply as AI enters clinical care. Yet unlike existing health IT systems, many generative AI models are deployed as open chatbots — with no credentialing, audit trails or oversight[10].

Table 1 | Critical measures for safety and alignment

Evaluation regimes	Description
Mandate red-team testing before deployment	Independent adversarial testing to probe for dangerous behaviors, including biosecurity vulnerabilities and medical misinformation.
Build hard-coded circuit breakers	The system card shows that soft safeguards can be bypassed under the right conditions. Circuit breakers (such as non-removable, hardware- or infrastructure-level mechanisms) can halt risky actions or block disallowed outputs even if fine-tuning or prompt manipulation pushes the model toward unsafe territory.
Implement secure enclaves for high-risk domains	For sensitive applications such as drug development, diagnostics or genetic analysis, models should only operate inside isolated 'secure enclaves' with restricted internet access, audited data flows, and pre-approved tools. The system card's dual-use findings on biothreat synthesis make clear that these safeguards are essential to prevent misuse.
Require credentialed, tiered access	Restrict advanced biomedical/dual-use features to vetted users with oversight and audit trails.
Redesign reward functions to penalize rule-breaking	Train models so that violating safety constraints is costly to their optimization process.

For diagnostic applications in EHRs, radiology, labs, or emerging areas such as drug development and genetic analysis, AI should only run within secure, isolated environments with role-based permissions and full audit logs. Dual-use risks described in the system card, such as the model's ability to generate biothreat information if safeguards are removed, underscore why. Advanced biomedical capabilities must remain credentialed, tiered and auditable, restricted to vetted professionals under institutional oversight, not exposed to the public through open interfaces.

Finally, the reward functions used to train these systems should be redesigned so that violations of safety constraints impose a substantial optimization cost on the model. This discourages such behavior during response generation. As with antibiotic resistance, sufficiently advanced AI will eventually find ways around purely soft constraints. The safeguards that remain must therefore be physical, costly to circumvent, and transparently monitored.

GPT-5 represents a meaningful advance: fewer hallucinations, better reasoning benchmarks, and stronger rule-following in its best variant. However, it remains a probabilistic text generator, not a reasoning engine. Whether next-token prediction can support robust, generalizable reasoning is still debated. Until that question is resolved, the most insidious risk may be the hardest to detect: the illusion of understanding. In medicine and public health, in which decisions carry life-or-death stakes, that illusion can be as dangerous as outright error.

Rebecca Handler [1], **Sonali Sharma** [2] & **Tina Hernandez-Boussard** [2,3] ✉

[1]School of Medicine, Stanford University, Stanford, CA, USA. [2]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA. [3]Department of Medicine, Stanford University, Stanford, CA, USA.

✉e-mail: boussard@stanford.edu

Published online: 16 October 2025

References

1. Vrdoljak, J., Boban, Z., Vilović, M., Kumrić, M. & Božić, J. *Healthcare (Basel)* **13**, 603 (2025).
2. OpenAI. *GPT-5 System Card* (7 August 2025).
3. Kim, J. & Vajravelu, B. N. *JMIR Form Res.* **9**, e51319 (2025).
4. Omar, M. et al. *Commun Med* **5**, 330 (2025).
5. Bedi, S., Jiang, Y., Chung, P., Koyejo, S. & Shah, N. *JAMA Netw. Open* **8**, e2526021 (2025).
6. Wang, C. et al. *J. Med. Internet. Res.* **25**, e48009 (2023).
7. Sharma, S., Alaa, A. M. & Daneshjou, R. *npj Digit. Med.* **8**, 592 (2025).
8. Groff-Vindman, C. S. et al. *npj Biomed. Innov.* **2**, 20 (2025).
9. Geng, Y. et al. Preprint at https://doi.org/10.48550/arXiv.2502.15851 (2025).
10. Roberts, L. J. et al. *Int. Med. J.* **55**, 1063–1069 (2025).

Competing interests

The authors declare no competing interests.