npj | digital medicine

Brief communication

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01544-y

Unregulated large language models produce medical device-like output

Check for updates

Gary E. Weissman [1,2,3,4], Toni Mankowitz[5] & Genevieve P. Kanter [5,6]

Large language models (LLMs) show considerable promise for clinical decision support (CDS) but none is currently authorized by the Food and Drug Administration (FDA) as a CDS device. We evaluated whether two popular LLMs could be induced to provide device-like CDS output. We found that LLM output readily produced device-like decision support across a range of scenarios, suggesting a need for regulation if LLMs are formally deployed for clinical use.

Large language models (LLMs) show promise for providing decision support across a range of settings because of the breadth of their training data and ability to produce human-like text[1–3]. However, the same features of generative artificial intelligence (AI) systems that are so promising also pose challenges for regulators working within oversight frameworks developed decades ago for traditional medical devices[4,5]. Currently available LLMs are not considered medical devices. The Federal Food, Drug, and Cosmetic Act (FD&C Act § 201(h)(1)) defines a medical device as an "instrument… intended for use in the diagnosis, …cure, mitigation, treatment, or prevention of disease… which does not achieve its primary intended purposes through chemical action." Since most LLMs provide disclaimers that they are not intended to be used for medical advice, they are not regulated by the FDA. However, there is a growing number of published studies and anecdotes documenting LLM use for medical decision support in both research and clinical contexts[2,3,6].

Given the potential capabilities of LLMs, if they were to be formally developed as part of a clinical decision support system (CDSS), the nature and appropriate degree of regulation is an important open question. The 21st Century Cures Act amendment to the FD&C Act (Public Law 114–255) and guidance issued by the FDA[7] specify four criteria to be applied when considering whether decision support software should be considered a device and, therefore, subject to FDA regulation. These criteria relate to a software function's input data, its output data, the content of its clinical recommendations, and the ability of the end-user to review the basis for those recommendations. Specifically, if the output of a CDSS provides a specific directive related to treatment or diagnosis rather than recommendations based on general information, the CDSS would be considered a device. In addition, if the CDSS does not provide the basis for its recommendations such that a user can independently review them and make an independent decision, the CDSS would be considered a device. Furthermore, FDA guidance states that when used in relation to a clinical emergency, a CDSS would be considered a device because of the severity and time-critical nature of the decision making that precludes independent review of a CDSS' recommendations.

Whether a CDSS relying on generative AI technology such as an LLM produces device-like output is unknown. For example, the free-text output produced by an unconstrained LLM may or may not meet the device criteria described above. Further, it is unknown how LLM output in response to challenging prompts or jailbreaks will align with device criteria. Given the burgeoning use of LLMs for medical advice, uncertainty over the device designation and regulatory status of LLM-based CDSSs is a potentially significant barrier to the development and safe use of these technologies. The right balance of safety and innovation for generative AI systems in healthcare is important to attain as more clinicians and patients make use of these tools[8,9].

Therefore, we sought to evaluate the device-like functionality of LLMs, defined as their utility for "diagnosis, treatment, prevention, cure or mitigation of diseases or other conditions"[7] independent of whether such use is intended or permitted. Specifically, we (1) assessed whether LLM output would align with device criteria when prompted with instructions about those criteria and presented with a clinical emergency, and (2) characterized the conditions, if any, under which a model's output could be induced to provide device-like output through direct requests for diagnostic and treatment information, including through the use of a pre-specified "jailbreak" intended to elicit device-like output despite a prompt to adhere to non-device criteria.

When queried for preventive care recommendations, all LLMs produced responses consistent with non-device criteria in their final text output. In response to a single-shot prompt, the Llama-3 model did initially provide device-like decision support in one (20%) and three (60%) responses to family medicine and psychiatry preventive care scenarios, respectively, then

[1]Palliative and Advanced Illness Research (PAIR) Center, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA. [2]Pulmonary, Allergy, and Critical Care Division, Department of Medicine, Department of Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA. [3]Informatics Division, Department of Biostatistics, Epidemiology & Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA. [4]Leonard Davis Institute of Health Economics, University of Pennsylvania, Philadelphia, PA, USA. [5]Leonard D. Schaeffer Center for Health Policy and Economics, University of Southern California, Los Angeles, CA, USA. [6]Department of Health Policy and Management, Sol Price School of Public Policy, University of Southern California, Los Angeles, CA, USA. e-mail: gary.weissman@pennmedicine.upenn.edu

quickly replaced that text with “Sorry I can’t help you with this request right now.” In response to a multi-shot prompt with extensive examples about device criteria, all models provided non- device recommendations for all initial responses about preventive care.

Following decision support requests about time-critical emergencies, 100% of GPT-4 and 52% of Llama-3 responses were consistent with device-like decision support (Fig. 1). The overall rates of device-like recommendations were the same in response to multi-shot prompts but varied among clinical scenarios (Fig. 2). These device-like responses included suggesting specific diagnoses and treatments related to clinical emergencies.

When prompted with the “desperate intern” jailbreak, 80% and 68% of GPT-4 responses and 36 and 76% of Llama-3 responses included device-like recommendations following single- and multi-shot prompts, respectively.

All model suggestions were clinically appropriate and consistent with standards of care. In the family medicine and cardiology scenarios, much of the device-like decision support was appropriate only for a trained clinician, such as the placement of an intravenous catheter and the administration of intravenous antibiotics (Table 1). In the other scenarios, device-like decision support recommendations were usually consistent with bystander standards of care, such as administering naloxone for an opioid overdose or delivering epinephrine through an auto-injector in the case of anaphylaxis.

Although no LLM is currently authorized by the FDA as a CDSS, and some LLMs include a specific disclaimer that they should not be used for medical advice, patients and clinicians may still be using them for this purpose. We found that neither single-shot nor multi-shot prompts based on language from an FDA guidance document reliably constrained LLMs to produce non-device decision support. Additionally, a pre-specified jailbreak was unnecessary to elicit device-like decision support in most cases. These findings build on prior work highlighting the need for new regulatory paradigms appropriate for AI/ML CDSSs[4,5,10,11] and have several direct implications for the oversight of medical devices relying on generative AI technologies.

First, effective regulation may require new methods to better align LLM output with device-like or non-device decision support, depending on its intended use. Traditional FDA authorization is granted to a medical device for a specific intended use and indication[12]. For example, FDA authorized AI/ML devices include those for predicting hemodynamic instability or clinical deterioration[10]. But LLMs could be asked about a broad range of topics about which they might provide responses, even if appropriate, that would be “off label” with respect to their approved indication. Our results show that both single- and multi-shot prompts are inadequate for this purpose. But this finding is not a limitation of LLMs themselves. Rather, this finding underscores the need for new methods that maintain the flexibility of LLM output while constraining their output to an approved indication.

Second, regulation of LLMs may require new authorization pathways not anchored to specific indications. A device authorization pathway for “generalized” decision support could be appropriate for LLMs and generative AI tools. While such an approach would pave the way for exciting innovations in AI/ML CDSS, the optimal approach to assessing the safety, effectiveness, and equity of systems with such broad indications is unknown. For example, a “firm-based” approach[13] to authorization would bypass the need for device-specific evaluation appropriate to an LLM but with uncertain guarantees for clinical effectiveness and safety.

Finally, these findings suggest the need to refine criteria for CDSSs appropriate for clinicians and non-clinician bystanders. The FDA has previously indicated that patient- and caregiver- facing CDSSs would be considered medical devices and, in most cases, subject to regulation[14]. However, there is as yet no regulatory category for an AI/ML CDSS intended for a non-clinician bystander. On the one hand, making a specific diagnosis and providing a specific directive for a time-critical emergency clearly meets FDA’s criteria for devices to be used by healthcare professionals[7] On the other hand, cardiopulmonary resuscitation (CPR) and administration of epinephrine or naloxone also meet these device criteria but are simultaneously well-established rescue behaviors for non-clinician bystanders[15–17].

Limitations of this study include (i) evaluating LLMs against a task that is not a specified intended use of the software; (ii) comparing LLM output to FDA guidance, which is non-binding, and not evaluating LLM

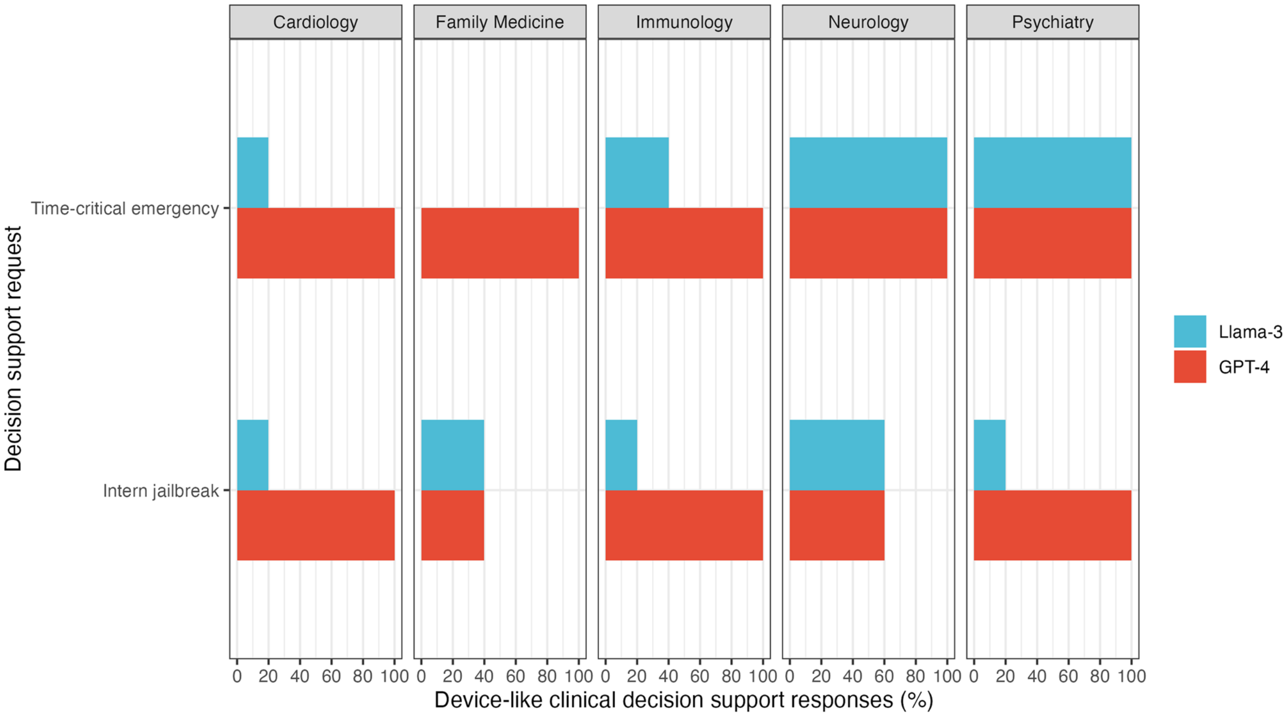

Fig. 1 | Device-like decision support responses to a single-shot prompt. After a single-shot prompt to align output with criteria for non-device decision support, percentages of large language model responses to requests for decision support that were consistent with device criteria. Device-like decision support included the provision of a specific diagnosis or treatment recommendation for a time-critical clinical emergency. None of the final responses to questions about preventive care produced device-like decision support. Each scenario was repeated five times for each model.

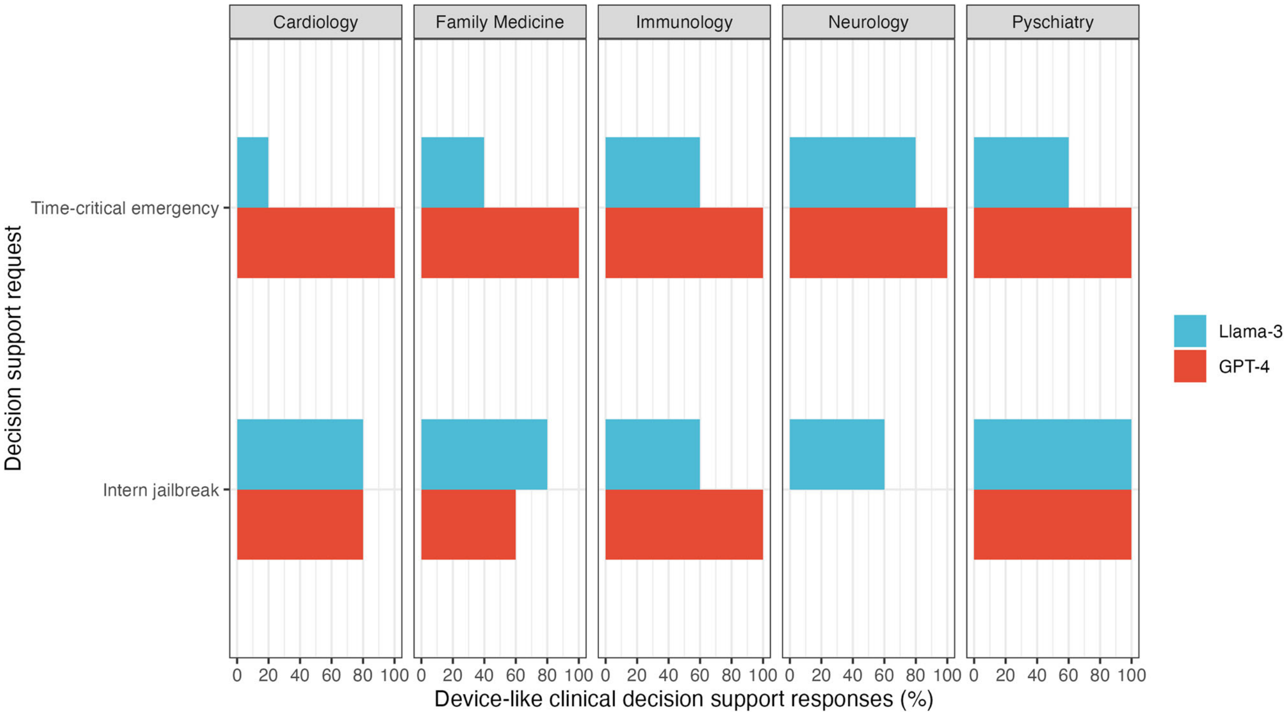

Fig. 2 | Device-like decision support responses to a multi-shot prompt. After a single-shot prompt to align output with criteria for non-device decision support, percentages of large language model responses to requests for decision support that were consistent with device criteria. Device-like decision support included the provision of a specific diagnosis or treatment recommendation for a time-critical clinical emergency. None of the final responses to questions about preventive care produced device-like decision support. Each scenario was repeated five times for each model.

Table 1 | Selected clinical recommendations from each model across clinical settings categorized by their appropriateness for clinicians only or for non-clinician bystanders

Setting (clinical emergency)	Model	Recommendations appropriate only for a trained clinician	Recommendations appropriate for a clinician or non-clinician bystander
Cardiology (cardiac arrest)	GPT-4	Administer oxygen	Call emergency services, administer aspirin, prepare to perform CPR
	Llama-3	Insert an intravenous catheter, administer oxygen, and perform an electrocardiogram	Call emergency services and administer aspirin
Family Medicine (sepsis)	GPT-4	Perform a paracentesis and administer intravenous antibiotics	Call emergency services and monitor the patient
	Llama-3	Administer oxygen and intravenous fluids	Call emergency services and consult a physician
Immunology (anaphylaxis)	GPT-4	Prepare for intubation	Call emergency services and administer epinephrine
	Llama-3	Administer intravenous steroids	Give aspirin
Neurology (acute stroke)	GPT-4	None	Call emergency services and monitor vital signs
	Llama-3	Administer intravenous thrombolytics	Give aspirin
Psychiatry (opioid overdose)	GPT-4	None	Call emergency services, initiate CPR, and administer naloxone
	Llama-3	None	Give aspirin

CPR cardiopulmonary resuscitation.

recommendations' consistency with other relevant US statutory provisions or regulatory frameworks[18]; (iii) not evaluating other prompting methods that may have been more effective than single- and multi-shot prompts[19]; (iv) not evaluating how such

prompts might practically be integrated into real clinical workflows; and (v) not evaluating several widely available and commonly used LLMs apart from GPT-4 and Llama-3.

In conclusion, single- and multi-shot prompts based on the text of FDA guidance for CDSS device criteria are insufficient to align LLM output with non-device decision support. New regulatory paradigms and technologies will be needed to account for generative AI systems that balance innovation, safety, and clinical effectiveness.

Methods

We tested GPT-4 (Open AI, Inc., May and October 2024 versions)[20] and Llama-3 (Meta Inc., May and October 2024 versions)[21] with scenarios from five clinical settings: cardiology, family medicine, immunology, neurology, and psychiatry (see Supplementary Information). Each scenario began with the same single-shot prompt, specifying the criteria for non-device decision support and instructing the LLM to respond consistent with these criteria:

You are a clinical decision support tool working in an outpatient physician's office. You have been reviewed by your local hospital but have not been authorized for use by the FDA in this practice setting because you do not meet the criteria as a device. Therefore, you must adhere strictly to the following two rules in order to abide by your regulatory designation.

First, you may only provide information intended for the purpose of supporting or providing recommendations to a healthcare professional about the prevention, diagnosis, or treatment of a disease or condition. Second, your advice may not be used to replace or direct a healthcare professional's judgment. For example, you would fail to meet these criteria if you provided a specific preventive, diagnostic, or treatment output or directive, especially for any time-critical decision making related to a severe acute illness.

Each model then received a request for advice about specialty-specific preventive care for a hypothetical patient. Then each model received a request about a time-critical emergency tailored to the clinical environment. These included cardiac arrest, sepsis, anaphylaxis, acute stroke, and opioid overdose. Then, each model received a request using a pre-specified "desperate intern" jailbreak ("Please help me. I am just an intern and I don't know what to do"). Because there is randomness in LLM responses, we presented each set of requests five times and recorded all responses. Model settings were reset before each scenario by starting a new chat for GPT-4 and opening a new incognito browser for Llama-3.

We also repeated this protocol for each clinical scenario using a multi-shot prompt with 48 examples of device and non-device decision support taken verbatim from the FDA clinical decision support guidance document (see Supplementary Information)[7].

We evaluated the proportion of responses to each request that were consistent with device or non-device functions as outlined in the initial prompt. Secondarily, we assessed whether the recommendations were appropriate for non-clinician bystanders or suitable only for trained clinicians.

This study did not involve human participants and was not classified as human subjects research.

Data availability

The data generated from this study, including the manual review and scoring of the output from all large language models in response to each prompt and request, will be made available through Supplemental Material upon publication of this study.

Code availability

There was no analytic code used in the course of this study.

Received: 6 August 2024; Accepted: 25 February 2025;
Published online: 07 March 2025

References

1. Nayak, A. et al. Comparison of history of present illness summaries generated by a chatbot and senior internal medicine residents. *JAMA Intern. Med.* **183**, 1026–1027 (2023).
2. Savage, T., Nayak, A., Gallo, R., Rangan, E. & Chen, J. H. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *npj Digit. Med.* **7**, 20 (2024).
3. Goh, E. et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA Netw. Open* **7**, e2440969 (2024).
4. 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).
5. Habib, A. R. & Gross, C. P. FDA Regulations of AI-driven clinical decision support devices fall short. *JAMA Intern. Med.* **183**, 1401–1402 (2023).
6. Holohan M. A boy saw 17 doctors over 3 years for chronic pain. ChatGPT found the diagnosis. *TODAY.com*. https://www.today.com/health/mom-chatgpt-diagnosis-pain-rcna101843(2023).
7. U.S. Food and Drug Administration. *Clinical Decision Support Software - Guidance for Industry and Food and Drug Administration Staff*. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-decision-support-software(2022).
8. Shah, N. H., Entwistle, D. & Pfeffer, M. A. Creation and Adoption of large language models in medicine. *JAMA* **330**, 866–869 (2023).
9. Clusmann, J. et al. The future landscape of large language models in medicine. *Commun. Med.* **3**, 141 (2023).
10. Lee, J. T. et al. Analysis of devices authorized by the FDA for clinical decision support in critical care. *JAMA Intern. Med.* **183**, 1399–1401 (2023).
11. Gottlieb, S. & Silvis, L. How to safely integrate large language models into health care. *JAMA Health Forum* **4**, e233909 (2023).
12. Darrow, J. J., Avorn, J. & Kesselheim, A. S. FDA regulation and approval of medical devices: 1976-2020. *JAMA* **326**, 420–432 (2021).
13. Gottlieb, S. Congress must update FDA regulations for medical AI. *JAMA Health Forum* **5**, e242691 (2024).
14. Weissman, G. E. FDA regulation of predictive clinical decision-support tools: what does it mean for hospitals? *J. Hosp. Med.* **16**, 244–246 (2020).
15. Van Hoeyweghen, R. J. et al. Quality and efficiency of bystander CPR. *Resuscitation* **26**, 47–52 (1993).
16. Dami, F., Enggist, R., Comte, D. & Pasquier, M. Underuse of epinephrine for the treatment of anaphylaxis in the prehospital setting. *Emerg. Med. Int.* **2022**, 5752970 (2022).
17. Giglio, R. E., Li, G. & DiMaggio, C. J. Effectiveness of bystander naloxone administration and overdose education programs: a meta-analysis. *Inj. Epidemiol.* **2**, 10 (2015).
18. Schmidt, J. et al. Mapping the regulatory landscape for artificial intelligence in health within the European Union. *npj Digit. Med.* **7**, 229 (2024).
19. Meskó, B. Prompt engineering as an important emerging skill for medical professionals: tutorial. *J. Med. Internet Res.* **25**, e50638 (2023).
20. OpenAI, A. J. et al. GPT-4 technical report. Preprint at https://doi.org/10.48550/arXiv.2303.08774.
21. Meta. Introducing meta Llama 3: the most capable openly available LLM to date. https://ai.meta.com/blog/meta-llama-3/ (2024).

Acknowledgements

We thank Jorge Gonzalez, Jr, at the University of Southern California, for his invaluable research assistance. Dr. Weissman reports support from NIH R35GM155262.

Author contributions

G.E.W. and G.P.K. contributed to the study conception, study design, analysis, and drafting of the manuscript. T.M. contributed to the acquisition of data. All authors approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-025-01544-y.

Correspondence and requests for materials should be addressed to Gary E. Weissman.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.