Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-026-02420-z

Regulation of clinical Artificial Intelligence (AI) in the Age of Agents: Unconfined Non-Deterministic Clinical Software (UNDCS) systems for healthcare

Check for updates

Caitlyn Tan[1,10], Dinesh Visva Gunasekeran[1,2,3,4,5,10] ✉, Cheng Ooi Low[5,6], Gabrielle Sze Yee Sim[7], Danella Yaoxin Foo[8], Robert J. T. Morris[1,2,11] & Tien Yin Wong[1,3,4,9,11]

In a recent article, Weissman et al.[1] examined the extent to which artificial intelligence (AI)-based large language models (LLMs) generate clinical decision support (CDS) outputs that meet the criteria of regulated medical devices[1] and called for new regulations for LLM-based CDS systems. In this manuscript, we respond to the proposed considerations highlighting those that have been addressed by existing guidelines and would not need new frameworks, as well as the need for new regulations of "generalized" CDSS that are not anchored to specific clinical indications. We contextualise this regulatory gap with an overview of the literature distinguishing between confined and unconfined AI systems. We also outline specific areas in which new regulations may be required, along with risk mitigation strategies that could be incorporated in new guidelines.

ARISING FROM **Weissman et al.** ***npj Digital Medicine*** **https://www.nature.com/articles/s41746-025-01544-y (2025)**

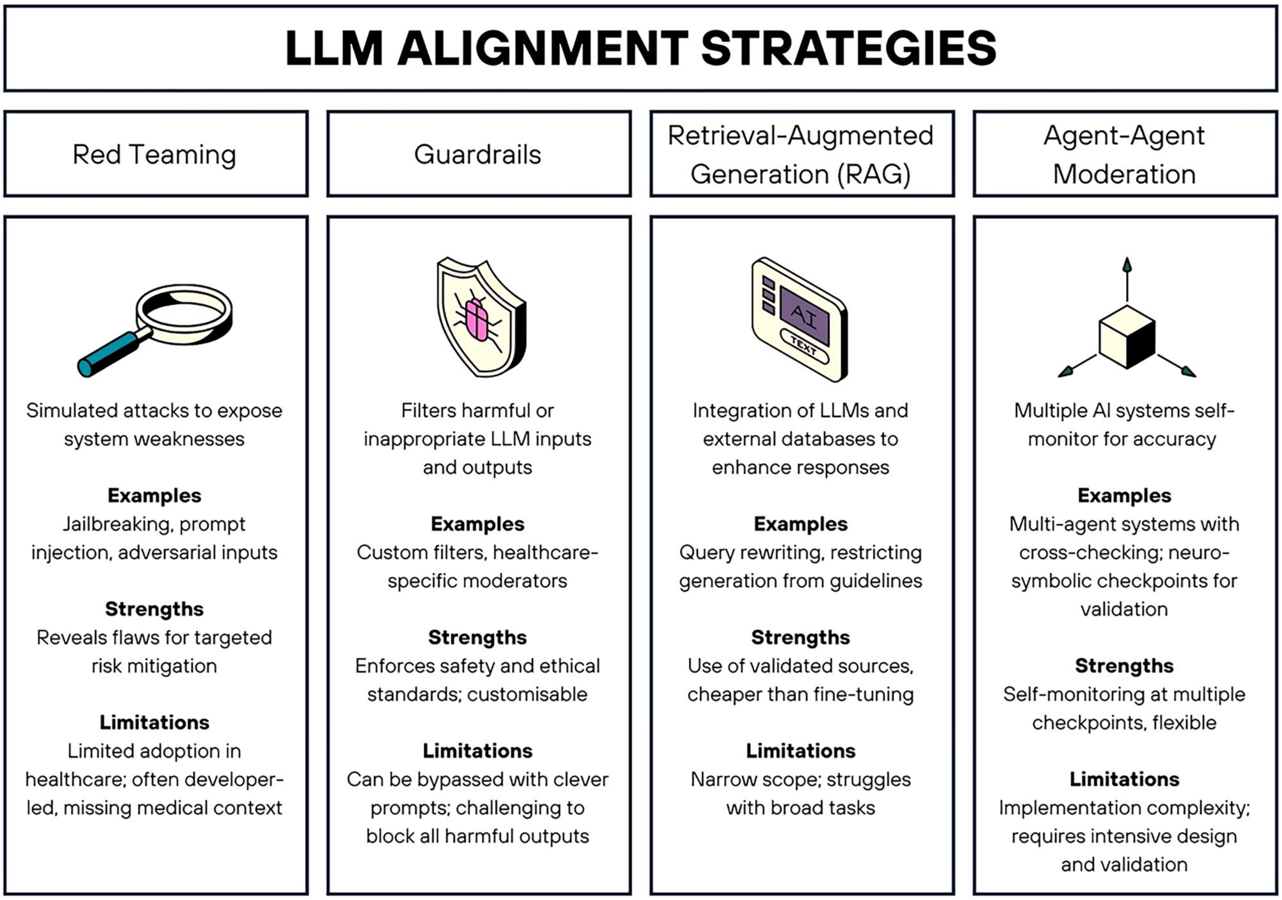
In their recent article, Weissman et al.[1] examined the extent to which artificial intelligence (AI)-based large language models (LLMs) generate clinical decision support (CDS) outputs that meet the criteria of regulated medical devices[1] and called for new regulations for LLM-based CDS systems. In this manuscript, we suggest that some proposed considerations have been addressed by existing guidelines and would not need new frameworks. We also outline specific areas in which new regulations may be required and risk mitigation strategies that could be incorporated.

Calls for Increased Scope of Regulations Given the Gaps Identified in This Research

First, the authors called for regulations to refine LLM CDS criteria pertaining to clinician or non-clinician end-users. Fortunately, this consideration has been embedded in the latest Food and Drug Administration (FDA) Software as a Medical Device (SaMD) guidelines, that are consistent with SaMD regulatory frameworks from the European Union, Australia, and Singapore[2–4]. CDS systems (CDSS) with device-like functionality are currently subject to strict regulatory approval calibrated to risk level across international frameworks[4] while those meeting all four non-device FDA criteria are exempt. Nonetheless, all CDSS (regardless of underlying technology) fall under the broader SaMD framework, with the 2025 update[5] refining explicit compliance requirements based on intended use and end-users (e.g. clinicians or non-clinicians).

Second, all CDSS regardless of their underlying technology, whether Generative AI (GenAI), LLM-based or otherwise, are subject to the current SaMD guidelines[2]. This has been accepted in the digital health field as reflected in publications by us and others. For example, in the APPRAISE study we gathered consensus from over 1000 international experts on clinician acceptance of various CDSS tools in ophthalmology based on

[1]Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore. [2]MOH Office for Healthcare Transformation, Singapore, Singapore. [3]Ophthalmology Academic Clinical Program (ACP), Duke-NUS Medical School, Singapore, Singapore. [4]Singapore Eye Research Institute, Singapore National Eye Centre, Singapore, Singapore. [5]Sheares Healthcare Group, Singapore, Singapore. [6]Chief Medical Informatics Officer (CMIO) Office, Ministry of Health (MOH), Singapore, Singapore. [7]Nanyang Technological University (NTU), Singapore, Singapore. [8]Department of Anaesthesia, National University Hospital Singapore (NUHS), Singapore, Singapore. [9]Tsinghua Medicine, Tsinghua University, Beijing, China. [10]These authors contributed equally: Caitlyn Tan, Dinesh Visva Gunasekeran.[11]These authors jointly supervised this work: Robert J. T Morris, Tien Yin Wong. ✉e-mail: mdcdvg@nus.edu.sg



LLM ALIGNMENT STRATEGIES

Red Teaming	Guardrails	Retrieval-Augmented Generation (RAG)	Agent-Agent Moderation
Simulated attacks to expose system weaknesses	Filters harmful or inappropriate LLM inputs and outputs	Integration of LLMs and external databases to enhance responses	Multiple AI systems self-monitor for accuracy
Examples Jailbreaking, prompt injection, adversarial inputs	**Examples** Custom filters, healthcare-specific moderators	**Examples** Query rewriting, restricting generation from guidelines	**Examples** Multi-agent systems with cross-checking; neuro-symbolic checkpoints for validation
Strengths Reveals flaws for targeted risk mitigation	**Strengths** Enforces safety and ethical standards; customisable	**Strengths** Use of validated sources, cheaper than fine-tuning	**Strengths** Self-monitoring at multiple checkpoints, flexible
Limitations Limited adoption in healthcare; often developer-led, missing medical context	**Limitations** Can be bypassed with clever prompts; challenging to block all harmful outputs	**Limitations** Narrow scope; struggles with broad tasks	**Limitations** Implementation complexity; requires intensive design and validation

Fig. 1 | Strengths and Weaknesses of Potential Safeguards to Facilitate the Alignment of Unconfined Non-Deterministic Clinical Software (UNDCS). Description: This figure outlines the strengths and weaknesses of potential safeguards for unconfined non-deterministic clinical software (UNDCS) including red teaming, guardrails, agent-agent moderation and confined retrieval-augmented generation (RAG) that can be used to facilitate their alignment with intended use.

FDA-SaMD criteria[6]. Therefore, new guidelines for LLM-based CDSS specifically may not be required.

Third, we agree with the authors on the need for new regulations of “generalized” CDSS that are not anchored to specific clinical indications. Contextualizing this regulatory gap requires distinguishing between confined and unconfined AI systems. Early CDSS that were popularised comprised of confined, deterministic clinical software (DCS) algorithms that have known, fixed input data-output label (IDOL) relationships. These systems generated outputs from predefined, bounded labels e.g., binary disease classifiers (present/absent) or categorical risk stratification (low/medium/high). They were well-addressed by existing regulatory guidelines based on evaluation using relatively small, well-characterized datasets. The next iteration of CDSS that followed were confined clinical software (CCS), using techniques such as deep learning (DL) to improve handling of IDOL pairings with unknown relationships. These CCS exhibited predictable variability due to a confined spectrum of output labels, and remained amenable to evaluation with expanded datasets based on existing FDA-SaMD guidelines. While the outputs were conformed in some ways, they were not necessarily safe[7] and so required extensive testing and assurances such as post-processing limits or guardrails.

In contrast, unconfined AI systems, such as general-purpose CDSS using transformer-based LLMs, operate across an open-ended semantic space in response to unstructured input prompts. This design introduces unique risks, including errors and outright “hallucinations”. Hallucinations are semantic errors and can be considered inherent to the engineering of LLMs, as the models are a small approximate representation of a large corpus of training material, using what could be considered a form of data compression[8]. Within unconfined systems, a further distinction may be made based on whether they incorporate non-deterministic components. Most LLMs are based on transformers which are *inherently deterministic* (the same input always results in the same output). However, *non-determinism* may be (sometimes intentionally) introduced via methods such as “temperature”, which involve random selection of the logits at the top layer of the transformer stack, or sometimes caused by inaccuracies in floating point calculations[9,10]. Some contemporary LLMs employ temperature to enhance naturalistic, human-like language generation. This stochasticity produces an unpredictably random spectrum of probabilistically-sampled outputs that are difficult to confine, as demonstrated in Weissman et al’s study. This limits the feasibility of traditional dataset-driven evaluation based on exhaustive testing with large datasets. Together with the massive compression of training data into a relatively small model, this creates a behaviour that could be referred to as non-determinism.

Therefore, we suggest that a new category of regulations may be required for novel general-purpose SaMD solutions developed using GenAI or other AI techniques for generalized CDS, which we term unconfined non-deterministic clinical software (UNDCS). This could apply across all health-related applications of such UNDCS technology, from healthcare administration to health promotion, not limited to CDSS. These regulations may set standards for the inclusion of potential safeguards such as red teaming, guardrails, agent-agent moderation and confined retrieval-augmented generation (RAG). Each approach offers distinct strengths and weaknesses in addressing UNDCS’ unique risks (Fig. 1) detailed in the next section. Unconfined deterministic systems may benefit from red teaming test cases specifically designed to target critical failure modes or underrepresented clinical scenarios, along with multi-agent system (MAS) implementations which offer the potential for *consensus* and thereby lower the frequency of errors. UNDCS can also be improved by extensive testing with repeated sampling and adjudication through LLM-as-a-Judge loops to score and verify aggregate output validity across multiple runs, albeit with some limitations[11].

Risk Mitigation for Unconfined Non-Deterministic Clinical Software (UNDCS) such as Clinical GenAI

First, red teaming involves stress-testing AI systems by simulating challenging scenarios under experimental conditions through techniques such as jailbreaking, prompt injection and adversarial attacks. This also presents an opportunity to involve clinicians early in traditionally developer-driven technical evaluation, providing them practical insight into UNDCS strengths and limitations[12].

Second, guardrails are algorithms that can be used to help filter inappropriate LLM output. Existing open-source frameworks include Llama Guard and Guardrails AI with healthcare-specific implementations demonstrating promise in addressing these risks[13]. However, these systems are underpinned by computationally-based methods that are not always able to consistently check the full spectrum of non-deterministic LLM outputs. Moreover, their susceptibility to jailbreaks[14] further highlights the need to co-implement this method with relevant defences against adversarial attacks[15].

Third, RAG may reduce risks by integrating information retrieval from additional trusted knowledge sources, grounding responses in validated sources. However, it trades versatility for specialization. It is highly effective within its source content domain, yet limited in broader contexts. Potential limitations of RAG include omissions where RAG materials overpower the local context of a query, leading to an incorrect response.

Fourth, the latest developments in AI have introduced software workflows and practical implementations that enable LLM-based agents to perform digital functions. Reinforcement learning with human feedback (RLHF) has emerged as a powerful tool that can both learn from other models (a process called *distillation*) and use human feedback (including aspects of goal and safety alignment) to catalyse a process of open-ended generation tasks[16]. However, limitations of single-agent evaluations include challenges with specialized domains and risk of biases including self-reference bias whereby LLMs have a positive bias towards models types similar to their own[17]. Agent-agent moderation can help address these limitations using MAS architectures. These systems may be further augmented through RAG integration across multiple checkpoints[18], and by incorporating neuro-symbolic models that reason deterministically from validated guidelines for improved reliability[19]. These approaches can help ensure CDSS outputs align with their intended use.

The Need for a New Regulatory Paradigm for UNDCS That Is Not Label-Driven

Given the enormous advances in GenAI techniques, another consideration is whether existing regulatory frameworks are even suitable for novel UNDCS. Current regulations are label-driven, with device classification based on manufacturer-designated intended use. These frameworks were effective for traditional medical devices whereby distribution was confined either to purpose-driven applications or licensed providers, themselves subject to certification requirements and ongoing quality audits. They were used effectively for earlier DCS and CCS SaMDs that had an AI core “wrapped” in a customized application layer bearing the appropriate labels, distributed by a regulated manufacturer.

However, today’s popular LLMs (e.g., ChatGPT, Grok, Claude etc.) are developed by technology providers that control the entire AI supply chain from base model to consumer-facing interface, and may not always detail their training sources. These direct-to-consumer models are not addressed by regulations tied to labelling and manufacturer registration. This regulatory void lacks protections for end users whereby LLM manufacturers have used general-purpose disclaimers while scaling distribution to a broad user base. For example, blanket statements prohibiting use of LLMs for clinical purposes[20] are unlikely to deter real-world use.

These LLMs are now widely accessible and lack the consumer protections afforded by traditional SaMD distribution pathways that ensure appropriate user selection (e.g., based on health and technology literacy), appropriate right-sighting of care (particularly for clinical emergencies) and adverse event monitoring. Weissman et al.[1] have demonstrated that in high-risk situations, LLMs may provide seemingly credible but inappropriate device-like recommendations based on incomplete clinical information, potentially leaving end-users at risk of serious medical harm.

With GenAI techniques advancing in sophistication, static regulations focused on present-day norms risk rapidly becoming outdated[21]. Recent work has demonstrated approaches to evaluate and deploy non-clinical LLM applications such as AI scribes[22]. However, even non-clinical, administrative tools applied in healthcare settings may have unforeseen clinical consequences, such as hallucinations causing flawed documentation or diagnosis labelling that compound errors in downstream clinical decisions, impact medical claims and potentially increase patients’ insurance premiums. Thus, future regulations for all UNDCS may need to set acceptable standards for risk mitigation. Particularly for direct-to-consumer UNDCSs, built-in safeguards may be needed to demonstrably restrict outputs to non-medical device use cases unless formally evaluated in clinical trials and deployed with ongoing quality controls.

In conclusion, while applications of UNDCS such as LLMs in healthcare could yield tremendous clinical benefits, appropriate safeguards are still required for consumer protections and patient safety. As UNDCS blurs the boundaries between intended uses and users, regulators have a challenging responsibility to adopt forward-looking frameworks as agile as the technologies they govern without stifling advances in healthcare transformation. Therefore, a new regulatory paradigm may be needed to encourage the safe use of UNDCS in healthcare, provide consumer protections for the public, and ensure manufacturers are accountable for the software solutions they monetise.

Data availability

No datasets were generated or analysed during the current study.

Received: 5 May 2025; Accepted: 29 January 2026;
Published online: 23 February 2026

References

1. Weissman, G. E., Mankowitz, T. & Kanter, G. P. Unregulated large language models produce medical device-like output. *NPJ Digital Med.* **8**, 148 (2025).
2. International Medical Device Regulators Forum. *Software as a Medical Device (SaMD): Key Definitions.* (2013).
3. Reddy, S. Global harmonization of artificial intelligence-enabled software as a medical device regulation: addressing challenges and unifying standards. *Mayo Clin. Proc.: Digital Health* **3**, 100191 (2025).
4. Health Sciences Authority. Regulatory Guidelines for Software Medical Devices - A Life Cycle Approach (2025).
5. International Medical Device Regulators Forum. Characterization Considerations for Medical Device Software and Software-Specific Risk. (2025).
6. Gunasekeran, D. V. et al. Acceptance and perception of artificial intelligence usability in eye care (APPRAISE) for ophthalmologists: a multinational perspective. *Front. Med.* **9**, 875242 (2022).
7. Seyyed-Kalantari, L., Zhang, H., McDermott, M. B., Chen, I. Y. & Ghassemi, M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat. Med.* **27**, 2176–2182 (2021).
8. Xu, Z., Jain, S. & Kankanhalli, M. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817* (2024).

9. Atil, B. et al. LLM Stability: A detailed analysis with some surprises. *arXiv preprint arXiv:2408.04667* (2024).
10. NVIDIA. Floating Point and IEEE 754. (2025).
11. Ye, J. et al. Justice or prejudice? quantifying biases in llm-as-a-judge. *13th International Conference on Learning Representations (ICLR)* https://scholar.google.com/citations?view_op=view_citation&hl=en&user=ZdgtY0EAAAAJ&citation_for_view=ZdgtY0EAAAAJ:Se3iqnhoufwC (2025).
12. Chang, C. T. et al. Red teaming ChatGPT in medicine to yield real-world insights on model behavior. *npj Digital Med.* **8**, 149 (2025).
13. Hakim, J. B. et al. The need for guardrails with large language models in pharmacovigilance and other medical safety critical settings. *Sci. Rep*. **15**, 27886 (2025).
14. 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).
15. Kraidia, I., Ghenai, A. & Belhaouari, S. B. Defense against adversarial attacks: robust and efficient compressed optimized neural networks. *Sci. Rep.* **14**, 6420 (2024).
16. Zheng, L. et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Adv. Neural Inf. Process. Syst.* **36**, 46595–46623 (2023).
17. Kocaman, V., Kaya, M. A., Feier, A. M. & Talby, D. Clinical Large Language Model Evaluation by Expert Review (CLEVER): Framework Development and Validation. *JMIR AI* **4**, e72153 (2025).
18. Nasim, I. in *AAAI 2025 Workshop on AI Governance: Alignment, Morality, and Law*.
19. Bougzime, O., Jabbar, S., Cruz, C. & Demoly, F. Unlocking the Potential of Generative AI through Neuro-Symbolic Architectures: Benefits and Limitations. https://neurosymbolic-ai-journal.com/paper/unlocking-potential-generative-ai-through-neuro-symbolic-architectures-%E2%80%93-benefits-and (2025).
20. OpenAI. Terms of Use. (2024).
21. Meskó, B. & Topol, E. J. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ digital Med.* **6**, 120 (2023).
22. Cain, C. H. et al. Quality Assurance during the Rapid Implementation of an AI-Assisted Clinical Documentation Support Tool. *NEJM AI* **2**, AIcs2400977, https://doi.org/10.1056/AIcs2400977 (2025).

Author contributions

All authors approve the final version of the manuscript, including the authorship list and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. Writing : All authors - Caitlyn Tan and Dinesh Visva Gunasekeran, Low Cheng Ooi, Sim Sze Yee Gabrielle, Danella Yaoxin Foo, Robert JT Morris and Wong Tien Yin.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Dinesh Visva Gunasekeran.

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.