Duty of Care: A Call for Open and Responsible AI Innovation in Healthcare

Jonathan S. Takeshita[1,2]

[1]Department of Computer Science and School of Cybersecurity, Old Dominion University, Norfolk, VA
[2] Department of Computer Science, Tokyo Institute of Technology, Meguro, Tokyo
jtakeshi@odu.edu

Abstract

Recent advances in AI, especially those of LLMs, bring the prospect of increased adoption of AI in medicine and medical education. In particular, many institutions responsible for medical treatment and education are rapidly aiming to increase AI use in practice and curricula. However, the potential downsides of overuse of AI in these fields are underdiscussed. In the rush to AI adoption, sources of healthcare risk such as LLM reliability, patient privacy, financial and environmental costs, vendor dependencies, and AI over-reliance are often not deeply considered. This paper discusses these recent trends and makes recommendations for healthcare institutions considering further adoption of AI.

Introduction

Recent advances in artificial intelligence (AI), especially the wide release of Large Language Models (LLMs), have the potential to completely reshape many industries and fields. Integration of LLMs has been proposed in fields including finance (Lee, Stevens, and Han 2025), software development (Xu et al. 2024), and cybersecurity (Zhang et al. 2025). Naturally, there has been significant interest in using AI and LLMs in healthcare and medical education (Lucas, Upperman, and Robinson 2024; Ullah et al. 2024; Shool et al. 2025; Alberts et al. 2023). LLMs can be applied for scenarios including diagnosis, treatment plans, education of medical professionals, and patient communication, making further research and development in these systems a topic of great interest.

The capabilities of AI are publicly discussed with a mix of truth and hyperbole, with much less attention given to possible downsides and caveats. The excellent conversational abilities of LLMs are often mistaken for the ability of, or potential for, true reasoning (Shojaee et al. 2025). In a medical context, security and privacy issues in LLMs can present a risk to patient privacy. Further, the basic reliability of LLMs is not even guaranteed (Huang et al. 2025; Xu, Jain, and Kankanhalli 2024).

Even independently of model reliability and security, use of LLMs may have dangerous implications in medicine. Physicians, nurses, and other health professionals who overrely on LLMs during their education or practice may fail to properly learn or relearn critical information (Harvey, Koenecke, and Kizilcec 2025; Lehmann, Cornelius, and Sting 2024), constituting a critical risk to the human expertise that underlies healthcare. There are also institutional risks to heavy reliance on LLMs, including financial cost and vendor lock-in.

Due to the overhyping of AI and LLMs in recent years, both the technical capabilities and business implications of AI are often greatly exaggerated, while potential downsides are dismissed or ignored (LaGrandeur 2024). Discourse on AI is frequently dominated and overhyped by corporate AI platforms – a clear conflict of interest (Floridi 2024). To provide a more correct perspective, a public discussion not influenced by profit motives or hype cycles is necessary.

In this position paper, these issues are discussed in detail, and offer recommendations to mitigate potential harms from the future adoption of AI and LLMs in medicine and medical education. Our overall position is that while AI in healthcare can be highly useful, at the present time the capabilities of AI, especially LLMs, are vastly overestimated, and that great caution should be taken to mitigate their risks. A cautious approach to these new technologies is imperative for the safe and ethical use of artificial intelligence in healthcare. Proposed recommendations are put forth as a strong foundation of specific approaches underlying a broader philosophy of AI-assisted, human-safeguarded medical care.

Recent Advances and Trends

Modern AI mimics human neural structures via a structure of neuron-like constructs. AI models such as LLMs are organized as a graph, where data flows between nodes in the graph. Developments in computational power in recent decades have made neural networks computationally feasible, leading to a renaissance of both research and application in AI (Toosi et al. 2021). For a full survey of recent technical advances in LLMs, the reader is refered to (Wang et al. 2025).

The recent AI craze has been jumpstarted by the public release of OpenAI's ChatGPT service in late 2022 (Wu et al. 2023), which facilitated access to a LLM as a chatbot. ChatGPT's popularity grew explosively, as it quickly reached 100 million users (Bowen and Watson 2024), and similar services were quickly offered by several other companies. LLMs have also been integrated into products such

as coding IDEs, corporate knowledge portals, and customer service chat windows.

The combination of greatly improved technical capabilities and public accessibility has lead to a massive surge of interest. AI and LLMs are now touted as a transformative tool for both healthcare (Weidener, Fischer et al. 2024) and education (Bowen and Watson 2024). Some commentators go as far as to claim that AI is "inevitable" (Armony and Hazzan 2024). In contrast, others point out that the current AI craze is a bubble (Zitron 2025b), and that the capabilities and applications of AI and LLMs are more limited than is discussed in the current hype cycle (Xu, Jain, and Kankanhalli 2024).

The healthcare sector has been using AI since well before the current LLM boom (Jiang et al. 2017). AI has been effectively applied to scenarios including cancer detection (Nassif et al. 2022), early warnings for sepsis (Yuan et al. 2020), and viral infection tracking during pandemics (Dananjayan and Raj 2020). LLMs' linguistic capabilities have the potential to greatly expand the use of AI in healthcare, with applications including documentation (Gebreab et al. 2024), patient communication (Yang et al. 2024), and translation (Ray et al. 2025).

Pitfalls of AI/LLM Over-Reliance

Industry proponents of AI products frequently and fervently put forth the possible benefits of their products (Floridi 2024), but they are not incentivized to temper public discourse with potential downsides of the platforms and systems they are selling. This section discusses some possible issues that can arise from incautious use of AI and LLMs in medicine and medical education. By examining these problems, we can glean recommendations for medical practitioners, researchers, and educators for the responsible use of AI in their work.

LLM Reliability

Even the basic reliability of LLMs is not guaranteed – LLMs notoriously "hallucinate", regurgitating false information (Huang et al. 2025). This is further compounded by LLMs' lack of actual understanding, making self-correction difficult (Saba 2023). **These limitations are fundamental to LLMs (Varela et al. 2025), and cannot be easily overcome by improving the designs or resources of existing models (Opus and Lawsen 2025).**

One critical subtype of reliability failures is that of bias. AI and LLMs simply reflect their training datasets without the ability to understand greater societal contexts (Dai et al. 2024); this may result in results that unfairly discriminate against some demographics, have incomplete information on some demographics, or further perpetuate biases (e.g., when LLM outputs are used for further LLM training or human education). These risks are not merely hypothetical, but can be seen in existing medical AI (Omar et al. 2025)! The rapid adoption of such flawed systems can thus pose a disproportionate risk to already-disadvantaged groups.

Failure to Learn

For trainees or new clinicians, relying on LLMs can reduce the actual learning of the material even if they can produce correct answers in the short term, which will lead to underdevelopment of the skills needed to become a proficient practitioner (Raihan et al. 2025). Thus, allowing unchecked use of AI platforms in healthcare, especially in educational contexts, can have the dangerous effect of producing professionals whose credentials do not actually reflect their skills.

LLMs are frequently compared to calculators (or slide rules, Wikipedia, spell checkers, etc.) as an educational tool, with the implication being that common use of LLMs in the classroom is inevitable and must be adapted to. However, this comparison makes several errors: first, LLMs used for writing are not comparable to the rote, mechanical calculations of a calculator. Writing does not merely convey a result; rather, the benefit in writing is in its *process*. When students use LLMs to skip this process, they harm their own learning (Rus and Kendeou 2025). Second, LLMs are prone to hallucination or other forms of incorrect responses, which a novice to a field will not be able to recognize, leading to the imbuing of incorrect information. Calculators and reputable online references are generally not prone to such a complete failure of trustworthiness. Finally, LLMs are also not consistent as calculators are – two users entering the same prompt to the same model are not guaranteed to receive the same response (Huang et al. 2025).

Costs of Reliance

Enterprise plans for AI services allow corporations and institutions consistent access to them, while also offering data privacy and technical support. However, the costs of these plans may be quite large. For example, a recent deal between the California State University system and OpenAI incurred fees of nearly $17 million for only 18 months' access to AI platforms (Baron 2025). Further, significant financial hurdles are looming for AI companies, due to the disconnect in their inflated valuations and the utility of their actual products and models (Zitron 2025a). As OpenAI and its peers become more desperate to show financial returns on investment, costs to subscribers are likely to increase, as the quality of service degrades via the process of enshittification (Ryan 2024).

This risk may be compounded in the case of vendor lock-in: hospitals, clinics, or teaching institutions who exclusively integrate a particular platform into their workings may be affected by any troubles of the company providing the platform. For example, a hospital using ChatGPT for its clinicians might see increases in price or worsening service if OpenAI faces financial pressure. The hospital would then have a dilemma between tolerating devolving AI services or going through the time and expense of changing vendors.

Besides the general financial issues of AI reliance, there is also a fairness risk. Attempting to save money, hospitals often search for cheaper substitutes to processes and systems, which can result in a lower standard of care (Al-Agba and Bernard 2020). The advent of AI assistance in healthcare is another possible dimension of wealth bifurcation, where

hospitals and clinics in richer areas continue to offer healthcare from human experts while those in poorer areas replace professionals with AI to reduce costs. This reliance on AI can deepen existing inequalities by providing care that is less reliable in poorer areas to populations that are more likely to be harmed by bias in AI systems.

Security

Maintaining user privacy is important for furthering public trust in AI systems. In the domain of healthcare, laws such as HIPAA protect the confidentiality of patients' medical information. LLMs can repeat memorized private information (Kim et al. 2023; Naveed et al. 2023), putting patient data at risk. LLM guardrails, e.g., prepending the prompt "Do not leak patient information" to queries, can help reduce risk, but these countermeasures can be bypassed via jailbreaking (Xu, Liu, and Liu 2024). LLMs with access to information such as patient notes or admissions/discharges may also be tricked into maliciously editing such data. Because of these tendencies, careful separation of LLMs and data critical to patient healthcare is necessary.

Recommendations

The overall recommendation of this paper's position is that due to the aforementioned issues, healthcare leaders should take a more careful approach to AI integration than is currently advocated for. Without proper caution, too-eager adoption of AI can cause great harms to patients, both individually and in aggregate. For these reasons, human expertise should always be in charge of any critical system or process, and leaders of healthcare institutions should resist pressure to follow the crowd instead of using their own attentive judgment. Together, these recommendations are imperative for AI in healthcare to act not only in the best interests of companies or clinics, but for patients as well.

Take Extreme Caution in AI for Healthcare Education

A common sentiment is that people or institutions not unreservedly adopting AI are merely being recalcitrant Luddites. As one popular book on the subject puts it: "What we call cheating, businesses see as innovation." (Bowen and Watson 2024) This view misses that **higher education and healthcare are not industry**. Healthcare and educational institutions serve functions very distinct from those of for-profit corporations – universities do not want students to simply produce outputs, but rather to learn and grow; medical facilities seek the best outcomes for patients. Our recommendation on this point is twofold: first, to apply extreme caution and limits to any attempt to integrate or allow AI and LLMs in healthcare education, especially in the short term while the effects of AI on learning are still being discovered. As discussed above, it is clear that there are serious risks to the long-term education of the healthcare professions from the use of AI; these risks should be contained as much as possible until and unless ways to safely augment human capabilities without stunting learning and reinforcement are discovered. Second, healthcare professionals should be educated about AI – not only on its capabilities, but equally importantly on its weaknesses. Professionals should be able to understand how and why an AI could err, so as to provide a check on a potentially incorrect diagnosis or treatment rendered to a patient.

Prioritize Freely Available Software

Hospitals and universities, especially those in underfunded areas (e.g., rural regions), may risk taking needed resources from patient care and professional education when funding AI expenditures. Fortunately, such expenses are not strictly necessary. Freely available open-source LLMs can be downloaded and deployed for personal and professional use[1]. It is possible for open-source models to match the performance of proprietary models with some tuning (Alizadeh et al. 2025). Utilizing open-source solutions also helps avoid vendor lock-in by avoiding the need for a vendor entirely.

Considering the high costs of healthcare (Statista 2024), it is imperative for leaders of hospitals and universities to be responsible stewards of financial resources funded by patients and the public. For this reason, healthcare leaders have a duty to seek out methods of innovating with AI that are cost-effective and not bound to corporate interests. Similarly, researchers and practitioners finding AI tools to use in their work should use freely available tools whenever possible, to allow more widespread access and reproducibility that is not dependent on for-profit corporations.

Responsible Advocacy

Persons promoting the expanded use of AI in healthcare should earnestly advocate for their views, but have a duty to do so responsibly. In order to avoid harm and maintain credibility, discussion around AI in healthcare must avoid common errors and always be cognizant of the issues discussed in this paper. The reliability, security, and costs of AI for healthcare should be at the center of public discussion, even if it is difficult or uncomfortable to force corporate executives or other AI advocates to speak candidly and precisely about these issues. Further, mystical thinking around the capabilities of AI should also be refuted: the strong abilities of LLMs in conversational tasks should not be mistaken for transferability to understanding or general reasoning. Similarly, it is irresponsible to assume that future innovations will be able to overcome fundamental limitations of AI as justification for the "inevitability" of AI. Advocacy of AI must be done from a perspective of technical proficiency and patient advocacy to be meaningful; anything less is simply putting personal gain over honest discussion.

While the current overhyping of LLMs has serious issues, this position should not be taken to discount the advances in the applications of AI in health and medicine. Discussion of shortcomings of LLMs and businesses selling LLM access should not be taken to be discounting the entire prospect of applying AI to health. In order to best promote these applications (e.g., medical imaging (Panayides et al. 2020)), advo-

[1]A list of some freely available LLMs can be found at https://github.com/Hannibal046/Awesome-LLM?tab=readme-ov-file#open-llm

cates of AI use should more clearly delineate these from new tools that are LLM chatbots or simple variations thereof. Healthcare leaders should also proactively and aggressively explore research into privacy solutions for their institutions' AI applications.

Acknowledgements

The author gratefully acknowledges the many helpful discussions that shaped this work with Ambrosio Valencia-Romero (Old Dominion University) and Joel Bock (Old Dominion University), as well as diligent proofreading by Christina Murray (Villanova University).

References

Al-Agba, N.; and Bernard, R. 2020. *Patients at risk: The rise of the nurse practitioner and physician assistant in healthcare*. Universal-Publishers.

Alberts, I. L.; Mercolli, L.; Pyka, T.; Prenosil, G.; Shi, K.; Rominger, A.; and Afshar-Oromieh, A. 2023. Large language models (LLM) and ChatGPT: what will the impact on nuclear medicine be? *European journal of nuclear medicine and molecular imaging*, 50(6): 1549–1552.

Alizadeh, M.; Kubli, M.; Samei, Z.; Dehghani, S.; Zahedivafa, M.; Bermeo, J. D.; Korobeynikova, M.; and Gilardi, F. 2025. Open-source LLMs for text annotation: a practical guide for model setting and fine-tuning. *Journal of Computational Social Science*, 8(1): 17.

Armony, Y.; and Hazzan, O. 2024. *Inevitability of AI Technology in Education*. Springer.

Baron, E. 2025. Beleaguered Cal State University's $17 million artificial intelligence initiative defended, attacked.

Bowen, J. A.; and Watson, C. E. 2024. *Teaching with AI: A practical guide to a new era of human learning*. JHU Press.

Dai, S.; Xu, C.; Xu, S.; Pang, L.; Dong, Z.; and Xu, J. 2024. Bias and unfairness in information retrieval systems: New challenges in the llm era. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 6437–6447.

Dananjayan, S.; and Raj, G. M. 2020. Artificial Intelligence during a pandemic: The COVID-19 example. *The International Journal of Health Planning and Management*, 35(5): 1260.

Floridi, L. 2024. Why the AI hype is another tech bubble. *Philosophy & Technology*, 37(4): 128.

Gebreab, S. A.; Salah, K.; Jayaraman, R.; ur Rehman, M. H.; and Ellaham, S. 2024. LLM-based framework for administrative task automation in healthcare. In *2024 12th International Symposium on Digital Forensics and Security (ISDFS)*, 1–7. IEEE.

Harvey, E.; Koenecke, A.; and Kizilcec, R. F. 2025. "Don't Forget the Teachers": Towards an Educator-Centered Understanding of Harms from Large Language Models in Education. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–19.

Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2): 1–55.

Jiang, F.; Jiang, Y.; Zhi, H.; Dong, Y.; Li, H.; Ma, S.; Wang, Y.; Dong, Q.; Shen, H.; and Wang, Y. 2017. Artificial intelligence in healthcare: past, present and future. *Stroke and vascular neurology*, 2(4).

Kim, S.; Yun, S.; Lee, H.; Gubri, M.; Yoon, S.; and Oh, S. J. 2023. Propile: Probing privacy leakage in large language models. *Advances in Neural Information Processing Systems*, 36: 20750–20762.

LaGrandeur, K. 2024. The consequences of AI hype. *AI and Ethics*, 4(3): 653–656.

Lee, J.; Stevens, N.; and Han, S. C. 2025. Large language models in finance (finllms). *Neural Computing and Applications*, 1–15.

Lehmann, M.; Cornelius, P. B.; and Sting, F. J. 2024. AI meets the classroom: When does ChatGPT harm learning? *Available at SSRN 4941259*.

Lucas, H. C.; Upperman, J. S.; and Robinson, J. R. 2024. A systematic review of large language models and their implications in medical education. *Medical education*, 58(11): 1276–1285.

Nassif, A. B.; Talib, M. A.; Nasir, Q.; Afadar, Y.; and Elgendy, O. 2022. Breast cancer detection using artificial intelligence techniques: A systematic literature review. *Artificial intelligence in medicine*, 127: 102276.

Naveed, H.; Khan, A. U.; Qiu, S.; Saqib, M.; Anwar, S.; Usman, M.; Akhtar, N.; Barnes, N.; and Mian, A. 2023. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*.

Omar, M.; Soffer, S.; Agbareia, R.; Bragazzi, N. L.; Apakama, D. U.; Horowitz, C. R.; Charney, A. W.; Freeman, R.; Kummer, B.; Glicksberg, B. S.; et al. 2025. Sociodemographic biases in medical decision making by large language models. *Nature Medicine*, 1–9.

Opus, C.; and Lawsen, A. 2025. The Illusion of the Illusion of Thinking. *arXiv preprint ArXiv:2506.09250*.

Panayides, A. S.; Amini, A.; Filipovic, N. D.; Sharma, A.; Tsaftaris, S. A.; Young, A.; Foran, D.; Do, N.; Golemati, S.; Kurc, T.; et al. 2020. AI in medical imaging informatics: current challenges and future directions. *IEEE journal of biomedical and health informatics*, 24(7): 1837–1857.

Raihan, N.; Siddiq, M. L.; Santos, J. C.; and Zampieri, M. 2025. Large language models in computer science education: A systematic literature review. In *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 1*, 938–944.

Ray, M.; Kats, D. J.; Moorkens, J.; Rai, D.; Shaar, N.; Quinones, D.; Vermeulen, A.; Mateo, C. M.; Brewster, R. C.; Khan, A.; et al. 2025. Evaluating a Large Language Model in Translating Patient Instructions to Spanish Using a Standardized Framework. *JAMA pediatrics*.

Rus, V.; and Kendeou, P. 2025. Are LLMs actually good for learning? *AI & SOCIETY*, 1–2.

Ryan, J. 2024. The Coming Enshittification of AI: Will AI follow internet search and e-commerce down the path of enshittification, or can we finally have nice things? *The Journal of Business and Artificial Intelligence*, 1(2).

Saba, W. S. 2023. Stochastic LLMs do not understand language: towards symbolic, explainable and ontologically based LLMs. In *International conference on conceptual modeling*, 3–19. Springer.

Shojaee, P.; Mirzadeh, I.; Alizadeh, K.; Horton, M.; Bengio, S.; and Farajtabar, M. 2025. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. *arXiv preprint arXiv:2506.06941*.

Shool, S.; Adimi, S.; Saboori Amleshi, R.; Bitaraf, E.; Golpira, R.; and Tara, M. 2025. A systematic review of large language model (LLM) evaluations in clinical medicine. *BMC Medical Informatics and Decision Making*, 25(1): 117.

Statista. 2024. Health expenditures in the US–statistics & facts.

Toosi, A.; Bottino, A. G.; Saboury, B.; Siegel, E.; and Rahmim, A. 2021. A brief history of AI: how to prevent another winter (a critical review). *PET clinics*, 16(4): 449–469.

Ullah, E.; Parwani, A.; Baig, M. M.; and Singh, R. 2024. Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology–a recent scoping review. *Diagnostic pathology*, 19(1): 43.

Varela, I. D.; Romero-Sorozabal, P.; Rocon, E.; and Cebrian, M. 2025. Rethinking the Illusion of Thinking. *arXiv preprint arXiv:2507.01231*.

Wang, Z.; Chu, Z.; Doan, T. V.; Ni, S.; Yang, M.; and Zhang, W. 2025. History, development, and principles of large language models: an introductory survey. *AI and Ethics*, 5(3): 1955–1971.

Weidener, L.; Fischer, M.; et al. 2024. Artificial intelligence in medicine: cross-sectional study among medical students on application, education, and ethical aspects. *JMIR medical education*, 10(1): e51247.

Wu, T.; He, S.; Liu, J.; Sun, S.; Liu, K.; Han, Q.-L.; and Tang, Y. 2023. A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, 10(5): 1122–1136.

Xu, K.; Zhang, G. L.; Yin, X.; Zhuo, C.; Schlichtmann, U.; and Li, B. 2024. Automated c/c++ program repair for high-level synthesis via large language models. In *Proceedings of the 2024 ACM/IEEE International Symposium on Machine Learning for CAD*, 1–9.

Xu, Z.; Jain, S.; and Kankanhalli, M. 2024. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*.

Xu, Z.; Liu, F.; and Liu, H. 2024. Bag of tricks: Benchmarking of jailbreak attacks on llms. *Advances in Neural Information Processing Systems*, 37: 32219–32250.

Yang, Z.; Xu, X.; Yao, B.; Rogers, E.; Zhang, S.; Intille, S.; Shara, N.; Gao, G. G.; and Wang, D. 2024. Talk2care: An llm-based voice assistant for communication between healthcare providers and older adults. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(2): 1–35.

Yuan, K.-C.; Tsai, L.-W.; Lee, K.-H.; Cheng, Y.-W.; Hsu, S.-C.; Lo, Y.-S.; and Chen, R.-J. 2020. The development an artificial intelligence algorithm for early sepsis diagnosis in the intensive care unit. *International journal of medical informatics*, 141: 104176.

Zhang, J.; Bu, H.; Wen, H.; Liu, Y.; Fei, H.; Xi, R.; Li, L.; Yang, Y.; Zhu, H.; and Meng, D. 2025. When llms meet cybersecurity: A systematic literature review. *Cybersecurity*, 8(1): 55.

Zitron, E. 2025a. Reality Check.

Zitron, E. 2025b. There is no AI Revolution.