



LLMs in medicine: The need for advanced evaluation systems for disruptive technologies

Yi-Da Tang,^{1,2,*} Er-Dan Dong,^{1,2,3} and Wen Gao⁴

¹Department of Cardiology and Institute of Vascular Medicine, Peking University Third Hospital, Beijing 100191, China

²State Key Laboratory of Vascular Homeostasis and Remodeling, Peking University, Beijing 100191, China

³University of Health and Rehabilitation Sciences, Qingdao 266000, China

⁴Peking University, Beijing 100871, China

*Correspondence: tangyida@bjmu.edu.cn

Received: December 8, 2023; Accepted: April 1, 2024; Published Online: April 2, 2024; <https://doi.org/10.1016/j.xinn.2024.100622>

© 2024 This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Citation: Tang Y.-D., Dong E.-D., and Gao W. (2024). LLMs in medicine: The need for advanced evaluation systems for disruptive technologies. *The Innovation* 5(3), 100622.

In an era of rapid advancements in artificial intelligence (AI), the potential role of large language models (LLMs) in the medical field is receiving increasing attention. The progression of this technology not only highlights upcoming revolutionary changes in medical services but also offers unprecedented opportunities for the health care sector.¹ LLMs' formidable capability to process natural language opens vast prospects for application in the medical field, including in the contexts of emergency triage, geriatric care, and digital medical workflows.

More notably, due to the continuous expansion of the types of data used in the medical field, the ability of LLMs to handle multimodal data, including images, sounds, and biomarkers, provides more precise and personalized suggestions for medical diagnosis and treatment.² This technological development not only improves the efficiency of medical work but also leads to better experiences and higher rates of treatment success for patients. For instance, LLMs can rapidly analyze patients' medical imaging data and combine them with textual data, thereby offering doctors more comprehensive and in-depth diagnostic suggestions. In telemedicine, LLMs enable patients to conveniently upload medical images, voice recordings, or symptom descriptions; interact with LLMs in real time; and obtain preliminary medical advice.³

Although LLMs have proven to be valuable in multiple fields, they still face numerous challenges in the context of medical applications, such as insufficient empirical research on this topic, issues related to model interpretability, and potential hallucination problems.⁴ In actual medical environments, the success of LLMs depends not only on their technical performance but also on whether they can gain trust from society, including patients and health care professionals, meet the professional needs of medical personnel, and provide better experiences for patients. First, with regard to the complexity of using LLMs to address medical data, ensuring data privacy and security is crucial. Second, model interpretability is another key issue. The complexity and seriousness of medical decisions entail that relying solely on the results of such models is insufficient. Both doctors and patients must understand the underlying logic and basis for these decisions. Therefore, developing LLMs that can provide clear, understandable outputs is essential. Additionally, technological biases and hallucination issues cannot be ignored. Hence, the fair auditing of data and the establishment of continuous bias correction mechanisms are indispensable. Moreover, to gain trust from patients and health care professionals, LLMs must not only provide professional and reliable medical advice but also consider humanization and empathy in the context of doctor-patient communication. When providing medical advice, LLMs should also be able to communicate in a way that is more closely related to human emotions and needs.

Therefore, a superior evaluation system that can be used to estimate whether LLMs and other AI technologies are suitable for application in medical scenarios is necessary. The American Medical Association's announcement on November 28, 2023, is particularly relevant in this context. These principles advocate for a balanced approach to the development of medical AI, which can emphasize ethical integrity, patient welfare, and equitable health care outcomes. Evaluating LLMs requires metrics that assess the ethical use of data, the clarity of decision-making processes, and the accessibility of model outputs to health care providers and patients. Metrics for evaluating the accuracy of medical recommendations and their impact on diagnostic processes and patient care outcomes ensure that LLMs can contribute meaningfully to medical practice.

The next key research direction in this context involves developing an evaluation system specifically for the use of LLM in the medical field and conducting relevant clinical trials.⁵ These efforts can provide deeper insights into the appli-

cation of LLMs in medicine and promote broader and more effective utilization of this technology. Developing such an evaluation system requires joint efforts on the part of medical experts, AI scientists, government regulatory bodies, and the public (Figure 1). Medical experts can provide profound insights into medical practice and patient care for the evaluation system. They are well aware of the complexity of the medical environment and the diversity of patient needs, thus ensuring that the evaluation system focuses not only on technical performance but also on the overall well-being of patients. AI scientists can provide knowledge regarding technical performance and potential technological limitations. Their expertise is crucial for understanding the operations of LLMs and the potential technical challenges with which they are associated. Government regulatory bodies can ensure that the evaluation system complies with relevant legal and ethical standards. These bodies are responsible for the formulation and enforcement of rules to protect patient privacy, prevent data misuse, and ensure that everyone can benefit equally from LLM technology. Moreover, they can provide a regulatory framework for the use of LLMs, ensuring that it is in line with the public interest. Public participation ensures the transparency and fairness of the evaluation system. Incorporating public opinions and concerns ensures that the evaluation system can comprehensively consider patients' experiences and expectations, thus allowing the necessary adjustments to be made.

Only by using this approach can the full potential of these technologies be harnessed, thereby ensuring improvements in medical quality and efficiency while simultaneously respecting and protecting fundamental human values and rights. This evaluation system should also be dynamic and capable of adapting and improving alongside the development of technology and changes in medical practice. Such adaptability is crucial for the system's ability to cope with a rapidly changing technological environment. As LLMs and other AI technologies undergo iterative updates, new technical features and application challenges may emerge. The evaluation system must be able to promptly identify and address these changes and emerging issues and to adjust the relevant evaluation criteria and processes to ensure continued safety and compliance. Medical practice itself is also continuously advancing and changing. Similarly, as new medical technologies and treatment methods emerge, LLMs may need to adapt to new medical environments and patient needs. The evaluation system should be able to reflect these changes, thus ensuring that LLMs and other AI technologies remain effective and applicable in new settings. Establishing a dynamic, flexible, and adaptable evaluation system is crucial.

When technology advances, it is imperative to ensure that these technologies benefit everyone, including individuals in economically or technologically less advanced regions and groups. Achieving this goal is not only a technical challenge but also a social responsibility. It is essential to ensure that the design and application of LLMs can overcome language and cultural barriers, thereby truly reaching and benefitting people from diverse backgrounds. Additionally, consideration must be given to issues related to technology access. For individuals in areas featuring lower levels of technological penetration or those who are less familiar with new technologies, such as elderly individuals, it is crucial to provide simpler access methods and more intuitive user interfaces, thereby increasing the ease of use of LLMs. To enhance the accessibility of technology, education and training are needed to help people understand and effectively utilize LLMs. Through these efforts, not only can the quality and efficiency of medical services be improved, but overall health and well-being in society can also be promoted.

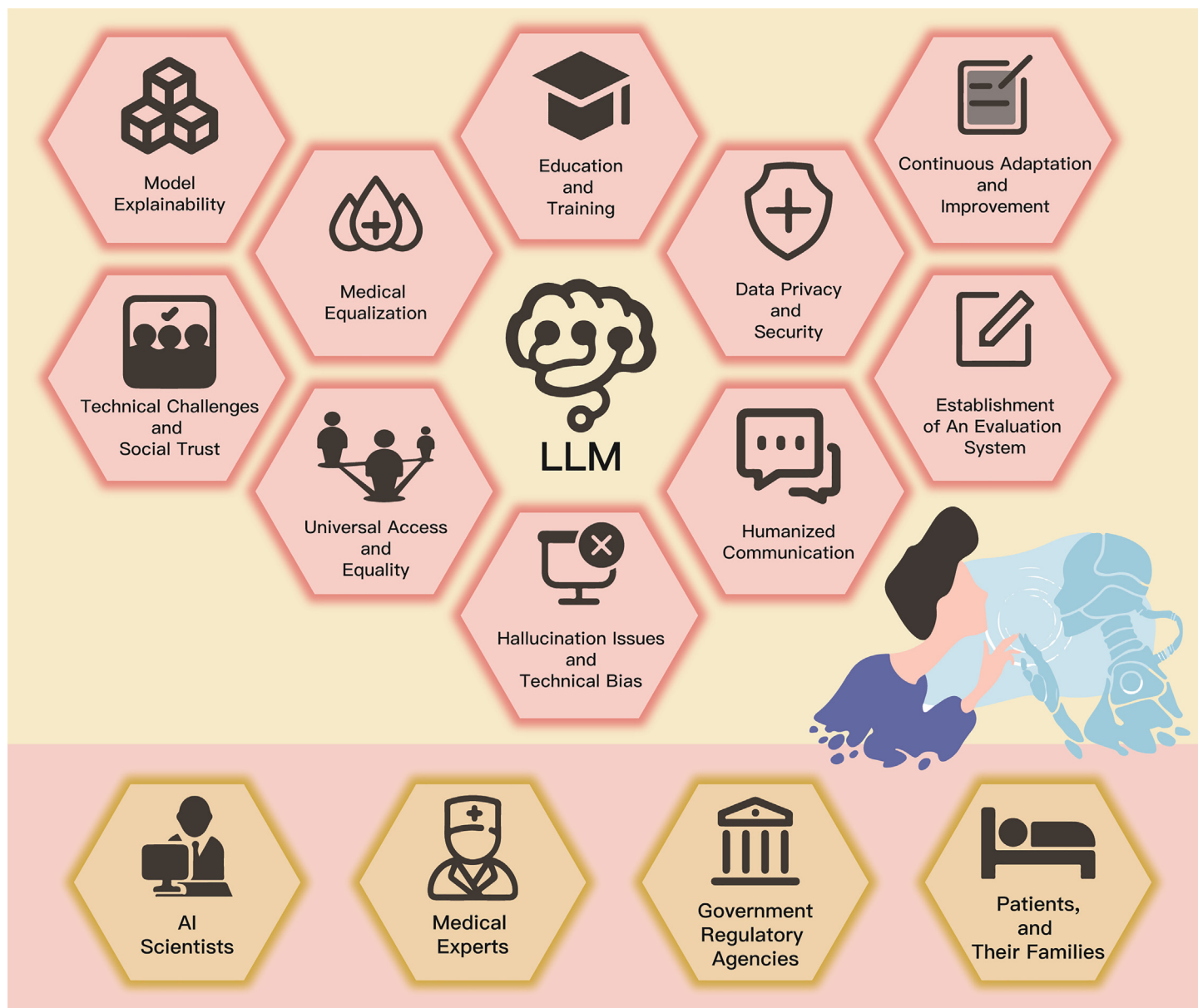


Figure 1. Interdisciplinary impact network of LLMs in health care The multidimensional impact of the application of large language models (LLMs) in the medical field involves core issues such as model interpretability, data privacy, and continuous improvement. Key groups that are relevant to the development of LLMs include medical professionals, AI scientists, government regulators, and patients and patients' families. To effectively apply LLMs in medical scenarios, interdisciplinary cooperation and multiparty participation among these groups are needed.

In conclusion, the application of LLMs in the medical field is promising, but this process must be implemented cautiously and steadily. Establishing and improving the relevant evaluation system is crucial to improve the supervision and guidance of technology application, ensuring that the use of such technology is in line with the unique and ethical standards of the medical field, thereby making significant contributions to human health and well-being. When exploring the limitless possibilities associated with the use of LLMs in the medical field, it is vital to remember the fundamental mission of medicine: respecting human life and improving quality of life. Therefore, it is necessary to ensure that the development of technology is not only efficient and innovative but also safe, fair, and inclusive. It is important to recognize that technology itself is not omnipotent; rather, its true value is determined by how it is used to serve humanity. This process requires joint participation and efforts on the part of medical experts, AI scientists, government regulatory bodies, patients, and patients' families to ensure that the application of technology not only achieves the expected medical effects but also respects and protects fundamental human values and rights. In this context, we hereby encourage experts in the fields medicine and AI to call for such collaborative efforts to transform the potential of LLMs and other AI technologies into

practical medical achievements, thus leading to revolutionary changes in the health and well-being of all humanity.

REFERENCES

1. Singhal, K., Azizi, S., Tu, T., et al. (2023). Large language models encode clinical knowledge. *Nature* **620**: 172–180. <https://doi.org/10.1038/s41586-023-06291-2>.
2. Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K., et al. (2023). Large language models in medicine. *Nat. Med.* **29**(8): 1930–1940. <https://doi.org/10.1038/s41591-023-02448-8>.
3. Xu, Y., Liu, X., Cao, X., et al. (2021). Artificial intelligence: A powerful paradigm for scientific research. *Innovation* **2**(4): 100179.
4. Shah, N.H., Entwistle, D., and Pfeffer, M.A. (2023). Creation and Adoption of Large Language Models in Medicine. *JAMA* **330**(9): 866–869. <https://doi.org/10.1001/jama.2023.14217>.
5. Peng, C., Yang, X., Chen, A., et al. (2023). A study of generative large language model for medical research and healthcare. *NPJ Digit. Med.* **6**(6): 210. <https://doi.org/10.1038/s41746-023-00958-w>.

DECLARATION OF INTERESTS

The authors declare no competing interests.