Technical roadmap towards trustworthy large-scale models in medicine

Jie Yang,[1] Qian Ding,[1] Jie Tian,[2,*] and Puxiang Lai[3,4,*]

[1]University of Electronic Science and Technology of China, Chengdu 611731, China
[2]CAS Key Laboratory of Molecular Imaging, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[3]Department of Biomedical Engineering, The Hong Kong Polytechnic University, Hong Kong 999077, China
[4]Hong Kong Polytechnic University, Shenzhen Research Institute, Shenzhen 518000, China
*Correspondence: jie.tian@ia.ac.cn (J. T.); puxiang.lai@polyu.edu.hk (P. L.)

Received: January 21, 2024; Accepted: March 7, 2024; Published Online: March 9, 2024; https://doi.org/10.59717/j.xinn-med.2024.100058

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

Citation: Yang J., Ding Q., Tian J., et al., (2024). Technical roadmap towards trustworthy large-scale models in medicine. The Innovation Medicine 2(1): 100058.

Large-scale models (LSMs) hold tremendous potentials for applications in medicine.[1,2] That said, generative LSMs inherently possess biases and errors, just like humans. Despite the successful commercial applications demonstrated by ChatGPT4, higher accuracy is demanded for the generative LSMs in medicine. This commentary aims to delve into the construction of trustworthy medical models by mitigating the hallucination of LSMs. We will revolve around four key areas: data, multimodal models, flexible AI agent, and dynamic evolution evaluation. By exploring these dimensions, one can deeply analyze the components required to build trustworthy medical models, generating a technical roadmap as illustrated in Figure 1.

DATA-CENTRIC LARGE-SCALE MODELS

Currently, medical LSMs that are directly pre-trained on medical data include BioBERT, PubMedBERT, MedSAM, etc.; those that are fine-tuned based on general models include DoctorGLM, HuatuoGPT, etc.; those that combine medical knowledge as prompts include MedPaLM, Dr Knows, ChatCAD, etc. Compared with simple finetuning and prompt-based methods, direct pretraining can more fundamentally address the biased probability generation in models. However, pretraining imposes higher requirements on the diversity and quality of the data. This data-centric artificial intelligence (DCAI) holds particular significance in the field of medicine.[3] However, collecting medical data still faces some unique challenges. Firstly, medical data exhibits high dimensionality due to the need for integrating multiple modalities such as imaging, genetic sequences, along with physicians' experiential knowledge to derive accurate conclusions. Thus, data collection should focus on two dimensions: deterministic knowledge and implicit experiential knowledge. Hybrid encoding of deterministic knowledge and empirical data can achieve performance improvement. Also, multi-modal data with strong correlation should be collected to reduce feature extraction difficulties. Secondly, due to rare diseases and imbalanced patient demographics, medical data often exhibits sample imbalance. To fundamentally address the issue, it is crucial to promote the participation of minority groups in data collection. Additionally, technological approaches to address imbalanced samples include small-sample oversampling, data augmentation, generative adversarial models, large-scale data synthesis, and specialized agents for small-sample cases. Thirdly, medical data often has significant noise, making

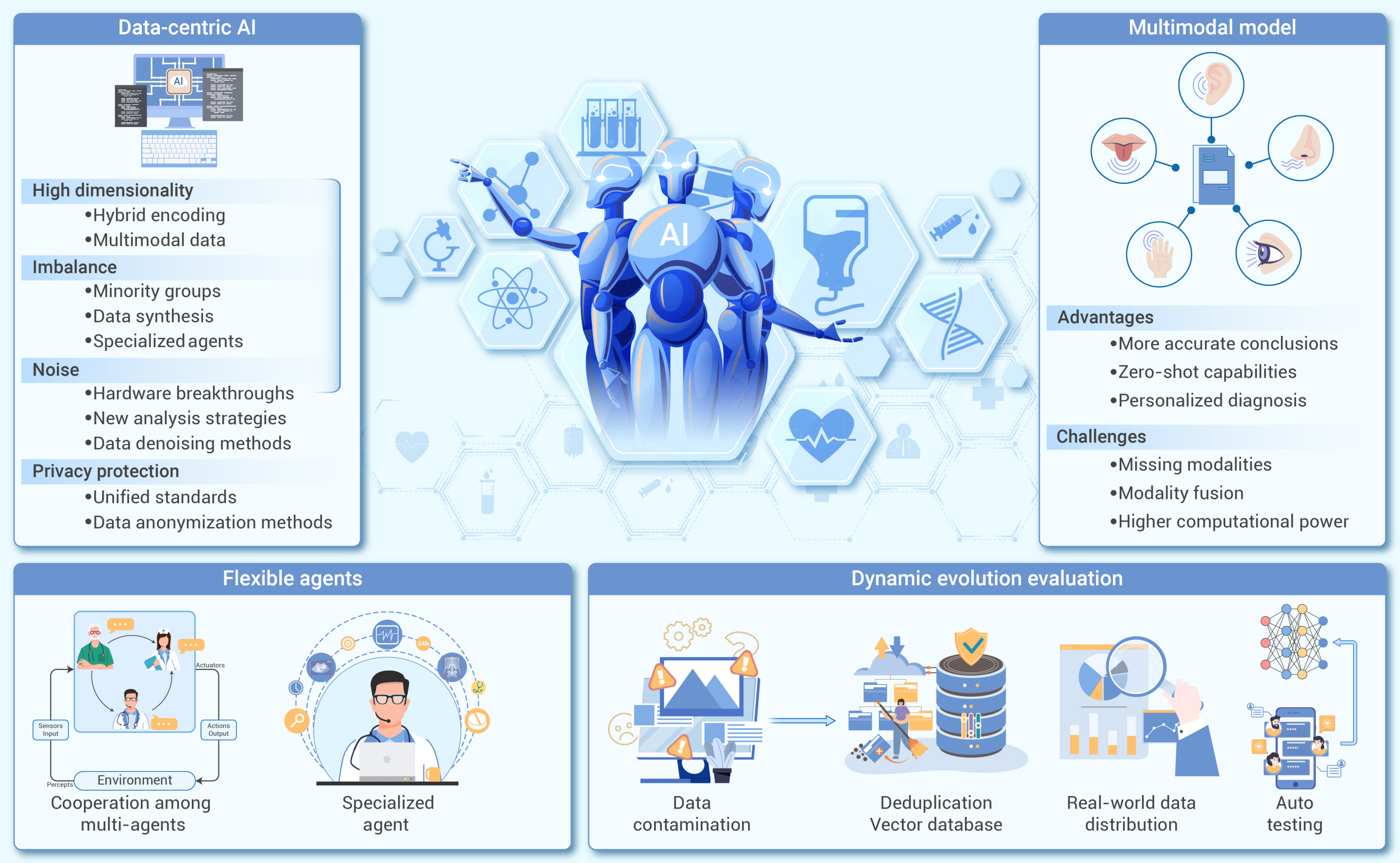

Figure 1. The way to overcome the hallucination of large-scale models in medicine.

it challenging to identify lesion signals. To enhance the signal-to-noise ratio, continuous hardware breakthroughs and new analysis strategies like arterial spin labeling are essential to improve data quality. Also, techniques such as super-resolution networks, noise estimation, and image reconstruction can be combined. Lastly, medical data involves sensitive personal information and requires privacy protection. As a result, methods like data encryption, de-identification, differential privacy, and federated learning should be further studied. Also, it is urgent for hospitals to collaborate and establish unified standards for data collection, storage, and transmission.

LARGE MULTIMODAL MODELS ALIGNED WITH REAL-WORLD NEEDS

The inherent multimodality of medical data stems from the diverse sources it encompasses, such as electronic health records, medical imaging, wearable biosensors, and genome and microbiome sequencing, presenting ample opportunities for large multimodal models (LMM). LMMs enable alignment of multimodal data in semantic space and are better equipped to handle complex medical cases.[4] Specifically, a comprehensive diagnosis often requires the integration of multiple modalities, including text (e.g., patient cases), images (e.g., CT scans, visual appearance), speech (e.g., patient narratives), and so on, to ensure more accurate conclusions. Also, the consistent representation of multimodal data enhances the zero-shot capabilities of models. By leveraging the shared encoding schemes (i.e., Transformer) and knowledge transfer across modalities, models can effectively generalize to unseen scenarios. This capability is particularly valuable in medical applications where data scarcity or the emergence of new conditions necessitate the model to adapt and make informed decisions based on limited or unseen data. In addition, by integrating multiple modalities clinical information, the model can provide personalized diagnosis and treatment plans for each patient. While a multimodal, human-like medical model is an inevitable trend in the future, there are still prominent challenges that need to be addressed. Collecting diverse, high-quality data across multiple modalities poses challenges, particularly in the medical field where stricter ethical, privacy, and legal restrictions apply. Effective solutions are also needed to address problems such as missing modalities and efficient modality fusion strategies. Finally, LMMs require higher computational power for both training and inference, necessitating the search for more energy-efficient solutions during training and deployment.

FLEXIBLE AI AGENT

Just like humans, to overcome these hallucinations, it is necessary to seek advice from external professionals. Similarly, guiding LSMs with the help of external tools can prevent them from generating inaccurate information and enhance their ability to handle complex problems. An AI agent refers to using a large-scale model as the central controller to actively decompose problems and invoke a series of external toolkits to solve problems. Establishing medical intelligent agents has great research prospects and challenges for alleviating model hallucinations. Firstly, medical diagnosis naturally involves decision-making from multiple perspectives. Therefore, building efficient cooperation among multi-agents contributes to achieving true intelligence. Establishing dynamic reinforcement learning methods based on feedback signals among agents facilitates continuous learning in multi-agent systems. There are also now mature software tools available to support interaction among multiple agents, such as the latest versions of AutoGen and LangGraph. However, it is important to ensure that the flow of information among the agents complies with ethical and legal requirements. Secondly, each agent should be capable of flexibly combining specialized knowledge graphs, Retrieval Augmented Generation (RAG), professional Computer-Assisted Diagnosis (CAD) systems, and other forms of knowledge. Thus, agents can access and integrate information from various medical literature, clinical guidelines, and patient records. For example, medical LSMs like ChatCAD and ChatCAD+ are established in this way to automatically generate diagnostic reports. On the one hand, prompts which combine these standard knowledge can significantly enhance the model's ability to reason and respond correctly based on context, alleviating the problem of insufficient sample size in sub-domains. On the other hand, it ensures the traceability and accountability of content generated by unreliable LSMs.

DYNAMIC EVOLUTION EVALUATION

The evaluation of LSMs still encounters many challenges. Firstly, existing medical evaluation datasets are susceptible to data contamination. Previous studies have evaluated data contamination through loss difference and pointed out models involving ranking manipulation to increase awareness.[5] To prevent unintentional data contamination, deduplication techniques can be used to prevent confusion between the training and testing sets. In addition to deterministic deduplication techniques, fuzzy deduplication methods such as MinHash and SimHash can also be combined in massive medical data. Also, utilizing vector databases such as Faiss and ElasticSearch, along with organizing pre-training corpora, enables the detection of data contamination in a transparent manner, allowing for the rewriting of erroneous data based on medical findings. Secondly, model evaluation requires datasets that can effectively simulate real-world distribution and demands. The current testing cases used to assess LSMs provide some reference but cannot fully represent the model's performance in practical applications. One feasible approach is to sample user online inputs and feedback to represent the real distribution of demands. However, this method is only applicable to well-known models such as ChatGPT4, as their user activity is higher and better represents real demands. Lastly, the real-world environment is dynamic and constantly changing, rather than static. A dynamic and real-time evaluation system can guide the efficient evolution of models. In the medical field, demands evolve with scientific advancements, the emergence of new diseases, and changes in medical practices. Therefore, an effective approach to enable LSMs to adapt and meet dynamic demands is to leverage reinforcement learning with human feedback and iterative training.[2] By using human feedback as a reward signal, models can be trained to optimize according to actual requirements and user preferences. In the future, more similar dynamic evolution evaluations based on feedback and new data will be studied and developed.

CONCLUSION

This commentary focuses on four dimensions, i.e., data, multimodal models, AI agents, and dynamic evaluation methods, to explore techniques for mitigating the hallucination of LSMs in the medical field. With enhanced data-driven capabilities to generate diverse and high-quality datasets, further explored potential of multimodal applications, integration with multiple toolchains, as well as the creation of an automated dynamic evaluation system, it is technically feasible to deploy more accurate, more intelligent, and more trustworthy large-scale medical models in the near future.

REFERENCES

1. Huang, T., Xu, H., Wang, H., et al. (2023). Artificial intelligence for medicine: Progress, challenges, and perspectives. The Innovation Medicine **1**(2): 100030. DOI: 10.59717/j.xinn-med.2023.100030.
2. Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K., et al. (2023). Large language models in medicine. Nat Med **29**(8): 1930–1940. DOI: 10.1038/s41591-023-02448-8.
3. Xu, Y., Wang, F., An, Z., et al. (2023). Artificial intelligence for science—bridging data to wisdom. The Innovation **4**(6): 100525. DOI: 10.1016/j.xinn.2023.100525.
4. Huang, H., Zheng, O., Wang, D., et al. (2023). ChatGPT for shaping the future of dentistry: the potential of multi-modal large language model. Int J Oral Sci **15**(1): 29. DOI: 10.1038/s41368-023-00239-y.
5. Wei, T., Zhao, L., Zhang, L., et al. (2023). Skywork: A more open bilingual foundation model. arXiv preprint arXiv: 2310.19341. DOI: 10.48550/arXiv.2310.19341.

ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation of China (81930048), Guangdong Science and Technology Commission (2019BT02X105), Hong Kong Innovation and Technology Commission (GHP/043/19SZ, GHP/044/19GD), Hong Kong Research Grant Council (15217721, C7074-21GF), Shenzhen Science and Technology Innovation Commission (JCYJ20220818100202005), and Hong Kong Polytechnic University (P0039517, P0043485, P0045762, and P0048314).

DECLARATION OF INTERESTS

The authors declare no competing interests.