The Current Development of Large Language Models in Medical Questions and Answers--Center on the Models of Med-PaLM and Med-PaLM2

Yueqi Zhu

Department of Information and Services, Sanya City College, Sanya NJ 572099, China
zhuyueqi681@outlook.com

Abstract. The large language models (LLMs) have brought a significant technological changes in the field of medical questions and answers. It can help medical personnel improve diagnostic efficiency and also help the public more easily understand their own physical health status. This article focuses on the Google Medical Model Med-PaLM and its iteratively updated model Med-PaLM2 to elaborate on its advantages, relevant application scenarios and the existing problems and optimization solutions. Med-PaLM can answer common disease diagnosis suggestions, medical knowledge inquiry and other related medical text responses. Med-PaLM2 was further upgraded and was capable of handling more complex scenarios. For instance, bedside consultation medical image interpretation and so on. But it can be seen from Med-PaLM that the three major problems it is facing are hallucination issue, data limitations issue and related ethical issues. Although the future prospects of the medical questions and answers model are extremely promising, still requires continuous optimization and improving the fine-tuning of professional medical data and achieve more comprehensive multimodal fusion. Strive to promote the transformation of LLMs from auxiliary tools to professional medical diagnosis systems.

Keywords: Artificial Intelligence, Large Language Models, Medical Questions and Answers, Med-PaLM, Med-PaLM2.

1 Introduction

With the development of science technology, artificial intelligence played many roles in the field of medical. Artificial intelligence has demonstrated great potential in improving diagnostic accuracy, treatment efficacy even reduces medical costs and other aspects. Having AI assist in the work not only reduces the workload of doctors but also significantly shortens the diagnosis and treatment time. Nowadays, many places have put AI into use include the construction of smart hospitals is also gradually underway. Smart healthcare will usher in a new era of human health management. Not only in the treatment of diseases but also plays an extremely crucial role in disease prevention, health management and the improvement of quality of life. In 2017, the introduction of Transformer has changed the structure of the traditional neural network language model. After its appearance, the researchers focused on pre-training the model. For example, Google launched BERT established the training mode of per-training models

K. Subramanian (ed.), *Proceedings of the International Workshop on Advances in Deep Learning for Image Analysis and Computer Vision (IWADIC 2025)*, Advances in Computer Science Research 128,
https://doi.org/10.2991/978-94-6239-648-7_49

and fine-tuning [1]. The emergence of the GPT series has played a leading role. Among them, GPT-3 expanded its model parameters to 175 billion [2,3]. Significantly enhanced the model's natural language processing capabilities. This also sparked a research craze for large language models. On the basis of GPT-3, the chatbot ChatGPT was born. After the release of ChatGPT, not only has it sparked widespread discussion among all sectors, but the ideal of using LLMs in medicine has also emerged. With the development and progress of LLMs, question-answering robots are becoming increasingly popular in various professional fields especially in the field of medicine. The aim is to develop tools to assist in decision-making and help medical experts in carrying out their daily work. In the field of medical of questions and answers. Due to the busyness of the medical staff. Sometimes they are unable to contact the patients in a timely manner and provide professional guidance. Therefore, LLMs play their important role. They can act as intermediaries between medical staff and the patients' families for communication such as post operation recovery guide and matters need attention for medication use. It has greatly saved the time for doctor-patient communication. At the same time, it can also provide 7*24-hour basic inquiries and triage services. It reduced the pressure of manual medical consultation. By answering medical questions to provide assistance to medical experts, this is a natural way to achieve human-machine interaction [4]. For the use of LLMs the researchers have proposed MedExpQA [5]. A multilingual benchmark test for medical question answering. Used to evaluate the performance of LLMs in medical question answering. Among them, Med-PaLM and Med-PaLM2 performed particularly well in the medical response task. In the United States Medical Licensing Examination, Med-PaLM is the first model to achieve a passing score [2]. Med-PaLM2 was further optimized. By applying multimodal technology, more complex scenarios can be handled. And the Med-PaLM2's responses were considered superior to those made by doctors in multiple aspects especially in the responses given at the bedside consultation. Although LLMs have received attention in the field of medical question answering, there is still a need for further improvement of various issues of the model in the future. This article will discuss the main application scenarios, existing problems and optimization solutions of LLMs and two core models Med-PaLM and Med-PaLM2. The main objective of this study is to analyze the current development status of LLMs in the medical field and to discuss the main issues of these models.

2 Used for Scenarios and Advantages

2.1 Core Application Scenarios

Clinical Decision Support. When doctors encounter complex or rare diseases during the diagnosis and treatment process, they can raise questions to Med-PaLM and Med-PaLM2 and receive diagnostic suggestions and treatment suggestions. For example, dealing with a complex patient, based on the patient's description, relevant medical literature and case studies can be searched to assist doctors in making diagnoses. After optimization of Med-PaLM2, it can be better integrated medical knowledge and

improve the accuracy of decision-making. It can also be used for bedside consultations and viewing medical images, etc [5], for instance, during bedside consultations, doctors can directly ask questions to obtain information about rare diseases symptoms. It can save doctors' time when they look up information. The model will also initially generate one or two sets of treatment recommendations based on the patient's condition. The doctor will then make further adjustments based on this. LLMs, after training, can also analyze medical images such as X-ray and ultrasound. It can detect abnormal images, then assist doctors in making diagnoses [6].

Patient Consultation. LLMs can better understand medical terms through natural language technology and deep learning [7]. In terms of medical consultation, LLMs can provide patients with health knowledge answers and medical guidance, as well as chronic disease management. Patients can ask questions to the model on the medical platform. Patients can describe their symptoms to the model. LLMs can perform simple diagonses based on the symptoms described by the patient. They can also consult some questions about medication precautions and how to prevent diseases. Taking ChatGPT as an example, among the responses from ChatGPT and doctors, 78.6% of people prefer to accept ChatGPT's answers [8]. It can recognize the patient's emotions when describing their condition and provide relatively empathetic responses. In terms of medical guidance, the patient may not know which clinic to go for their condition. At this point, they can ask LLMs. It will provide patients with priority department selection and possible examination materials during medical visits. And for patients with chronic diseases such as high blood pressure and diabetic patients. It can generate dietary recommendation tables and daily medication dosages for patients. However, Med-PaLM2 is better able to understand the patient's intentions when providing answers. Given that the professional medical issues generated by the model can sometimes be difficult for patients to understand. Med-PaLM2 uses simpler and more understandable language to answer patients' questions and solve their problems.

Medical Education. LLMs has a broad prospect for development in medical education. It can offer personalized knowledge assistance and make complex medical knowledge easy to understand. LLMs often assist people in preparing for various medical licensing exams and generate different types of multiple-choice questions [6]. LLMs can also act as virtual teaching assistants. They can analyze students' weaknesses based on their answers and generate more targeted questions. At the same time, LLMs can simulate real patient-doctor conversations and provide realistic scenarios to help train students' clinical skills. Research shows that this learning method is more likely to stimulate students' enthusiasm for learning [9]. In the future, LLMs will still be able to drive the development of medical education and cultivate more talents in the field of medicine.

2.2 Advantages

Med-PaLM is an artificial intelligence chatbot that can address various text-related issues in the medical field, such as providing diagnosis and treatment suggestions for common diseases and conducting medical knowledge inquiries. Med-PaLM proposed

the MultiMedQA benchmark evaluation to cover areas such as medical examinations, patient health, and medical research. And a set of manual evaluation criteria was established to promote the standardized development of the medical question-answering model. The advantages of Med - PaLM2 in the medical field are quite obvious. After upgrading based on the first-generation model. The accuracy rate of its responses in the dataset is 86.5% [5]. And its responses are generally regarded as being better than those of human doctors in many aspects. This can provide doctors with better reference for making clinical decisions, and at the same time can reduce the pressure on medical workers and enhance their work efficiency. As shown in Table 1:

Table 1. Summary of Advantages Table

Advantageous Area	Embody
Efficiency and Automation	Can automatically and quickly generate medical documents and medical reports It enables 24/7 basic inquiries and triage, thereby reducing the pressure on manual inquiries.
knowledge management	Be able to quickly search for literature and be capable of summarizing the content of the literature and generating an abstract.
Personalization and Communication	Instead of relying on medical staff to conduct preoperative and postoperative communication with patients and their families, this measure has improved the communication methods between doctors and patients.

3 The Existing Problems of Large Language Models

3.1 Accuracy and Illusion Issues

The hallucinations of LLMs refer to the generation of meaningless or inconsistent results with the provided source content. It is prone to generating "pseudo-professional" content. And these contents may seem reasonable, but in fact, they are answers that do not match the facts. For instance, LLMs might misinterpret the symptoms and thus propose incorrect treatment plans. It may also lead to problems in improper medical management. The next issue is the lack of genuine clinical experience. The model cannot perform diagnoses like a human doctor does. It is unable to take into account the non-textual information such as the patient's facial expressions and body language. It is also impossible to conduct examinations such as auscultation and palpation. Their judgment is limited to the patient's own description only.

3.2 Data Limitations

The main reason for the poor clinical adaptability of Med-PaLM is that the lack of clinical data support results in a relatively low accuracy rate of its responses for specialized diseases. Secondly, the medical data used for training the pre-training models of LLMs may become outdated. Additionally, Med-PaLM is limited to providing textual responses. Therefore, it is difficult to handle complex medical scenarios, such as being unable to process multimodal medical data including images. And the access to clinical data is scarce and costly. This has led to the fact that the datasets used for training LLMs are not sufficiently extensive, which has severely restricted their development. The access to clinical data is subject to strict ethical, legal and privacy issues. It can only be used after going through a complicated approval process. Next, medical data must be manually labeled by experienced medical experts to ensure its accuracy. This process is time-consuming and not very efficient. So how to safely obtain and utilize medical data has become a key issue for the future development of LLMs.

3.3 Privacy and Ethical Issues

When using LLMs to provide initial diagnosis or post-operative rehabilitation information, there is no doubt that there is a risk of patient medical records being leaked. Health information is the most sensitive type of personal data. During the process of model training and application, ensuring that patient data is not leaked is the primary challenge. In many regions, the processing of medical data is subject to strict legal regulations. When applying LLMs in medical work, it is essential to ensure their legality and compliance. This is also a major challenge in terms of ethics. When using LLM to assist in clinical decision-making, it is necessary to determine the responsibility for any incorrect diagnoses or treatments provided by the model.

4 The Solutions to the Existing Problems

In response to the above issues, Med-PaLM2 has made technological innovations. Firstly, adopt retrieval-enhanced generation (RAG) [4]. It can incorporate external relevant knowledge into the input of the large language model to enhance the quality of the final answer. The specific implementation method is as follows: the retrieval chain [5]. The retrieval chain will conduct individual retrievals based on the initial answers provided by the model, based on the search results, further optimization is carried out to obtain the final answer. This method can alleviate the problem of hallucinations. In addition, the use of fine-tuning methods has also effectively improved the accuracy of the model and reduced the occurrence of hallucination phenomena [10]. Furthermore, optimizing the model training layer can significantly reduce the probability of hallucinations occurring from the very beginning. Using higher-quality and more precise training data to filter out incorrect information can also significantly improve the accuracy of the model. Secondly, integrate the visual encoder with the genetic data processing module to enhance the multi-modal medical task capabilities

and cover more clinical scenarios. Thirdly, conduct specialized fine-tuning based on multi-source medical data, and introduce the "ensemble optimization (ER)" prompt strategy [5]. In the first stage, it can provide different solutions and multiple answers based on the questions and examples. In the second stage, it can optimize the final answer through various solutions and answers. Its accuracy rate in the MedQA dataset has increased to 86.5% [5,11]. In addition, a standardized data sharing model should be established to obtain more accurate and reliable data. Expand the sources of data, establish data alliances for hospitals and departments in various regions and even countries, carry out resource integration, and share the data after obtaining the informed consent of patients. It can also integrate data from different departments, inpatient records, and imaging tests to address the issue of data singularity. Besides, PubMed, as a medical literature database, has included a large number of medical research papers [11]. The researchers can also obtain high-quality literature from it as a reference. Regarding ethical issues, it is necessary to conduct in-depth research and establish sound ethical guidelines and regulatory mechanisms to protect patient privacy and prevent data abuse. At the same time, rigorous clinical trials are necessary to verify its safety and efficacy. Finally, as medical assistance tools, LLMs can adopt a "human-machine collaboration" mode in medical question answering, which is a cooperative approach combining "human intervention" and "AI technology". LLMs can conduct the initial diagnosis based on the patient's description. Then, the doctor will select the treatment plan based on the initial diagnosis made by the LLMs. However, they must not directly provide treatment and medication plans to the patient.

5 Conclusion

This article analyzes the current development status of large language models centered around Med-PaLM and Med-PaLM2. It raised its main issues and discussed the solutions. Currently, LLMs are still developing rapidly. However, the challenges they face are equally significant and cannot be ignored. Improving model accuracy, addressing data limitations, and resolving ethical issues remain the main challenges that need to be overcome in the future. Although LLMs have performed well in the medical field, they still cannot replace human doctors, especially in terms of clinical decision-making. Looking ahead, the development of medical LLMS requires the joint efforts of experts in the medical field and experts in AI. Develop multimodal LLMs and conduct more clinical experiments to carry out personalized medical services. Without violating ethical and safety standards, LLMs will be one of the most promising applications in future medicine.

References

1. Wang, B., Xie, Q., Pei, J., et al.: Pre-trained language models in biomedical domain: A systematic survey. ACM Computing Surveys, 56(3): 1-52 (2023)

2. Li, H Y., Fu, J F.: Python A. Implementing large language models in health care: clinician-focused review with interactive guideline. Journal of medical Internet research, 27: e71916 (2025)
3. Omiye, J A., Gui, H., Rezaei, S J., et al.: Large language models in medicine: the potentials and pitfalls: a narrative review. Annals of internal medicine, 177(2): 210-220 (2024)
4. Alonso, I., Oronoz, M., Agerri, R. Medexpqa: Multilingual benchmarking of large language models for medical question answering. Artificial intelligence in medicine, 155: 102938 (2024)
5. Singhal, K., Tu, T., Gottweis, J., et al.: Toward expert-level medical question answering with large language models. Nature Medicine, 31(3): 943-950 (2025)
6. Wang, D., Zhang, S.: Large language models in medical and healthcare fields: applications, advances, and challenges. Artificial intelligence review, 57(11): 299 (2024)
7. Zheng, Y., Gan, W., Chen, Z., et al.: Large language models for medicine: a survey. International Journal of Machine Learning and Cybernetics, 16(2): 1015-1040 (2025)
8. Tan, S., Xin, X., Wu, D. ChatGPT in medicine: prospects and challenges: a review article. International journal of surgery, 110(6): 3701-3706 (2024)
9. Yu, E., Chu, X., Zhang, W., et al.: Large Language Models in Medicine: Applications, Challenges, and Future Directions. International Journal of Medical Sciences, 22(11): 2792 (2025)
10. Qiu, P., Wu, C., Zhang, X., et al.: Towards building multilingual language model for medicine. Nature Communications, 15(1): 8384 (2024)
11. Xiao, H., Zhou, F., Liu, X., et al.: A comprehensive survey of large language models and multimodal large language models in medicine. Information Fusion, 117: 102888 (2025)