Proceedings of CONF-MSS 2025 Symposium: Automation and Smart Technologies in Petroleum Engineering
DOI: 10.54254/2755-2721/149/2025.KL22365

Large Language Models in Healthcare: A Review of Current Applications

Yonghao Shi

School of Electronic and Information Engineering, Tongji University, Shanghai, China
incredibletony@tongji.edu.cn

Abstract: The application of the Large language model(LLM) in the medical field has obtained initial achievements, attracting widespread excitement and concern. This paper generalizes the development history of LLM and demonstrates the advantages of LLM and its good applicability in the medical field. In the aspects of medical dialogue systems, discharge summaries and combination with computer vision, this paper introduces the cutting-edge applications of LLMs in the medical field and elaborates on the problems that this application solves and the benefits it brings to healthcare and clinical medicine. This paper also discusses the problems that faces the full application of LLMs. The ethical problem includes biased datasets and incomplete risk management, while the technical problems cover the inaccurate information caused by the out-of-date database and model reliability. For the suggestions, real-time internet searching is recommended to be combined with the database, and fine tuning and model reconstruction may be the solution to unsatisfying model reliability. Considering the potential of LLMs, their application in medical field may also provide a great reference value for other industries, bringing significant development and convenience to human society.

Keywords: Natual Language Processing, Large Language Models, Healthcare

1. Introduction

With the development of artificial intelligence technology, Large language model(LLM) is profoundly changing the landscape of various industries. As a powerful and productive tool, LLM has also fully demonstrated its potential in medicine and the healthcare industry. At present, the medical and healthcare field still confronts many problems that urgently need improvement, such as the lack of human resources of medical professionals and untimely medical advice in certain periods or regions. However the innovative applications of LLM point out a promising path for improving current situation.

Medical dialogue system provides the patient with a far more convenient means to obtain professional medical advice [1]. Relying on the dataset of patient-doctor dialogues, patient can get credible answers from the dialogue system without the need to go to hospital [2]. Besides, the next-generation capabilities of LLMs can also save time for doctors. MEDISCHARGE can generate discharge summaries automatically based on the information of patients, reducing the workload of the doctors and redirecting medical human resources to areas of greater need [3]. LLMs combined with computer vision models can further enhance its practicality. For example, ChatCAD such a model that combines ChatGPT and a lesion detector to diagnose radiology images [4].

DOI: 10.54254/2755-2721/149/2025.KL22365

This paper explores the current applications of LLMs in the medical field. In Section 2, the origin of LLM is generalized, outlining the development history of language models which shows an evolving natural language processing capability. In Section 3 the cutting-age applications of LLM in medical field are introduced, focusing on doctor-patient communication, discharge summary and the combination of LLMs with computer vision. In section 4, the suggestions and prospects are proposed according to the current medical environment and the state of LLM technology application. Afterwords, the paper concludes with a summary.

2. Introduction of LLM

Natural Language Processing (NLP) is a field of research that intersects with artificial intelligence and linguistics, including the understanding, analysis, generation, and interaction of text. To solve the NLP task, there has been growing interest in exploring how to make machines master language intelligence. Initially, NLP underwent a transition from statistical to neural language modeling. During the process, the traditional language model(LM) only concentrates on specific tasks. Subsequently, due to the emergence of Transformer architecture, solving the long-term dependencies problems in Recurrent Neural Networks (RNNs), the pre-trained language model (PLM) has been proposed. Pre-trained on large corpus pf text in a self-supervised setting, PLMs can catch the basic features and patterns of natural language and perform, having good task versatility and better performance gains than conventional LMs after fine-tuning for downstream tasks [1]. Moreover, researchers found that by increasing the model parameters and the size of the data set to a certain extent, PLM not only significantly improved its performance, but also gained some capabilities that did not appear on small-scale models, such as contextual learning. Therefore, in order to distinguish NLPs of different sizes to reflect this phenomenon, the research community coined the term LLM to describe large-scale PLM [2].

With powerful natural language processing capabilities, LLM can recognize, interpret, and generate texts. Hence, LLM can achieve the generative goal of simulating human capabilities and creating content according to user needs, such as chatbot and text prediction. As a LLM chatbot, Chatbot Generative Pre-Trained Transformer (Chatgpt) possess strong cross-field cognitive abilities, reaching to humans' performance. This disruptive capability brings endless possibilities to various fields including the medical field, where the processing of text and images is highly empirical, as it often requires professional experience for accurate judgment.

Additionally, LLM can also be combined with computer vision to develop visual language models for multimodal dialogues, further expanding its application in medicine. Nowadays, in addition to ChatGPT, there are also other excellent LLMs such as the Gemini series and the Llama series, bringing revolutionary changes to the medical field, such as medical dialogue systems, discharge summaries and so on.

3. Application of LLM in medicine

In recent years, the powerful LLMs have opened up an era of possibilities in the medical field. Thanks to researchers' continuous exploration of how to effectively apply LLMs in medical work, some LLM-based tools have already taken shape and demonstrated great potential.

3.1. Medical dialogue system

Via online diagnosis, Telemedicine can save patients' time and give immediate medical advice, avoiding unnecessary waste of resources or neglect of symptoms that require proper attention. Meanwhile, after the breakthrough of COVID-19, the advantage of remote medical care has been more pronounced since people are not willing to contact other patients in the hospital. Medical

DOI: 10.54254/2755-2721/149/2025.KL22365

dialogue system based on LLMs provides a new approach to telemedicine, making the advantage of timeliness more prominent. The medical dialogue system aims to play the role of a doctor and give corresponding treatment suggestions based on the patient's symptoms, showing great potential in collecting patient information as well. Douglas et al conducted research, showing that ChatGPT without any specialized training has already obtained a relatively satisfactory outcome in responding to medical query. 33 physicians grated GPT-generated answers to 284 medical questions, reaching a median accuracy score of 5.5 in a 6-point Likert scale and a mean score of 4.8,garanteeing the basic reliability of the medical dialogue system [3].

The methods of Medical dialogue system based on LLMs can be summarized into two ways. One way is prompt word method, using Zero/Few-shot Prompting, Chain-of-Thought Prompting, Self-consistency Prompting and other techniques to optimize model performance and improve its application in medical question answering. The other one is fine-tuning the basic model with medical database, allowing models a better performance in specific fields [4].

For example, ChatDoctor is a medical chat model adapted from a large language model meta-AI (LLaMA), using a large text dataset including 100,000 patient-doctor dialogues for fine-tuning [5]. Besides, ChatDocter implemented a self-directed information retrieval mechanism, bringing a comprehensive answering ability with the help of extra knowledge. Table 1 shows the quantitative comparison between ChatDoctor and ChatGPT with BERTScore.

Table 1: Comparison between ChatDoctor and ChatGPT with PERTScore

	ChatGPT	ChatDoctor
Precision	±0.0188	0.8444±0.0185
Recall	±0.0164	0.8451±0.0157
F1 Score	±0.0143	0.8446±0.0138

Huatuo is another tuned LLaMA model tailored to Chinese medicine by being trained on the Chinese medical knowledge graph [6]. The researchers introduced a new evaluation metric - SUS (Safety, Usability, Smoothness), which measures the safety (whether it misleads users), practicality (medical expertise) and language fluency of the generated responses respectively. Table 2 shows the comparison between Huatuo and other three baseline models with the SUS scale ranging from 1 (not acceptable) to 3 (good).

Table 2: Comparison between Huatuo and other baseline models with SUS

	Safety	**Usability**	**Smoothness**
LLaMA	**2.93**	1.21	1.58
Alpaca	2.64	2.05	2.30
ChatGLM	2.59	1.93	2.41
Hua Tuo	2.88	**2.12**	**2.47**

3.2. Discharge summaries

Discharge summary is another potential application field for LLMs. A discharge summary is a comprehensive summary of the patient's health condition, recording the diagnosis results, treatment status, etc. during hospitalization. It is the basis for the patient to understand his or her own condition and ensure the continuity of treatment, While a clear and comprehensive discharge summary can effectively reduce information omissions in subsequent medical decision-making and improve the quality of medical services. However, writing a high-quality discharge report containing essential

DOI: 10.54254/2755-2721/149/2025.KL22365

information often consumes a lot of doctors' time and energy. This situation may further lead to a pressing healthcare issue: excessive workload for doctors and understaffing. Moreover, insufficiently written discharge reports may invite potential risks.

Facing these challenges, a system that can automatically generate discharge summaries is urgently needed. Since the discharge summary has a largely standardized template, NLP tools such as ChatGPT provide new possibilities for automatically generating summaries. By inputting some necessary information and instructions, doctors can use ChatGPT to generate a formal discharge summary in a few seconds. This greatly reduces the workload of doctors. Meanwhile, research also shows that ChatGPT can complement the details that often overlooked in traditional writing methods [7].

MEDISCHARGE is a MEditron-7B-based medical summary generation system which can be used in discharge summary generation [8]. Instead of inputting basic information by doctors, this system can automatically read the data from a patient's Electronic Health Record (EHR), filter and summarize the important information to meet the size of the context window of the model and finally generate the summary. With a larger context window and precisely extracted text, MEDISCHARGE outperforms its base model in all aspects, including the similarity between the generated text and ground truth and the consistency of clinical concepts. On the other hand, MEDISCHARGE still faces some limitations. For example, this system is only applicable to text-only summaries and lacks the ability to process medical images, which can be a crucial component of patient data. Examples of such images include anatomical diagrams.

3.3. Multimodal diagnosis based on computer vision

The combination of computer vision and LLMs also holds great potential for application in the field of medicine. The main purpose of this combination is to address historical challenges concerning the manual interpretation of medical images. With the development of medical processing technology, the evolving imaging modalities have brought increasingly precise medical image data. However, the efficiency of manual image analysis fails to match the explosive growth of data, leading to instability in diagnostic accuracy, reporting speed and quality. The application of CV and LLM can greatly reduce the workload of imaging departments and, to a certain extent, curb the adverse effects caused by doctors' subjective judgment, improving the analysis quality [9].

ChatCAD is a framework that combines an LLM with a medical image computer-aided diagnosis (CAD) network [10]. It converts the multimodal analysis results of medical images into natural language text, and uses the medical knowledge and logical reasoning capabilities of LLM to generate more accurate and easier-to-understand diagnostic reports. With ChatCAD, a diagnose-assistive interactive chatbot can explore greater possibilities in assisting doctors with making decisions.

SkinGPT-4 is such a dermatology diagnostic system supported by an interactive vision - language model [11]. Firstly, as a fine-tuned version of MiniGPT-4SkinGPT-4 uses natural language to describe the medical features in the images, Identifying the characteristics and categories of skin conditions. After that, it diagnoses skin diseases and provides corresponding treatment recommendations. By evaluating 150 cases and comparing the results with manual diagnoses, it was shown that SkinGPT-4 could consistently make correct judgments. Although SkinGPT-4 cannot completely replace human doctors, It allows patients to understand their own conditions more conveniently, facilitates their communication with doctors, and brings constructive changes to regions where people who have limited access to medical services.

DOI: 10.54254/2755-2721/149/2025.KL22365

4. Suggestions and Prospects

As an emerging research direction, the application of LLMs in the medical field still faces many problems such as ethical issues, out-of-date database and model reliability.

In the aspect of ethical considerations, the training data of LLMs may transmit biases, including those related to gender, race, culture, etc. Besides, as LLM's body of knowledge mainly comes from high-income, English-speaking countries, The medical paradigms of the suggestions given by the LLM may not be suitable for regions with lower economic levels [12]. In addition, when LLM is put into actual medical applications on a large scale, risk management issues cannot be avoided. Once an accident based on LLM misjudgment occurs, there must be a transparent and clear responsibility system to protect the rights and interests of patients.

Besides, the lack of recency in the database may contribute to the reduction of reliability and accuracy of LLM-generated text. For example, GPT is trained on the text before September 2021. In the medical field, the untimeliness of this database will bring significant disadvantages, because there are often new academic terms or research methods, especially when there is an epidemic virus. One corresponding solution is to implement real-time internet searching to ensure the access to latest medical information. However, internet materials like medical notes certainly contain unavoidable errors and misguiding information. The following improvement on database can focus on secondary verification. Since it seems impossible to validate such a large database manually, a machine learning method can be adapted to automatically carry out data assessment based on the initial manual grading by experts.

Since LLM is trained to capture the associations between words rather than truly understanding the meaning of the query of the given information, it can output plausible-sounding but not-accurate answers, which is called 'hallucinations'. In order to promote large-scale practical application of LLMs in the medical field, further research is needed to eliminate or surpass this kind of hallucinations as far as possible. One way is to eliminate the inaccurate information through fine-tunning, while the other is to reconstruct the model and training method, enabling the model to develop semantic knowledge [13].

5. Conclusions

From the evolution of LLMs to the modern applications of LLMs, this paper mainly discussed the history, applications in medicine, current challenges and corresponding suggestions of LLMs. The emergence of LLMs has fundamentally transformed natural language processing technologies. The subsequent advent of advanced models such as GPT has brought infinite possibilities for the application of LLMs in the medical field. In addition to the derivative medical applications of chatbots and text generation, LLMs can also combined with technologies such as computer vision, exhibiting a strong potential and vitality. Nevertheless, the potential risks are still causing concerns over the ethical and safety issues, indicating that the application of LLMs in this field is not yet mature and cannot be widely applied at this stage. The biased dataset contributes to outputs that are unsuitable for all populations. The lack of recency in the database may generate unreliable outputs, while the misjudgment management is not yet perfected. The 'hallucinations' are still inevitable, which is fateful for clinical medicine. To solve these problems, researchers and medical institutions should collaborate closely, while the voice from patients also matters. Further researches are needed to precisely validate the reliability of the models. Once the ethical and technology issues are addressed, LLMs may revolutionize the medical field with its convenience and efficiency. The successful model will also provide a great reference value for other industries, triggering a chain reaction to further explore the potential of LLMs.

DOI: 10.54254/2755-2721/149/2025.KL22365

References

[1] *Shi, X., Liu, Z., Du, L., Wang, Y., Wang, H., Guo, Y., Ruan, T., Xu, J., Zhang, X., and Zhang, S. (2024) Medical Dialogue System: A Survey of Categories, Methods, Evaluation and Challenges. Findings of the Association for Computational Linguistics: ACL 2024, Bangkok, Thailand, 2840–2861.*

[2] *Li, Y., Li, Z., Zhang, K., Dan, R., Jiang, S., and Zhang, Y. (2023) ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge. Cureus, 15(6), e40895. doi: 10.7759/cureus.40895.*

[3] *Wu, et al. (2024) EPFL-MAKE at "Discharge Me!": An LLM System for Automatically Generating Discharge Summaries of Clinical Electronic Health Record. BioNLP 2024.*

[4] *Wang, S., Zhao, Z., Ouyang, X., et al. (2023) Chatcad: Interactive Computer-Aided Diagnosis on Medical Image Using Large Language Models. arXiv preprint, arXiv:2302.07257.*

[5] *Naveed, H., Khan, A., Qiu, S., Saqib, M., Anwar, S., Usman, M., Barnes, N., and Mian, A. (2023) A Comprehensive Overview of Large Language Models. arXiv preprint, arXiv:2307.06435.*

[6] *Zhao, W., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., and Wen, J.-R. (2023) A Survey of Large Language Models. arXiv preprint, arXiv:2303.18223.*

[7] *Johnson, D., Goodman, R., Patrinely, J., Stone, C., (2023) Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model. Res Sq [Preprint], rs.3.rs-2566942. doi: 10.21203/rs.3.rs-2566942/v1.*

[8] *Wang, H., Liu, C., Xi, N., et al. (2023) Huatuo: Tuning LLaMA Model with Chinese Medical Knowledge. arXiv preprint, arXiv:2304.06975.*

[9] *Patel, S. B., and Lam, K. (2023) ChatGPT: The Future of Discharge Summaries? The Lancet Digital Health, 5(3), e107-e108.*

[10] *Tian, D., Jiang, S., Zhang, L., Lu, X., and Xu, Y. (2024) The Role of Large Language Models in Medical Image Processing: A Narrative Review. Quant Imaging Med Surg, 14(1), 1108-1121. doi: 10.21037/qims-23-892.*

[11] *Zhou, J., He, X., Sun, L., et al. (2023) SkinGPT-4: An Interactive Dermatology Diagnostic System with Visual Large Language Model. arXiv preprint, arXiv:2304.10691.*

[12] *Li, H., et al. (2023) Ethics of Large Language Models in Medicine and Medical Research. The Lancet Digital Health, 5(6), e333-e335.*

[13] *Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., et al. (2023) Large Language Models in Medicine. Nat Med, 29, 1930–1940. https://doi.org/10.1038/s41591-023-02448-8.*