

Application, Challenges and Prospects of Large Language Models in the Medical Field

Dongjie Wang^{1,a,*}

¹*School of Computer and Information, Fujian Agriculture and Forestry University, Fuzhou, China*
a. wuyan@asu.edu.pl

**corresponding author*

Abstract: Artificial intelligence has become a new trend in the new era. Within its field, large language models (LLM) have unique advantages in today's massive data due to their excellent text understanding and generation capabilities. This article briefly outlines the development background and architecture of LLM. In view of the rapid development of large language models, this article analyzes the practical applications, challenges and future development prospects of LLM from a medical perspective. The advantages of LLM in medical application were analyzed, and the application of LLM in various medical fields was reviewed by taking BioGPT, Zhongjing (CMLM-ZhongJing) large language model and HuatuoGPT as examples. Secondly, the paper analyzes the challenges that LLM currently faces in the medical field, such as the "hallucination" problem, the explainability problem, etc., and provides corresponding solutions. Finally, the future development prospects of LLM in the medical field are proposed, which is to transform from single modality to multimodality, from large-scale to lightweight, strengthen the integration with medical equipment, and better provide services for relevant medical staff, patients, and students in the medical field.

Keywords: Artificial Intelligence, Transformer, Large Language Models, Medical.

1. Introduction

The large language model (LLM) is an important branch in the field of artificial intelligence. It is a natural language processing model that can process massive amounts of text content and has good generalization capabilities. The concept of artificial intelligence was first proposed in 1956, and its development has undergone many changes and innovations. Initial artificial intelligence research focused on rule programming, hoping to allow machines to complete designated tasks through clear instructions. In the 1990s, machine learning became a hot topic in artificial intelligence research. After the 21st century, deep learning has become the main content of artificial intelligence. LLM is based on deep learning technology. It imitates the human brain in processing data by building a deep artificial neural network model. Without the need for external programming instructions, it analyzes massive data to understand text and generate natural human language. This ability has unique advantages in today's information age. Large language models are gradually becoming an important part of the development of various fields, especially in the medical field. The potential it demonstrates is unpredictable. Research in the medical field includes the development of medical big language models and why big language models can perform well in the medical field. Zhang Zhicheng, Wang Jing, Zhang Yang and others developed a large language model OrthoGPT for precision diagnosis

and treatment[1]. The model integrates the alignment and fusion technology of multimodal features, incorporates the orthopedic simulation world based on parallel theory and multi-agent intelligent interaction strategy, and can provide preliminary disease diagnosis and emergency treatment guidance in non-hospital environments. In addition, the model can simulate and predict individualized postoperative outcomes, assist doctors in surgical planning, and provide personalized rehabilitation treatment recommendations. Chen Zijia, Peng Wenxi, Zhang Dezheng and others conducted an in-depth analysis of the advantages of big language models in the field of traditional Chinese medicine, and analyzed a variety of big language models applied in the field of traditional Chinese medicine [2]. This article is divided into five parts. It will introduce and explain the architecture of the big language model, analyze why the big language model has unique advantages in the medical field, introduce the medical big language model developed by many scholars today based on existing technologies, the many challenges that the medical big language model will encounter in the future and the corresponding reasonable solutions, and the model development prospects given by combining multiple considerations. It summarizes the application, challenges and prospects of the big language model in the medical field.

2. The development and architecture of large language models

The Large Language Model (LLM) is a neural network model formed by using large-scale data and complex network structure, and a language model formed through large-scale text pre-training. The Transformer neural network architecture proposed by the Google team in 2017 laid the foundation for the mainstream algorithm architecture in the current large model field. In 2023, Open AI launched ChatGPT-4, which set off a wave of "artificial intelligence" craze[3]. At present, in addition to ChatGPT, the more popular large language models include BERT, Wenxin Yiyan, Tongyi Qianwen (QWen), Meta's LLaMA, etc. Its development process from the 1950s to the present is shown in Figure 1:

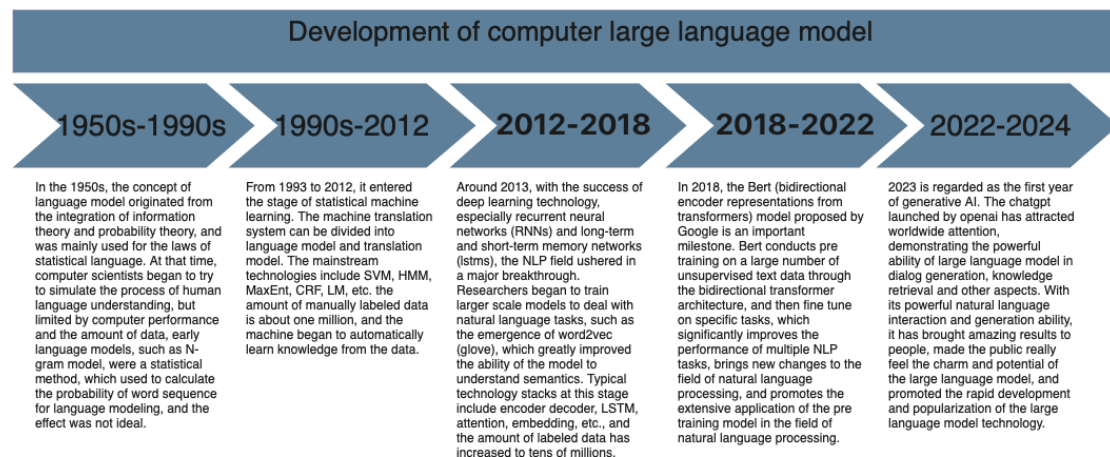


Figure 1: Development of large language model (Picture credit: Original)

As shown in Figure 2, the core of the Transformer architecture consists of two parts: the encoder and the decoder[4]. After the input sequence (a piece of text), the encoder learns its features, converts it into a hidden fixed-length vector representation, namely the context representation, and sends it to the decoder, which generates the output sequence. Through Transformer's unique self-attention mechanism, different attention weights are assigned to different parts of the input sequence, decomposing long text into a series of independent tokens. When outputting, the probability of each token is used to predict what the next output value will be, commonly known as text chain. The

Transformer architecture solves the problem that the Recurrent Neural Network (RNN) architecture cannot handle tasks that require long-term memory and the Long Short-Term Memory (LSTM) architecture cannot handle parallel computing. In 2023, OpenAI launched ChatGPT4, which has demonstrated excellent problem-solving capabilities in all walks of life. This has prompted more and more scientific research institutions and companies to increase their research and development of LLM, especially in the medical field, and has completed the development of many large language models in the medical field. For example: DoctorGPT designed for medical knowledge question answering[5], OrthoGPT a large multimodal orthopedic model for precision diagnosis and treatment[1], and the MedPaLM model designed specifically for medical dialogue systems[6].

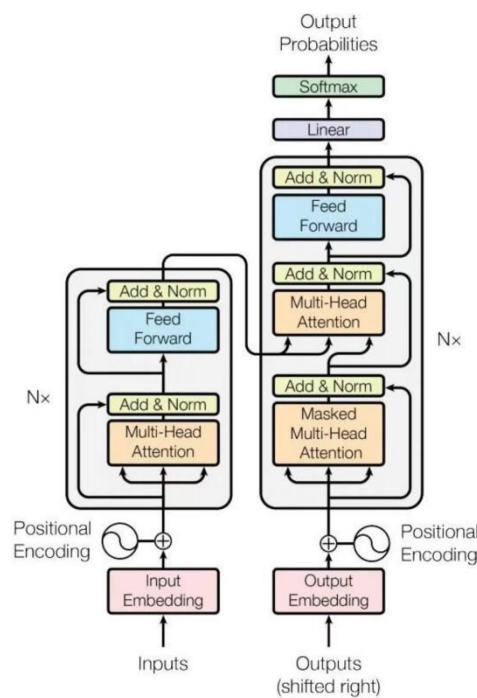


Figure 2: The Transformer - model architecture[4].

3. Application advantages of large language models in the medical field

3.1. Advantages of large language models in the medical field

Nowadays, there are many data sets available for pre-training and learning of large language models, as shown in Table 1, which enables large language models to have a strong understanding of specific knowledge and can quickly learn, absorb, and digest massive amounts of medical information. It can also be used to perform natural language processing analysis on specific paper journals and medical literature, and to mine large-scale medical and drug data to discover potential patterns, trends, and associations, effectively helping researchers in research and development. When patients use it, the large language model can also analyze their historical cases and patient information to assist doctors in making diagnoses. Its advantages are reflected in the following aspects:

3.1.1. Large model assisted therapy

Before diagnosis, the big model can use guided questions and free questions to collect patient information in advance and complete medical record generation, which can be directly referenced by doctors to improve the patient's medical experience. After the diagnosis, the big model can act as a virtual doctor to answer various questions for patients: medication consultation, operation guidance,

report analysis, etc. Doctors can also use large models to assist patients in understanding the entire recovery process in detail and better follow up on patient health assessments. It can provide support in clinical decision-making, simplify clinical decision-making, improve the efficiency and accuracy of medical treatment, analyze massive clinical case data, and provide theoretical support for medical personnel[7,8]. In an emergency, it can quickly analyze the patient's various indicators and provide doctors with detailed health data to ensure that the patient can receive timely.

3.1.2. Large model online consultation

Before the launch of the big model, people generally searched for disease information by entering their medical conditions through search engines such as Baidu. This method had extremely low interactivity, was difficult to meet needs, and could not obtain accurate results. However, after the launch of LLM, online medical consultation has undergone a transformation. LLM can not only maintain online communication with patients 24 hours a day, but also answer some of the patients' privacy issues. A study compared the performance of ChatGPT with that of doctors answering patients' questions in online clinics. The results showed that ChatGPT performed as well as doctors in online clinics. This result also shows the good performance and application potential of LLM in online consultations [9].

3.1.3. Large model to enhance knowledge of paramedics

After pre-training, LLM can provide a stress-free platform for medical students and interns to train, improving their clinical practice, experience and ability to communicate with doctors and patients. Relevant scientific researchers can rely on LLM's powerful literature search capabilities and only need to enter keywords to search for relevant literature for reference, thereby improving the efficiency of academic research.

Table 1: Common big data sets in medical field

Data Set	Introduction	Link
PubMed	It has collected more than 30 million medical literatures from MEDLINE, journals of life sciences and other fields, covering clinical medicine, pharmacy and other fields.	https://pubmed.ncbi.nlm.nih.gov
cMedQA2	The cMedQA2 dataset is a Chinese medical question-answering dataset designed to support natural language processing tasks in the medical field (such as medical question-answering systems, medical dialogue systems, etc.). The dataset contains multiple questions and answers from the medical field, covering different areas of medical knowledge, such as clinical symptoms, treatment plans, drug use, etc.	https://github.com/zhangsheng93/cMedQA2

Table 1: (continued).

Huatuo-26M	It contains more than 26million Chinese medical question and answer pairs from the medical knowledge base and encyclopedia.	https://github.com/FreedomIntelligence/Huatuo-26M
UK Biobank	This is a large-scale biomedical database and research resource, containing detailed health information from 500000 British volunteers aged 40-69, including genome data, lifestyle information, medical images and long-term health records.	https://www.ukbiobank.ac.uk
MIMIC-CXR	This is a large-scale chest X-ray image data set, which contains X-ray films of 227943 patients from 371920 imaging examinations. This data set not only provides high-quality medical images, but also contains corresponding radiology reports, providing researchers with rich clinical information.	https://mimic.mit.edu
webMedQA	The problems in the data set cover a variety of clinical departments, including internal medicine, surgery, gynecology, pediatrics, etc., with a total of 63284 problems. These questions come from real user consultation, which are representative and practical.	https://github.com/hejunqing/webMedQA
NHANES	The national health and Nutrition Examination Survey (NHANES) database, It is a cross-sectional survey based on people at all levels in the United States	https://www.cdc.gov/nchs/nhanes/

3.2. Specific applications of large language models in the medical field

In view of the superiority and adaptability of LLM in the medical field, the following summarizes several vertical models in the medical field:

3.2.1. BioGPT

BioGPT is a generative pre-trained model designed for biomedical text generation and mining based on the Transformer architecture by Renqian Luo and other scholars [10]. By pre-training large-scale biomedical texts, efficient natural language processing tasks in professional fields are achieved. It can help medical researchers generate research summaries and speed up the literature review process. BioGPT can help researchers identify the connection between diseases and various compounds, and assist in the development of drugs for related diseases.

3.2.2. CMLM-ZhongJing

The ZhongJing Large Language Model (CMLM-ZhongJing) was developed by Professor Wang Haofen of Tongji University, Professor Zhang Wenqiang and Dr. Wang Yan of Fudan University under the guidance of ROI members of Fudan University for application in the field of traditional Chinese medicine [11]. Data support and manual evaluation were provided by Dr. Wu Sunsi, Dr. Zhao Xue, Dr. Ma Qingshan, Dr. Fang Yide, Dr. Chen Yue, Dr. Jiao Yuqi, and Dr. Liu Xiyu from Longhua Hospital of Shanghai University of Traditional Chinese Medicine. Inspired by the ancient Chinese medicine scientist Zhang Zhongjing, the model uses deep learning technology to pre-train Chinese medicine knowledge. The model refers to the process of human memory knowledge, uses professional tables and the language representation ability of large language models, sets strict prompt templates, and enables the model to generate multiple scenarios in the field of Chinese medicine based on the table data of Chinese medicine gynecology prescriptions. For example: prescription function, tongue and pulse, diagnosis plan formulation, prescription, follow-up, etc., so that the model can improve the diagnostic thinking logic and reasoning ability of TCM prescription data. Instructions based on the content of ancient TCM books, noun explanations, symptom synonyms, antonyms, symptoms, and other instructions have been added.

Through preliminary comparison with domestic large language models, it was found that the model has excellent generalization ability on a diversified diagnosis and treatment decomposition instruction data set based on 300 TCM prescription data. In terms of medical care, it can analyze and understand people's needs in more detail and provide more understandable explanations and relief methods. This vertical medical model is superior to the general model in the medical field. Figure 3 is a comparison chart. Preliminary testing found that the large model also has certain diagnostic and prescription capabilities in clinical fields of traditional Chinese medicine other than gynecology, and has good generalization ability. It is shown that using a million-level guidance dataset and a multi-task decomposition treatment strategy are effective in enhancing the diagnostic thinking ability of the model. Researchers also show the potential of large language models with 7B parameters in areas where expertise has a low tolerance for errors, such as medical and legal scenarios. This also shows that the vertical state of the large language model is not only applicable in the medical field, but also in other fields.

Question	ChatGPT-3.5	Cici	CMLM-Zhongling
<p>Zhou XX, male, 29 years old. Due to sudden deafness for nearly a month, he was treated with Yiqi Congming Decoction and other medical treatments, but his deafness was as good as ever and he was in great pain, so he delayed the diagnosis and treatment. Careful inquiry knows that the disease starts after a cold and sweating. Diagnosis: Bai complained of deafness, palpitations and fatigue, slight chills, reddish tongue and thin white fur. The pulse is thin and weak. Give TCM diagnosis and prescription suggestions</p>	<p>The patient suffered from sudden deafness after cold and sweating, accompanied by palpitations, fatigue, chills, reddish tongue, thin white fur, weak pulse, and was diagnosed as deafness due to Qi deficiency and yang deficiency. The treatment should be beneficial to Qi and spleen, warming yang and unblocking orifices. The prescription was modified with Bazhong Yiqi Decoction, combined with Ramulus Cinnamomi and asarum Wenyang powder to relieve cold and unblock the orifices. For patients with significant tinnitus, magnet and Acorus tatarinowii were added to open the orifices. For patients with severe chills, aconite was added to warm and replenish Yang Qi. Supplemented by Moxibustion at Shenshu, Guanyuan and Zusanli to help Yang and replenish qi, diet should be warm and tonic, and avoid cold stimulation. Pay attention to regulating qi and blood, observe the changes of the disease, and adjust the treatment plan when necessary.</p>	<p>Zhou XX, male, 29 years old, has sudden deafness for nearly a month. His disease started after a cold and sweating, and was invalid after treatment with Yiqi Congming decoction. The diagnosis showed deafness, palpitations, fatigue, chills, reddish tongue, thin white fur, and weak pulse. Traditional Chinese medicine diagnosis is heart kidney yang deficiency, ear orifices loss of nourishment syndrome. It is suggested to use ephedra Fuzi Asarum decoction combined with Guizhi Gancuo Decoction, 6G for medicinal ephedra, 9g for prepared aconite (decocted first), 3G for asarum, 9g for Guizhi, 6G for licorice, 9g for Acorus tatarinowii, 15g for magnetite (decocted first). When decocting, decoct aconite and magnetite for 30-60 minutes, soak the remaining drugs for 20 minutes and decoct them together with the decocted drugs first, boil them in high heat and turn them into low heat for 20-30 minutes and 15-20 minutes respectively, mix the two solutions and take them in warm twice, one dose a day. During treatment, observe the changes of symptoms and adjust the medication. Patients should pay attention to rest, and the prescription should be applied under the guidance of professional traditional Chinese medicine practitioners.</p>	<p>According to the symptoms of the patient, it can be preliminarily judged as white complaint deafness syndrome. To further understand the etiology of the patient is cold sweating, and the pathogenesis is Qi stagnation and blood stasis. Therefore, traditional Chinese medicine should be used to clear Qi and blood, moisten lung and relieve cough, relieve phlegm and thirst, and promote blood circulation and remove blood stasis in order to treat white v. deafness syndrome. Among them, traditional Chinese medicine prescriptions, such as clearing Qi and blood, moistening lung and relieving cough, resolving phlegm and thirst, and promoting blood circulation and removing blood stasis, have the effects of moistening lung and relieving cough, promoting blood circulation and removing blood stasis, and can regulate qi and blood, promoting blood circulation and removing blood stasis, so as to treat white deafness syndrome. It should be noted that overdose should be avoided during use to avoid side effects such as drug allergy. At the same time, patients should pay attention to maintain good living habits, avoid overwork and emotional fluctuations, so as to promote the recovery of the disease.</p>

Figure 3: Question and answer comparison (Picture credit: Original)

3.2.3. HuatuoGPT

To meet the growing demand for rapid online medical consultation, the Chinese University of Hong Kong developed HuatuoGPT[12], an LLM specifically designed for medical consultation. The core of this model is to utilize the refined data of ChatGPT and real data from doctors in the supervised fine-tuning (SFT) stage. In order to coordinate the advantages of the two data sources, reinforcement learning from mixed feedback (RLMF) was cited, in which a reward model was trained to align the advantages of the two data sources (ChatGPT and doctors), so that the model can chat as fluently as ChatGPT and behave like a doctor.

The research team analyzed three Chinese medical datasets, cMedQA2, webMedQA, and Huatuo-26M, and compared HuatuoGPT with ChatGLM and ChatGPT using multiple evaluation metrics such as BLEU, ROUGE, GLEU, and DISTINCT. The results show that HuatuoGPT outperforms these two in many indicators. Even in cMedQA2, the performance of HuotuoGPT is better than that of the fine-tuned T5[13]. It shows that HuatuoGPT performs well in extended conversation contexts, has strong generalization capabilities, and has more outstanding interactive diagnostic capabilities in patient consultation scenarios, and can effectively handle a wide range of medical question-answering tasks.

4. Challenges of large language models in the medical field

The wide application of LLM in the medical field has brought modern medicine to a new stage. However, large language models still face a series of problems and challenges, such as "illusion" problems, interpretability problems, technical limitations, moral and ethical issues, and privacy question. This section will explore the challenges of LLM in the medical field and give possible solutions.

4.1. "Illusion" problem

The "illusion" problem of large language models means that when the model handles long sequence problems, the coherence of the output sequence does not match the logic and the real world. Mainly manifested in three aspects: logical contradictions in the content before and after the answer; the answer is not relevant to the input content; the answer seriously deviates from the facts. If a similar problem occurs in the medical large language model, it may endanger the life and health of users. This problem seriously affects the reliability and accuracy of the large language model.

The causes of "hallucinations" are diverse. The repetitiveness of pre-training data may cause the model to overly favor certain high-frequency words and cause "hallucinations". In the process of

model decoding, text is generated through probability and is random. Even if the patient inputs some content that the model has not been trained on, the model will predict the content and continue writing.

There are several ways to alleviate the "hallucination" problem. Manually clean up and design training content during training to reduce training repeatability and unreliable data. Adopt different decoding strategies and perform knowledge injection with the help of external knowledge bases. However, human tutoring requires a lot of manual operations, and the incorporation of external knowledge may conflict with the original knowledge. Therefore, how to design a large model that does not cause "illusion" problems still needs to be continuously explored in the future.

4.2. Interpretability problem

Interpretability refers to whether humans have sufficient understanding of the working principles, decision-making processes, and output results of the model [14]. Nowadays, most LLMs basically use the Transformer model. Its unique self-attention mechanism and deep learning make it difficult to explain, like a "black box". In the medical field, LLM's interpretability has higher requirements, which is not only related to the lives of patients but also the public's trust in the model. One study studied the explainability of artificial intelligence in the medical field from four aspects: technology, medical care, patients, and law [15]. Finding such models difficult to interpret can have serious consequences in the medical field. Nowadays, many researchers are already working hard to improve the interpretability of models. Biomedical researchers try to explain decision-making through T-PLMs [16], Some studies have proposed the open source library Ecco, a method that can analyze the Transformer architecture model mechanism and improve decision-making transparency [17].

4.3. Moral and ethical issues

When LLM is applied to the medical field, it may be due to the inherent bias of the input training content data. The bias will be amplified during training, and the model will output inappropriate remarks that are contrary to social values, causing harm to social moral concepts [18]. This content involves multiple considerations, such as: race, occupation, region, gender, etc.

Solving moral and ethical issues requires the public and relevant teams to increase their ethical awareness and integrate ethical principles into every aspect of design, development, and training. Artificial intelligence-related industries have developed sound laws, regulations and supervision systems to ensure that LLM's moral and ethical concepts are consistent with society. Improving interpretability is also an important step, which will help medical and scientific personnel make timely adjustments and corrections after discovering that the model outputs sequences that are not ethical. Accelerate explainable artificial intelligence technology breakthroughs and artificial intelligence ethical calculations [19], and strengthen cooperation among medical staff, artificial intelligence researchers, and legal experts to jointly develop a healthy development mechanism for LLM.

4.4. Privacy Issues

Medical data is a privacy-sensitive issue, and the opaque and difficult-to-explain working mechanism of LLM puts medical data at great risk of leakage. Privacy risks mainly involve four aspects: Data collection: The training of large language models requires massive text data; Model training: During the model training process, the user's private information may be learned and memorized by the model, resulting in easy leakage of privacy in future output; Model inference: During the process of model inference, someone with ulterior motives may be able to obtain the user's private information by entering specific queries or prompts; Model storage: Most large models are usually stored on cloud servers. If the server is attacked, user privacy can easily be obtained.

For the protection of private information, privacy environments or computing technologies, such as TEE, a trusted execution environment and secure multi-party computation (MPC), can be combined with large models [20, 21]. Large model privacy protection technology based on secure multi-party computation mainly uses polynomials to replace nonlinear activation functions so that methods such as homomorphic encryption or similar homomorphic protocol large model secure reasoning can be applied to large models [22]. Since the key training data of large models are distributed in different organizations and regions, cross-organizational collaboration is required to achieve comprehensive language understanding. The large model training paradigm of federated learning and vertical federated learning can be considered [23]. These technologies can be mixed and the advantages of different technologies can be absorbed to achieve more efficient and safer model training and inference.

Although these methods play a certain role in protecting privacy, they still encounter the problems of high communication cost and long time consumption, and there are also gaps in actual production applications. In actual production applications, efficient model inference is very important. Researchers still need to continue to further explore new methods and technologies to reduce the cost of large model privacy protection technology, improve efficiency, and make it closer to actual applications.

5. The development direction of large language models in the medical field

5.1. Multimodal large language model

When users use medical large language models, they cannot fully describe their condition using only text. Therefore, the integration of images, videos, and audio, that is, multi-modality, is an imperative step to make large language models more capable of meeting human intentions. Using multi-modal data input such as speech and pictures for training, multi-modal output under different needs can be obtained to meet diverse application scenarios. At present, a variety of big data models have integrated multimodality into the LLM system, such as GPT-4 and OrthoGPT launched by Microsoft, which are multimodal orthopedic large models [3,5]. Moreover, the quantity and quality of training data content will also affect the performance of large language models. Pre-training multi-modal large models means multi-modal generation tasks [24, 25]. With strong cross-modal problem-solving capabilities and stronger performance, it can provide patients with more complete services and expand the application scope of LLM.

5.2. Models tend to be lightweight

In actual development, in addition to improving the capabilities of LLM, the computing and storage requirements for large-scale data are also increasing. Large-scale costs lead to insufficient algorithm resources, and the scale of language models continues to expand, making some resource-constrained devices brings huge challenges. With continued development in the future, only a handful of leading technology companies may be able to afford the huge training costs in the future. Therefore, while maintaining model performance, it is necessary to transform large language models into small language models and use compression model parameter technology to make large models closer to lightweight.

Compression scale parameter technologies currently include Quantization [26], Pruning [27], and Knowledge Distillation [28]. The use of these technologies can simplify the model structure, reduce the amount of calculation and storage requirements, occupy less memory, make reasoning faster, reduce debugging difficulty and increase generalization capabilities.

However, the large scale of large language models also has advantages. Compared with lightweight "small-scale" models, large models can handle more complex tasks, can accommodate

more parameter data, and have strong fitting capabilities. Therefore, in actual production applications, each enterprise and user should choose the model size based on the complexity of the task, the size of the data set, and the actual application scenario.

5.3. Integrate medical equipment

The powerful capabilities of large language models combined with medical equipment can monitor and analyze patients' health conditions in real time, and can effectively provide doctors and patients with current body data in real time and analyze potential health risks. Integrated medical equipment can also allow doctors to understand patients' conditions 24 hours a day and provide remote medical services. This can not only save medical resources, but also provide patients with specialized medical services and obtain a better medical experience. Therefore, in the future development of medical equipment, LLM should be consciously combined with large language models to integrate LLM into the medical system.

6. Conclusion

The article reviews the great value and potential of large language models in medical services and medical research. They can alleviate the problem of tight medical resources, improve medical quality and efficiency, reduce medical costs, and promote medical innovation and progress. At the same time, the challenges that will be encountered and the corresponding reasonable solutions are analyzed, and the future development direction is discussed.

The development of LLM has become a general trend, and the combination of the medical field and artificial intelligence is an imperative step for the medical level to move to a new stage. Large language models have shown great potential in auxiliary treatment, disease diagnosis, clinical decision-making, and education and research and development in the medical field. This not only shows that they can improve the quality of medical services, but also optimize the allocation of medical resources and improve the overall level and efficiency of all aspects of the industry. At the same time, researchers have found that despite the huge potential of LLM development, there are still various challenges in its practical application. LLM cannot completely replace professional medical personnel for the time being. Due to the difficulty of interpreting large language models and its "illusion" problems, patients cannot trust their health conditions to a large model. At the same time, users will also worry about the leakage of privacy.

In the future, relevant scientific researchers will need to continue to improve LLM and devote themselves to solving various challenges encountered by LLM in the medical field, including but not limited to "hallucinations", interpretability, ethics, and privacy issues, and to provide guidance to many Modal and lightweight models are brought closer together to strengthen integration with medical equipment to better serve users.

References

- [1] Zhang, Z. C., Wang, J., Zhang, Y., Tian, Y. L., Zhang, M. M., & Lu, Y. S. et al. (2024). Orthogpt: a multi-modal orthopedic large model for precise diagnosis and treatment. *Journal of Intelligent Science and Technology* (3).
- [2] gChen, Z. J., Peng, W. Q., Zhang, D. Z., Liu, X. & Wang, Z. F. (2024). Application, challenges and prospects of large language models in the field of traditional Chinese medicine. *Journal of Peking Union Medical College* 1-8.
- [3] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- [4] Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.
- [5] Li, W., Yu, L., Wu, M., Liu, J., Hao, M., & Li, Y. (2023, December). DoctorGPT: A Large Language Model with Chinese Medical Question-Answering Capabilities. In *2023 International Conference on High Performance Big Data and Intelligent Systems (HDIS)* (pp. 186-193). IEEE.

- [6] Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., ... & Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature*, 620(7972), 172-180.
- [7] Wang, B., Gao, Z., Lin, Z., & Wang, R. (2023). A disease-prediction protocol integrating triage priority and BERT-based transfer learning for intelligent triage. *Bioengineering*, 10(4), 420.
- [8] Olang, O., Mohseni, S., Shahabinezhad, A., Hamidianshirazi, Y., Goli, A., Abolghasemian, M., ... & Shaker, P. (2024). Artificial Intelligence-Based Models for Prediction of Mortality in ICU Patients: A Scoping Review. *Journal of Intensive Care Medicine*, 08850666241277134.
- [9] Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., ... & Smith, D. M. (2023). Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA internal medicine*, 183(6), 589-596.
- [10] Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., & Liu, T. Y. (2022). BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6), bbac409.
- [11] https://gitcode.com/gh_mirrors/cm/CMLM-ZhongJing. 2024/12/26.
- [12] Zhang, H., Chen, J., Jiang, F., Yu, F., Chen, Z., Li, J., ... & Li, H. (2023). Huatuogpt, towards taming language model to be a doctor. *arXiv preprint arXiv:2305.15075*.
- [13] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140), 1-67.
- [14] Ye, H., Liu, T., Zhang, A., Hua, W., & Jia, W. (2023). Cognitive mirage: A review of hallucinations in large language models. *arXiv preprint arXiv:2309.06794*.
- [15] Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4), e1312.
- [16] Farruque, N., Goebel, R., Zaïane, O. R., & Sivapalan, S. (2021, December). Explainable zero-shot modelling of clinical depression symptoms from text. In *2021 20th IEEE international conference on machine learning and applications (ICMLA)* (pp. 1472-1477). IEEE.
- [17] Alammari, J. (2021, August). Ecco: An open source library for the explainability of transformer language models. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing: System demonstrations* (pp. 249-257).
- [18] Xiao, Y. N. (2023). Ethical risks and governance of artificial intelligence in the digital economy era [J]. *Journal of Shanxi University of Finance and Economics*, 2023, 45(S2):77-79.
- [19] Gao, Y. L., Zhang, R., Li, X. L. (2024). Artificial Intelligence Ethical Computing[J/OL]. *Chinese Science: Information Science*, 1-31. <http://kns.cnki.net/kcms/detail/11.5846.TP.20240712.1253.002.html>.
- [20] Huang, W., Wang, Y., Cheng, A., Zhou, A., Yu, C., & Wang, L. (2024, April). A Fast, Performant, Secure Distributed Training Framework For LLM. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4800-4804). IEEE.
- [21] Li, D., Shao, R., Wang, H., Guo, H., Xing, E. P., & Zhang, H. (2022). Mpcformer: fast, performant and private transformer inference with mpc. *arXiv preprint arXiv:2211.01452*.
- [22] Lu, W. J., Huang, Z., Gu, Z., Li, J., Liu, J., Hong, C., ... & Chen, W. (2023). Bumblebee: Secure two-party inference framework for large transformers. *Cryptology ePrint Archive*.
- [23] Fan, T., Kang, Y., Ma, G., Chen, W., Wei, W., Fan, L., & Yang, Q. (2023). Fate-llm: A industrial grade federated learning framework for large language models. *arXiv preprint arXiv:2310.10049*.
- [24] Wang, H. R., Li, X. H., Li Z., Ma, C. M., Ren, Z. Y., & Yang, D. (2023). Review of multi-modal pre-training models. *Computer Applications*, 43(4), 14.
- [25] Hu, M. F., Zuo, X., & Liu, J. W. (2022). Review of deep generative models. *Acta Automata*, 48(1), 35.
- [26] Tao, C., Hou, L., Zhang, W., Shang, L., Jiang, X., Liu, Q., ... & Wong, N. (2022). Compression of generative pre-trained language models via quantization. *arXiv preprint arXiv:2203.10705*.
- [27] He, Y., Zhang, X., & Sun, J. (2017). Channel pruning for accelerating very deep neural networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 1389-1397).
- [28] Gou, J., Yu, B., Maybank, S. J., & Tao, D. (2021). Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6), 1789-1819.