Application and technology of an open source AI large language model in the medical field

Key points

- Detailed comparison of the advantages and disadvantages of open-source LLM applied in the medical field.

- Listed three technical solutions for deploying large models in medical scenarios.

- Analyzed the potential application scenarios of large models in the medical field.
Application and technology of an open source AI large language model in the medical field

Tairui Zhang, Tianyi Feng

© 2023 The Authors. Attribution-NonCommercial-NoDerivatives 4.0 International

Abstract

To explore the application prospects of an open source artificial intelligence (AI) large language model (LLM) in the medical field, we conducted an analysis from multiple dimensions, including the introduction of LLM, the classification of model types, and the status quo of the open source ecosystem development. The development of an open source LLM is currently in the rapid expansion phase, and there are many types of models and related tools. After analyzing the advantages and disadvantages of the models, we expounded feasible technical solutions for the application of an LLM in the medical field and made corresponding predictions. At present, LLMs in the medical field are still in the early stages, and there are still many problems related to ethics, technology, legal issues, and medical use.

Keywords: open source, large language model, artificial intelligence, medical application

1. INTRODUCTION

Medical and healthcare are important parts of China’s national economy and key industries that protect people's lives and health. Given the spread of the coronavirus and other diseases, numerous countries have encountered many problems, such as a shortage of medical resources and medical personnel. Combining artificial intelligence (AI) with medical care can assist physicians in virus screening [1] and disease diagnosis [2], thereby reducing the misdiagnosis rate and improving the efficiency of diagnosis and treatment. Recently, large language models represented by ChatGPT [3] and GPT-4 [4] have attracted attention from academia and industry. Many Chinese technology companies have also launched large language models (LLMs) to compete internationally at an advanced level. With powerful communication fluency, semantic understanding, inductive reasoning, and other abilities, LLMs have rapidly penetrated all walks of life.

In this context, the combination of LLM and medicine has established a new direction in the medical field. Because ChatGPT, GPT-4, and other models require high computing power and labor costs, many companies and research teams have launched a variety of open source LLMs. Indeed, the initiative promotes the rapid development of LLMs. In the current study we analyzed and discussed the advantages and disadvantages, technical solutions, and application scenarios of open source LLMs in the medical field by sorting out the technological development status, therefore aiming to promote the mutual integration and development of open source models and medicine.

2. AI LLMs

AI LLMs are also referred to as foundation models [5]. AI LLMs are trained on massive, diverse datasets and can handle a variety of downstream tasks [6]. The LLMs have multiple rounds of dialogue and the ability to understand user intentions. The LLMs have better versatility and generalization, which overcomes the traditional model problem of poor versatility.

Transformer was proposed by Vaswani et al. [7] in 2017. With excellent scalability and parallel computing capabilities, Transformer quickly replaced recurrent neural network (RNN) and long short-term memory (LSTM) to become the mainstream architecture in natural language processing (NLP). It has also been extended to the computer vision (CV). It is possible to design and train a model with a parameter scale exceeding 100 billion based on Transformer, and the models have good
generalization. Figure 1 shows the AI LLMs with parameters > 10 billion that have emerged since 2019.

With the release of GPT-3 [9], ChatGPT, and GPT-4, prompt learning [10], instruction learning [11], reinforcement learning from human feedback (RLHF) [12] have become common training methods. Prompt learning unifies downstream tasks into pre-training tasks and converts downstream tasks into natural language with specific templates. Instruction learning can better motivate model comprehension ability compared to prompt learning. Instruction learning uses instructions to guide the model to take the correct action, which makes model generalization ability stronger. RLHF refers to evaluating the output of the models in the form of human feedback and using the feedback as a loss to optimize the model. Indeed, this approach can make the output more innocuous.

LLMs be divided into decoder-only, encoder-only, and decoder-encoder structures [13]. Models with different structures are suitable for different downstream tasks (Table 1). Most of the early LLMs are open source, such as BERT [28], ERNIE [29], T5 [30], and BART [31]. These models use encoder or encoder-decoder as the main structure and have better encoding capabilities. In recent years, GPT-3, ChatGPT, and GPT-4 have adopted the decoder-only structure. Indeed, decoder-only is the most popular structure due to its excellent generation ability. With the high research cost of LLMs, many decoder-only models are not open source.

Although ChatGPT and GPT-4 can be used at no cost to the user, many companies have not announced the implementation details of the models. There are insurmountable technical barriers. It is difficult for individual

Figure 1 | Large models with parameters greater than 10 billion since 2019 [8].

Table 1 | Summary of mainstream large language models.

<table>
<thead>
<tr>
<th>Structure</th>
<th>Publisher</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Encoder-only</td>
<td>Google</td>
<td>BERT, ALBERT [14]</td>
</tr>
<tr>
<td></td>
<td>Baidu</td>
<td>ERNIE, ERNIE2.0 [15]</td>
</tr>
<tr>
<td></td>
<td>Meta</td>
<td>RoBERTa [16]</td>
</tr>
<tr>
<td></td>
<td>Microsoft</td>
<td>DeBERTa [17]</td>
</tr>
<tr>
<td>Encoder-decoder</td>
<td>Google</td>
<td>T5, Flan-T5 [18]</td>
</tr>
<tr>
<td></td>
<td>Tsinghua University</td>
<td>GLM [19], GLM-130B [20]</td>
</tr>
<tr>
<td>Decoder-only</td>
<td>OpenAI</td>
<td>GPT-1 [21], GPT-2 [22], GPT-3, InstructGPT, ChatGPT, GPT-4</td>
</tr>
<tr>
<td></td>
<td>Google</td>
<td>XLNet [23], LaMDA [24], Bard, PaLM [25]</td>
</tr>
<tr>
<td></td>
<td>Meta</td>
<td>LLaMA [26], Galactica [27]</td>
</tr>
</tbody>
</table>
developers, small companies, and research institutions to develop more innovative models, and the technical barriers hinder the promotion and application of LLMs in more fields.

In February 2023, Meta open sourced the LLaMA. The LLaMA derivatives, Alpaca [32] and Vicuna [33], can be trained at a lower cost. These models can even achieve the ability of ChatGPT, which promotes the wave of a LLM open source. At present, a number of open source LLMs for the medical field have been released. BioMedLM is a domain-specific LLM for biomedical text released by the Center for Research on Foundation Models (CRFM) in January 2023. BioMedLM uses a dataset that includes 16 million medical abstracts and 5 million studies. BioMedLM achieved state of the art results on the USMLE medical question and answer test. In April 2023, Tsinghua University open-sourced BioMedGPT [34]. The training data includes multi-scale and cross-modal biomedical data. The BioMedGPT model has the ability to predict drug properties and natural language processing. Visual Med-Alpaca, which was released in April 2023 by the Language Technology Laboratory at the University of Cambridge, recognizes and analyzes chest X-rays, and generates diagnostic conclusions. The research team of health intelligence (HIT) constructed a Chinese medical instruction dataset based on the knowledge map and application programming interface (API) of InstructGPT. The research team of HIT trained HuaTuo, a LLM of intelligent consultation based on LLaMA [35], which overcame the LLM limited language problem in the Chinese context.

Technology open source promotes the rapid development of LLMs in the vertical field, medical models with low deployment costs, high professionalism, and a strong understanding ability. Compared with traditional medical models, the capabilities of the LLMs have improved.

3. SUMMARY OF LLM OPEN SOURCE ECOSYSTEM

The term, open source, was officially proposed by the open system interconnect in 1988. After decades of development, open source has become the main driving force for innovation in emerging technologies. Open source can minimize repetitive labor, save development resources, promote technological breakthroughs, lower development thresholds, and accelerate the promotion and application of new technologies. The term, ecosystem, originated from the field of biology and refers to the natural system formed by organisms and the environment [36]. We believe that the LLM open source ecosystem is centered on open source models, and supported by AI technology, training platforms, and datasets. Together, the open source model support elements constitute a technical ecosystem.

3.1 Classification

The LLMs are classified based on different modalities and fine-tuning methods. We will introduce the development of two new LLMs.

When classified based on modality, open source LLMs can be divided into single modality, bimodal, and multimodal models. Single modality models can only handle NLP, CV, or audio tasks, such as Alpaca, BLOOM [37], ChatGLM, and GPT-2. The language model can be subdivided according to the output or the language, such as the code generation model (StarCoder [38]), the Chinese dialogue model (Chinese-Vicuna), the multilingual dialogue model (ChatGLM-6B), and the medical advice generation models (MedicalGPT-zh and Chat Doctor [39]). The bimodal models can handle two types of data and can be divided into text-to-image (CogView [40] and consistency models [41]), text-image mutual generation (UniDiffuser [42]), image-text matching (BrIVL [43]), text-to-speech (Massively Multilingual Speech [44]), speech-to-text (Whisper [45]), and text-speech mutual generation (AudioGPT [46]) models. The multimodal LLMs can process data involving three or more modalities (text, image, and speech). For example, ImageBind can achieve an arbitrary understanding and conversion between six modalities (text, image, audio, depth, inertial measurement unit and thermal) [47].

Fine-tuning models can also be divided into models that have not been fine-tuned (LLaMA), models that have been fine-tuned by instructions (WizardLM [48], Dolly2.0, and Chinese-LLaMA-Alpaca) and RLHF models (StableVicuna, ChatYuan-large-v2, and OpenAssistant [49]). Fine-tuning refers to initializing the target network with the obtained parameters and training the target network with a dedicated dataset. Instruction tuning uses supervisory signals to guide the model to perform tasks described in the form of instructions so that the models can respond correctly to new tasks. WizardLM-7B uses the evol-instruct to automatically generate open-domain instructions with various levels of difficulty and skill ranges. A part of the WizardLM-7B output content achieves an effect similar to ChatGPT. RLHF relies on manually labeled data and the support of open source frameworks. StableVicuna uses Vicuna as the basic model, follows the three-stage RLHF training proposed by OpenAI, and has the ability to communicate.

In addition to the above types of LLMs, autonomous AI and large language models with plug-in systems are two new types of AI products. Autonomous AI is represented by AutoGPT, AgentGPT, and BabyAGI. This product can use the GPT-4 interface and other models to independently complete tasks given by humans, making up for the GPT-4 shortcomings that cannot be searched online. The NLP Group at Fudan University released the MOSS in April 2023, which can use plug-ins, such as search engines and calculators, to complete specific tasks. The plug-in system makes the models more flexible, enhances the expertise, and improves model interpretability and robustness.

After several years of development, open source LLMs have shown the advantages derived from different types, comprehensive functions, and wide usage...
scenario coverage. Fine-tuning based on the above models has become the most popular method for developing large language models in the medical field. For example, Huatuo, PMC-LLaMA [50], and ChatDoctor are based on LLaMA for fine-tuning, MedicalGPT-zh, DoctorGLM, and ChatGLM-Med are based on ChatGLM, and BioMedLM are based on GPT-2. While open sourcing the model code, most research institutions also provide models with different parameter scales to assist developers in reproducing the model under different hardware resources and publish relevant training data, which lowers the entry threshold for LLMs.

3.2 Open source framework
The open source framework encapsulates the commonly used training paradigms (instruction tuning and RLHF) into services or interfaces, which greatly reduces the amount of manually written code and saves graphics memory. The open source framework decreases the difficulty of training and achieves the unity of high efficiency and economy.

Instruction tuning frameworks include OpenGPT and LMFlow. OpenGPT can create samples based on domain data and the NHS-LLM trained with this framework has achieved more accurate results than ChatGPT based several tests. The RLHF frameworks include triX, DeepSpeed-Chat, ColossalAI, and Lamini. This type of framework realizes the popularization of RLHF training. For example, DeepSpeed-Chat can train a model with more than 13 billion parameters under the support of a single GPU, which enables researchers to create more powerful models under limited conditions. Lamini can package time-consuming and complex fine-tuning as a service.

In addition to optimizing, integrating, and encapsulating the training process of LLMs in the framework, there are also a number of new research projects. The self-instruct released by the University of Washington generates instructions autonomously by the models. This method effectively reduces the cost of manually labeled data and improves the ability of the model to follow instructions [51]. LoRA is a fine-tuning method proposed by Microsoft that can reduce the trainable parameters of the model without sacrificing performance [52]. The Alpaca-Lora uses this method to finetune the LLaMA 7B and achieves the same effect as Alpaca with few training parameters.

With the support of open source frameworks and new methods, the hardware resource requirements and development difficulties have been continuously reduced, and model performance has continued to improve.

3.3 Open source dataset
The capabilities of LLMs arise from datasets. LLM training relies on sufficiently large and complex training data. For example, GPT-1 is trained with BookCorpus (a corpus of unpublished free books by the authors). The model has acquired important world knowledge and the ability to manage long-term dependencies. When institutions open source their models, institutions usually open source the training data as well. For example, CRFM released the self-instruct dataset generated by text-davinci-003 while open sourcing Alpaca. Dataset open sourcing improves the utilization rate of resources and has a positive impact on academic research pertaining to LLMs.

The medical data include clinical datasets (MIMIC-II Clinical Database), doctor-patient dialogue datasets (HealthCareMagic-100k, icliniq-10k, GenMedGPT-5k, and alpaca-52k used in ChatDoctor training), Chinese medical dialogue datasets (data used by DoctorGLM), and self-built datasets consisting of medical paper abstracts and texts, and medical image datasets (DDSM, MIAS, and MURA). Medical open source datasets in the Chinese field are relatively scarce and rely on the instructions generated by ChatGPT, but this method is inaccurate and uncertain. To build a healthy and high-quality Chinese medical open source model field, there is an urgent need to gather real and reliable medical data at a higher level to improve data quality. At the same time, evaluation sets that assess the capabilities of LLMs are also necessary.

In summary, the open-source ecologic development of LLMs is in a rapid growth phase. Various models, frameworks, and methods emerge in an endless stream, which provides a broad range of models and technology selection for the researcher to use. A highly versatile model comparable to GPT-4, however, is lacking. The limitation of a model’s capability remains a problem that cannot be dismissed, and the gap between close and open source models still exists. There is also a lack of a unified framework that simultaneously integrates instruction tuning and RLHF. There are no professional and systematic evaluation indicators in the construction of datasets. An open source ecosystem of LLMs needs to develop in the direction of generalization, specialization, and systematization.

4. OPEN SOURCE LLMS IN THE MEDICAL FIELD
The following will describe the application of open source LLMs in the medical field from three aspects: advantages and disadvantages analysis, feasible technical solutions, and application scenarios.

4.1 Advantages of open source models in the medical field
The advantages of open source LLMs in the medical field can be summarized as low-cost deployment, variety of functions, and diverse interactions.

First, LLMs can be solved in clusters. The WebLLM project move the reasoning process to the client and runs in the browser, which minimizes server overhead and is more friendly to users (i.e., no need to use the complex command to run the model). Localized
development is also more suitable for application scenarios, such as hospitals with limited hardware resources and high data security levels.

Second, the open source LLMs have a wide variety of functions and there are mature open source products in medical image processing and text generation. ChatDocter can conduct consultations in text form and ImpressionGPT can summarize and optimize radiology reports \[53\]. To eliminate the defect of LLM information lag, WebCPM was released by Tsinghua University in May 2023 and can interact with search engines and collect answers \[54\]; the generated content is more real-time.

Third, as the most common application scenario for LLMs, online medical consultation requires the model to have a high level of Chinese dialogue ability. Linly-Chinese-LLAMA, BELLE, Chinese-Vicuna, and Bai Ze are trained on Chinese datasets and have reached a high level in Chinese communication.

### 4.2 Disadvantages of open source models in the medical field

Due to the low-fault tolerance of the healthcare industry, most open source models are trained based on the community open corpus, and the content was not manually corrected. At the same time, open source models are limited by parameter scale and hardware resources. The model may generate biased, toxic, and inaccurate content, which will pose a threat to the safety of patients. Beaver, a highly modular RLHF training framework open sourced by the Peking University team, significantly reduces the biased and discriminatory content output of the model through constrained value alignment (CVA). This type of method is currently still in the development stage. In medical scenarios, physicians are also required to evaluate and give feedback on the professionalism of the output to reduce the errors and inaccurate information.

### 4.3 Feasible technical solutions

Deploying LLMs in medical scenarios can be divided into the following three technical solutions: 1) The capabilities of ChatGPT and GPT-4 should be used to solve professional tasks in the medical field with API. This approach is similar to the AutoGPT and HuggingGPT technical solutions \[55\]. Medical institutions can use the interface provided by LangChain or similar frameworks to effectively utilize the capabilities of multiple models to complete a large number of tasks. This method is easier to develop and easy to deploy. The disadvantage is that frequent usage of the ChatGPT and GPT-4 API may incur a large amount of expenses. Moreover, the degree of customization of the model is low, and the risk of data security is high. 2) Due to the sensitivity of medical data, it is difficult for cloud services to guarantee data security. Medical institutions or teams can rely on open source or medical field datasets to independently develop medical LLMs. The advantage of this solution is that the model fits perfectly with medical purposes and has a high level of customization. The disadvantage is that the independent development of LLMs will consume a lot of manpower and financial costs, which only the top institutions can afford. 3) Pre-training and fine-tuning for the medical use on the open source models is a compromise between the above two solutions. Researchers can choose the more popular decoder-only structure, which has stronger generation capabilities. The steps of pre-training, supervised fine-tuning, and RLHF should be followed. The datasets include open source data, artificially generated data, and self-instruct data. The open source model can be customized and developed at a controllable cost; however, some popular models (LLaMA and Alpaca) do not support commercial use, and it is necessary to avoid the risk of infringement when using models. Currently, the powerful open source models that allow commercial applications include ChatGLM2, Baichuan2, and LLaMA2, each of which has different parameter sizes.

Faced with so many models, frameworks, and technologies, some basic steps for building a medical large model are provided for reference: 1. Clarify the types of model users, including patients, medical institutions, physicians, and medical regulatory departments. 2. Clarify the requirements and objectives based on the first step. For example, the requirements and objectives can be divided according to the breadth of coverage (functional enhancement, process intelligence, and intelligence across multiple processes). The requirements and objectives are also divided into single modal and multimodal requirements. 3. Collect, filter, and standardize training data to form a high-quality supervised fine-tuning (SFT) dataset, including a high-quality physician-patient dialogue dataset, medical knowledge question and answer-related data, and a human preference dataset. The second and third steps interact with each other, such as generally clarifying the requirements of the current stage by considering the available data. 4. Select model versions of different sizes based on the range of data trained by the model and the prepared data, hardware resources, and funding situation among many open source models. 5. Train or fine tune the model. 6. Evaluate the capability of the new model based on the publicly available dataset.

### 4.4 Potential application

The development of medical AI began in 1972 with the AAPHelp system released by the University of Leeds \[56\]. Entering the era of LLMs, the computing power and comprehensive performance of models have been continuously improved and have reached the same level as humans in many fields. LLMs will play an increasingly important role in the medical field. Some typical application scenarios are as follows: 1. In scenario 1, open source LLMs can be used as analytic tools for medical images. Unlike traditional CV models that can only label and recognize images in a single domain, LLMs are more
versatile. The SAM has strong generalization ability and can achieve zero-shot transfer on new tasks [57]. At the same time, the LLMs can also output the disease information of the medical image in the form of text, which can achieve a rapid diagnosis. 2. In scenario 2, open source LLMs can be used as daily medical assistants that provide medical consultation and drug recommendation services for patients. Patients can input their symptoms and medical history into the models, and it can search and summarize based on existing medical knowledge or search engines to form diagnostic recommendations. Finally, based on medical and clinical data, the best treatment drugs are recommended. 3. In scenario 3, using open source LLMs to generate or retrieve medical reports can reduce physician workload. The generation of medical reports is often performed manually by physicians. Because the reports are highly formatted with systematic text, the LLM has a strong ability to generate the reports. 4. In scenario 4, open source LLMs can be applied to clinical research to improve the efficiency of data analysis and problem investigation. Researchers can use autonomous AI products, such as AutoGPT, to complete preliminary research work by independently generating tasks and searching online. For text writing, researchers can use BioGPT [58] or similar tools to complete the classification, summarizing, and text generation. 5. In scenario 5, using a small amount of data or even no labeled data for training will comprise a general medical model capable of various medical tasks in future research. Generalist medical artificial intelligence (GMAI) was proposed by Topol and Rajpurkar in 2022 [59]. Idealized GMAI can be trained on large and diverse datasets, and the model can flexibly handle multimodal tasks. GMAI will have advanced medical reasoning abilities that can support clinical decision-making and generate protein amino acid sequences.

The various capabilities currently displayed by the LLMs have many potential applications in the medical field; however, the risks in multiple dimensions (ethics, harmlessness, and public acceptance) need to be considered. We need more mature technical support to continuously improve the reliability of the models.

5. SUMMARY

As one of the most important technical branches of AI, LLMs have penetrated all aspects of our society in < 1 year, but we need to be cautious when applying this technology in the medical field. The laws and regulations in China related to medical AI have not been well-established and the LLM open source ecosystem is still in the early stage. Medical LLMs should be based on open source products, and continuously deepen the research on the professionalism, humanistic care, and accuracy of model output. We also need to complement the supporting tools and promote the healthy development of this field. In the current study, by sorting out and analyzing the status quo of open source ecosystem development in LLMs, we hope to provide reference for promoting the application of LLMs in the medical field.

ACKNOWLEDGEMENTS

This study did not receive any specific grants from funding agencies in the public, commercial, or non-profit sectors.

CONFLICT OF INTEREST

None.

ABBREVIATIONS

AI, artificial intelligence; LLM, large language model; GPT, generative pre-trained transformer; RNN, recurrent neural network; LSTM, long short-term memory; NLP, natural language processing; CV, computer vision; RLHF, reinforcement learning from human feedback.

REFERENCES


