WEP Warwick Evans Publishing

International Journal of Computer Science and Information Technology

ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 2, Number 3, Year 2024
DOI: https://doi.org/10.62051/ijcsit.v2n3.24
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Large Language Models in the Medical Field: Principles and Applications

Xi Chen

College of Computer Engineering, Jimei University, Xiamen, China

ABSTRACT

Large language models (LLMs) have emerged as powerful tools in various fields, including healthcare. This paper explores the transformative role of LLMs in healthcare quality enhancement, their applications in medical decision-making, and their potential to drive healthcare innovation. Adopting a method of case study, the present study demonstrates how LLMs streamline medical processes, assist in diagnosis and treatment, and enable personalized healthcare solutions. Additionally, the principles of LLMs in medicine were discussed, including pre-training, fine-tuning, and prompt engineering. By leveraging LLMs, healthcare professionals can enhance patient care, optimize workflows, and make more informed decisions, ultimately leading to better healthcare outcomes.

KEYWORDS

Large Language Models; Healthcare; Medical Decision-Making; Healthcare Innovation; Principle

1. INTRODUCTION

With the continuous advancement of large language models, the medical field is undergoing a transformation. These advanced technologies are not only playing a crucial role in patient care and medical processes but also demonstrating significant potential in assisting medical decision-making and improving healthcare efficiency. Particularly in the current landscape of scarce medical resources and high pressure on healthcare professionals, the application of large language models brings new innovative solutions to existing problems faced by the healthcare industry.

This paper aims to explore how large language models are applied in the medical field to drive healthcare innovation. Firstly, it will focus on the transformative application of large language models in improving healthcare quality, optimizing medical processes, assisting medical decision-making, and other areas. Secondly, it will discuss the leadership role of artificial intelligence and large language models in healthcare innovation, especially in the prospects of personalized treatment and medical imaging diagnosis. Finally, it will introduce the basic principles and technical routes of large language models in the medical field, laying the foundation for further exploration of their applications in healthcare.

Content from this work may be used under the terms of CC BY-NC 4.0 licence (https://creativecommons.org/licenses/by-nc/4.0/).
Published by Warwick Evans Publishing.

2. TRANSFORMATIVE ROLES OF LARGE LANGUAGE MODELS IN HEALTHCARE

2.1. Enhanced Healthcare Quality

As large language models progress towards higher sophistication, the medical field emerges as one of the primary beneficiaries. LLMs have shown significant potential in enhancing patient care and simplifying medical procedures. With the increasing public emphasis on life and health, there comes a dramatic surge in the workload for doctors. However, doctors' energy is limited. Large language models can effectively streamline the workflow for doctors. Upon arriving at the hospital, individuals first engage in a conversation with robots trained by large language models, conveying information on symptoms, needs, and concerns. The robot processes this information, and extracts the most concise and intuitive parts to present to the doctor. Consequently, doctors can quickly assess the condition, diagnose the disease, and take the next steps in treatment.

A research team from NYU Langone Health, in collaboration with the NYU Center for Data Science and the NYU Department of Electrical and Computer Engineering, developed an AI-based large language model called NYUTron[1]. This model assists healthcare systems in decision-making processes. Leveraging a powerful large language model, the team inputted various raw materials such as unstructured patient records and diagnostic reports into the model to test its reliability in providing decision support for healthcare practitioners. The research team collected 7.25 million clinical reports from four hospitals under the NYU Langone Health system in New York City, spanning from January 2011 to May 2020 and covering over 380,000 individuals. This formed a corpus containing 4.1 billion words. Additionally, they gathered clinical reports for patients hospitalized within 1-10 years (ranging from 55,791 to 413,845 patients) for specific downstream tasks, constructing a fine-tuning dataset. The application of this model in practice yielded excellent results, saving considerable time and effort.

In the field of oncology, a study utilized LLM models for the identification of four common cancer types: lung cancer, breast cancer, prostate cancer, and colorectal cancer[2]. Results showed that using the XGBoost model in conjunction with SBERT and SimCSE feature extraction, the accuracy for lung cancer was 73% and for breast cancer was 75%. Another study evaluated the effectiveness of large-scale language models in accurately inferring cancer disease response from imaging reports[11]. They collected 10,602 computed tomography scan reports from cancer patients and categorized them into different disease response categories. The GatorTron transformer model exhibited an accuracy of 0.8916 on the test set and 0.8919 on the RECIST validation set. In skeletal imaging, the combination of deep learning models with ChatGPT-3.5 and Python demonstrated an impressive 88.7% accuracy (specificity) and 56.0% sensitivity across all images after 10-fold data augmentation. This figure is remarkable, achieving an accuracy rate of over 70%.

Utilizing a large pre-trained language model to train artificial intelligence in summarizing alert comments and determining if the AI-generated summaries can be used to improve Clinical Decision Support (CDS) alerts[3]. The final experimental results show that artificial intelligence (such as GPT-4) can distill alert comments into summaries with high clarity, accuracy, and completeness. These summaries are known for their speed and accuracy, sometimes even surpassing those generated manually. The AI-generated summaries provide a new method for CDS experts to review user comments rapidly, facilitating the swift optimization of both online and offline CDS alerts. This application demonstrates the use of large language models to summarize clinical alert summaries effectively, thereby aiding healthcare professionals in making faster and more accurate decisions.

2.2. Healthcare Innovation

Presently, the diagnosis and treatment of Parkinson's disease primarily depend on clinical observations by physicians[12], responses to medication, and subjective assessments. However, this

conventional approach is characterized by significant subjectivity and relies heavily on the clinical experience of physicians. Consequently, different physicians may offer varying assessments for the same patient, thereby impacting the accurate evaluation of disease progression and trajectory. Furthermore, patients often encounter challenges in accessing tailored treatments, support, or caregiving. Therefore, employing artificial intelligence to discern Parkinson's disease subtypes holds the potential to offer personalized treatment options for patients of diverse backgrounds, representing a notable advancement in the realm of medicine.

In the field of medical imaging (including pathology, radiology, ultrasound, etc.), the application of artificial intelligence has played a significant role in advancing precise, consistent, and efficient imaging diagnosis and treatment. This can primarily be divided into two aspects: image recognition and diagnostic assistance.

In terms of image recognition, there are often crucial pieces of information in medical images that are vital for disease diagnosis and treatment but are difficult for the human eye to discern. Through artificial intelligence technology, medical images can be analyzed to extract and consolidate potentially indicative information of lesions, thereby achieving accurate identification of affected areas.

As for diagnostic assistance, artificial intelligence utilizes imaging big data and model training, acquiring certain evaluation and diagnostic capabilities. It can provide clinicians with assistance in diagnosis and treatment based on existing data and patterns[6]. Particularly in the diagnosis of pulmonary, ocular, cerebral, neurological, and cardiovascular diseases, artificial intelligence can offer more accurate and rapid diagnostic results, assisting clinicians in making better treatment decisions.

3. PRINCIPLES OF LARGE LANGUAGE MODELS IN MEDICINE

As shown in Figure 1, from the development of large language models in 2019 to the proliferation of various models in 2021, the advancement of large language models has become increasingly comprehensive, with their application domains expanding significantly.

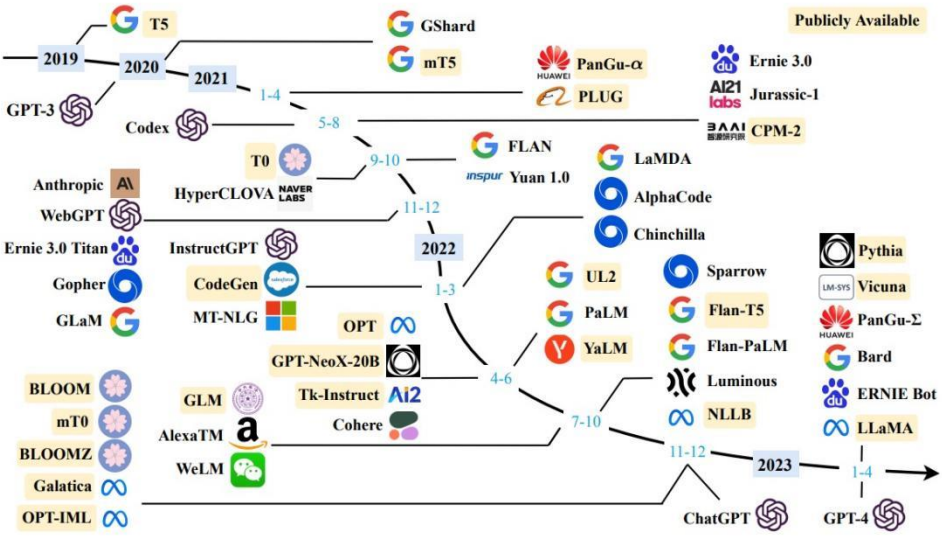

Figure 1. A time map for the development of LLMs (Zhao, et al.,2023)

The existing medical LLMs mainly encompass three types: (1) pre-training from scratch, (2) fine-tuning from existing general LLMs, and (3) aligning a general LLM with the medical domain through prompts to obtain medical-specific capabilities directly[8].

3.1. Pre-training

Pre-training serves as a pivotal stage in the development of medical large-scale language models (LLMs), enabling them to absorb and comprehend extensive medical knowledge. This process entails training the model on vast medical text corpora, encompassing structured and unstructured data like electronic health records (EHRs), clinical notes, DNA sequences, and medical literature. Major medical corpora utilized for pre-training include PubMed, MIMIC-III clinical notes, and PubMed Central (PMC) literature. Models can be pre-trained on a single corpus or a combination of multiple corpora, such as PubMedBERT trained on PubMed and ClinicalBERT on MIMIC-III. BlueBERT integrates both, while BioBERT encompasses PubMed and PMC. Furthermore, pre-training datasets like the University of Florida's Health EHR and clinical practice guidelines (CPGs) for GatorTron and MEDITRON respectively enrich the model's knowledge base.

During pre-training, common objectives from general LLMs are refined, encompassing masked language modeling, next sentence prediction, and next token prediction. BERT-based models like BioBERT, PubMedBERT, ClinicalBERT, and GatorTron primarily focus on masked language modeling and next sentence prediction, while GPT-based models like BioGPT and GatorTronGPT emphasize next token prediction.

Following pre-training, medical LLMs undergo fine-tuning and evaluation on diverse biomedical AI tasks to gauge their comprehension and text generation capabilities. Once pre-trained, these models possess comprehensive language representations applicable to various downstream tasks. Fine-tuning on task-specific datasets further adapts their language representations, ensuring optimal performance. The combination of large-scale pre-training and fine-tuning consistently yields state-of-the-art results.

3.2. Fine-tuning

Training medical LLMs from the ground up demands substantial computational resources and time, often spanning several days or even weeks. To mitigate these challenges, researchers advocate for fine-tuning general language models using medical data. This approach leverages existing resources and accelerates model development. Three prominent fine-tuning methods have emerged: Supervised Fine-Tuning (SFT), Instruction-based Fine-Tuning (IFT), and Parameter-efficient Tuning.

Supervised Fine-Tuning (SFT) extends pre-training data with high-quality medical corpora like medical dialogues, Q&A sets, and knowledge graphs. By integrating these sources, general LLMs acquire specialized medical knowledge, transforming into tailored medical LLMs.

Instruction-based Fine-Tuning (IFT) centers on constructing instruction-based training datasets, typically featuring instruction input-output triplets such as instruction Q&A pairs. Its objective is to enhance the model's capacity to follow various artificial/task instructions, thereby fostering the development of medical LLMs aligned with specific medical contexts.

Parameter-efficient Tuning aims to reduce computational and memory demands by fine-tuning only a minimal subset of parameters (or additional parameters) in LLMs, while maintaining the bulk of parameters unchanged from pre-training. This strategy optimizes efficiency without compromising performance, facilitating the fine-tuning process for medical LLMs.

3.3. Prompt engineering

While fine-tuning reduces computational costs compared to pre-training, it still entails additional model training and the acquisition of high-quality datasets, necessitating computational resources and manual effort. In contrast, the "prompt" approach effectively tailors general LLMs (like PaLM) to specific domains (e.g., MedPaLM) without the need for parameter training. Popular prompt methods include zero/few-shot prompts, chain-of-thought prompts, self-consistency prompts, and prompt tuning.

Zero/few-shot prompts offer direct instructions to LLMs for effective task execution. Zero-shot prompts omit examples, while few-shot prompts provide a small number of task demonstrations before task execution.

Chain-of-thought prompts enhance output accuracy and logic by guiding LLMs to generate intermediate steps or reasoning paths when tackling complex problems.

Self-consistency prompts build upon chain-of-thought prompts to bolster response robustness. They prompt the model to generate multiple answers to the same question and select the most consistent one, enhancing model performance.

Prompt tuning enhances downstream model performance by combining prompt and fine-tuning techniques. Learnable prompts, such as trainable continuous vectors, are optimized during fine-tuning to better adapt to different tasks and scenarios.

Figure 2 illustrates an evolutionary tree of medical large language models, showcasing the advancements made by various companies in this domain. It highlights the dynamic and rapidly progressing landscape of medical large language models.

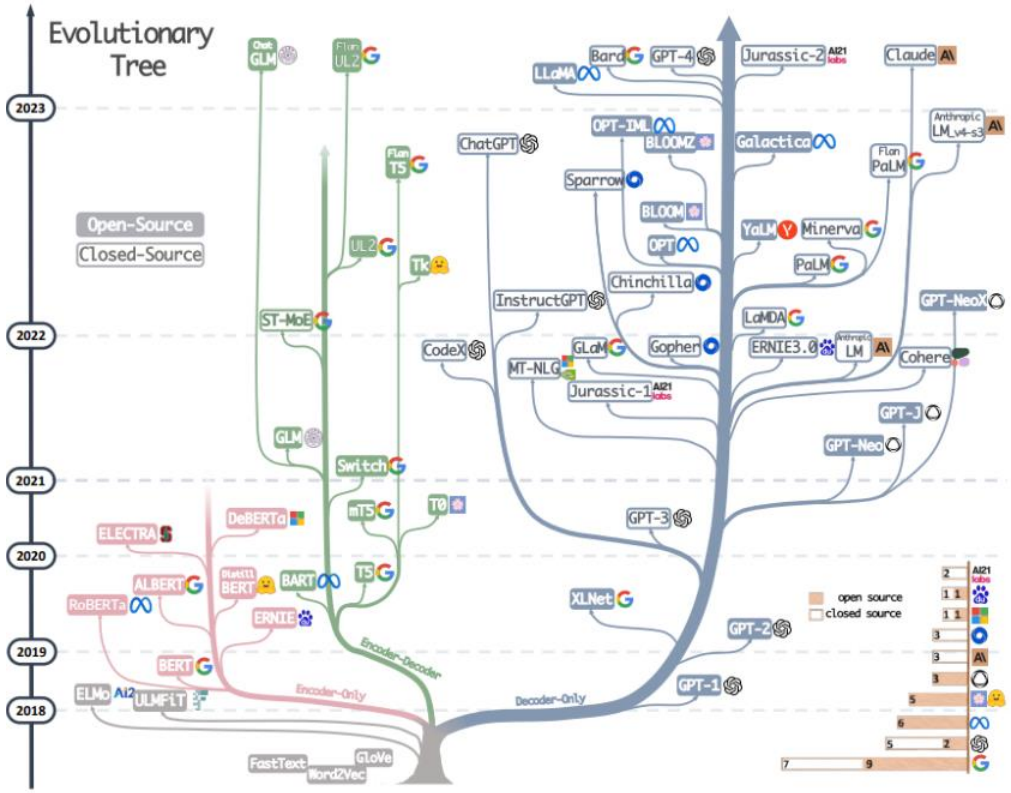

Figure 2. Evolutionary tree (Yang, et al.,2024)

4. SUMMARY

Large language models (LLMs) are revolutionizing healthcare by streamlining processes, aiding diagnosis and treatment, and driving innovation. Through case studies and discussions, this paper highlights how LLMs improve patient care and optimize workflows. Their leadership role in fostering healthcare innovation, such as personalized treatment options and precise diagnostic capabilities, is emphasized. Understanding LLM principles, including pre-training, fine-tuning, and prompt engineering, is crucial for maximizing their potential in healthcare. Overall, integrating LLMs into healthcare systems promises to enhance patient care and drive positive outcomes.

CONFLICTS OF INTEREST

The authors declare that they have no conflict of interest.

REFERENCES

[1] Jiang, L.Y., Liu, X.C., Nejatian, N.P. et al. Health system-scale language models are all-purpose prediction engines. Nature 619, 357–362 (2023). https://doi.org/10.1038/s41586-023-06160-y.

[2] Mumtaz, Ummara, Awais Ahmed, and Summaya Mumtaz. "LLMs-Healthcare: Current applications and challenges of large language models in various medical specialties." Artificial Intelligence in Health 1.2 (2024): 16-28.

[3] Liu, Siru, et al. "Why do users override alerts? Utilizing large language model to summarize comments and optimize clinical decision support." Journal of the American Medical Informatics Association (2024): ocae041.

[4] Kung, Tiffany H., et al. "Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models." PLoS digital health 2.2 (2023): e0000198.

[5] Ayers, John W., et al. "Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum." JAMA internal medicine 183.6 (2023): 589-596.

[6] Wang, Sheng, et al. "Chatcad: Interactive computer-aided diagnosis on medical image using large language models." arXiv preprint arXiv:2302.07257 (2023).

[7] Kim, Kiduk, et al. "Updated Primer on Generative Artificial Intelligence and Large Language Models in Medical Imaging for Medical Professionals." Korean Journal of Radiology 25.3 (2024): 224-242.

[8] Zhou, Hongjian, et al. "A survey of large language models in medicine: Progress, application, and challenge." arXiv preprint arXiv:2311.05112 (2023).

[9] Singhal, Karan, et al. "Towards expert-level medical question answering with large language models." arXiv preprint arXiv:2305.09617 (2023).

[10] Li, Yunxiang, et al. "Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge." Cureus 15.6 (2023).

[11] Tan, Ryan Shea Ying Cong, et al. "Inferring cancer disease response from radiology reports using large language models with data augmentation and prompting." Journal of the American Medical Informatics Association 30.10 (2023): 1657-1664.

[12] Armstrong, Melissa J., and Michael S. Okun. "Diagnosis and treatment of Parkinson disease: a review." Jama 323.6 (2020): 548-560.