Proceedings of CONF-MLA 2025 Symposium: Applied Artificial Intelligence Research
DOI: 10.54254/2755-2721/2025.BJ27258

Decoding Health: Large Language Models Transforming Healthcare

Jiaming Liang

School of Computer Science and Technology, Xidian University, Xi'an, China
3498753515@qq.com

Abstract. With the development of the Transformer architecture and large-scale unsupervised pre-training, large language models (LLMs) have demonstrated unprecedented application potential in healthcare. This article first reviews the key technological evolution of LLMs and their scaling effects. It then, focusing on five representative scenarios—assisted consultation, decision support, personalized treatment, disease prevention and control, and medical research—systematically synthesizes the practical achievements and application outcomes of LLMs. Our comprehensive analysis indicates that LLMs can markedly improve diagnostic and therapeutic efficiency, optimize treatment regimens, and accelerate medical innovation. Nonetheless, significant challenges remain in data security and privacy protection, access to high-quality multidimensional clinical data, model interpretability, and the harmonization of evaluation standards. In response, this paper proposes feasible countermeasures to guide the safe, reliable, and sustainable deployment of LLMs in the medical domain.

Keywords: Large Language Models, Medical Artificial Intelligence, Clinical Decision Support, Personalized Medicine, Multimodal LLMs

1. Introduction

Modern healthcare is confronted with the dual challenges of an exponential increase in medical data and limited resources. How to leverage artificial intelligence to enhance diagnostic efficiency and accuracy has become a critical issue [1]. The rapid development of large language models (LLMs) has opened up new opportunities in this regard. Since the introduction of the Transformer architecture, its self-attention mechanism has significantly improved performance in natural language processing (NLP) tasks and, due to its strong parallelism, has become a core technology for training large-scale medical models [2].

As representatives of pre-trained language models (PLMs), the performance of LLMs is largely influenced by three factors: model parameter size (N), dataset scale (D), and computational resources (C) [1,3]. In recent years, Transformer-based LLMs have expanded rapidly from billions of parameters to hundreds of billions and even trillions. Proprietary models such as GPT-4 have achieved breakthroughs in zero-shot and few-shot learning as well as complex reasoning [4], while open-source models such as LLaMA and BLOOM have substantially lowered research barriers [5,6]. Techniques such as retrieval-augmented generation (RAG) and chain-of-thought prompting (CoT) have greatly strengthened models' abilities in professional knowledge acquisition and reasoning. Meanwhile, efficient fine-tuning methods such as low-rank adaptation (LoRA) have enabled the development of customized medical models [7]. In addition, reinforcement learning with human

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.BJ27258

feedback (RLHF) has been widely adopted in mainstream models, continuously improving their ability to reduce harmful outputs and minimize hallucinations.

Overall, continuous progress in scale, efficiency, and customization has driven the diversified application of LLMs in healthcare. Through unsupervised pre-training on massive datasets such as electronic health records, medical literature, and multimodal data, and further enhanced by strategies such as RAG and CoT, LLMs have demonstrated remarkable potential in medical knowledge understanding and generation. Against this backdrop, this paper systematically reviews representative applications and development trends of LLMs in the medical field, with the aim of providing insights for future research and practical deployment. The main contributions of this study are as follows: 1. For the first time, this paper provides a comprehensive review of studies on representative scenarios including assisted consultation, clinical decision support, personalized treatment, risk assessment and prediction, patient management and monitoring, and medical research. 2. Based on a synthesis of relevant literature, the paper identifies six potential research directions regarding the application of LLMs in healthcare: risks to data privacy and security, ethical concerns, issues of trustworthiness and controllability, high costs of fine-tuning and continual learning, and others.

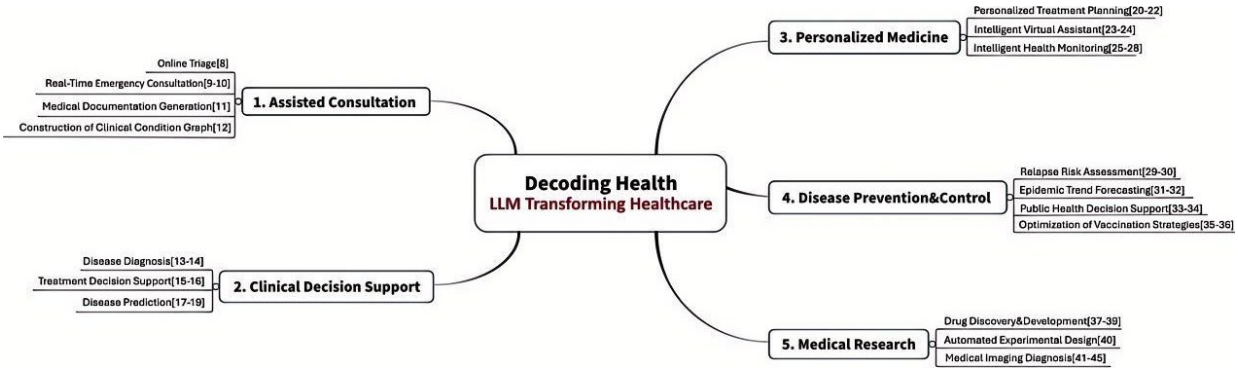

Figure 1. Five major application scenarios for LLM in the medical field

2. Application scenarios of LLMs in healthcare

2.1. Assisted consultation

2.1.1. Online consultation

Traditional online consultations usually depend on search engines. The information on these platforms is scattered and unnecessary, and it does not have the features of personalization and interaction. ChatDoctor is a medical LLM that is built on LLaMA-7B [8]. This model combines the databases of diseases, medical examinations and drugs together. What's more, it is carefully adjusted through the use of about 100,000 real and genuine dialogues between doctors and patients. This adjustment makes the model much better at understanding what patients actually need. It also helps the model give advice that is made for individual patients and offer support for getting information right away in real time. The results from related experiments tell us clearly that the accuracy, recall and F1 score of ChatDoctor are all better than those of ChatGPT. At the same time, Carl Preiksaitis and his research partners did a systematic analysis of 43 different studies [9]. After the analysis, they pointed out that LLMs play an extremely important part in real-time medical triage and emergency medicine (EM). There is a specific example for this point. Saitama Prefecture in Japan created and started using the world's very first emergency consultation system that is based on AI [10]. People who use this system can type in their physical symptoms through a chat window on the platform. Then the AI system does part of the analysis work for the symptoms. It also looks for and gets related

DOI: 10.54254/2755-2721/2025.BJ27258

information from special databases, and with these information, it can give basic medical suggestions or practical guidance to the users. It is worth paying special attention to that the study also brought in a new technique called the mixed-frequency vector autoregression (MF-VAR). This technique can put together data of different frequencies into a single model, and this integration can make the model's ability of prediction and response much stronger. Up to now, this AI-based emergency consultation system has already been used in the field of pediatric care. Its application has made the efficiency of emergency response improve in an effective way, and it has also lessened the work pressure on various healthcare institutions.

2.1.2. Offline consultation

In offline medical scenes, LLMs are mainly used to help doctors gather clinical history information, make preliminary diagnoses for patients and write medical documents. Zhuang Yan and his research partners put forward an intelligent consultation system [11]. This system combines LLMs, digital-human technology and 3D visualization technology together. The structure of this system is designed in the following way: on the patient's side, a multimodal LLM based on Transformer and combined with automatic speech recognition (ASR) can analyze all kinds of input information in real time and do the initial analysis of patients' cases; on the server and doctor's side, retrieval-augmented generation (RAG) and prompt engineering play a guiding role for the model. They help the model pick out key medical entities, and then these entities are arranged into a disease graph by means of dependency parsing, which can give effective help to doctors in the diagnosis work. LLMD is a large-scale medical language model that PicnicHealth has put into use [12]. This model gets continuous pre-training on real electronic health records (EHRs) and common general medical knowledge, and it has more than 350 million labels that are marked by human hands carefully. LLMD takes in RAG and chain-of-thought (CoT) strategies in its operation, and it also uses a multi-layer verification method in the process. All these methods make the output results of the model more reliable. LLMD shows extremely good performance on benchmark tests like PubMedQA. It is better than many large-parameter models in the accuracy of text generation, and it also keeps stable performance in different tasks such as structured information extraction, abstract comprehension and temporal reasoning. What's more, it is worth mentioning that LLMD has outstanding consistency and adaptability when it deals with those long-tail medical concepts. The system applies a dual review framework of "LLM + expert". This framework not only makes the diagnostic efficiency and accuracy get improved, but also reduces the situations of misdiagnosis and missed diagnosis for patients. In the end, it successfully achieves the core goal that "LLMs provide assistance in the treatment process while doctors can focus all their attention on the patients themselves".

2.2. Decision support

2.2.1. Disease diagnosis

Disease diagnosis is one of the most important parts of clinical decision-making. As electronic health records (EHRs) and medical literature increase quickly, it has become a key point to improve diagnostic efficiency by using LLMs to get useful information from unstructured text. General-purpose models often have semantic bias when they are used in medical situations, so recent research has focused on combining LLMs with medical knowledge in a deep way to make them work better in clinical diagnostic tasks. To solve the problem that general language models perform badly on biomedical text, researchers developed BioBERT, a pre-trained language model made specially for biomedical uses [13]. BioBERT is built on the basis of BERT, and it has received pre-training on large-scale biomedical corpora including PubMed and PMC. It did much better than BERT in three key tasks: it raised the performance by 0.62% in named entity recognition, 2.80% in relation

DOI: 10.54254/2755-2721/2025.BJ27258

extraction and 12.24% in QA tasks. These results clearly show that domain-adaptive pre-training works well in helping people understand medical text. By contrast, Med-PaLM and its updated version Med-PaLM 2 are based on the PaLM architecture. They use medical fine-tuning and reasoning-enhancement methods, and they focus on dealing with complex medical QA problems and providing clinical decision support. Med-PaLM 2 reached an accuracy rate of 86.5% on the MedQA dataset, which is more than 19% higher than its earlier version [14]. Its answers were better than those from physicians in 8 clinical aspects, and 90.6% of its answers were judged as low-risk. In real-world QA scenarios, 65% of experts chose to accept its answers more than others. Different from BioBERT, which pays main attention to information extraction work, Med-PaLM is designed for generative responses, clinical reasoning analysis and safety evaluation of medical content.

2.2.2. Treatment decision support

LLMs can help doctors pick out suitable medicines and make practical treatment plans that are made according to the specific health conditions of each patient. By keeping on analyzing the latest medical research and related reports, LLMs can also offer doctors the most up-to-date medical ideas and understandings. In a certain research project, five clinical doctors made an evaluation on two types of suggestions: 36 AI-generated recommendations from seven alerts and 29 suggestions created by humans [15]. When all the suggestions were ranked according to the survey scores, nine of the top 20 suggestions were from ChatGPT. This research result shows that LLMs are able to provide new and valuable medical insights, and these insights are easy for doctors to understand and are also highly useful in actual clinical practice. All over the world, the number of people who go to the emergency department (ED) for medical treatment keeps increasing. The generative AI models which are based on LLMs can do a lot of important work [16]. By analyzing key medical data like patients' physical indicators, medical imaging results and electronic health records (EHRs), these models can find out critically ill patients in an effective way. They can also quickly make personalized treatment plans for patients, and in this way, they can make sure that patients receive timely, accurate and high-quality medical care.

2.2.3. Risk assessment and prediction

In clinical risk assessment and disease prediction tasks, large language models (LLMs) are showing more and more potential. The reason is that they can deal with multimodal information and complex textual information at the same time. In the past few years, a lot of research has looked into how well these models work when they use electronic health record (EHR) data. MedGPT is a predictive model based on Transformer technology. It is combined with MedCAT, and this combination can do named entity recognition and normalization work on more than 15 million EHR documents [17,18]. In the task of "next disease prediction", MedGPT did much better than the traditional LSTM and SVM models. This good performance could be seen in the Top-1, Top-3 and Top-5 accuracy rates of the model. Philip Chung and his research partners did more research to prove a related point. They found that GPT-4 Turbo could get good results in many risk prediction tasks. The model only used preoperative descriptions and clinical notes from EHRs to finish these tasks, which include ASA classification, ICU admission prediction and in-hospital mortality prediction. For the prediction of ICU admission and in-hospital mortality, the model got F1 scores of 0.81 and 0.86 separately [19]. Researchers applied few-shot and CoT prompting strategies to GPT-4 Turbo in the research. After the application, the accuracy and interpretability of the model both got a big improvement. This result also shows that general-purpose LLMs have great potential in perioperative risk assessment. In general, how well LLMs perform in medical risk prediction is getting closer and closer to the performance of traditional methods. It is worth noticing that LLMs do not need to depend a lot on

DOI: 10.54254/2755-2721/2025.BJ27258

large-scale structured data. Even so, they can still produce results with high quality and good interpretability. This good performance of LLMs builds a good base for the development of intelligent clinical decision support systems.

2.3. Personalized medicine

2.3.1. Personalized treatment plans

The main goal of personalized treatment plans is to create treatment strategies made just for each patient's special needs. This kind of plan is especially important for patients who need special medical care or who have cancer diseases. One research project used a large language model (LLM) to analyze 10 simulated medical cases. The study found a key point: LLMs cannot yet make cancer treatment decisions all by themselves, but they can be helpful in other ways. They can quickly make treatment suggestions and find treatment choices that medical experts might miss [20]. But this study had some clear limitations. The number of cases it used was small, it only relied on one expert's judgment, and AI models are developing very fast all the time. Because of these factors, the conclusions of the study could not stay valid for a long time. To fix these problems, another research was carried out later. This new study created MEREDITH, a large language model system that is specially designed for the medical field. MEREDITH includes RAG and CoT mechanisms in its design. It also brings together different kinds of medical information sources, such as medical literature, clinical trials, drug details and treatment guidelines. With these features, MEREDITH can help doctors make personalized treatment decisions for patients [21,22]. This new study also used 10 simulated cases for research, but it added a new important step. It used a multidisciplinary molecular tumor board (MTB) and let the board do a two-stage blind review and give feedback on the research results. This step made the study more scientifically strict and more authoritative. When compared with the earlier study, MEREDITH had obvious advantages. It could effectively stop the appearance of false reference information in its suggestions. Its treatment recommendations also covered a wider range of content and had better semantic similarity with professional medical requirements. What's more, 94.7% of its suggestions were approved by medical experts and could be supported by medical literature that can be checked and verified. All these results show that MEREDITH has real practical potential in the field of precision oncology treatment.

2.3.2. Creating virtual assistants

LLMs have been used widely in medical conversations, but most of the current models ignore the differences between individual patients. They also cannot offer non-stop, personalized health care and companionship for patients [23]. To fill this gap, researchers developed the MaLP framework. This framework adds Dual-Process Enhanced Memory (DPeM) and Parameter-Efficient Fine-Tuning (PEFT) to its design. The DPeM includes both short-term and long-term memory functions. With these designs, the MaLP framework can build smart virtual assistants. These assistants can remember users, understand their needs and develop their service abilities all the time. Test results prove that the MaLP framework works much better than traditional LLMs and models with a single memory structure. It shows better performance in many medical tasks, such as answering medical questions, classifying users' health preferences and creating medical dialogues. These results confirm that the MaLP framework is effective and practical to use as a personalized health care assistant. For future development, combining the MaLP framework with other technologies will bring more possibilities. These technologies include federated learning, knowledge graphs and multimodal sensing. The combination can help develop smart health companion systems that can adjust to different groups of people. Such systems can collect all kinds of health data from users non-stop, including their health behaviors, reactions to medicines and how their diseases develop. At the same time, the systems can

DOI: 10.54254/2755-2721/2025.BJ27258

make dynamic health reports and risk assessments for doctors on a regular basis. In this way, doctors can get timely and accurate information about patients' real-time health conditions. It can also make sure that patients receive long-term medical follow-up and effective communication with medical staff. All these can help make clinical decision-making more scientific and optimize the overall management of patients [24].

2.3.3. Intelligent health monitoring

LLMs have great potential when they are used together with wearable sensors to analyze health data and give personalized feedback to people. PHIA is the first agent system based on LLMs, and it combines the ReAct reasoning framework, code generation function and real-time retrieval tools all in one. This system can read and analyze the complex behavioral health data collected by wearable devices on its own, and then make personalized health suggestions for users [25,26]. It is an innovative system that overcomes the problems LLMs have when dealing with time-series data. But up to now, it has not been tested in clinical environments to prove if it can work well as a health intervention method. PhysioLLM [27] is different from PHIA, for PHIA pays more attention to building agent architecture and integrating toolchains while PhysioLLM puts more effort into helping users know better about their own health data and encouraging them to change their daily behaviors. PhysioLLM uses offline statistical analysis and multi-turn dialogues together, and it makes clear and practical health recommendations according to the data from wearable devices. What's more, it has been proved to be more effective in personalization and behavioral guidance during field experiments. These systems have achieved some initial results in personalized health management, yet the common methods like fine-tuning, RAG and agent frameworks still meet big difficulties when they deal with wearable data, which has the features of high dimensionality, obvious temporal characteristics and heterogeneity. These problems make it hard for LLMs to create health recommendations that are well-structured and helpful for gaining health insights. To solve this problem, researchers have put forward the Graph-Augmented LLMs [28]. This method builds graph structures that cover cross-individual and temporal aspects, and it also combines the modeling of feature importance. In this way, it can provide LLMs with input information that is more related to specific contexts and has a clear structure, which in turn greatly improves how personalized the health insights are and how easy they are to understand.

2.4. Disease prevention and control

LLMs have a key part to play not only in treating diseases but also in stopping diseases from coming back and backing up the strategies for primary disease prevention. In the management of diseases, the chance of a disease recurring has a close link with the risks of treatment plans, the changes patients make to their daily lifestyles and how well patients follow their treatment requirements. For example, Health-LLM, a disease prediction model that brings together language understanding, knowledge retrieval and medical reasoning, has been proved to cut down the risk of disease recurrence in an effective way by doing accurate data analysis and making personalized intervention strategies [29]. The research carried out at present points out that putting LLMs into clinical work processes can not only lower the number of diagnostic mistakes but also make patients follow treatment plans better and help doctors make treatment plans that are more suitable for each individual patient [30]. As we talked about before, virtual assistants powered by LLMs can combine patients' past health data with their present physical condition, and in this way offer non-stop health management and personalized prevention strategies for them. These tools can help patients stay away from disease recurrence and keep their health in a good state for a long time. Also, LLMs have special advantages when it comes to dealing with the complex connections between different kinds of data and the detailed contextual

DOI: 10.54254/2755-2721/2025.BJ27258

information. Their strong abilities of generalization and reasoning let predictive models catch and analyze the trends of epidemics with great accuracy. By collecting and analyzing information from different data sources and making logical judgments based on that, LLMs can make epidemic early-warning systems work better and produce accurate forecasts that are based on useful insights [31,32]. Such predictive abilities make LLMs important tools in disease prevention and epidemic monitoring, for they can not only help find the signs of an epidemic outbreak at an early stage but also give support to public health strategies. In the control of major infectious diseases, LLMs can help the public health departments with resource allocation and emergency response work. For example, they can make supply chains work more efficiently, arrange medical workers in a reasonable way and hand out medical resources properly to slow down and stop the spread of epidemics. Vaccination, as a basic and important intervention measure for disease prevention, has also become a key field where LLMs are applied. In recent practical applications, LLMs have been used in vaccine supply chain management (VSCM). By analyzing the data about vaccine demand and distribution, these systems can make the storage, transportation and allocation of vaccines more optimized. This can help raise the vaccination rates among people in high-risk groups and reduce the spread of diseases [33-36]. All these practical applications show that LLMs have huge potential to push forward the work of disease prevention and control in many aspects.

2.5. Medical research

2.5.1. Drug discovery and development

LLMs offer solid technical support throughout the drug development process, and they work especially well in the two key aspects of literature mining and the design of new molecular structures. Relying on NLP techniques, LLMs can process and summarize huge volumes of biomedical literature in an efficient way. They can help researchers gain in-depth understandings of disease biology, find out potential targets for developing new drugs and keep up with the latest progress in scientific research, and this in turn speeds up the whole process of drug discovery [37]. At the drug discovery stage, LLMs that have been trained with large chemical and biological datasets can give a hand to scientists when they design new drug molecules. The special models which are trained to understand and create molecular SMILES representations make research work much more efficient and also raise the success rates of related studies [38]. A good example is ChemCrow, a system built on GPT-4 that combines 18 professional tools at the expert level. This system has shown outstanding effects in making various chemical tasks automatic, and it has cut down the time and money needed for drug development to a large extent [39]. These LLM models can do a lot of work well: they can make experimental designs more optimized, predict the results of chemical reactions and make plans for synthesis pathways. Besides, they can also analyze electronic health records and the protocols of clinical trials with high efficiency [40]. By making the matching of patients and clinical trials, the design of trials and the prediction of trial results an automatic process, LLMs not only make clinical trials more accurate and efficient but also lighten the work load of researchers. All these new developments show that LLMs will keep playing a bigger and bigger role in drug development, and they will bring brand-new and unmatched opportunities to the research work of modern pharmaceutical industry.

2.5.2. Medical imaging analysis and research

In medical imaging, the advanced frameworks that mix LLMs with deep learning algorithms [41] and computer vision technologies have shown great potential for practical use. These frameworks make use of methods like image segmentation, edge detection and feature extraction. They can find and analyze abnormal parts in medical images with high accuracy, spot possible lesions in the human

DOI: 10.54254/2755-2721/2025.BJ27258

body, and create detailed diagnostic reports for doctors to use in their work [42]. LLaVA-Med is a typical example here—it is a large multimodal language-vision assistant that is specially developed for biomedical uses. By combining natural language processing with visual analysis work, this assistant can understand various complex questions about biomedical imaging and give targeted responses to them. What's more, when LLMs are integrated with computer-aided diagnosis (CAD) systems, the medical reasoning abilities of LLMs and the visual recognition advantages of CAD can work well together. This combination makes the clinical decision support provided for medical staff more effective [43,44]. More sophisticated LLM models can combine different types of data together, including text data, imaging data such as CT scans, and patients' clinical records like genomic data. With these combined data, the models can form a complete framework for assessing patients' diseases. In a new research project, Claude 3.5 Sonnet was used to assess the diagnostic reports written by physicians all on its own. The research results proved that large multimodal models can do better than humans in diagnostic work for some specific medical tasks. It is worth noting that in 85.27% of case analyses, the Llama 3.2-90B model reached a higher diagnostic accuracy than professional physicians [45]. This important research finding further proves that LLMs are effective in the analysis of medical imaging, and it also highlights that LLMs have the potential to take on a key role in clinical practice in the future.

2.6. Discussion

Based on a systematic review of existing literature and representative application scenarios, several limitations remain in current research on the use of LLMs in healthcare:

2.6.1. Data privacy and security risks

Medical data are highly sensitive, encompassing diagnostic information, genetic profiles, and personal identifiers. Improper management of LLMs during training or inference could result in privacy breaches, with risks such as "reverse inference" of patient attributes including gender, age, and geographic location [46]. Furthermore, the lack of transparency in model reasoning processes makes it difficult to trace data flows.

2.6.2. Ethical concerns

LLM training corpora are often biased toward data from developed countries or dominant language groups, resulting in predictive disparities when applied to minority or underserved populations [47]. For example, studies have found that predictions of hospitalization duration and treatment costs vary significantly across different demographic groups, highlighting potential risks of inequitable healthcare delivery.

2.6.3. Limited trustworthiness and controllability

Most existing LLMs function as black-box models with limited interpretability. When hallucinations or erroneous recommendations occur, clinicians cannot directly intervene to correct them, which may lead to misdiagnosis or inappropriate prescriptions, raising concerns about patient safety and liability. The opacity of model outputs also undermines clinician trust in AI-generated recommendations.

2.6.4. High cost of fine-tuning and continuous learning

To achieve domain-specific deployment in healthcare, the costs of fine-tuning and continual learning must be addressed. While low-rank adaptation (LoRA) techniques can reduce trainable parameters by approximately 10,000-fold and decrease memory requirements threefold [7], each new task still

DOI: 10.54254/2755-2721/2025.BJ27258

requires separate training and storage of adapter modules, resulting in significant computational and storage overhead. Moreover, updating models with continuously emerging clinical data is challenged by LLM context length limitations, often necessitating replay strategies or regularization methods, which further increase training time and maintenance complexity.

2.6.5. Scarcity of high-quality data and limitations in multimodal processing

Comprehensive medical data are fragmented across institutions and restricted by privacy regulations. Electronic health records often contain unstructured text, handwritten notes, and heterogeneous report formats, lacking standardized structures. These barriers make it difficult to assemble sufficiently large and well-annotated real-world clinical datasets. Although multimodal models such as *LLaVA-Med* and *PhysioLLM* have been explored, integrating high-dimensional medical imaging (e.g., CT, MRI) and continuous monitoring data (e.g., ECG, glucose levels) remains challenging. Such integration often requires external feature extractors or dimensionality reduction techniques, which risk losing critical information, reducing real-time applicability, and limiting clinical interpretability.

2.6.6. Lack of unified evaluation standards

Most studies rely on relatively small benchmarks such as *MedQA* and *PubMedQA*, or use self-constructed datasets, without large-scale, standardized, and comprehensive evaluation frameworks to validate the true clinical utility of research outcomes.

These shortcomings reveal that LLMs in the medical field are still at the stage of validation and pilot testing, and have not yet formed a mature system that can be deployed on a large scale. To enhance the efficiency, quality, and accessibility of healthcare when transitioning LLMs from laboratory research to large-scale application, the following technical approaches and measures can be taken. Build a risk control platform that incorporates multiple LLM safety evaluation methods, strengthen real-time monitoring and auditing, especially for promptly capturing and intercepting potential hallucinations or harmful outputs; re-evaluate the impact of common datasets on LLM output bias, taking into account data collection channels, demographic distribution, and annotation quality; adopt generative models such as GANs and VAEs to enrich high-quality medical data, with a particular focus on information extraction and heterogeneous integration of multimodal data; improve laws and regulations that safeguard patient data privacy and security, and conduct regular safety evaluations and updates of LLMs to ensure that the models remain secure, controllable, compliant, and reliable, while being able to respond promptly to evolving regulatory environments and societal expectations.

3. Conclusion

This paper provides a comprehensive review of the research progress, current applications, and challenges of large language models (LLMs) in the medical domain. Building on an overview of the principles and current state of LLM applications in healthcare, we examined in detail their specific roles in assisted consultation, clinical decision support, personalized treatment, disease prevention and control, and medical research. By systematically summarizing representative case studies and practical implementations, this review highlights the significant advantages of LLMs in improving diagnostic efficiency, optimizing treatment strategies, and accelerating medical innovation. At the same time, it also underscores the key challenges that remain, including data security risks, scarcity of high-quality datasets, and limitations in multimodal data processing.

DOI: 10.54254/2755-2721/2025.BJ27258

References

[1] Cascella, Marco, et al. "The breakthrough of large language models release for medical applications: 1-year timeline and perspectives." Journal of Medical Systems 48.1 (2024): 22.
[2] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).
[3] Kaplan, Jared, et al. "Scaling laws for neural language models." arXiv preprint arXiv:2001.08361 (2020).
[4] Brown, Tom, et al. "Language models are few-shot learners." Advances in neural information processing systems 33 (2020): 1877-1901.
[5] Touvron, Hugo, et al. "Llama: Open and efficient foundation language models." arXiv preprint arXiv:2302.13971 (2023).
[6] Workshop, BigScience, et al. "Bloom: A 176b-parameter open-access multilingual language model." arXiv preprint arXiv:2211.05100 (2022).
[7] Hu, Edward J., et al. "Lora: Low-rank adaptation of large language models." ICLR 1.2 (2022): 3.
[8] 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).
[9] Preiksaitis, Carl, et al. "The role of large language models in transforming emergency medicine: scoping review." JMIR medical informatics 12 (2024): e53787.
[10] Amagasa, Shunsuke, and Takashi Moriya. "Emergency consultation system with automatic response software using artificial intelligence." Health Policy and Technology 11.3 (2022): 100629.
[11] Zhuang, Y., et al. (n.d.). Research on the construction of an intelligent medical consultation assistant based on large medical models. Journal of PLA Medical College.
[12] Porter, Robert, et al. "LLMD: A Large Language Model for Interpreting Longitudinal Medical Records." arXiv preprint arXiv:2410.12860 (2024).
[13] Lee, Jinhyuk, et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining." Bioinformatics 36.4 (2020): 1234-1240.
[14] Singhal, Karan, et al. "Toward expert-level medical question answering with large language models." Nature Medicine 31.3 (2025): 943-950.
[15] Liu, Siru, et al. "Using AI-generated suggestions from ChatGPT to optimize clinical decision support." Journal of the American Medical Informatics Association 30.7 (2023): 1237-1245.
[16] Arslan, B., et al. "Evaluating LLM-based generative AI tools in emergency triage: A comparative study of ChatGPT Plus, Copilot Pro, and triage nurses." The American Journal of Emergency Medicine 89 (2025): 174-181.
[17] Kraljevic, Zeljko, et al. "MedGPT: Medical concept prediction from clinical narratives." arXiv preprint arXiv:2107.03134 (2021).
[18] Kraljevic, Zeljko, et al. "Multi-domain clinical natural language processing with MedCAT: the medical concept annotation toolkit." Artificial intelligence in medicine 117 (2021): 102083.
[19] Chung, Philip, et al. "Large language model capabilities in perioperative risk prediction and prognostication." JAMA surgery 159.8 (2024): 928-937.
[20] Benary, Manuela, et al. "Leveraging large language models for decision support in personalized oncology." JAMA Network Open 6.11 (2023): e2343689-e2343689.
[21] Lammert, Jacqueline, et al. "Expert-guided large language models for clinical decision support in precision oncology." JCO precision oncology 8 (2024): e2400478.
[22] Ng, Karen Ka Yan, Izuki Matsuba, and Peter Chengming Zhang. "RAG in health care: a novel framework for improving communication and decision-making by addressing LLM limitations." Nejm Ai 2.1 (2025): AIra2400380.
[23] Zhang, Kai, et al. "Llm-based medical assistant personalization with short-and long-term memory coordination." arXiv preprint arXiv:2309.11696 (2023).
[24] ZHANG, Xuan, et al. "Application and Challenges of Intelligent Robots in Grassroots Chronic Disease Management." Chinese General Practice 28.01 (2025): 7.
[25] Merrill, Mike A., et al. "Transforming wearable data into health insights using large language model agents." arXiv preprint arXiv:2406.06464 (2024).
[26] Aksitov, Renat, et al. "Rest meets react: Self-improvement for multi-step reasoning llm agent." arXiv preprint arXiv:2312.10003 (2023).
[27] Fang, Cathy Mengying, et al. "Physiollm: Supporting personalized health insights with wearables and large language models." 2024 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI). IEEE, 2024.
[28] Subramanian, Ajan, et al. "Graph-augmented llms for personalized health insights: A case study in sleep analysis." 2024 IEEE 20th International Conference on Body Sensor Networks (BSN). IEEE, 2024.

DOI: 10.54254/2755-2721/2025.BJ27258

[29] Kim, Yubin, et al. "Health-llm: Large language models for health prediction via wearable sensor data." arXiv preprint arXiv:2401.06866 (2024).
[30] Suraj, Timothy. "The Potential of Large Language Models in Treating Recurrent Medical Conditions through Enhanced History-Taking." Available at SSRN 5160698 (2025).
[31] Su, Jing, et al. "Large language models for forecasting and anomaly detection: A systematic literature review." arXiv preprint arXiv:2402.10350 (2024).
[32] Shah, Chaitya, et al. "Infectious Disease Forecasting in India using LLM's and Deep Learning." arXiv preprint arXiv:2410.20168 (2024).
[33] Duijzer, Lotty Evertje, Willem Van Jaarsveld, and Rommert Dekker. "Literature review: The vaccine supply chain." European Journal of Operational Research 268.1 (2018): 174-192.
[34] De Boeck, Kim, Catherine Decouttere, and Nico Vandaele. "Vaccine distribution chains in low-and middle-income countries: A literature review." Omega 97 (2020): 102097.
[35] Joshi, Sudhanshu, and Manu Sharma. "A literature survey on vaccine supply chain management amidst COVID-19: literature developments, future directions and open challenges for public health." World 3.4 (2022): 876-903.
[36] Sharma, Manu, et al. "Developing a framework for enhancing survivability of sustainable supply chains during and post-COVID-19 pandemic." International Journal of Logistics Research and Applications 25.4-5 (2022): 433-453.
[37] Arif, Taha Bin, Uzair Munaf, and Ibtehaj Ul-Haque. "The future of medical education and research: Is ChatGPT a blessing or blight in disguise?." Medical education online 28.1 (2023): 2181052.
[38] Honda, Shion, Shoi Shi, and Hiroki R. Ueda. "Smiles transformer: Pre-trained molecular fingerprint for low data drug discovery." arXiv preprint arXiv:1911.04738 (2019).
[39] M. Bran, Andres, et al. "Augmenting large language models with chemistry tools." Nature Machine Intelligence 6.5 (2024): 525-535.
[40] Wornow, Michael, et al. "Zero-shot clinical trial patient matching with LLMs." NEJM AI 2.1 (2025): AIcs2400360.
[41] Ker, Justin, et al. "Deep learning applications in medical image analysis." Ieee Access 6 (2017): 9375-9389.
[42] Chen, Xuxin, et al. "Recent advances and clinical applications of deep learning in medical image analysis." Medical image analysis 79 (2022): 102444.
[43] Li, Chunyuan, et al. "Llava-med: Training a large language-and-vision assistant for biomedicine in one day." Advances in Neural Information Processing Systems 36 (2023): 28541-28564.
[44] Wang, Sheng, et al. "Interactive computer-aided diagnosis on medical image using large language models." Communications Engineering 3.1 (2024): 133.
[45] Ruan, Cailian, Chengyue Huang, and Yahe Yang. "Comprehensive evaluation of multimodal ai models in medical imaging diagnosis: From data augmentation to preference-based comparison." arXiv preprint arXiv:2412.05536 (2024).
[46] Staab, Robin, et al. "Beyond memorization: Violating privacy via inference with large language models." arXiv preprint arXiv:2310.07298 (2023).
[47] Suenghataiphorn, Thanathip, et al. "Bias in Large Language Models Across Clinical Applications: A Systematic Review." arXiv preprint arXiv:2504.02917 (2025).