Large Language Models in Healthcare: A Review

2023 7th International Symposium on Computer Science and Intelligent Control (ISCSIC) | 979-8-3503-4298-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISCSIC60498.2023.00038

Shun Zou*
College of Information and Communication
National University of Defense Technology
Wuhan, China
zoushun@nudt.edu.cn

Jun He
College of Information and Communication
National University of Defense Technology
Wuhan, China
hejun_nudt@nudt.edu.cn

***Abstract*— This paper examines the potential of large language models (LLMs) in the healthcare sector, delving into their prospective applications, challenges, and future trajectories. LLMs have demonstrated encouraging results in various healthcare-related domains, including the development of clinical decision support systems, natural language processing in electronic health records, healthcare question/answer systems, and healthcare education. However, integrating these models into healthcare practice also raises several concerns, such as data privacy and security issues, the requirement for vast amounts of training data, model biases, and the limited interpretability of model predictions. Overcoming these hurdles necessitates a collaborative effort from experts across multiple disciplines. Despite these obstacles, the deployment of LLMs in healthcare holds the potential to transform the industry and significantly enhance patient outcomes.**

Keywords—Large Language Models, ChatGPT, Applications in Healthcare;

I. INTRODUCTION

The rise of large language models (LLMs) such as GPT [1] [2] , Llama[3] , Alpaca[4] , and others has revolutionized various sectors, including healthcare. These models, powered by machine learning and artificial intelligence, have the potential to reshape healthcare delivery, research, and education.

Large language models have been instrumental in the development and application of Clinical Decision Support Systems (CDSS). These systems leverage the power of these models to analyze vast amounts of data and provide evidence-based treatment recommendations [5] [6] [7] [8] [9] . In addition to CDSS, large language models have found application in Natural Language Processing (NLP) in Electronic Health Records (EHR). EHRs contain tons of valuable patient information, but much of this data is unstructured and difficult to analyze. Large language models can process this data, extract relevant information, and present it in a structured format that is easy to understand and analyze [10] [11] [12] [13] [14] .

Large language models are also being used in healthcare Q/A systems. These systems can answer patient queries, provide information about diseases and treatments, and even guide patients through the healthcare system [15] [16] [17] [18] .In the field of healthcare education, large language models can be used to develop intelligent tutoring systems. These systems can provide personalized learning experiences, adapt to the learning pace of individual students, and provide instant feedback[19] [20] [21] [22] . Large language models also have great potential to revolutionize medical research and drug discovery. They enables rapid in-depth analysis of medical record data to identify typical patterns or trends, and generate hypotheses. This can accelerate the pace of medical research and lead to the discovery of new treatments and cures [23] [24] [25] .

Despite these potential benefits, there are also several challenges related to the implementation of large language models in healthcare. These include issues related to data privacy and security, the requirement for large training dataset, the risk of model bias, and insufficient interpretability of model results. Moreover, the successful implementation of these models requires a multidisciplinary collaboration, involving not only computer scientists and engineers, but also doctors, patients, and policymakers [6] .

Large language models have the potential to reshape healthcare, but their implementation requires careful consideration of both the advantages and challenges. In following sections, an overall review about large language models in healthcare will be made from several aspects: the applications of LLMs in healthcare, the challenges of using LLMs in healthcare, and the future directions. This will give us a more comprehensive understanding of the current development, application, limitations, and future direction about LLMs in healthcare.

II. APPLICATIONS OF LARGE LANGUAGE MODELS IN HEALTHCARE

The application of LLMs in healthcare has been a topic of interest in recent years. These models have the potential to revolutionize various aspects of healthcare, including clinical decision support systems, natural language processing in electronic health records, healthcare Q/A systems, healthcare education, and medical research and drug discovery.

A. Clinical Decision Support Systems

Clinical Decision Support Systems (CDSS) have long been at the forefront of integrating technology into healthcare to improve patient outcomes and healthcare efficiency. With the rise of LLMs like GPT-3 and its successors or variations, there has been a paradigm shift in how these systems assist healthcare professionals in making clinical decisions. Large language models, trained on large amounts of corpus, have the capability to generate human-like answer based on their received input. This ability can be harnessed in a CDSS to provide real-time, evidence-based recommendations to healthcare professionals. For instance, when a physician inputs patient symptoms and medical history, the LLM can quickly scan through the latest medical literature and provide potential diagnoses,

979-8-3503-4298-7/23/$31.00 ©2023 IEEE
DOI 10.1109/ISCSIC60498.2023.00038

treatment options, and even predict patient outcomes based on similar cases [5] .

The potential impact of integrating LLMs into CDSS on patient outcomes cannot be understated. Compared with human beings, large language models will not be affected by experience, emotion, and physical strength, and will only consider practical factors, therefore these advanced systems can reduce diagnostic errors, especially in complex cases where symptoms might be ambiguous [6] .Moreover, they can assist in personalizing treatment plans by considering a broader range of factors than a human could feasibly analyze in a short time. Furthermore, the efficiency of healthcare delivery can be significantly improved. For example, in emergency situations where time is of the essence, an LLM-enhanced CDSS can provide rapid insights, helping medical professionals make quicker decisions.

A two-stage recommendation framework for assisted medical decision making in was proposed in study [7] based on the publicly available MIMIC dataset, where a pre-trained language model was used to extract relevant and useful information from the clinical records. Their findings indicated LLMs' potential to enhance clinical decision-making capabilities and reduce the burden of information processing. In study [8] , ChatGPT demonstrated strong performanceassessing its suitability for radiologic decision making. For breast cancer screening prompts, the model averaged an OE score of 1.83 out of 2 and a SATA percentage correct of 88.9%. Similarly, for breast pain prompts, it achieved an average OE score of 1.125 out of 2 and a SATA percentage correct of 58.3%. These results suggest that ChatGPT has the potential to aid radiologists in making accurate decisions. In study [8] , Clinicians evaluated both human-generated and AI-generated suggestions for enhancing clinical decision support (CDS) alerts. Interestingly, among the top 20 rated suggestions, 9 were created by ChatGPT. This outcome indicates that AI-generated ideas can contribute to optimizing CDS alerts and support experts in developing their own recommendations for CDS enhancement.

However, it's crucial to note that while LLMs offer immense potential, they are not without limitations. Their recommendations are based on the data they were trained on, and they lack the intuitive reasoning that human professionals possess. Thus, while they can be a valuable and powerful tool, they should be utilized judiciously alongside human intuition and proficiency.

B. Natural Language Processing in Electronic Health Records

Electronic Health Records (EHRs) offer a rich source of data regarding patients' medical backgrounds, treatment paths, and health outcomes. However, the majority of this information is stored in unstructured text formats, making it difficult to analyze and extract meaningful insights manually. Natural Language Processing (NLP) techniques, especially LLMs, offer a solution to this problem by enabling the automatic processing and analysis of EHR data. These models can analyze unstructured data in electronic health records, such as clinical notes, and extract relevant information. This can help healthcare professionals understand a patient's medical history and current health status more quickly and accurately, facilitating the creation of large datasets for analysis and research [11] .By automate coding, information extraction, and data mining, these models can significantly reduce the workload of healthcare professionals and improving the efficiency of healthcare services [10] .

LLMs have been proven to be a powerful tool for processing EHRs. In study [12] , a novel clinical language model called GatorTron was developed and assessed on five distinct clinical natural language processing (NLP) tasks. These tasks included extracting clinical concepts, identifying medical relations, determining semantic textual similarity, inferring natural language (NLI), and responding to medical queries (MQA). The notable improvements in performance across all five tasks suggest that GatorTron can be effectively integrated into medical AI systems, ultimately enhancing healthcare provision. In study [13] , a med-7 model was trained for named-entity recognition, encompassing drug names, route of administration, frequency, dosage, strength, form, and duration. With a micro-averaged F1 score of 0.957 across all seven categories, the model demonstrated its potential for identifying medical concepts and extracting relevant information. In study [14] , a cancer domain-specific language model, CancerBERT, was developed to extract breast cancer phenotypes from electronic health records. The model outperformed all other models on this task, suggesting its potential to support clinical decision-making.

In conclusion, NLP, particularly large language models, holds great potential in enhancing the use of EHRs. Through automated coding, information extraction, and data mining, it can streamline healthcare processes and contribute to research and policy-making. As the growing of the volume and complexity of EHRs, and the boosting of the capability and applications of LLMs, the importance of LLMs in healthcare will only increase. Therefore, further investment in LLMs research and development is warranted to fully realize its potential in improving patient care and health outcomes.

C. Healthcare Q/A system

Healthcare Q/A systems aim to provide patients and healthcare professionals with quick and accurate answers to their medical queries. Large language models can play a key role in future healthcare Q/A systems. These models can understand and answer complex medical questions, thereby improving the efficiency and accuracy of healthcare services. Unlike traditional systems that often rely on hardcoded knowledge bases, LLMs are trained on vast amounts of data, enabling them to generate human-like responses in real-time and handle a wide array of medical inquiries. These models can swiftly provide detailed explanations or recommend further reading based on the context of the question, potentially streamlining the information retrieval process for medical professionals.

The accuracy of these models, especially when combined with domain-specific data, can rival or even surpass that of manual searches. Large language models (LLMs) have significantly advanced medical question answering, with Med-PaLM achieving a milestone score of 67.2% on the MedQA dataset, surpassing the "passing" mark in US Medical Licensing Examination (USMLE) style questions. Its successor, Med-PaLM 2, has shown even greater promise, scoring up to 86.5%. Notably, a

comprehensive human evaluation of 1066 consumer medical questions revealed that Med-PaLM 2's answers were preferred over those provided by physicians on eight of nine criteria pertinent to clinical utility, highlighting its potential for practical application in medicine. [17] . In another study, the Codex model with 175B parameters demonstrated an answering accuracy of 60.2% on the USMLE test, 59.7% on the MedMCQA validation, and 78.2% on the PubMedQA test [18] . These results are comparable to the performance of human experts.

Furthermore, the implementation of LLMs in healthcare Q/A systems can lead to significant cost reductions. The traditional model, involving manual searches or consultations with specialists, can be time-consuming and expensive. In contrast, once an LLM is trained, it can reduce the need for human intervention, thereby its marginal cost for answering additional questions is almost negligible.

D. Healthcare education

Healthcare education is another area where large language models can have a significant impact. In healthcare education, large language models can be leveraged to create customized learning experiences for students. By analyzing a student's individual learning style, these models can tailor the educational content to better meet their needs, providing a more effective and engaging learning experience. Moreover, these models can also provide instant feedback, identifying important concepts, highlighting knowledge gaps, thereby helping students improve their learning outcomes [20] .

In a recent study [22] , the capabilities of the large language model ChatGPT were assessed on the United States Medical Licensing Exam (USMLE). The pretrained ChatGPT delivered impressive results, performing at or near the passing threshold for all three exams and showcasing a high degree of consistency and depth in its explanations. These findings hint at the possibility of ChatGPT serving as a valuable tool in medical education. Another study discussed the potential use of large language models in healthcare education [21] . LLMs could analyze medical texts and generate quizzes based on the most important concepts. LLMs could also generate a virtual patient simulation, Traditional medical simulations often rely on mannequins or actors to mimic patient scenarios. With the advent of LLMs, these simulations can be enhanced with realistic patient dialogues, allowing medical students to practice their communication skills in a controlled environment.

E. Medical research and drug discovery

LLMs can play a crucial role in medical research, particularly in drug discovery. LLMs have shown promise in analyzing vast amounts of literature and data to identify potential drug candidates. The traditional drug discovery process is time-consuming and resource-intensive, often taking years and significant financial investments to bring a drug from concept to market. However, with the computational capabilities of LLMs, researchers can expedite the initial stages of drug discovery by rapidly sifting through existing literature to identify potential drug candidates or mechanisms [23] . This can significantly speed up the drug discovery process and reduce the cost of drug development.

For example, researchers have explored the use of large language models (LLMs) in developing new treatments for COVID-19, including drug repurposing, which could inform prevention strategies for future pandemics [24] . Additionally, customized domain-specific LLMs, such as chemical language models, have shown promising results in accelerating de novo drug design by generating new molecules with desired properties [25] .

In conclusion,, LLMs hold great potential in transforming various areas of healthcare, though further investigation is necessary to thoroughly understand their advantages and limitations in this field.

III. Challenges of Using LLM in Healthcare

The use of large language models (LLMs) in healthcare has shown great promise in various applications, including clinical text analysis, patient communication, and medical decision-making. However, there are several challenges associated with the adoption of LLMs in healthcare, which must be addressed to fully realize their potential.

A. Need for Large Amounts of Training Data

One of the primary challenges in developing and deploying LLMs in healthcare is access to large amounts of high-quality training data. Deep learning models require vast amounts of data to learn patterns and relationships within the data, and healthcare is no exception. The larger the training dataset, the better the model's ability to understand context, make predictions, and provide relevant information. Pre-trained models usually performs poorly on specific problems, which makes it essential to create domain-specific datasets. However, obtaining such data can be difficult due to privacy concerns, data fragmentation, and the time-consuming process of manual annotation.

Moreover, collecting data from diverse sources, such as electronic health records (EHRs), clinical notes, and medical imaging reports, poses additional challenges in terms of data heterogeneity and quality variability. Medical data can vary in format, structure, and language, especially when sourced from different healthcare systems or countries. This diversity requires additional preprocessing and standardization efforts before it can be used for training LLMs.

B. Lack of Interpretability of Model Predictions

Another significant challenge in using LLMs in healthcare is the lack of interpretability of model predictions. Unlike traditional machine learning models, deep learning models are often criticized for their opaqueness, making it difficult to understand the reasoning behind their predictions. In healthcare, where decisions can have life-altering consequences, understanding the rationale behind a model's prediction is crucial. When an LLM suggests a particular diagnosis or treatment plan, clinicians need to understand the basis for that recommendation to trust and act upon it.

The lack of transparency in large language models (LLMs) poses a significant challenge in healthcare, where understanding the thought process behind a recommendation or suggestion is critical. The so-called "black box" nature of these models makes it difficult to comprehend how they reach their conclusions, leading to mistrust among healthcare professionals who may be

reluctant to rely on a tool that doesn't offer clear reasoning for its decisions. Furthermore, the absence of interpretability hinders the detection and correction of potential biases within the model, which can result in inaccurate recommendations and further undermine confidence in the technology. To tackle this issue, researchers have put forth various methods, such as attention mechanisms, feature importance scores, and visualizations, aimed at offering insights into the decision-making process of LLMs. Nevertheless, the ability to interpret model predictions remains an essential problem that requires resolution.

C. Ethical Implications

A key ethical consideration regarding the utilization of large language models in healthcare revolves around privacy and data security. These models necessitate extensive sets of personal health data for training, giving rise to worries about patient confidentiality and data protection. The integration of artificial intelligence in healthcare presents substantial hurdles in safeguarding patients' privacy. Hence, it becomes imperative to establish and uphold adequate measures to protect patient data and uphold confidentiality standards.

Another ethical aspect to consider involves the risk of bias influencing the decision-making process. If large language models are trained on biased data or influenced by a specific worldview, they can inadvertently perpetuate existing health disparities. For instance, a language model created for predicting patient mortality may exhibit bias against certain racial groups. To tackle this challenge, it is essential to prioritize the use of diverse and representative datasets when training these models.

D. Legal Considerations

The use of large language models in healthcare also raises several legal considerations. One of the most pressing issues is liability and accountability. Who is responsible when an AI system makes a mistake or provides suboptimal advice? Is it the developer, the clinician who used the system, or the patient who consented to its use? These questions highlight the need for clearer regulations and guidelines regarding the use of AI in healthcare.

Another legal concern relates to intellectual property rights. Who owns the intellectual property rights to the data used to train these models? Is it the patients who provided the data, the clinicians who collected it, or the companies that developed the models? Clarifying ownership rights is essential to prevent disputes and ensure that data are used responsibly.

E. Biases and Limitations

Despite their promise, large language models in healthcare are not without their biases and limitations. One limitation is that they may not perform well outside the dataset used for training. This means that the models may not generalize well to new populations or situations, leading to poor performance or incorrect recommendations.

Another limitation is that these models rely heavily on structured data, which may not capture the full complexity of patient experiences. Unstructured data, such as clinical notes, can provide valuable insights into patient symptoms and experiences, but they are often difficult to analyze using machine learning techniques.

IV. Future Directions

It can be seen from above that ensuring data privacy and security, mitigating bias, establishing accountability, and evaluating performances critically are vital steps towards integrating LLMs into healthcare safely and effectively.

To overcome these challenges, potential research directions include improving data quality and integrity through better data governance and curation practices, developing more advanced techniques for data preprocessing and normalization, and implementing robust validation methods to assess the performance and reliability of LLMs in healthcare applications. Moreover, there is a growing interest in developing hybrid approaches that combine LLMs with domain-specific knowledge and expertise to improve the interpretability and trustworthiness of AI models in healthcare.

References

[1] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

[2] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

[3] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

[4] Stanford alpaca: an instruction-following llama model. (2023). Accessed: April 3, 2023:https://github.com/tatsu-lab/stanford_alpaca.

[5] Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. New England Journal of Medicine, 380(14), 1347-1358.

[6] Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nature medicine, 25(1), 44-56.

[7] Raza, S. (2023). Improving Clinical Decision Making with a Two-Stage Recommender System Based on Language Models: A Case Study on MIMIC-III Dataset. medRxiv, 2023-02.

[8] Rao, A., Kim, J., Kamineni, M., Pang, M., Lie, W., & Succi, M. D. (2023). Evaluating ChatGPT as an adjunct for radiologic decision-making. medRxiv, 2023-02.

[9] Liu, S., Wright, A. P., Patterson, B. L., Wanderer, J. P., Turer, R. W., Nelson, S. D., ... & Wright, A. (2023). Using AI-generated suggestions from ChatGPT to optimize clinical decision support. Journal of the American Medical Informatics Association, 30(7), 1237-1245.

[10] Pons, E., Braun, L. M., Hunink, M. M., & Kors, J. A. (2016). Natural language processing in radiology: a systematic review. Radiology, 279(2), 329-343.

[11] Ford, E., Carroll, J. A., Smith, H. E., Scott, D., & Cassell, J. A. (2016). Extracting information from the text of electronic medical records to improve case detection: a systematic review. Journal of the American Medical Informatics Association, 23(5), 1007-1015.

[12] Yang, X., Chen, A., PourNejatian, N., Shin, H. C., Smith, K. E., Parisien, C., ... & Wu, Y. (2022). A large language model for electronic health records. NPJ Digital Medicine, 5(1), 194.

[13] Kormilitzin, A., Vaci, N., Liu, Q., & Nevado-Holgado, A. (2021). Med7: A transferable clinical natural language processing model for electronic health records. Artificial Intelligence in Medicine, 118, 102086.

[14] Zhou, S., Wang, N., Wang, L., Liu, H., & Zhang, R. (2022). CancerBERT: a cancer domain-specific language model for extracting breast cancer phenotypes from electronic health records. Journal of the American Medical Informatics Association, 29(7), 1208-1216.

[15] Miner, A. S., Milstein, A., Schueller, S., Hegde, R., Mangurian, C., & Linos, E. (2016). Smartphone-Based Conversational Agents and Responses to Questions About Mental Health, Interpersonal Violence, and Physical Health. JAMA Internal Medicine, 176(5), 619-625.

[16] Miner, A. S., Laranjo, L., & Kocaballi, A. B. (2020). Chatbots in the fight against the COVID-19 pandemic. NPJ digital medicine, 3(1), 65.

[17] Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., ... & Natarajan, V. (2023). Towards expert-level medical question answering with large language models. arXiv preprint arXiv:2305.09617.

[18] Liévin, V., Hother, C. E., & Winther, O. (2022). Can large language models reason about medical questions?. arXiv preprint arXiv:2207.08143.

[19] Woolf, B. P. (2010). A roadmap for education technology.

[20] Wartman, S. A., & Combs, C. D. (2018). Medical education must move from the information age to the age of artificial intelligence. Academic Medicine, 93(8), 1107-1109.

[21] Eysenbach, G. (2023). The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. JMIR Medical Education, 9(1), e46885.

[22] Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., ... & Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health 2 (2): e0000198.

[23] [9] Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., & Blaschke, T. (2018). The rise of deep learning in drug discovery. Drug Discovery Today, 23(6), 1241-1250.

[24] Liu, Z., Roberts, R. A., Lal-Nag, M., Chen, X., Huang, R., & Tong, W. (2021). AI-based language models powering drug discovery and development. Drug Discovery Today, 26(11), 2593-2607.

[25] Grisoni, F. (2023). Chemical language models for de novo drug design: Challenges and opportunities. Current Opinion in Structural Biology, 79, 102527.