Standards for Use of LLM in Medical Diagnosis

Mihailo Svetozarević
Clinic for Neurology
University Clinical Center Niš
Niš, Serbia
mihailo.svetozarevic@gmail.com

Isidora Janković
Center for Radiology
University Clinical Center Niš
Niš, Serbia
isidora_jankovic@yahoo.com

Sonja Janković
Center for Radiology
University Clinical Center Niš
Niš, Serbia
sonjasgirl@gmail.com

Stevo Lukić
Clinic for Neurology
University Clinical Center Niš
Niš, Serbia
srlukic@gmail.com

Abstract

Artificial intelligence, particularly large language models (LLMs), is increasingly being recognized for its potential to revolutionize medical diagnosis by mimicking human cognitive functions in clinical decisionmaking. Despite promising developments, such as the ability to pass medical exams and assist in complex diagnostic processes, LLMs still face significant hurdles, including issues with accuracy, bias, and safety. This paper critically consider evaluation of LLMs performance across various criteria to ensure they meet the required standards for clinical use. Several dimensions of evaluations such as accuracy, calibration, and robustness are used. While LLMs and generative AI more broadly show real potential for healthcare, these tools are not ready yet. The medical community and developers need to develop more rigorous evaluation, analyze across specialties, train on real-world data, and explore more useful types of GenAI beyond current models. But ultimately, we believe these tools can help in improving both physician workload and patient outcomes. We urgently need to set up evaluation loops for LLMs where models are built, implemented, and then continuously evaluated via user feedback.

Keywords

large language models, artificial intelligence, clinical AI implementation, AI in clinical practice, AI safety in healthcare

1 Introduction

Artificial intelligence (AI) by its definition, and in the broadest of terms, represents intelligence exhibited by computer systems. The main goal of AI is to enable computers and machines to mimic human cognitive function. In other words, it aims to simulate human learning, comprehension, problem solving and critical decision making. AI approaches human cognition in two distinct ways, the symbolic and the connectionist approach [1]. The symbolic approach aims to replicate human intelligence by analyzing cognition independent of the biological structure of the central nervous system while the connectionist approach aims to create neural networks that imitate the brains's structure. To realize the potential of AI in healthcare, we believe that the systematic approach to evaluation and benchmarking can get us to a place where AI can be a net positive for health systems.

2 LLM's in Medicine

The rapid advancements in AI, particularly in the realm of large language models (LLM's), have transformed various sectors, including healthcare [2,3]. LLM's and Chat GPT in particular has earned much attention in recent years due to its ability to complete tasks previously considered completable by humans alone as in passing United States Medical Licensing Examination [4]. The ability of LLM's to accurately answer questions, provide advice and even triage patients based on clinical input exceeds that of the everyday person. However, the accuracy of these systems to resolve real world medical issues is yet to exceed that of a fully trained physician. Also, a finite percentage of LLM answers to patients had safety errors, and in one instance the advice given to a patient could have been fatal [5]. In order to avoid this error in the future it is essential to assess these models through rigorous comparative benchmarks. One of the most critical aspects of benchmarking medical LLM's is comparing their performance with existing clinical decision support systems (CDSS) and other AI models. Traditional CDSS, often rule-based or statistical, have been used in healthcare for decades to assist clinicians in making evidence-based decisions. By comparing LLMs to these systems, researchers can determine whether the new models offer significant improvements in accuracy, speed, and comprehensiveness [6]. For example, a comparative benchmark might involve evaluating the diagnostic accuracy of an LLM against a well-established CDSS in predicting outcomes for specific conditions, such as sepsis or diabetes. The LLM's ability to incorporate a broader range of data, including unstructured text from electronic health records (EHRs), could be a key factor in outperforming traditional systems [7]. However, it is also crucial to consider scenarios

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.chtm.9

where traditional systems might still have an edge, particularly in specialized tasks where they have been finely tuned over many years of clinical use [8]. Outside of primary care, radiology is perhaps the medical branch that has been the most upfront and welcoming to the use of new technology [2,3]. The concept of computer-assisted diagnosis (CAD) is well known. AI's provide substantial aid by labeling abnormal or most often borderline exams or simply by quickly excluding negative exams in computed tomographies, X-rays, magnetic resonance images especially in high volume settings like the emergency room where human resources might be less available. AI-driven diagnostic tests have the potential to overcome several current limitations in the clinical approach to patient care [9]. Namely the clinical review, time to diagnosis, diagnostic accuracy and consistency. In tandem with AI, diagnosticians of all medical branches are capable of improving measures of diagnostic accuracy (mainly sensitivity and specificity) as well as minimizing observer variability in specific patient interpretation. This proves most useful in settings where the clinical diagnosis is in question – such as with complex patient presentation or in patients with long histories and various comorbidities. Currently not many prospective studies and randomized trials exist in medical AI application. Most are not prospective, are at high risk of bias and deviate from existing report standards. Data availability is lacking and human cooperator groups are more often small and inadequate. LLM's, in particularly GPT-3, has shown promise in various clinical applications, ranging from creation of patient notes to helping healthcare providers diagnose rare conditions. However, it is important to recognize the inherent limitations of these systems.

3 Standardized Evaluation Framework for Assessing LLM's Clinical Utility for Future Clinical Practice

Medical diagnosis involves a complex process in which a practitioner uses objective data from a clinical exam, as well as data collected from medical tests along with self-described subjective symptoms to conclude the most likely health problem. This kind of approach relies heavily on the synthesis and individual interpretation of a vast amount of information from various sources. These most often include available patient histories, clinical exam data correlated with current medical literature. In this setting LLM's open up new opportunities for enhancing the diagnostic process. In order to better evaluate the LLMs clinical utility a direct comparison must be made between LLMs and human clinicians. This approach is essential to gauge how well AI models can replicate or even enhance the decision-making process of experienced healthcare professionals. Studies often involve presenting both clinicians and LLMs with the same clinical cases and comparing their diagnoses, treatment recommendations, and reasoning [10]. Human clinician benchmarking can reveal important insights into the strengths and limitations of LLMs. For instance, while LLMs might excel at processing and synthesizing vast amounts of data quickly, they may struggle with nuanced cases that require deep contextual understanding or ethical considerations that a human clinician might naturally account for [11]. Furthermore, these benchmarks can highlight areas where LLMs might support clinicians, such as providing second opinions or identifying potential errors in human judgment, rather than replacing them [12]. Randomized controlled trials (RCTs) are considered the gold standard in clinical research for evaluating the efficacy of innovations. Comparative benchmarking of LLMs can also involve assessing how well these models predict or align with outcomes from RCTs. For example, an LLM could be tested on its ability to recommend treatments for stroke prevention based on patient data, and its recommendations could be compared with those validated by RCTs [7]. However, this approach presents a set of challenges, as RCTs often involve highly controlled environments that might not fully capture the complexities of real-world clinical settings. Currently LLMs are most often tested on small datasets acquired for a specific research study or large public benchmark dataset, both of which are usually collected on a limited number of very similar sites with consistent diagnostic techniques. This does not reflect the substantial differences in manufacturer, quality and clinical practices often found in real-world hospitals. As an example, the UK Biobank, a widely employed public imaging benchmark dataset includes brain magnetic resonance images (MRI) for a total of 100,000 patients and more. It restricts image acquisition to four sites each of which has identical equipment in terms of hardware and software and performs regular quality check to ensure the harmonization of data. In contrast most medical centers, including our own in Serbia, extracts data from clinical archives over a period of 20 years which reflects the much more diverse array of available data in everyday settings. Another point of interest is a lack of consensus on which dimensions of evaluation to consider and prioritize for various healthcare tasks. While accuracy is the most often examined dimension when evaluating LLM performance, other dimensions such as fairness, bias and toxicity, robustness, and deployment considerations need to be considered as well [13]. Therefore, while alignment with RCT outcomes is a strong indicator of an LLMs clinical relevance, it is also important to test these models in more varied and less controlled environments to ensure their robustness [11]. Unlike traditional systems or statistical models that remain relatively static once developed, LLMs can be continuously updated and refined. This raises the question of how implement models that are constantly evolving. Development of standardized benchmarks that can be applied across different versions of a model are essential to address this challenge [14]. These benchmarks help identify areas where LLMs can enhance clinical practice and highlight the potential risks or limitations that need to be addressed [6]. By rigorously comparing LLMs against existing systems, human clinicians, and traditional models, we can ensure that these advanced AI systems are integrated into healthcare in a way that maximizes their benefits while minimizing potential harms [10]. In general, there is a lack of consensus on what to consider and prioritize for various healthcare tasks. Several dimensions of evaluations such as accuracy, calibration, and robustness are used [13]. While accuracy is the most often examined when evaluating LLM performance, other aspects such as fairness, bias and toxicity, robustness, and deployment considerations need to be considered as well. A list of possible aspects are presented on Table 1. Comparative benchmarks can guide the development of future AI models. Insights gained from these evaluations can inform

model improvements, such as enhancing interpretability, reducing bias, or improving performance on specific tasks. As the field of AI in healthcare continues to evolve, comparative benchmarking will remain a crucial tool for ensuring that new models are both safe and effective for clinical use [8].

Table 1. Comparative benchmarks for evaluation of LLG performances in healthcare (adapted and modified from Bedi et al. 2024)

Dimension of Evaluation	Definition	Metric Examples
Accuracy	Measures how close the LLM output is to the true or expected answer	Human evaluated correctness, ROUGE, MEDCON
Calibration and Uncertainty	Measures how uncertain or underconfident an LLM is about its output for a specific task	Human evaluated uncertainty, calibration error, Platt scaled calibration slope
Robustness	Measures the LLMs resilience against adversarial attacks and perturbations like typos	Human evaluated robustness, exact match on LLM input with intentional typos, F1 on LLM input with intentional use of word synonyms
Factuality	Measures how an LLMs output for a specific task originates from a verifiable and citable source. It is important to note that it is possible for a response to be accurate but factually incorrect if it originates from a hallucinated citation	Human evaluated factual consistency, citation recall, citation precision
Comprehensiveness	Measures how well an LLMs output coherently and concisely addresses all aspects of the task and reference provided	Human evaluated comprehensiveness, fluency, UniEval relevance
Fairness, bias and toxiticy	Measures whether an LLMs output is equitable, impartial, and free from harmful stereotypes or biases, ensuring it does not perpetuate injustice or toxicity across diverse groups	Human evaluated toxicity, counterfactual fairness, performance disparities across race
Deployment considerations	Measures the technical and parametric details of an LLM to generate a desired output	Cost, latency, inference runtime

4 Conclusion

Comparative benchmarking is a critical process in the development and deployment of medical large language models. By comparing LLMs to existing clinical decision support systems, human clinicians, traditional statistical models, and outcomes from randomized controlled trials, we can gain a comprehensive understanding of their strengths, limitations, and potential impact on healthcare. As AI continues to play an increasingly prominent role in medicine, rigorous comparative benchmarks will be essential for ensuring that these models deliver on their promise of improving patient care while adhering to the highest standards of safety and effectiveness.

Acknowledgments

Views and opinions expressed in this paper are those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor any other authority can be held responsible for them. All authors contributed equally in the final version of this paper. This project is funded by the European Union under Horizon Europe (project ChatMED grant agreement ID: 101159214).

References

[1] Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., et al., 2019. A guide to deep learning in healthcare. Nature Medicine, 25(1), pp.24-29.

[2] Thirunavukarasu, A. J. et al., 2023. Large language models in medicine. Nature Medicine, 29, 1930–1940.;

[3] Thirunavukarasu, A.J., et al., 2023. Trialling a large language model (ChatGPT) in general practice with the applied knowledge test: Observational study demonstrating opportunities and limitations in primary care. JMIR Medical Education, 9, p.46599.

[4] Gilson, A., Safranek, C.W., Huang, T., Socrates, V., Chi, L., Taylor, R.A., et al., 2023. How does ChatGPT perform on the United States Medical Licensing Examination? The implications of large language models for medical education and knowledge assessment. JMIR Medical Education, 9, p.e45312.

[5] Chen, S., Guevara, M., Moningi, S., Hoebers, F., Elhalawani, H., Kann, B.H., Chipidza, F.E., Leeman, J., Aerts, H.J.W.L., Miller, T., Savova, G.K., Gallifant, J., Celi, L.A., Mak, R.H., Lustberg, M., Afshar, M., & Bitterman, D.S., 2024. The effect of using a large language model to respond to patient messages. The Lancet Digital Health, 6(6), pp.e379-e381.

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

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

[8] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S., 2019. Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.

[9] Fletcher, E., Burns, A., Wiering, B., Lavu, D., Shephard, E., Hamilton, W., et al., 2023. Workload and workflow implications associated with the use of electronic clinical decision support tools used by health professionals in general practice: A scoping review. BMC Primary Care, 24(1), p.23.

[10] Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I., 2020. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Medical Informatics and Decision Making, 20(1), 1-9.

[11] Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Webster, D. R., 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA, 316(22), 2402-2410.

[12] Ribeiro, M. T., Singh, S., Guestrin, C., 2016. "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135-1144)

[13] Bedi S, Liu Y, Orr-Ewing L. et al. (2024) A Systematic Review of Testing and Evaluation of Healthcare Applications of Large Language Models (LLMs). MedRvix August 16. 2024.

[14] Goodfellow, I., Shlens, J., & Szegedy, C., 2014. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572