ASEE AMERICAN SOCIETY FOR ENGINEERING EDUCATION

Paper ID #44893

Large Language Models in Healthcare: Bridging the Gap between Performance Evaluation and Socio-Ethical Implications

Mr. Abdullah Aldwean, University of Bridgeport

Abdullah Aldwean is a Ph.D. candidate in Technology Management and innovation at the University of Bridgeport with 15+ years of professional experience in healthcare industry. His ongoing research explore the application of Generative Artificial Intelligence in healthcare, with particular interest in Generative large Language Models evaluation analysis. Abdullah holds a Master of Business Administration from Saudi Electronic University in association with Colorado State University global campus.

Dr. Dan Tenney, University of Bridgeport

Dr. Tenney is an Assistant Professor at the University of Bridgeport in the Technology Management Department as part of the Engineering School. Dan Tenney worked in various Quality, Technical, and Operational positions in manufacturing divisions of HJ Heinz Company, 3M Company and Nile Spice Foods (acquired by Quaker Oats). For more than 25 years Dan was a member of the executive teams that directed and managed these divisions. Dan's current focus is strategic technical and business management, application and research. Dan is a Board member on a Child's Mental Health nonprofit agency where he has facilitated strategic planning and operational management training and guidance. He has published numerous publications on strategic, technology, and business management topics.

Large Language Models in Healthcare: Bridging the Gap between Performance Evaluation and Socio-Ethical Implications

Abstract

Utilizing large language models (LLMs), such as the Bidirectional Encoder Representations Transformer (BERT), presents an opportunity to revolutionize the healthcare experience by enhancing patient engagement, facilitating medical education, and improving the overall healthcare service outcomes. However, integrating large language model solutions in a highly regulated industry such as healthcare poses many challenges to healthcare decision-makers due to the high level of uncertainty, the complexity, and the potential social and ethical implications. Therefore, conducting thorough evaluations of LLM-based systems to ensure their ability to achieve intended goals securely, ethically, and safely is critical for healthcare organizations. In this paper, we reviewed the recent advancements in LLM evaluation fronts, mainly focusing on the performance evaluation of medical LLMs in the healthcare domain. We highlighted the potential opportunities and limitations of utilizing these advanced technologies in the context of clinical services. Additionally, we propose a comprehensive framework that integrates various evaluation aspects to better meet the unique requirements of LLMs adoption in healthcare. This framework aims to facilitate the adoption decision-making process by ensuring the utilization of the LLMs potential while holding high standards of safety, security, and ethical practice. This paper contributes to the knowledge by providing researchers, decision-makers, and healthcare practitioners with valuable insights into important aspects that should be considered in LLMs adoption decisions in the healthcare domain.

Introduction

Technological innovation has always been an essential element of healthcare sector development. This is due to the ability of these innovative technologies to drive significant improvements in the Quadruple Aim dimensions of healthcare services, which include enhancing patient experience, improving population health, reducing costs, and improving the work life of healthcare providers [1]. Emerging technologies such as Artificial intelligence (AI) have the potential to transform the healthcare sector by improving patient experience, increasing operational efficiency, and advancing medical research. Researchers in the medical field widely acknowledge the potential role of AI applications in revolutionize traditional healthcare models and shift the service toward data-driven, patient-focused care [2],[3]. However, among many AI applications, the Large Language Model (LLM) has been recognized as one of the most promising AI applications in the healthcare sector [4]. Unlike many traditional AI systems, which often require significant investment and complex implementation processes, LLMs offer highly accessible solutions with low deployment requirements that make their benefits available to a broader range of healthcare users.

Large Language Models are advanced systems designed to process human natural language and generate responses without being specifically trained for the tasks. These models are trained on large amounts of text data from different sources to understand and effectively generate human-like language [5]. In the healthcare context, LLMs can provide many advantages in a wide range of clinical, non-clinical, and educational tasks. In clinical settings, LLMs can improve diagnosis accuracy, support clinical decisions, and extract essential information from clinical data [5]. In non-clinical tasks, LLMs can play a crucial role in reducing the burden on healthcare professionals by streamlining administrative processes. It can also improve the patients' experience and outcomes by enhancing the documentation accuracy. In the area of medical education, the advantages of LLMs are massive and impactful. Models such as GPT-4 have the capability to pass well-known medical exams such as the United States Medical Licensing Examination (USMLE) [6]. These Models can provide an augmented learning experience for medical students allowing them to gather and analyze data in a faster and more efficient manner. Nevertheless, while LLMs can provide promising advantages in the healthcare sector, they raise several issues and concerns. Several issues and concerns surrounding the application of LLM in healthcare share similarities with those associated with the application of other artificial intelligence applications in healthcare, such as data privacy, transparency, and algorithmic bias. However, specific concerns are uniquely distinct LLM applications due to their ability to mislead users by generating convincing yet incorrect or harmful text [7].

The need for a comprehensive evaluation process to assess different aspects associated with LLM in healthcare is a crucial step to ensure safe, efficient, and ethical deployment. The healthcare sector is by far one of the most critical sectors for society; thereby, the evaluation frameworks of LLMs should incorporate societal and ethical aspects in addition to traditional technical performance metrics. Understanding the social and ethical implications of LLMs in healthcare is critical not only for healthcare decision-makers but for broader stakeholders in society. With these ideas in mind, the main driver behind this search comes from an apparent gap in the comprehensive understanding of the evaluation requirement of LLMs within the healthcare landscape. With a focus on the technology management domain, this research aims to identify LLMs' potential opportunities and limitations in clinical settings and highlight the related evaluation practices and frameworks. As shown in Fig 1. This research is concerned with the intersection area between three broader domains: technology management, healthcare services, and artificial intelligence. Additionally, this research provides a review of the LLMs evaluation frameworks in the literature. The main query this research proposed is: What are the key evaluation criteria and frameworks that guide the effective assessment of LLMs in the healthcare environment? By identifying the criteria that drive the successful evaluation of LLMs in healthcare, this research concludes by proposing a comprehensive human-based evaluation model that blends technical performance with social and ethical aspects.

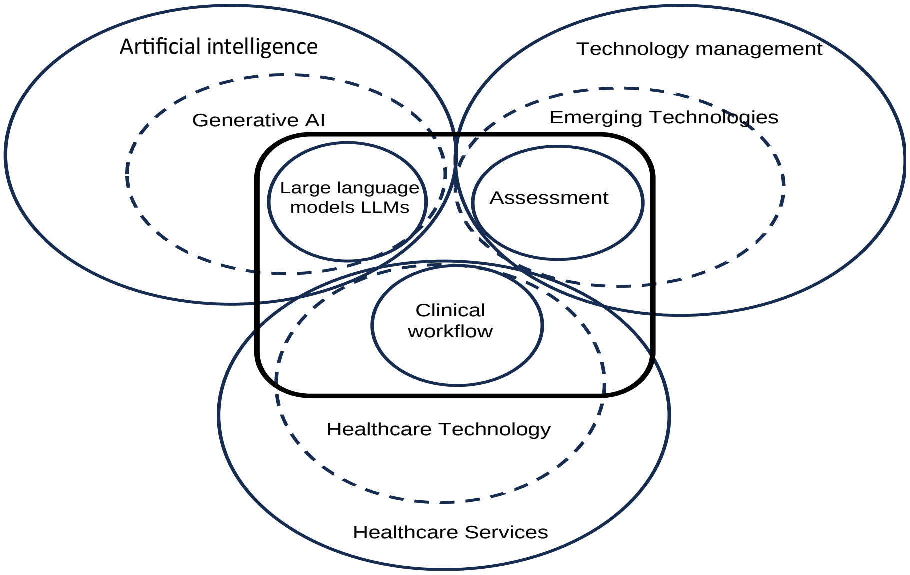

Figure 1: Research area of interest.

Literature review

There is a growing body of literature on the useability of large language models (LLMs) in healthcare. This expanding interest from researchers reflects the importance of this technology in the medical domain. Recent research has emphasized the potential of LLMs to tackle different challenges facing the healthcare sector [5],[6],[7],[8]. Several challenges placed enormous stress on the healthcare sector and threatened the system's future stability. For example, the cost of health services is rising in many countries around the world. In the USA, the National Health Expenditure (NHE) increased by 2.7% to $4.3 trillion or $12,914 per person in 2021, accounting for 18.3% of the Gross Domestic Product (GDP) [9]. The continued rise in health costs is considered a leading factor influencing health outcomes and intensifying health disparities, especially for underserved populations [10]. Another significant challenge facing the healthcare system is medical error, which refers to any incidence in medical practice that leads to or could lead to unintended results [11]. Medical error can occur at any stage of the healthcare service process, from pre-examination to diagnosis to treatment follow-up. Medical errors are considered the third cause of death in the USA after heart disease and cancer [12]. The complexity of healthcare is increasing, and ensuring safe, high-quality, affordable, and efficient services become more challenging over the years.

However, multidiscipline efforts continue to explore how LLM could tackle the current challenges facing healthcare. For example, a study by Karabacak and Margetis [4] discussed the potential of LLMs in enhancing diagnostic accuracy and supporting clinical decisions. It emphasized the importance of education, robust evaluation frameworks, clinical validations, and adherence to health regulatory standards. In another study, Karttunen [13] reviewed 44 LLMs model performance in healthcare decision support by focusing on dataset and model architectures. His study showed a promising future for LLMs in medical data analysis, contextual understanding, and automation. Moreover, a study by Klang, et al. [14] explored the feasibility of LLMs in the cardiology field. Their study highlighted the benefits of LLMs in this medical field, such as saving time, accelerating research, and improving accuracy.

LLM is an advanced application of artificial intelligence under natural language processing (NLP). Some examples of LLMs include the Generative Pre-Trained Transformer (GPT) by OpenAI and the Bidirectional Encoder Representations Transformer (BERT) by Google [15], [16]. LLM can be considered an extension to NLP's early rule-based and statistical models, such as long-short-term memory (LSTM). These models leverage complex deep learning techniques, such as convolutional neural networks (CNN) and a vast amount of text data to learn the patterns and structures of natural language. LLM is designed to generate logical and fluent text to perform different tasks, such as answering questions, summarizing, analyzing sentiments, and translating language [17]. The architecture of LLMs is complex and consists of many layers that work to gather and manipulate text using a revolutionized mechanism called attention [18]. Despite the LLM technical characteristics and sophisticated capabilities, the introduction of conversational-based models such as ChatGPT has established a new paradigm in human-AI interaction. The wide adoption and the fast diffusion of these models reflect a deep interest in AI based products by users. According to the OpenAI company, the ChatGPT has 100 million weekly active users [15].

Efficient evaluation is a crucial step in ensuring the applicability of LLMs in the healthcare sector. While automated evaluation methods are more cost and time-efficient, human evaluation remains the golden standard for determining the safety and usefulness of LLM in healthcare. Experts-based evaluation is critical in a sensitive domain such as healthcare [19]. However, some organizations have developed semi-automated evaluation pipelines to assess new LLMs; these pipelines include several stages and become more time-consuming. The low programming barriers have led to an explosion in the number of new LLMs. As of the time of writing this paper, the hugging face platform listed more than 400 thousand LLM for different purposes [20]. These growing numbers of models challenge the continuous monitoring and evaluation using human experts and increase the popularity of automated methods such as Perplexity and ROUGE.

Results

This research concerns the evaluation of Large Language Models (LLMs) in healthcare. It explores the possible application of LLMs in the field and the current evaluation approaches. The opportunities that LLM holds cannot be realized without a careful evaluation process. The literature review reveals several use examples of LLM in the clinical context, which include:

- Clinical decision support: LLMs can be used for real-time, evidence-based decision support by retrieving medical research [21].
- Medical predictive analysis and risk assessment: LLMs can be used to analyze large datasets for predicting health outcomes, such as readmission risks and potential complications [22].
- Medical image analysis: LLMs can be used to detect abnormalities in various medical images, such as mammograms, radiological scans, pathology slides, and dermatological images.
- Personalized treatment: LLMs can be used to generate personal recommendations based on individual health conditions, lifestyle patterns, and treatment plans [23].

LLMs can be utilized for non-clinical tasks. Non-clinical tasks refer to supporting activities that are essential in healthcare, such as administration, research, and education. Some examples of the potential role of LLMs in this context include:

- Patients' engagement: LLMs can be used to answer patients' questions, remind them of appointments, and provide medical educational content.

- Administrative tasks: LLMs can be used to optimize administration workflows such as scheduling, billing, and coding.
- Research and training: LLMs can be used in data mining and knowledge discovery, such as population epidemiology, infection disease patterns, and resource allocations. It also can be helpful in creating interactive, dynamic models for training purposes.

Despite the potential of LLMs in healthcare, evaluating these models is the most crucial step in their utilization. The most common metrics for assessing LLMs are designed to evaluate natural language processing (NLP) and machine learning models. Some examples include:

- Accuracy: Accuracy metrics measure the percentage of correct predictions of the model to the total number of predictions. It can be used for tasks with clear right or wrong answers, such as classification.
- Perplexity: Perplexity metrics measure the probability of predicting the next token (e.g., character or word) in a certain sequence. However, these metrics are easy to apply but hard to interpret [19]. The lower the score, the better the model at predicting the next token, which reflects better performance in response.
- BLEU (Bilingual Evaluation Understudy): BLEU metrics used for evaluating translated text against defined references. The high BLEU score reflects the higher translation quality of the LLM model [24].
- ROUGE (Recall-Oriented Understudy for Gisting Evaluation): ROUGE metrics are used to evaluate automatic summarization and translation quality by comparing the overlap between generated text and reference text. It mainly focuses on precision (how much of the output summary is in the reference summary) and recall (how much of the references are included in the output). High ROUGE scores indicate better quality, reflecting a higher degree of text overlap [25].

Discussion

Researchers determine the type of metrics when evaluating Large Language Models (LLMs) based on the specific use cases. As discussed in the results section, most of the used metrics are widely established in the field of machine learning. These metrics are helpful in measuring the quality of LLM in the developing phase of the product life cycle. However, other metrics are critically important in extending the assessment of LLMs to non-technical aspects. For instance, the users of LLM are more concerned with the language proficiency, context understanding, coherence, relevance, and diversity nature of the output. [26]. Additionally, healthcare researchers are more concerned with multiple evaluation aspects such as human alignments, autonomy, and ethical implications. Therefore, to comprehensively determine the quality of LLMs, especially in the healthcare sector, the LLM should be evaluated through different methods using human expertise, such as human annotation, A/B questions, and expert judgment. Human evaluation is critical for assessing the LLMs outputs quality and their alignment with human needs and values. Although different research has shown that the results of automated LLM evaluation are consistent with the results acquired by expert humans [27], the evolving nature of LLM models and their ability to improve over time using reinforcement learning required continuous monitoring and assessments.

Technology assessment is one of the most important research areas in the technology management domain. It mainly focuses on the outcomes and impact of technology. The concept of technology assessment was first developed in the United States during the 1960s, led by the Office of Technology Assessment [28]. Over the years, scholars have developed different techniques to assess the quality of technological innovations, for example, structural modeling, impact analysis, scenario analysis, risk assessment, and decision analysis. However, since emerging technologies, including LLMs, have relatively fewer years in production, determining the impact of these technologies can be challenging. Therefore, this research proposes a fuzzy decision analysis approach to evaluate LLMs in healthcare. As shown in Fig 2, the framework includes different aspects that are essential to measure the quality of different LLMs. The proposed model consists of sixteen criteria for evaluating different LLMs based on human experience. These criteria and their explanation are illustrated in Table 1.

Table 1: Evaluation Model perspectives and criteria

Aspect	Criteria	Definition	Example
Social-Ethics	Safety	The extent to which LLMs are designed to prevent mental and physical risks to users and beneficiaries, such as hallucination,	Cyber security measurement. Human oversite. Error responses

		misinformation, harmful content, and malicious cyber breaches.	
	Social Acceptance	The extent to which the LLM is perceived by potential users/beneficiaries as a trustworthy source of information in the medical environment.	Public attitude, image, personal opinions, beliefs, social force.
	Transparency	The extent to which LLM has clear, understandable information about how the model works and what are the features and limitations.	Model strengths, limitations, potential biases, data usage,
	Liability	Refer to the determination of where responsibility and accountability lies among end users, developers, and other parties in case of medical errors or system malfunctions	Auditability measurement. Monitoring mechanisms. Disclaimer notes.
Technology	System Quality	Refers to LLM's efficiency, effectiveness, output quality, and ability to maintain robust performance and provide consistent, accurate information.	Consistency, relevancy, fluency, usefulness.
	Relative advantage	Refer to the extent to which LLMs are perceived as superior to similar existing technologies.	Technical advantage, medical benefits Production Speed,
	Complexity	Refers to the degree of difficulty in understanding, learn, and to use the LLM in clinical settings.	Perceived level of effort Required technical skills. Required training.
	Clinical Validity	Refer to the model's accuracy and reliability in interpret and generate medical information or advice that is consistent with clinical knowledge and practices.	Medical knowledge benchmarks. Evidence strength and quality. Error rate.
	Adaptability	Refer to degree that to which the LLM can be customized, tailored, refined to meet specific needs.	Fine tuning.
Organization	Stability	Refer to the LLM provider position in the market, financial stability, and record of	Technological infrastructure, workers' skills, financial resources,

		innovation and developing of advanced AI technologies.	innovation record, and number of patents.
	Data Governances	Refers to the LLM provider policies, procedures, and standards implemented to manage, protect, and ensure the quality and security of user's data.	Data sharing policies. data handling, storage
	Customer support	Refer to the LLM provider's level of commitment to customer support that aimed to facilitate the integration and the use of the technology.	Technical assistance, training resources, support response time.
Regulation	Compliance	Refers the LLM provider's commitment to the set of laws, regulations, guidelines, and best practices that govern the development, deployment, and use of AI products.	The draft of the EU AI act.
	data protection	Refers to the LLM provider's ability to comply with specific regulations that govern the medical practices and medical data handling protocols.	GDPR (General Data Protection Regulation) in the European Union. HIPAA (Health Insurance Portability and Accountability Act) in the US.
Economic	Entail Costs	Refers to the financial investment and obligation associated with the implementation expenses.	licensing fees, hardware requirements, and system upgrades.
	Operation cost	Refer to the ongoing financial expenses required to maintain and use the technology.	subscription, maintenance, technical support, experts' consultation, and energy consumption

The selected criteria can also be used in comparing different models to select the most suitable solution for specific tasks. The needs of each healthcare organization are unique and may differ with respect to the type of desired information. Nevertheless, standard requirements such as data protection, privacy, and clinical evidence are shared regardless of the location or size of the organization. The framework attempts to measure these criteria in a fuzzy environment. The purpose of using fuzzy numbers instead of crisp numbers to assess criteria is to deal with the uncertainty of human judgment. The decision analysis quality heavily relies on human judgment accuracy. Like many emerging technologies, LLMs are relatively new, and measuring their impact is difficult to be determined on crisp numbers. The Fuzzy theory can provide a solution for decision analysis under uncertainty[29]. It allows evaluators to assess specific criteria with a degree of hesitation without compromising the quality of overall judgment.

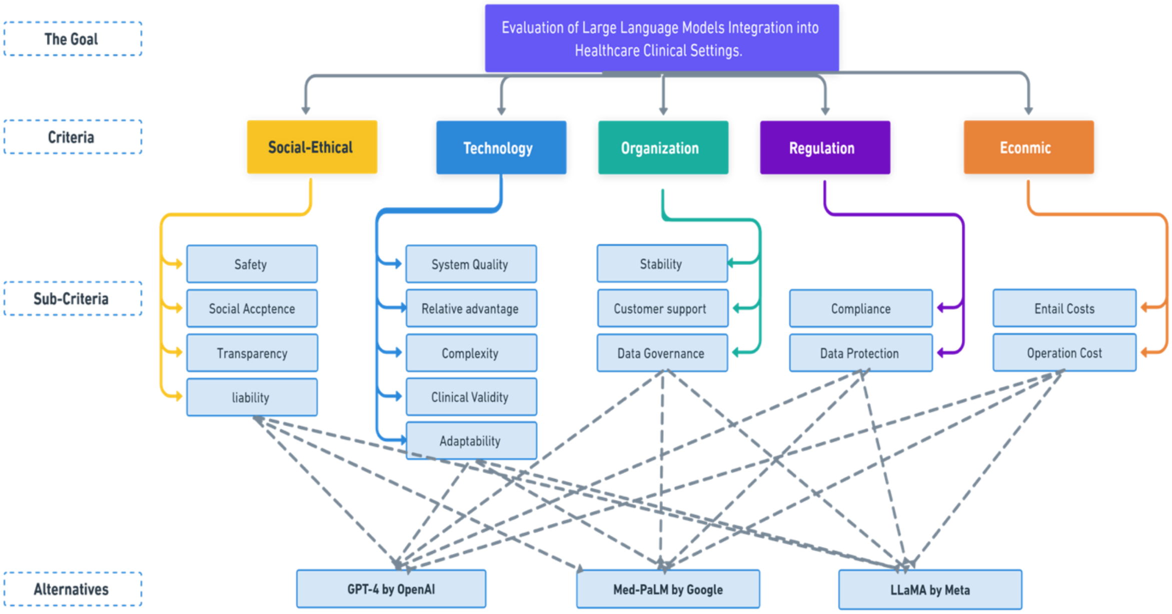

Figure 2: LLM Evaluation Model

An illustrative example

This section provides an illustration example to demonstrate the application of the proposed framework in healthcare. A healthcare organization that desires to leverage Large Language Models (LLM) for clinical support application should consider different perspectives to evaluate and select the most appropriate model. While different stakeholders in the healthcare domain share similar concerns with respect to LLM implementation, their level of concerns may differ based on their role, for instance, technical complexity by IT professionals compared to clinical validity by the clinicians. The proposed framework provides a comprehensive evaluation process encompassing important perspectives such as Social-Ethics, Technological integration, organizational factors, regulations, and economic considerations, which are further decomposed into specific evaluation metrics, such as system quality, transparency, and safety protocol.

The LLM evaluation framework utilizes the Fuzzy Analytic Hierarchy Process (FAHP). The FAHP decision analysis technique is an extension of the Analytic Hierarchy Process (AHP) [30]. It leverages fuzzy logic to address subjective judgment and decision uncertainty. The application of the proposed framework includes several steps to evaluate three different LLMs: GPT-4, Med-PaLM, and LLaMA.

step 1: form the expert's panel.

The evaluation process will depend on the input from a diversified panel of healthcare decision-makers, typically clinicians, IT professionals, and management executives. The assembled panel judgment will be used to validate the framework criteria and drive their relative weights (priority level).

step 2: validate framework criteria and sub-criteria.

The framework includes five main criteria and 16 sub-criteria. Each panel member will be asked to provide feedback on framework elements through a survey or workshop. This step is mainly concerned with the relevance of each defined criterion to the evaluation objective. The panel feedback will be used to modify the framework, if necessary, for instance, adding or removing certain criteria.

step 3: construct fuzzy pairwise comparison matrices.

Each panel member will compare the main criteria and sub-criteria pairwise using linguistic terms (e.g., strong importance, extreme importance) that can be converted into Fuzzy numbers. The term's corresponding values represent the Triangular Fuzzy Number parameters (l, m, u), which donate to the low, middle, and upper values. These weights reflect the relative importance of each main criterion with respect to the main objective and the relative importance of each sub-criterion with respect to the main criterion.

step 4: calculate the fuzzy weights and check consistency.

Apply the FAHP methodology to calculate the final weights of each criterion and sub-criterion to the decision framework. Additionally, the consistency ratio is calculated to check the reliability of comparison results. The pairwise comparison step will be repeated if the consistency ratio is not within the acceptable range.

step 5: rate alternatives against each criterion.

Evaluate each LLM alternative against each criterion using the same linguistic terms in step 3. This step involves assessing each model with respect to the framework elements. It generates an overall fuzzy score for each LLM that reflects its suitability as a clinical support application from different perspectives.

step 6: defuzzification.

Covert the obtained fuzzy scores from the previous step into crip values using the centroid method to facilitate the comparison between different LLMs models. The LLM alternative with the highest score is considered the most appropriate option based on the collective group judgment.

The above evaluation process steps underscore the importance of human judgment in navigating the complex dimension of LLM evaluation in healthcare. The proposed framework can facilitate a rigorous evaluation process aligned with multi-perspective criteria essential to ethical and safe adoption decisions.

Conclusion

In conclusion, this research provides an overview of the potential of utilizing Large Language Models (LLMs) within the healthcare sector. It shows potential use cases within clinical and non-clinical tasks, such as clinical decisions, personalized treatment, and administrative tasks. The current LLMs' performance quantitative assessment tools and metrics such as Perplexity, BLEU, ROUGE, and Accuracy have been discussed, and the need for supporting assessment tools has been justified. Furthermore, this research proposed a human-based analysis evaluation framework that incorporates social and ethical criteria to measure the broader implications of deploying LLMs in the healthcare landscape. The findings of this research highlight the urgent need to develop assessment tools with high capabilities that match LLMs and other artificial intelligence products' evolving nature. It also emphasizes the critical role of a balanced approach between automated and augmented evaluation practice by incorporating human-based judgment and expertise. By evaluating social and ethical impacts, even with a degree of uncertainty, we can ensure that the deployment of LLMs aligns with the values and needs of patients, healthcare providers, and society.

References:

[1] T. Bodenheimer and C. Sinsky, "From triple to quadruple aim: care of the patient requires care of the provider," *Ann Fam Med,* vol. 12, no. 6, pp. 573-6, Nov-Dec 2014, doi: 10.1370/afm.1713.

[2] S. Sunarti, F. Fadzlul Rahman, M. Naufal, M. Risky, K. Febriyanto, and R. Masnina, "Artificial intelligence in healthcare: opportunities and risk for future," *Gac Sanit,* vol. 35 Suppl 1, pp. S67-S70, 2021, doi: 10.1016/j.gaceta.2020.12.019.

[3] E. Sezgin, "Artificial intelligence in healthcare: Complementing, not replacing, doctors and healthcare providers," *DIGITAL HEALTH,* vol. 9, p. 20552076231186520, 2023, doi: 10.1177/20552076231186520.

[4] M. Karabacak and K. Margetis, "Embracing Large Language Models for Medical Applications: Opportunities and Challenges," *Cureus,* vol. 15, no. 5, 2023.

[5] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, "Large language models in medicine," *Nature medicine,* vol. 29, no. 8, pp. 1930-1940, 2023.

[6] M. Sallam, "ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns," (in eng), *Healthcare (Basel),* vol. 11, no. 6, Mar 19 2023, doi: 10.3390/healthcare11060887.

[7] P. Lee, S. Bubeck, and J. Petro, "Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine," *New England Journal of Medicine,* vol. 388, no. 13, pp. 1233-1239, 2023.

[8] Y. Wang, Y. Zhao, and L. Petzold, "Are Large Language Models Ready for Healthcare? A Comparative Study on Clinical Language Understanding," *arXiv preprint arXiv:2304.05368,* 2023.

[9] C. f. M. M. Services. "National health expenditure data " https://www.cms.gov/data-research/statistics-trends-and-reports/national-health-expenditure-data/nhe-fact-sheet (accessed 28/09/2023.

[10] J. S. Williams, R. J. Walker, and L. E. Egede, "Achieving equity in an evolving healthcare system: opportunities and challenges," *The American journal of the medical sciences,* vol. 351, no. 1, pp. 33-43, 2016.

[11] E. D. Grober and J. M. Bohnen, "Defining medical error," *canadian Journal of Surgery,* vol. 48, no. 1, p. 39, 2005.

[12] M. A. Makary and M. Daniel, "Medical error—the third leading cause of death in the US," *Bmj,* vol. 353, 2016.

[13] P. Karttunen, "LARGE LANGUAGE MODELS IN HEALTHCARE DECISION SUPPORT," Tampere University, 2023.

[14] E. Klang, M. Cohen-Shelly, and F. Lopez-Jimenez, "Leveraging Large Language Models to Enhance Digital Health in Cardiology: A Preview of a Cutting-Edge Language Generation Model," *Mayo Clinic Proceedings: Digital Health,* vol. 1, no. 2, pp. 105-108, 2023/06/01/ 2023, doi: https://doi.org/10.1016/j.mcpdig.2023.03.003.

[15] OpenAI. "Introducing ChatGPT." https://openai.com/blog/chatgpt#OpenAI (accessed July 12, 2023).

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805,* 2018.

[17] A. Toma, P. R. Lawler, J. Ba, R. G. Krishnan, B. B. Rubin, and B. Wang, "Clinical Camel: An Open-Source Expert-Level Medical Language Model with Dialogue-Based Knowledge Encoding," *arXiv preprint arXiv:2305.12031,* 2023.

[18] A. Vaswani *et al.*, "Attention is all you need," *Advances in neural information processing systems,* vol. 30, 2017.

[19] Z. Guo *et al.*, "Evaluating large language models: A comprehensive survey," *arXiv preprint arXiv:2310.19736,* 2023.

[20] Y. Chang *et al.*, "A survey on evaluation of large language models," *ACM Transactions on Intelligent Systems and Technology,* 2023.

[21] L. Caruccio, S. Cirillo, G. Polese, G. Solimando, S. Sundaramurthy, and G. Tortora, "Can ChatGPT provide intelligent diagnoses? A comparative study between predictive models and ChatGPT to define a new medical diagnostic bot," *Expert Systems with Applications,* vol. 235, p. 121186, 2024.

[22] K. Huang, J. Altosaar, and R. Ranganath, "Clinicalbert: Modeling clinical notes and predicting hospital readmission," *arXiv preprint arXiv:1904.05342,* 2019.

[23] G. P. Patrinos, N. Sarhangi, B. Sarrami, N. Khodayari, B. Larijani, and M. Hasanzad, "Using ChatGPT to predict the future of personalized medicine," *The Pharmacogenomics Journal,* 2023/09/19 2023, doi: 10.1038/s41397-023-00316-9.

[24] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311-318.

[25] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74-81.

[26] J. Ilicki, "A Framework for Critically Assessing ChatGPT and Other Large Language Artificial Intelligence Model Applications in Health Care," *Mayo Clinic Proceedings: Digital Health,* vol. 1, no. 2, pp. 185-188, 2023/06/01/ 2023, doi: https://doi.org/10.1016/j.mcpdig.2023.03.006.

[27] U. Daiju, L. W. Shannon, M. Toshimasa, D. Ryo, T. Hiroyuki, and M. Yukio, "Evaluating GPT-4-based ChatGPT’s Clinical Potential on the NEJM Quiz," *medRxiv,* p. 2023.05.04.23289493, 2023, doi: 10.1101/2023.05.04.23289493.

[28] T. A. Tran and T. Daim, "A taxonomic review of methods and tools applied in technology assessment," *Technological Forecasting and Social Change,* vol. 75, no. 9, pp. 1396-1405, 2008/11/01/ 2008, doi: https://doi.org/10.1016/j.techfore.2008.04.004.

[29] R. E. Bellman and L. A. Zadeh, "Decision-making in a fuzzy environment," *Management science,* vol. 17, no. 4, pp. B-141-B-164, 1970.

[30] T. L. Saaty, "Axiomatic foundation of the analytic hierarchy process," *Management science,* vol. 32, no. 7, pp. 841-855, 1986.