Meeting Cultural and Linguistic Demands to Accommodate Fine-Tuned LLMs to Local Medical Customs and Patient Communication

Gordana Petrovska Dojchinovska[1],Monika Simjanovska Misheva[1], Kostadin Mishev[1]
[1]Ss. Cyril and Methodius University,
Faculty of Computer Science and Engineering,
Skopje, N. Macedonia
gordana.petrovska.dojchinovska@students.finki.ukim.mk
{monika.simjanoska, kostadin.mishev}@finki.ukim.mk

ABSTRACT

Integrating advanced open-source large language models (LLMs), such as LLaMA and GatorTron, into healthcare offers a novel approach to enhancing communication between physicians and patients. This paper provides a comprehensive review of the potential of these models to improve patient-provider interactions, focusing on their ability to process and generate human-like language in real-time clinical settings. The review outlines the methodology used to evaluate LLMs, which includes a detailed comparison based on qualitative factors such as linguistic adaptability, cultural sensitivity, and context-awareness, alongside quantitative metrics such as accuracy rates, error margins, and patient satisfaction scores from clinical studies. Key ethical considerations are explored, particularly concerning data privacy, patient consent, and accountability. The paper delves into how adopting specific ethical frameworks or guidelines can help mitigate risks associated with bias, misinformation, and patient autonomy. Additionally, the potential for LLMs to perpetuate biases or cultural misunderstandings is discussed, emphasizing the need for fine-tuning these models to align with clinical guidelines and patient needs across different geographies and medical customs. While the paper acknowledges the gaps in current research, such as the emotional sensitivity of LLMs and their ability to understand nuanced patient concerns, it also proposes structured future research directions. This includes the development of LLMs that are more contextually aware, emotionally intelligent, and capable of operating in diverse healthcare settings. By synthesizing current studies and real-world applications, this paper aims to provide a transparent, reproducible framework for evaluating the effectiveness of fine-tuned LLMs in transforming healthcare communication, ultimately improving patient outcomes and satisfaction.

KEYWORDS

large language models, healthcare, patient-doctor communication, ethics in AI, cultural sensitivity

1 INTRODUCTION

Effective communication has always been at the forefront of successful interaction between physicians and patients. It has already been purported that integrating cultural and linguistic competency into healthcare policies, provider training, and patient care strategies will improve the quality of care for diverse populations[5]. Conversely, language barriers, cultural misunderstandings, and a lack of cultural awareness among healthcare providers can lead to miscommunication, misdiagnosis, medication errors, and other safety risks[11]. Large Language Models (LLMs) have shown immense potential in various fields due to their proficiency in generating and understanding natural language. In healthcare, models such as LLaMA and GatorTron present an exciting frontier for improving communication between physicians and patients. These models can process and generate human-like language, which could address significant challenges in clinical communication, such as linguistic, cultural, or emotional barriers. However, significant challenges remain, particularly regarding ethical implications, technical hurdles, local customs, linguistic demands, and the critical aspect of patient-doctor communication. This review examines current applications of LLMs in healthcare, the associated challenges, ethical concerns, and potential gaps, incorporating a range of recent research studies in the field from 2023 and early 2024.[1]

1.1 Purpose and Scope

This paper aims to critically examine the potential of integrating LLMs into healthcare settings, emphasizing their role in enhancing communication. The scope of this review extends to a comparative evaluation of LLaMA, GPT-4, L2M3, and GatorTron, focusing on their potential impact, ethical considerations, and future research needs. The goal is to provide healthcare practitioners, researchers, and developers with a transparent and reproducible framework for adopting LLMs.

2 METHODOLOGY

This review systematically analyzes the current applications of LLMs in healthcare by conducting a structured comparison of models such as GPT-4, Llama, GatorTron, and L2M3. The evaluation criteria include both qualitative and quantitative metrics, focusing on model accuracy, error rates, patient satisfaction scores, and the ability to generate culturally sensitive and clinically appropriate responses. The sources for this review consist of peer-reviewed studies published between 2023 and 2024, covering applications of LLMs in real-time clinical settings, mental health, and patient-doctor communication. Each model's performance is assessed based

.

[1]https://doi.org/10.70314/is.2024.chtm.11

on specific clinical tasks, adaptability to local medical customs, and ethical considerations. This methodology ensures transparency and reproducibility, allowing future researchers to replicate and validate the findings.

3 CURRENT APPLICATION OF LLMS IN HEALTHCARE

LLMs are being adopted for various tasks within healthcare, with their most prominent use in medical communication, clinical decision support, and facilitating patient-doctor interactions. Conversational agents like ChatGPT assist patients in understanding medical terminology and provide immediate answers to healthcare-related queries, often enhancing patient engagement with care providers. LLMs also bridge communication gaps between patients and healthcare providers, improving clarity and comprehension in patient education[25]. In patient-doctor communication, LLMs serve as intermediaries to help patients articulate their symptoms and concerns more clearly. This enhances clinical encounters, as doctors can quickly grasp patient issues without the noise of miscommunication. They can improve the precision of patient-doctor dialogue, particularly when addressing complex conditions and explaining treatment plans[21]. Similarly, mental health applications have seen LLMs being used to summarize counseling sessions, allowing healthcare providers to focus on therapeutic interventions rather than administrative tasks[1]. Beyond patient communication, LLMs assist clinicians with diagnoses and medical research. GPT-4 and LLaMA are increasingly used as clinical assistants, offering support in diagnosis and treatment planning[29]. Large language models enhance decision-making in surgical care by answering patients' questions, thereby improving pre-surgical communication and reducing patient anxiety[18]. Figure **??** shows the areas of the medical domain where LLMs are currently being applied.

Figure 1: Applications of large language models in healthcare

4 IMPROVING PATIENT-DOCTOR COMMUNICATION

Effective communication between patients and doctors is a cornerstone of healthcare. Miscommunication can lead to misunderstanding, reduced adherence to treatment plans, and dissatisfaction with care. LLMs offer a promising solution to enhance communication by simplifying medical information into layman's terms. The role of LLMs in improving patient-doctor interactions is significant, as they ensure the translation of medical jargon into easily understandable language during consultations, ultimately leading to more informed decision-making and patient compliance[19]. Recent research explores how generative AI tools, including LLMs, have been applied to reduce misunderstandings in clinical services. They show that patients often feel overwhelmed by medical language, and LLMs can mitigate this issue by acting as interpreters, promoting clearer communication between patients and their healthcare providers[28]. A pictorial demonstration of the ways large language models can improve the communication between patients and their providers is given in Figure 2.

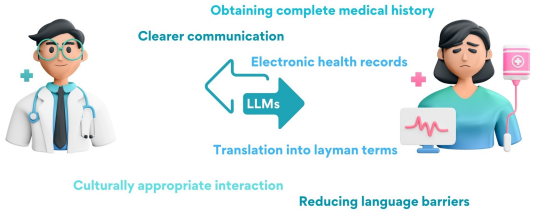

Figure 2: Patient-doctor communication facilitated by the use of LLMs

There is also research that addresses how LLMs can cater to multilingual settings, facilitating better communication in regions with diverse linguistic backgrounds. This ensures that patients from low-resource, multilingual regions receive equitable healthcare without language barriers compromising the patient-doctor relationship[3]. However, the risks of over-reliance on LLMs in direct patient communication must also be acknowledged. Some research argues that cultural sensitivity and local medical customs play a crucial role in healthcare communication. LLMs that fail to adapt to these factors may risk undermining trust between patients and healthcare providers. For instance, culturally inappropriate language or failure to understand local medical norms could weaken the patient-doctor bond and compromise care[11]. The diagram in Figure 3 shows a basic workflow of the integration of LLMs in healthcare.

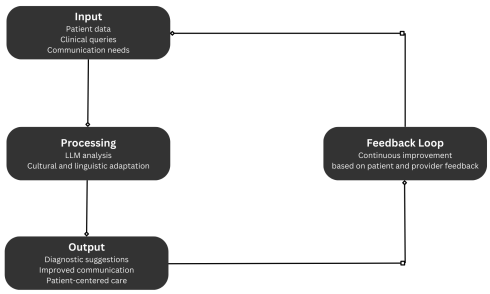

Figure 3: Workflow of LLM integration in healthcare

5 COMPARATIVE ANALYSIS OF THE MOST PROMINENT LLMS IN HEALTHCARE

Our review would be incomplete if we failed to mention the most prominent large language models that have been fine-tuned for

Model Name	Primary Applications	Cultural and Linguistic Adaptations	Current Limitations	Future Research Needs
GPT-4	Diagnostic support, patient communication	Multilingual capabilities, cultural sensitivity	Bias, accuracy issues	Domain-specific adaptations, ethical frameworks
L2M3	Health equity, multilingual support	Designed for multiple languages, cultural context integration	Limited domain-specific fine-tuning	Research on impact in low-resource regions
LlamaCare	Healthcare knowledge sharing	Tailored for healthcare terminology, multilingual support	May lack emotional sensitivity	Enhanced cultural adaptation, emotional sensitivity
GatorTron	Clinical decision-making, patient interaction	Adapted for diverse clinical contexts	Potential bias, limited contextual awareness	Improving contextual understanding and bias reduction
Additional Models (e.g., Me-LLaMA)	Specific clinical domains, patient interaction	Customizable for local languages and cultures	Challenges with generalizability	Further development for specialized clinical needs

Table 1: Comparison of LLMs in Healthcare Applications

applications in the medical domain. Therefore, this section will highlight the strong points and then perform a comparative analysis between GPT-4, LlamaCare, GatorTron, and L2M3. A graphic display of these models and their most prominent features in the medical domain can be found in Figure 4, while Table 1 below summarizes the models' strong points and current areas that can be improved, as well as areas where research can be focused in the future.

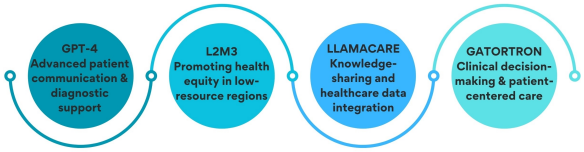

Figure 4: Overview of the key LLMs that are applied in healthcare

5.1 GPT-4

General Application in Medicine. As one of the most versatile LLMs, GPT-4 has been widely studied for its role in clinical decision-making, medical education, and patient engagement. GPT-4's broad applications, from summarizing medical records to assisting with diagnosis and patient interaction. However, GPT-4's lack of specialized medical training means it faces challenges when compared to models like LlamaCare and GatorTron[23].

Patient-Doctor Communication. GPT-4 has also demonstrated potential in improving patient communication by generating empathetic, human-like responses. The model can generate emotionally intelligent text, potentially enhancing patient trust in AI-generated advice[17]. Despite this, GPT-4 struggles with more complex medical queries where detailed clinical context is needed[24].

5.2 L2M3

Multilingual Healthcare Accessibility. One of L2M3's greatest strengths is its support for multilingual healthcare environments. L2M3 is a model designed specifically for low-resource healthcare settings where language barriers and cultural diversity pose significant challenges. By offering support for multiple languages, L2M3 has the potential to increase access to healthcare for marginalized communities globally[8].

Cultural Sensitivity. L2M3 goes beyond just linguistic adaptability by embedding cultural knowledge into its model[23]. This enables L2M3 to provide more contextually appropriate advice that aligns with local medical customs. For instance, it can recommend treatments or healthcare guidelines that resonate with the cultural practices of the patient's region, something that more generalized models like GPT-4 may fail to do effectively.

Consideration for Health Equity. L2M3 plays a pivotal role in reducing healthcare disparities by making culturally and linguistically appropriate care accessible in underdeveloped and diverse regions. It is particularly adept at filling gaps left by monolingual or culturally neutral models, such as GPT-4[23].

5.3 LlamaCare

Knowledge Sharing in Healthcare. LlamaCare was developed to facilitate knowledge sharing among healthcare professionals. Unlike GPT-4, which is designed for general applications, LlamaCare is fine-tuned specifically for healthcare, giving it an edge in clinical decision support[29]. LlamaCare's training on specialized medical datasets makes it particularly useful for knowledge-intensive tasks such as diagnosing complex conditions or synthesizing information from clinical trials[12].

Fine-tuned for Medical Data. LlamaCare's ability to provide accurate, context-specific information gives it a significant advantage over more general models. Its precision comes from training on large volumes of healthcare-specific datasets, allowing it to outperform models like GPT-4 when it comes to specialized clinical decision-making[26].

Patient Communication. Although LlamaCare's primary role is to assist healthcare providers, it can also be used to improve patient communication by offering detailed and reliable medical information. However, unlike GPT-4, which is more conversational, LlamaCare's focus remains on delivering precise medical knowledge rather than generating empathetic dialogue[4].

5.4 GatorTron

Focus on Medical Records. GatorTron is being specifically designed for the U.S. healthcare system, where it excels in processing vast amounts of electronic health records (EHRs). Its ability to rapidly synthesize and analyze patient data has made it an essential tool for improving diagnostic accuracy and reducing administrative

burden[2]. GatorTron outperforms models like GPT-4 in this regard, owing to its fine-tuning on clinical records[7].

Enhancing Clinical Workflows. One of GatorTron's key strengths is its ability to streamline clinical workflows by accurately summarizing medical histories, diagnoses, and treatment plans.[14]. GatorTron enhances clinical efficiency by reducing the amount of time doctors spend on administrative tasks like reviewing patient charts, allowing them to focus more on patient care.[16].

Adaptation to U.S. Healthcare. GatorTron's design is optimized for the U.S. healthcare system, making it particularly effective in this context. However, this focus on English-language records means it may not be as adaptable in multilingual or international healthcare environments. GatorTron's success illustrates the need for more research on how to adapt models to diverse healthcare systems[27].

5.5 Comparative Analysis

Local Medical Customs and Linguistic Demands. When comparing L2M3 and GatorTron, L2M3's strength in multilingual settings is highlighted, which makes it ideal for global healthcare applications. GatorTron, by contrast, is more specialized for English-speaking environments. GPT-4 offers broader utility but lacks the cultural and linguistic specificity of L2M3 and the clinical precision of GatorTron and LlamaCare[23].

Patient Communication. GPT-4 and LlamaCare both demonstrate potential in patient-doctor communication, but they serve different purposes. GPT-4 excels in generating empathetic responses, while LlamaCare offers more medically precise information. L2M3's focus on multilingual communication gives it an edge in culturally diverse settings, but GatorTron's role remains more data-focused[17][9][26].

Customization and Accuracy. Both GatorTron and LlamaCare excel in accuracy due to their fine-tuning on medical data[30]. GPT-4, while highly versatile, does not have the same level of specialization[23]. L2M3, on the other hand, is a standout for global healthcare, particularly in low-resource regions[3].

6 CHALLENGES

6.1 Technical Challenges of Implementing Medical LLMs

Despite promising applications, adapting LLMs to specific medical tasks presents technical challenges. A major issue is the need for domain-specific training data. The models need considerable fine-tuning for clinical natural language processing (NLP) tasks, making zero-shot learning an emerging solution[20]. When it comes to the application of zero-shot learning in preventive healthcare, it has an ability to provide accurate decision support even in niche medical contexts[13]. In addition, fine-tuning LLMs to accommodate local medical customs and linguistic variations is critical for effective patient communication across diverse healthcare settings. That is why multilingual models like L2M3, which cater to low-resource regions where local dialects and cultural practices significantly influence healthcare delivery, have such importance[3]. Without incorporating local linguistic demands and medical traditions, LLMs risk misinterpretation, leading to misdiagnosis or poor patient outcomes.

6.2 Ethical Implications of LLMs in Healthcare

The ethical implications of deploying LLMs in healthcare extend beyond technical challenges, touching on critical issues such as data privacy, patient consent, and accountability. are substantial, particularly in terms of biases, transparency, and patient autonomy. While current studies briefly mention these concerns, this review delves deeper into specific ethical frameworks that should guide the development and deployment of LLMs. One such framework is the "Data Protection by Design" principle, which emphasizes the need to integrate privacy safeguards into AI systems from the outset. In addition, patient consent must be a core element when implementing LLMs in clinical settings, ensuring that patients are fully aware of how their data is used and stored. Accountability in AI decision-making also needs to be addressed, particularly in high-stakes scenarios like diagnostics and treatment planning, where errors can have life-altering consequences. Finally, the risk of bias and perpetuating healthcare inequities through LLMs necessitates stricter ethical oversight, with transparent mechanisms for identifying and mitigating bias. Racial and ethnic biases in GPT-4 were explored for medical diagnosis and triage, uncovering disparities in the model's responses. These concerns underscore the need for ethical oversight when using LLMs in culturally diverse settings[10]. Furthermore, LLMs could reinforce healthcare inequities if not properly adapted to the specific needs and practices of various cultures[6]. Additionally, LLMs must account for linguistic diversity in global healthcare contexts. For example, healthcare systems in multilingual nations, such as India, require models that can function across multiple languages while understanding the nuances of local medical customs. There is an innate link between culture, language, and patient safety, making linguistic competency critical for patient-centered care[11]. The LlamaCare model provides a framework for sharing healthcare knowledge across diverse linguistic groups, showing how LLMs can facilitate cross-cultural knowledge sharing in healthcare[22]. Figure 5 below focuses on the challenges, as opposed to the benefits, of applying LLMs in the healthcare domain.

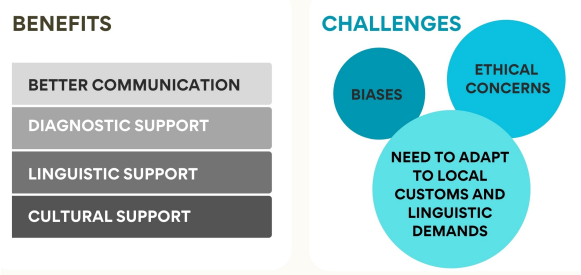

Figure 5: The benefits and challenges of applying large language models in the medical domain

7 EVALUATING THE EFFECTIVENESS OF LLMS IN HEALTHCARE

The evaluation of LLMs in healthcare requires a combination of qualitative and quantitative metrics to ensure a balanced assessment of their performance. In addition to the qualitative analysis

of their language generation capabilities, this review incorporates quantitative metrics such as accuracy rates, error rates in specific tasks (e.g., diagnosis or patient communication), and patient satisfaction scores derived from recent studies. These metrics provide a robust framework[9] for comparing LLMs across various clinical applications. For example, models like GPT-4 and LlamaCare have been evaluated for their diagnostic accuracy, while GatorTron has shown promise in reducing administrative burdens through improved EHR management. Quantitative evidence, such as the accuracy rates of LLMs in diagnosing rare conditions or the error rates in patient-facing applications, strengthens the argument for their continued refinement and deployment in healthcare.

Evaluating LLMs in medical applications, focusing on adaptability to local linguistic demands, accuracy, and transparency requires a detailed framework[9]. There are METRICS in place that serve as a tool for standardizing LLM evaluations, ensuring they are tailored to local healthcare systems and relevant cultural needs[19]. In multilingual and multicultural settings, the evaluation of LLMs should prioritize their ability to maintain cultural and linguistic appropriateness. The L2M3 model, designed for regions with limited healthcare resources, demonstrates how multilingual and culturally aware LLMs can improve healthcare outcomes in underserved communities[3].

8 GAPS IN THE EXISTING RESEARCH

As LLMs become more prevalent in clinical environments, the focus on improving patient-doctor communication and addressing cultural and linguistic considerations will be crucial. Overreliance on LLMs without adequate human oversight, particularly in sensitive clinical areas like diagnostics and treatment planning is dangerous[29]. Additionally, there is a need for stringent ethical guidelines to ensure that LLMs do not perpetuate healthcare inequities[6]. In the papers that were evaluated for this review, we have identified the most notable gap in the limited focus on multicultural and linguistic diversity. There is a need for more comprehensive studies and models specifically designed for non-Western and linguistically diverse healthcare settings. LLMs must be fine-tuned to local medical practices and languages to truly serve global healthcare needs. The second gap identified is the bias that exists in LLM healthcare recommendations[10][6]. There is a lack of clear, actionable frameworks for systematically identifying and reducing bias in LLMs. Addressing this will require a multidisciplinary approach combining AI ethics, clinical expertise, and sociocultural considerations. The third gap is the emotional sensitivity in patient-doctor communication[1][21][18]. Finally, there is the danger of over-reliance on LLMs for decision support[6][29]. Still, the field is quite promising, so exploring future directions for research is worthwhile.

9 FUTURE DIRECTIONS FOR RESEARCH

The potential of LLMs in healthcare remains largely untapped, with significant gaps in current research that need addressing. One critical area for future investigation is the development of LLMs that are emotionally sensitive and capable of handling high-stakes, emotionally charged patient interactions. For example, models could be designed to recognize emotional cues in patient language and adjust their responses to provide more empathetic care. Another vital research direction is improving the contextual awareness of LLMs, particularly in culturally diverse settings where understanding local customs, languages, and medical practices is essential. Furthermore, future research should explore ways to integrate LLMs with other healthcare technologies, such as electronic health records (EHRs), to streamline clinical workflows while ensuring data privacy. Finally, the creation of robust, standardized evaluation frameworks[19][9] will be crucial in assessing the long-term effectiveness of LLMs in healthcare, focusing on patient outcomes, model accuracy, and ethical and regulatory compliance[6][15].

10 CONCLUSION

LLMs hold immense potential to revolutionize healthcare by enhancing diagnostic support, improving patient-doctor communication, and facilitating equitable care. Their success, however, hinges on addressing challenges such as local medical customs, linguistic diversity, and ethical concerns. Proper adaptation of LLMs can significantly enhance patient-centered care by making communication between patients and healthcare providers more culturally sensitive and linguistically appropriate. Prioritizing inclusivity and transparency is essential for improving healthcare outcomes and equity. Future research must focus on addressing cultural and linguistic diversity, improving emotional sensitivity in patient-doctor interactions, mitigating biases, and establishing ethical and legal frameworks for AI in healthcare. Specialized research is also needed to tailor LLMs for specific clinical domains, such as mental health and surgery, to ensure these tools are safe, reliable, and contextually aware.

11 ACKNOWLEDGEMENT

This project is funded by the European Union under Horizon Europe (project ChatMED grant agreement ID: 101159214).

Funded by the European Union

REFERENCES

[1] Prottay Kumar Adhikary, Aseem Srivastava, Shivani Kumar, Salam Michael Singh, Puneet Manuja, Jini K Gopinath, Vijay Krishnan, Swati Kedia Gupta, Koushik Sinha Deb, and Tanmoy Chakraborty. Exploring the efficacy of large language models in summarizing mental health counseling sessions: Benchmark study. *JMIR Mental Health*, 11:e57306, 2024.

[2] Xi Chen. Large language models in the medical field: Principles and applications. *International Journal of Computer Science and Information Technology*, 2(3):219–224, 2024.

[3] Agasthya Gangavarapu. Introducing l2m3, a multilingual medical large language model to advance health equity in low-resource regions, 2024.

[4] Marius Geantă, Daniel Bădescu, Narcis Chirca, Ovidiu Cătălin Nechita, Cosmin George Radu, Stefan Rascu, Daniel Rădăvoi, Cristian Sima, Cristian Toma, and Viorel Jinga. The potential impact of large language models on doctor–patient communication: A case study in prostate cancer. In *Healthcare*, volume 12, page 1548. MDPI, 2024.

[5] Tawara D Goode, M Claire Dunne, and Suzanne Bronheim. *The evidence base for cultural and linguistic competency in health care.* Commonwealth Fund New York, NY, 2006.

[6] Joschka Haltaufderheide and Robert Ranisch. The ethics of chatgpt in medicine and healthcare: a systematic review on large language models (llms). *NPJ Digital Medicine*, 7(1):183, 2024.