A Systematic Review on Evaluating Bias and Equity in Large Language Model (LLM) Applications for Patient Communication in Healthcare

Gihantha Kavishan
Software Engineering Teaching Unit
University of Kelaniya
Dalugama,Sri Lanka
gihanthakavishan@gmail.com

Nimasha Arambepola
Software Engineering Teaching Unit
University of Kelaniya
Dalugama,Sri Lanka
nimasha@kln.ac.lk

Keywords—large language models (llms), patient communication, bias mitigation, equity in healthcare, systematic review.

A. Introduction

The rapid advancement of AI has transformed healthcare, with Large Language Models (LLMs) like ChatGPT and Med-PaLM enhancing patient communication by automating responses, summarizing medical documents, and aiding clinical decision-making, particularly in resource-limited settings. However, biases and inequities remain as critical challenges, undermining their effectiveness and fairness. Systemic biases can be culturally inappropriate, exacerbate healthcare disparities, and marginalize certain groups. Additionally, structural and economic factors contribute to these issues, necessitating urgent attention to demographic, cultural, and linguistic discrimination. This research aims to systematically evaluate these biases, their consequences, and potential solutions, offering recommendations to enhance AI driven healthcare communication systems, making them more just, efficient, and reliable.

B. Literature Review

This section describes the existing body of knowledge related to research. The integration of Large Language Models (LLMs) in healthcare has enhanced patient communication through real time, personalized interactions, improved education, and support in clinical decision-making [1]. Despite these benefits, significant challenges persist, particularly in ensuring fairness, inclusivity, and ethical use. Biases in training datasets can continue health disparities and negatively affect marginalized communities [2]. Demographic and algorithmic biases have been shown to result in inaccurate diagnoses and treatments [5], while language and cultural limitations reduce LLM effectiveness in diverse and multilingual settings [3], [4]. Ethical concerns, including lack of transparency, accountability, and informed consent, create additional barriers to safe implementation [6], [8], emphasizing the need for ethical governance and stakeholder involvement [7]. To mitigate these issues, researchers advocate for fairness aware AI practices, inclusive data training, and regular dataset audits [11], [15]. Additional strategies such as promoting data diversity, involving clinicians in oversight, and applying privacy preserving methods like federated learning can further reduce bias [12]. Ethical frameworks like BE FAIR offer guidance for equitable AI development [16], while continuous stakeholder engagement is key to fostering trust and ensuring responsible deployment [7], [13]. These measures can help healthcare systems deploy LLMs more effectively and equitably.

C. Materials and Methods

This study adopts the PRISMA framework to systematically review bias and equity in the application of Large Language Models (LLMs) for patient communication in healthcare. A thorough literature search was conducted across platforms including Google Scholar, PubMed, ResearchGate, Springer, Scopus, and IEEE Xplore, supplemented by bounds tracking to ensure completeness. The review focused on English language publications from 2017 to 2024 that examined bias and equity in LLM use within healthcare and included proposals aimed at mitigating these challenges. The keyword selection process was guided by a rationale that balanced scope with focus, ensuring the inclusion of relevant studies. This method ensures a comprehensive, fact based evaluation of current efforts to address bias and promote equity in healthcare communication through LLMs.

D. Results and Discussion

Large Language Models (LLMs) like ChatGPT and Med-PaLM are transforming healthcare communication by simplifying complex medical language, enabling personalized responses, and supporting services such as mental health care, education, and asynchronous consultations [1], [2]. Despite these advancements, key challenges limit widespread adoption particularly around fairness, accessibility, and ethics. Bias in training data leads to unequal outcomes, especially for underrepresented groups, deepening existing healthcare disparities [2], [3]. LLMs often

https://doi.org/10.31705/ADScAI.2025.04

lack adequate multilingual and cultural support, working best in dominant languages like English, which marginalizes non dominant language speakers [4], [5]. Barriers such as low digital literacy, limited internet access, and socioeconomic constraints affect equitable usage [3], [6]. Ethical gaps, including inadequate transparency, consent, and data protection, further complicate responsible deployment [7], [8]. Technical limitations, such as inconsistent responses and lack of real-world validation, can also pose risks to patient safety [9], [10].

To overcome these limitations, researchers are implementing several strategies. Fairness aware algorithms, rebalancing techniques, and adversarial training can mitigate bias during model development [2], [11], while regular dataset audits promote demographic inclusivity [11]. Participatory design, involving marginalized users, has led to more accurate, context aware systems, as seen in multilingual chatbot projects [7]. Enhancing multicultural competence through diverse, multilingual training data and demographic specific evaluation metrics ensures broader effectiveness [4], [5], with real world examples like a culturally adapted LLM in Southeast Asia improving health literacy among rural populations [1]. Clinical pilot testing and user feedback loops also improve reliability; one such loop reduced inaccurate responses in a Med-PaLM mental health assistant by 27% [2], [10]. Ethical and governance measures, such as the BE FAIR framework [8], privacy preserving methods like federated learning [12], and active stakeholder engagement [7], [13], are essential for trust and accountability. Moving forward, the focus should be on inclusive model design, improved accessibility, thorough clinical validation, and transparent governance to ensure LLMs benefit all patients equitably and safely.

E. Conclusion

This study highlights significant biases in Large Language Models (LLMs) used for patient mediation in healthcare, particularly in training datasets that generate inappropriate responses to bereaved patients, use poor English, and provide harmful healthcare recommendations. It also identifies barriers to equitable use, including a lack of contextualization for diverse groups, low digital literacy, and limited access in underserved regions. To address these issues, the study proposes enhancing dataset diversity, incorporating fairness evaluation criteria, and developing user-friendly interfaces for patients with low digital skills. Additionally, it emphasizes the need for real-world validation of LLMs, expanding multilingual options, and conducting multi-site studies to improve cultural and linguistic sensitivity. Future research should focus on ethical and legal frameworks for data protection and privacy to ensure safer and more inclusive LLM-assisted healthcare services.

REFERENCES

[1] S. Aydin, J. Zhou, and L. Sun, "The integration of Large Language Models (LLMs) in healthcare: Transforming patient communication," *J. Health Technol.*, vol. 22, no. 1, pp. 15–30, 2024.

[2] M. Poulain, A. Ghosh, and J. Li, "Biases in training datasets and their impact on healthcare LLMs," Health Equity J., vol. 18, no. 3, pp. 210–225, 2024.

[3] F. Norori, S. Velichkovska, and M. Yang, "Language and cultural barriers in LLM applications for healthcare," *Int. J. Med. Inform.*, vol. 109, pp. 12–19, 2021.

[4]] M. Yang, L. Sun, and S. Aydin, "The cultural limitations of LLMs in global healthcare settings," *Med. Inform. Decis. Mak.*, vol. 26, no. 4, pp. 334–343, 2024.

[5] S. Velichkovska, M. Yang, and Y. Zhou, "Algorithmic biases in healthcare LLM systems and their impact on patient care," *Artif. Intell. Med.*, vol. 43, no. 5, pp. 556–565, 2023.

[6] C. Goh, C. Okonji, and V. Reddy, "Ethical concerns in the application of LLMs in healthcare," *AI Ethics Rev.*, vol. 10, no. 1, pp. 45–52, 2023.

[7] R. Rodriguez, R. Nazer, and K. Okamoto, "Ethical AI governance and stakeholder engagement in healthcare LLM deployment," *AI Soc.*, vol. 11, no. 4, pp. 247–257, 2024.

[8] C. Okonji, V. Reddy, and C. Goh, "Transparency, accountability, and informed consent in healthcare LLMs," *J. Ethical AI Med.*, vol. 5, no. 2, pp. 130–141, 2024.

[9] M. Yang et al., "Med-PaLM in practice: AI communication support for mental health," *Med. Informatics Res.*, vol. 21, no. 3, pp. 210–219, 2023.

[10] A. Fraser et al., "Limitations of AI in clinical deployment: An empirical review," *Clin. AI J.*, vol. 10, no. 2, pp. 100–112, 2023.

[11]] C. Pfohl, A. Tejani, and M. Robinson, "Fairness-aware AI development and inclusive training data for healthcare LLMs," *Healthc. Technol. Ethics*, vol. 14, no. 2, pp. 91–102, 2024.

[12] K. Okamoto, "Federated learning for privacy-preserving healthcare applications," *J. AI Health Secur.*, vol. 6, no. 1, pp. 25–34, 2023.

[13]] R. Nazer, R. Rodriguez, and K. Okamoto, "Transparency and trust in LLM applications in healthcare," *J. Ethics AI*, vol. 9, no. 2, pp. 185–200, 2023.

[14] T. Chinta et al., "Challenges in scalability of bias mitigation methods," *J. AI Biomed.*, vol. 17, no. 2, pp. 44–58, 2024.

[15] B. Chen et al., "Dataset audits and fairness metrics in AI health systems," *AI Med. Ethics*, vol. 19, no. 3, pp. 77–89, 2024.

[16] J. Cary Jr. et al., "BE FAIR: An ethical framework for inclusive AI," *Ethical AI Rev.*, vol. 8, no. 1, pp. 19–32, 2024.