Exploring the Potential of Large Language Models (LLMs) in Healthcare

Subin Pulliyil Santhosh
The University of Adelaide
Adelaide, SA
a1917668@adelaide.edu.au

Abstract

Large Language Models (LLMs) are gaining attention in the healthcare field because of their ability to understand and generate human like text. Some studies show that LLMs like Chatgpt, Llama, etc can pass medical exams and help answer patient questions. But there are still many concerns about their accuracy, their safety and whether they are reliable enough to use in real healthcare situations. This project aims to see how useful LLMs are in healthcare. We will test different models like ChatGPT, Grok, Llama and Gemini on healthcare tasks like answering questions and summarising medical notes and we will also evaluate their performance. The goal of this project is to identify where they can help the most and what are their potential risks.

1. Introduction

Artificial Intelligence (AI) has become more common in many industries and healthcare as well. One area of AI that has really advanced is Large Language Models (LLMs). These models can understand and write text in a way that feels natural which is why people are excited about their use in many industries especially healthcare. These are tools like ChatGPT, Llama, Grok, etc that can write, summarise and answer questions almost like a human. In healthcare, LLMs have been used to explain medical terms to patients, help doctors with writing notes and even pass medical licensing exams (Nori et al., 2023).

Even though the LLMs have advanced, the LLMs are not still always reliable especially in healthcare. One of the biggest issues is that these models can 'hallucinate' which means they can sometimes give wrong or made up answers but still sound correct (Duong & Solomon, 2023) and in healthcare, that is a major problem because even small mistakes can become very serious for patients. Studies like Ayers et al. (2023) showed that ChatGPT gave useful and correct responses on public discussion, it also gave some answers that were somewhat inaccurate. Tam et al. (2024) also mentioned that most of the existing studies does not really test LLMs using real clinical tasks or datasets which means there is still a lot we do not know about how well these LLMs actually perform in real world healthcare situations. This project addresses a clear gap between the technical abilities of LLMs and their practical and safe application in healthcare.

This project focuses on how LLMs perform on practical healthcare tasks such as answering medical questions and summarising clinical notes. The goal is to test different LLMs using real or public datasets and evaluate how accurate, safe and useful are their responses. Models like ChatGPT, LLaMA and Grok will be examined through specific use cases to find out where these tools can be most helpful and where they may or may not be accurate. The findings will help provide a insight of how LLMs will be useful in healthcare and how they can be improved in the future.

Research Questions:
1. What are the current strengths and limitations of LLMs in healthcare?
2. How can LLMs be used in tasks like summarising notes or answering medical questions?
3. What are the risks and how can we evaluate and reduce them?

Key Contributions:
- A literature review of existing research on LLMs in healthcare.
- Testing and evaluation of open source LLMs using public healthcare datasets.
- Identification of performance gaps and areas for improvement in current LLMs healthcare applications.

2. Literature Review

LLMs have advanced in healthcare sector due to their ability to do natural language processing which is a main thing in healthcare. The use of ChatGPT in passing medical licensing exams showed that these models have some medical knowledge (Kung et al., 2023). Also, some of the domain specific models such as ChatDoctor and BioGPT have shown improvement in accuracy in generating med-

ical dialogues and literature summaries (Li et al., 2023). Ayers et al. (2023) found that ChatGPT responses to patient queries on public discussions were rated more accurate than a normal physician responses. This shows that the LLMs may be useful for inital patient engagement and for the telemedicine chatbots. Similarly, Szabo et al. (2025) evaluated LLMs in teaching medical history and found that their responses were helpful in medical education. Yang et al. (2023) demonstrated 'Zhongjing' the first Chinese medical Llama-based LLM in responding to clinical queries in Chinese which shows the the multilingual potential of LLMs. Duong and Solomon (2023) found that LLMs underperform in topics requiring precise genetic knowledge. Bedi et al. (2025) showed how hallucinations can mislead people if they are not properly checked. Also, Tam et al. (2024) showed that most studies lack real world testing and they only rely on the old datasets. Recent work has also explored multimodal models integrating text and image inputs, such as Radiology-LLaMA (Liu et al., 2023). While these would help in imaging and diagnostic support, they also have many technical difficulties. Study by Mumtaz et al. (2024) shows the importance of proper dataset selection and the need for human interaction when deploying LLMs in healthcare environment.

Literature Review Research Questions:

1. What kinds of LLMs are currently being developed and evaluated for healthcare?
2. What tasks have LLMs been successfully applied to and what are the limitations?
3. What are the research gaps in evaluating LLMs in real-world healthcare scenarios?

3. Methodology

This project will adopt a five-phase methodology for implementation of LLM applications in healthcare. The complete process is shown in the methodology flowchart (Figure 1). The approach is to get good insights for both practical testing and conceptual analysis.

Phase 1: Identification of Tasks
We will first identify specific healthcare related tasks where LLMs could be useful. These may include medical question answering (QA) and summarisation of clinical notes.

Phase 2: Dataset Selection
Publicaly available datasets such as MedMCQA (Agrawal et. al. 2022), Pub-MedQA (Qiao et. al. 2019) and synthetic clinical notes will be used. These datasets will help evaluate LLMs in medical scenarios and reduce ethical risks with the patient data.

Phase 3: Model Setup
We will use both general purpose like ChatGPT, LLaMA etc and domain specific like ChatDoctor, Grok, etc LLMs. These models will be tested using structured prompts similar to the tasks identified.

Phase 4: Evaluation Design
Each models outputs will be assessed using standard NLP metrics like BLEU, ROUGE and F1-score. We will also do human evaluation like using the public health dataset to check the clinical relevance and safety.

Phase 5: Analysis and Interpretation
The results will be analysed to identify trends, limitations and opportunities. We will check how the LLMs responded under different tasks and do comparisons between models and extract lessons for future implementation.

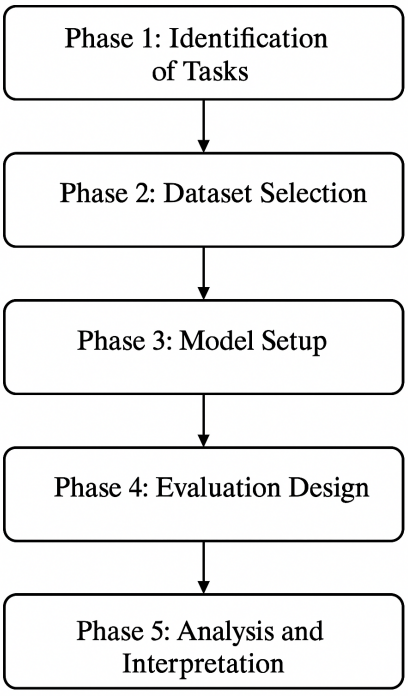

Figure 1. Methodology

4. Experimental Evaluation

In experimental phase, we will test the LLMs in two tasks such as summarisation of clinical notes and patient query answering. Each output will be compared for correctness and usability. Comparison and review with trusted sources like Medline (U.S. National Library of Medicine, 2025) or Mayo Clinic (Mayo Clinic, 2025) will be conducted. Each model will be tested using the same input datasets to get a fair comparison.

For QA tasks, we will check if the LLM can provide accurate responses to common symptoms or common conditions. Model outputs will be evaluated using objective metrics such as ROUGE and BLEU scores for summarisation and accuracy. For summarisation, we will check that all key points are summarised. For chat tasks, we will check the tone and clarity. This testing will help in understanding how LLMs might behave in healthcare sector.

5. Expected Outcomes

We expect LLMs to be useful in basic healthcare related tasks like note summarisation and patient communication. These models could help doctors by automating repetitive paperwork and making hard to understand language easier for patients. However, we may also see challenges such as inconsistent output, hallucinations and performance issues in complex queries. Domain specific models like ChatDoctor would outperform general purpose models in accuracy and understanding. The project aims to identify when and how LLMs can be used in healthcare and provide recommendations for improving safety and effectiveness.

6. Timeline

The timeline for this project is shown in the Gantt chart (Figure 2). It outlines when each part of the project will happen like literature review, proposal writing, implementation, testing, analysing results and final report writing. Milestones like proposal submission, presentation and progress report are also included.

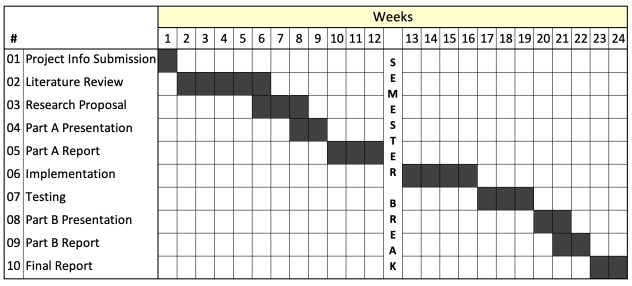

Figure 2. Timeline

7. Ethics

There is no formal ethics requirement for this project as we will only use open source models and publicaly available datasets. No real patient data will be used. However, we will still check that the inputs and outputs are handled correctly and do not include personal or sensitive content.

8. Cybersecurity

This project is not focused on cybersecurity. It is important to consider potential risks as the project does not involve sensitive data. However, basic precautions will still be considered. Even though no sensitive or personal data will be used, there is always a small chance that a system could be breached or misused. If the LLMs or systems used are breached or hacked, there is a chance of misuse or a chance manipulation of the models outputs. We will also document any unusual or potentially harmful responses from the models. These steps will help make sure the project stays safe and responsible even though cybersecurity is not the main focus.

References

[1] Agrawal, A., Patil, D., Rajagopal, R., & Goyal, P., 2022. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. arXiv preprint arXiv:2203.14991.

[2] Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., Faix, D. J., Goodman, A. M., Longhurst, C. A., Hogarth, M., & Smith, D. M. (2023). Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. JAMA internal medicine, 183(6), 589–596. https://doi.org/10.1001/jamainternmed.2023.1838

[3] Cascella, M., Montomoli, J., Bellini, V., & Bignami, E. (2023). Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios. Journal of medical systems, 47(1), 33. https://doi.org/10.1007/s10916-023-01925-4

[4] Duong, D., & Solomon, B. D. (2024). Analysis of large-language model versus human performance for genetics questions. European journal of human genetics : EJHG, 32(4), 466–468. https://doi.org/10.1038/s41431-023-01396-8

[5] Yang, Songhua & Zhao, Hanjia & Senbin, Zhu & Zhou, Guangyu & Xu, Hongfei & Jia, Yuxiang & Zan, Hongying. (2023). Zhongjing: Enhancing the Chinese Medical Capabilities of Large Language Model through Expert Feedback and Real-world Multi-turn Dialogue. 10.48550/arXiv.2308.03549.

[6] Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, Madriaga M, Aggabao R, Diaz-Candido G, Maningo J, Tseng V. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023 Feb 9;2(2):e0000198. doi: 10.1371/journal.pdig.0000198.

PMID: 36812645; PMCID: PMC9931230.

[7] Li, Y., Li, Z., Zhang, K., Dan, R., Jiang, S., & Zhang, Y. (2023). ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge. Cureus, 15(6), e40895. https://doi.org/10.7759/cureus.40895

[8] Bedi, S., Liu, Y., Orr-Ewing, L., Dash, D., Koyejo, S., Callahan, A., Fries, J. A., Wornow, M., Swaminathan, A., Lehmann, L. S., Hong, H. J., Kashyap, M., Chaurasia, A. R., Shah, N. R., Singh, K., Tazbaz, T., Milstein, A., Pfeffer, M. A., & Shah, N. H. (2025). Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review. JAMA, 333(4), 319–328. https://doi.org/10.1001/jama.2024.21700

[9] Liu, Z., Zhong, A., Li, Y., Yang, L., Ju, C., Wu, Z., Ma, C., Shu, P., Chen, C., Kim, S., Dai, H., Zhao, L., Sun, L., Zhu, D., Liu, J., Liu, W., Shen, D., Li, X., Li, Q., & Liu, T. (2023). Radiology-GPT: A large language model for radiology (arXiv:2306.08666). arXiv. https://doi.org/10.48550/arXiv.2306.08666

[10] Tam, T.Y.C., Sivarajkumar, S., Kapoor, S. et al. A framework for human evaluation of large language models in healthcare derived from literature review. npj Digit. Med. 7, 258 (2024). https://doi.org/10.1038/s41746-024-01258-7

[11] Ummara Mumtaz, Awais Ahmed, Summaya Mumtaz. LLMs-Healthcare: Current applications and challenges of large language models in various medical specialties. Artificial Intelligence in Health 2024, 1(2), 16–28. https://doi.org/10.36922/aih.2558

[12]Nori, Harsha & King, Nicholas & McKinney, Scott & Carignan, Dean & Horvitz, Eric. (2023). Capabilities of GPT-4 on Medical Challenge Problems. 10.48550/arXiv.2303.13375.

[13] Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., Agüera Y Arcas, B., … Natarajan, V. (2023). Large language models encode clinical knowledge. Nature, 620(7972), 172–180. https://doi.org/10.1038/s41586-023-06291-2

[14] Szabo, Aniko & Dolatkhah, Ghasem. (2025). Comparative Evaluation of Large Language Models for Medical Education: Performance Analysis in Urinary System Histology.. 10.21203/rs.3.rs-6186253/v1.

[15] Al Nazi, Zabir & Peng, Wei. (2024). Large Language Models in Healthcare and Medical Domain: A Review. Informatics. 11. 57. 10.3390/informatics11030057.

[16] Yang, R., Tan, T. F., Lu, W., Thirunavukarasu, A. J., Ting, D. S. W., & Liu, N. (2023). Large language models in health care: Development, applications, and challenges. Health care science, 2(4), 255–263. https://doi.org/10.1002/hcs2.61

[17] Omiye, J. A., Gui, H., Rezaei, S. J., Zou, J., & Daneshjou, R. (2024). Large Language Models in Medicine: The Potentials and Pitfalls : A Narrative Review. Annals of internal medicine, 177(2), 210–220. https://doi.org/10.7326/M23-2772

[18] Al-Garadi, Mohammed & Mungle, Tushar & Ahmed, Abdulaziz & Sarker, Abeed & Miao, Zhuqi & Matheny, Michael. (2025). Large Language Models in Healthcare. 10.48550/arXiv.2503.04748.

[19] Yu, P.; Xu, H.; Hu, X.; Deng, C. Leveraging Generative AI and Large Language Models: A Comprehensive Roadmap for Healthcare Integration. Healthcare 2023, 11, 2776. https://doi.org/ 10.3390/healthcare11202776

[20] AlSaad, R., Abd-Alrazaq, A., Boughorbel, S., Ahmed, A., Renault, M. A., Damseh, R., & Sheikh, J. (2024). Multimodal Large Language Models in Health Care: Applications, Challenges, and Future Outlook. Journal of medical Internet research, 26, e59505. https://doi.org/10.2196/59505

[21] Nassiri, K., & Akhloufi, M. A. (2024). Recent Advances in Large Language Models for Healthcare. BioMedInformatics, 4(2), 1097-1143. https://doi.org/10.3390/biomedinformatics4020062

[22] Haltaufderheide, J., & Ranisch, R. (2024). The ethics of ChatGPT in medicine and healthcare: a systematic review on Large Language Models (LLMs). NPJ digital medicine, 7(1), 183. https://doi.org/10.1038/s41746-024-01157-x

[23] Wang, L., Wan, Z., Ni, C., Song, Q., Li, Y., Clayton, E., Malin, B., & Yin, Z. (2024). Applications and Concerns of ChatGPT and Other Conversational Large Language Models in Health Care: Systematic Review. Journal of medical Internet research, 26, e22769. https://doi.org/10.2196/22769

[24] Peng, C., Yang, X., Chen, A., Smith, K. E., PourNejatian, N., Costa, A. B., Martin, C., Flores, M. G., Zhang, Y., Magoc, T., Lipori, G., Mitchell, D. A., Ospina, N. S., Ahmed, M. M., Hogan, W. R., Shenkman, E. A., Guo, Y., Bian, J., & Wu, Y. (2023). A study of generative large language model for medical research and healthcare. NPJ digital medicine, 6(1), 210. https://doi.org/10.1038/s41746-023-00958-w

[25] Karabacak, M., & Margetis, K. (2023). Embracing Large Language Models for Medical Applications: Opportunities and Challenges. Cureus, 15(5), e39305. https://doi.org/10.7759/cureus.39305

[26] Tian, Shubo & Jin, Qiao & Lai, Po-Ting & Zhu, Qingqing & Chen, Xiuying & Yang, Yifan & Chen, Qingyu & Kim, Won & Comeau, Donald & Dogan, Rezarta & Kapoor, Aadit & Gao, Xin & lu, Zhiyong. (2023). Opportunities and Challenges for ChatGPT and Large Language Models in Biomedicine and Health. ArXiv.

[27] Qin, H., & Tong, Y. (2025). Opportunities and Challenges for Large Language Models in Primary Health Care. Journal of primary care & community health, 16, 21501319241312571. https://doi.org/10.1177/21501319241312571

[28] Raghu Subramanian, C., Yang, D. A., & Khanna, R. (2024). Enhancing Health Care Communication With Large Language Models-The Role, Challenges, and Future Directions. JAMA network open, 7(3), e240347. https://doi.org/10.1001/jamanetworkopen.2024.0347

[29] Busch, F., Hoffmann, L., Rueger, C., van Dijk, E. H., Kader, R., Ortiz-Prado, E., Makowski, M. R., Saba, L., Hadamitzky, M., Kather, J. N., Truhn, D., Cuocolo, R., Adams, L. C., & Bressem, K. K. (2025). Current applications and challenges in large language models for patient care: a systematic review. Communications medicine, 5(1), 26. https://doi.org/10.1038/s43856-024-00717-2

[30] U.S. National Library of Medicine. (2025). MEDLINE: Description of the database. National Institutes of Health. https://www.nlm.nih.gov/bsd/medline.html

[31] Mayo Clinic. (2025). Patient care and health information. Mayo Foundation for Medical Education and Research. https://www.mayoclinic.org

[32] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A Dataset for Biomedical Research Question Answering. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.