THE UNIVERSITY OF ADELAIDE

Research Project Report on

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

Subin Pulliyil Santhosh

A1917668

Supervisor: Dr. Hussain Ahmad

Master of Computer Science

Contents

List of Figures

List of Tables

Abstract

Large Language Models (LLMs) are gaining significant attention in the healthcare field because of their ability to understand and generate human-like text also making them useful for various medical tasks. Recent studies have shown that models such as ChatGPT, LLaMA, Grok and Gemini can pass parts of medical licensing exams and can also assist in making patient documentation and even respond to health-related queries in such a way that feels natural and helpful. These capabilities have helped for reducing the workload on healthcare professionals. Despite their strengths, there are still many concerns about their accuracy, potential biases, safety and reliability in real-world healthcare settings where decisions can have serious consequences. This project focuses on testing different LLMs on tasks such as summarising patient notes and answering medically-relevant questions. Publicly available datasets like MedMCQA and PubMedQA will be used to evaluate their performance using standard metrics such as ROUGE, BLEU and accuracy. By comparing the outputs of different models, this research aims to highlight where LLMs can support healthcare professionals and where they still fall short and need assistance. The main goal is to understand where these models perform well and where they struggle and what risks or limitations need to be considered before they can be safely integrated into everyday healthcare settings.

1 Introduction

1.1 Background

Artificial Intelligence has been rapidly advancing in recent years and one of its most powerful development is the creation of Large Language Models (LLMs). These models which are trained on big amounts of text data can generate good and optimised and human-like responses. Tools like ChatGPT, LLaMA, Grok and Gemini have become popular for a wide range of applications including writing assistance, coding, customer service and in healthcare as well.

In the medical field, LLMs have shown potential to support doctors, researchers and patients by simplifying complex medical information and generating patient summaries and even assisting in diagnosis or clinical decision-making. As shown in Figure 1, LLMs are being applied across multiple areas of healthcare including patient care, pharmaceutical research, education and clinical research. Their use cases range from summarising patient notes and supporting mental health communication to helping with drug information, research assistance and regulatory documentation. For example, models like ChatGPT have successfully passed the United States Medical Licensing Examination (USMLE) and answered health-related queries with a reasonable level of accuracy (Kung et al., 2023; Nori et al., 2023). Domain-specific models such as ChatDoctor and Radiology-GPT have been developed to further improve performance by focusing on medical conversations and radiological tasks respectively.

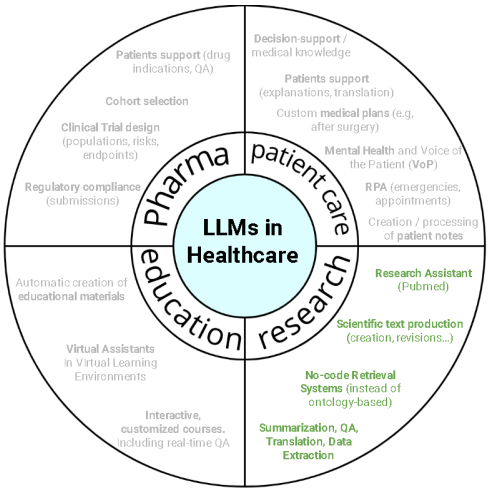

Figure 1: Applications of Large Language Models (LLMs) in Healthcare

Despite these developments, there are concerns about their real-world implementation. Issues such as hallucinations (generating incorrect or misleading information), lack of ethical

and privacy concerns and limited evaluation in healthcare still has challenges to their safe deployment in healthcare environments. Most existing studies focus on technical performance, leaving a gap in understanding how these tools can be responsibly used in real-world healthcare settings.

1.2 Motivation

The rise of Large Language Models (LLMs) like ChatGPT, Grok, LLaMA and Gemini has opened new possibilities for improving healthcare communication, clinical documentation and patient education. These models are already being used to summarise information, answer health-related queries and assist medical professionals with administrative tasks. However, despite their progress, LLMs are still not always reliable especially in healthcare. One of the biggest concerns is that these models can 'hallucinate' which means they sometimes provide wrong or fully made-up answers while sounding very confident (Duong & Solomon, 2023). There are several types of hallucinations as shown in Figure 2 that can occur when using LLMs including factual errors, contradictions and random insertions. Understanding these types is important when evaluating LLM performance in medical settings. In medicine, even small mistakes can have serious consequences for patients.

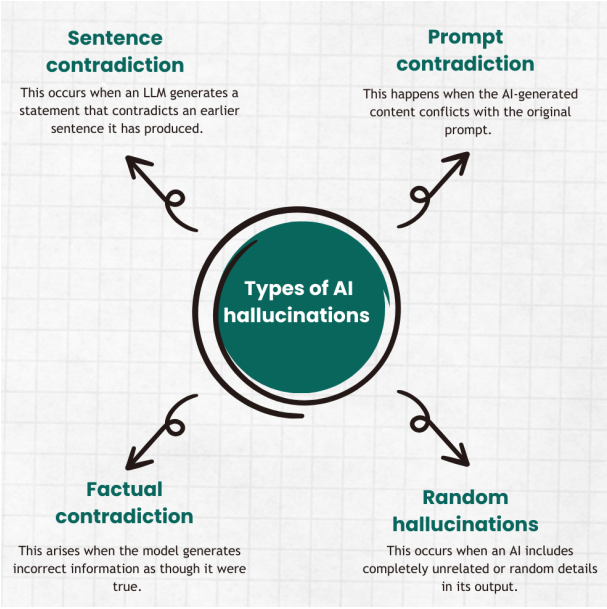

Figure 2: Common types of hallucinations in AI language models.

This project aims to fill that gap by evaluating the performance of different LLMs on real or publicly available healthcare datasets. The focus is on practical tasks such as summaris-

ing clinical notes and answering medical questions. By testing models like ChatGPT, LLaMA, Grok and Gemini, the goal is to understand how accurate, safe and useful their outputs really are in healthcare settings. The project also aims to find out in what areas these tools can be most helpful and where they might still fall short or need improvement.

Based on all these, the research questions are:

1. How do LLMs like ChatGPT, LLaMA, Grok, and Gemini perform well in specific healthcare tasks such as summarising patient records and answering clinical questions?
2. What are the strengths and weaknesses of these models when applied to healthcare datasets like MedMCQA and PubMedQA?
3. What are risks with the use of LLMs in real-world healthcare and how can their performance and safety be evaluated?

1.3 Contribution

This project makes several important contributions to the growing field of LLMs in healthcare. Firstly, this project reviews a wide range of recent research studies more than 26 academic papers to understand how these language models are currently being used in medical tasks. It highlights the progress and the ongoing challenges reported by different researchers. Secondly, the project tests a group of popular LLMs including ChatGPT, Grok, LLaMA and Gemini on real medical datasets like MedMCQA and PubMedQA. These datasets include multiple-choice medical questions and health-related texts which helps to assess how well each model can handle tasks like answering questions or summarising medical notes. The project compares how each model performs and identifies which ones are more accurate, consistent or useful in different healthcare tasks by using standard evaluation tools like BLEU and ROUGE scores. Finally, this project also looks at potential risks like wrong or biased answers and discusses where LLMs might safely be used and where more caution is needed or any assistance is needed. The main goal is to see how reliable and helpful these models are in healthcare and to give simple and practical ideas on how to make them better and safer for future use.

2 Literature Review

2.1 Methodology

To conduct a comprehensive literature review for this project, a structured and systematic approach was followed. The goal was to gather a wide range of studies that see how LLMs are applied in healthcare and their limitations and improvements. This was done to make sure that the selected literature was relevant enough and covered different aspects of LLM use in healthcare. This helped build an understanding of current trends, challenges and opportunities in applying LLMs to healthcare.

2.1.1 Research Questions

A set of research questions was made to make sure that the literature review remained focused and aligned with the objectives of the study. These questions are designed to guide the selection and evaluation of relevant academic literature. The selection of these research questions was made using the identified gaps in recent LLM-related healthcare studies and the main aim of evaluating the effectiveness, reliability and risks of LLM applications in healthcare.

The research questions were designed to cover three things: current applications, limitations and risks, and research gaps. Table 1 below summarises each research question along with the specific motivation behind it.

Table 1: Research Questions and Their Motivations

Research Question	**Motivation**
RQ1: How are LLMs currently being used in healthcare?	To understand the current practical uses of LLMs and how they support healthcare tasks like diagnosis, documentation and communication.
RQ2: What are the strengths and weaknesses of using LLMs in medical tasks?	To explore both the strengths (eg. summarisation, decision support) and challenges (eg. hallucinations, safety concerns) faced in healthcare.
RQ3: What are the current gaps, risks or challenges in using LLMs in real healthcare scenarios?	To identify areas where LLMs need improvement or further testing before being safely used in healthcare.

2.1.2 Search Strategy

A systematic search strategy was used to identify relevant academic papers. The main databases used were:

- Google Scholar
- PubMed
- IEEE Xplore
- SpringerLink
- Scopus
- ACM Digital Library

Keywords such as "Large Language Models in healthcare," "ChatGPT medical", "LLMs patient care", and "AI hallucinations in healthcare" were used. The boolean operators like 'AND' and 'OR' were used to refine the searches. To make sure no relevant paper was missed, a 'snowballing' method was also used. Snowballing is basically checking the reference lists of selected papers to find more useful studies.

2.1.3 Inclusion and Exclusion Criteria

A set of inclusion and exclusion criteria was applied to make sure the quality and relevance of this review was good. Only those studies that directly addressed the research questions and focused on the use of LLMs in healthcare were selected. Peer-reviewed publications and research papers involving actual implementation, evaluation or analysis of LLMs in clinical and biomedical contexts were considered suitable. Studies published from 2020 onwards were included to get up-to-date insights. Meanwhile, articles not written in English, papers that lacked experimental validation and those that were only opinion-based like discussion reviews were excluded from the final review. The detailed criteria are shown in Table 1 below.

Table 2: Inclusion and Exclusion Criteria

Inclusion	**Exclusion**
Peer-reviewed articles and arXiv preprints	Non-English language studies
Studies from 2020 onwards	Articles not related to LLMs or healthcare
Focus on LLM evaluation or application in healthcare	Opinion pieces without systematic evaluation

2.1.4 Paper Selection

More than 100 articles were found initially. After going through the titles and abstracts, 45 were shortlisted. These were read fully and then finally 26 research papers were selected that matched all the criteria and were relevant to the research questions. These include studies on models like ChatDoctor, BioGPT, Med-PaLM, Grok and domain-specific applications like Radiology-GPT and ChatGPT-MedQA.

2.1.5 Data Extraction and Analysis

Each paper was reviewed to extract details about:

- The type of LLM used (eg. general-purpose vs domain-specific).
- The datasets applied (eg. MedMCQA, PubMedQA).
- The task (eg. question answering, summarisation).
- Evaluation methods (BLEU, ROUGE, accuracy, F1-score, etc.).
- Key findings, benefits, risks and limitations.

The findings were grouped based on their use case categories (eg. clinical documentation, decision support, patient communication). The results are shown in detail in the next section.

2.2 Findings

Recent research highlights the growing role of LLMs in various healthcare applications as seen in response to the first research question, how LLMs are currently being used in healthcare. These models such as ChatGPT, LLaMA, Grok and Gemini are being deployed in diverse use cases like from patient interaction to medical document summarisation. For instance, Ayers et al. (2023) compared ChatGPTs responses with physicians answers to patient questions and found that ChatGPT offered good and detailed replies which was preferred by users. Also, Cascella et al. (2023) noted the models usefulness in explaining clinical terms and improving patient communication. This suggests that LLMs can assist healthcare providers in reducing communication gaps and handling routine queries.

In terms of documentation, LLMs like ChatDoctor and BioGPT have shown value in summarising patient histories and simplifying discharge summaries and even producing draft reports from electronic health records. Tools such as Radiology-GPT demonstrate domain-specific strength, especially in radiology, by accurately interpreting image findings based on

reports and helping in diagnosis documentation. This aligns with RQ1, showing a application of LLMs in supporting both clinical decision-making and backend documentation.

However, addressing RQ2, which checks the benefits and limitations of using LLMs in medical tasks and many challenges have also been identified. One of the most widely cited limitations is the phenomenon of 'hallucination' where models generate inaccurate or generate fake or made up information. Duong and Solomon (2023) showed that LLMs may sound confident even when producing incorrect responses, which is particularly risky in medical contexts. This sentiment is showed by Tam et al. (2024) who said that despite promising results, real-world testing of these models is still minimal and few have been evaluated using real patient data.

Studies such as Nori et al. (2023) and Kung et al. (2023) showed that ChatGPT and similar models can pass medical licensing exams. This success does not means safe clinical practice especially when nuanced patient contexts are involved. Evaluation metrics like BLEU, ROUGE and F1-scores used in many of these studies focus on surface-level linguistic similarity rather than the clinical validity or accuracy of responses. As a result, a performance score that appears high may still mask serious content errors when applied in a healthcare context.

In addressing RQ3, which says about the gaps, risks or challenges in using LLMs in real healthcare scenarios, it becomes clear that ethical, regulatory and practical concerns form a major barrier. For example, the lack of transparency in how LLMs derive their outputs like the 'black box problem' which makes it difficult for healthcare professionals to trust the recommendations these tools provide. Also, existing models have not been trained on diverse demographic data which raises concerns of racial and gender bias in medical outcomes. Bedi et al. (2025) warn that biased or stereotypical outputs in clinical narratives could lead to harm, especially for underrepresented populations .

A number of papers have called for more rigorous real-world evaluation. Agrawal et al. (2022) researched 'MedMCQA' a large-scale dataset focused on medical multiple-choice questions to improve results across various LLMs. PubMedQA similarly provides domain-specific QA tasks derived from PubMed abstracts which offers a semi-realistic environment for evaluating the relevance of the answers.

Mumtaz et al. (2024) propose fine-tuning general-purpose LLMs with medical corpora to reduce hallucinations and Li et al. (2023) showed that incorporating human in the loop systems to ensure model outputs are validated before being shown to clinicians or patients. Several studies including Yang et al. (2023) and Wang et al. (2023) shows that collaboration between medical professionals and AI developers is essential to create a better evaluation frameworks.

2.3 Comparison and Limitations in Existing Literature

To better understand how this study differs and improves upon these earlier efforts, Table 3 presents a side-by-side comparison of key parameters between previous research and this project.

Table 3: Comparison of Previous Studies vs This Study

Parameter	**Previous Studies**	**This Study (2025)**
Tasks Addressed	Narrow focus: either QA or simulated case studies (Kung et al., 2023; Cascella et al., 2023)	Evaluates both QA and summarisation tasks on real/public datasets, reflecting practical use cases.
Datasets Used	Patient forum data, synthetic questions, or domain-specific datasets (often not standardised)	Uses benchmark datasets like MedMCQA and PubMedQA which are widely accepted for medical LLM evaluation.
Evaluation Methods	Mostly qualitative reviews, human feedback or basic accuracy	Has standard NLP metrics like BLEU, ROUGE, Accuracy and F1-score for better quantitative comparison.
Risk Assessment	Limited or no focus on hallucinations, biases, or misuse potential (Duong & Solomon, 2023)	Includes detailed risk assessment, highlighting hallucination risks, reliability concerns and model limitations.
Scope and Depth	Narrow, model-specific, or domain-specific insights (e.g., only ChatDoctor for PubMedQA)	Broader scope with cross-model analysis, multi-task evaluation and practical recommendations across general healthcare domains.
Output and Contribution	Observational insights across models without any improvement	Offers performance comparison, practical insights and proposes improvement areas for LLMs in healthcare using a systematic approach.

While the research highlights the use of LLMs in healthcare, it is important to recognise the common limitations found across previous studies. Many existing works tend to focus on a single model such as ChatGPT or ChatDoctor or specific domains like radiology or public health queries. Also some studies is based on qualitative analysis or simulated dialogues rather than

real medical datasets which reduces their use for clinical applications. The evaluation methods are also inconsistent in which most of them lack a unified approach to get the performance and only few studies have researched on question answering and summarisation together. Risk assessment related to hallucination, bias and reliability is also not researched enough.

Most previous studies looked for single LLMs on limited datasets using narrow or observational evaluation methods. Studies like Ayers et al. (2023) and Kung et al. (2023) focused on either qualitative reviews or narrow MCQ-based approach that ignored some metrics like F1-score or BLEU or ROUGE. Others like Liu et al. (2023) and Li et al. (2023) centered on domain-specific tasks without testing general models or comparing across multiple tasks.

So, this project offers a broader, more comprehensive review, combining the datasets (MedMCQA, PubMedQA), multiple evaluation metrics and a proper comparison of LLMs. This project checks for both benefits and risks including hallucinations and reliability concerns which many past studies do not mention at all or ignore. By evaluating multiple models across diverse healthcare tasks, this project fills a critical gap in the literature which will offer not just performance comparison but also many insights for real world use. This makes the current literature review both deeper and more practically relevant than previous studies.

3 Research Progress

3.1 Research Methodology

The research methodology for this project follows a structured five-phase approach to study how well LLMs like ChatGPT, LLaMA, Grok and Gemini work in healthcare. The aim is to go step-by-step from picking the right tasks and data to testing the models and understanding how they perform. The flowchart for this methodology is shown in Figure 3.

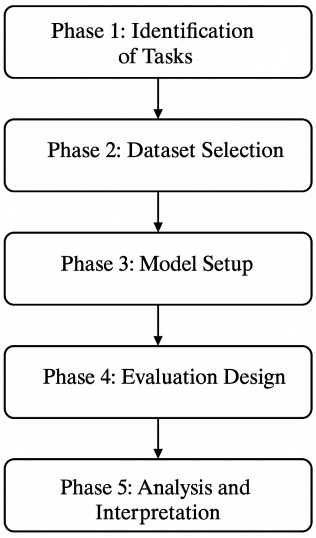

Figure 3: Flowchart of Five-Phase Methodology

- Phase 1: Identification of Tasks
  The first step was to figure out what kind of healthcare tasks LLMs could help in healthcare settings. After reading through recent research, two common tasks were chosen:
  - Question Answering (QA) – where the model answers medical questions like those found in exams or patient queries.
  - Summarising Medical Notes – where the model takes a long medical report or discharge note and turns it into something shorter and easier to understand.

  These tasks are useful in real-life healthcare settings and also have good datasets available for testing.

- Phase 2: Dataset Selection
  Next, we needed real healthcare data to test the models. Two public datasets were selected:

  - MedMCQA: a large collection of multiple-choice medical questions used in Indian medical entrance exams.
  - PubMedQA: a dataset with medical research questions based on real abstracts from the PubMed database.

  These were chosen because they are reliable, open-source and not used in many LLM healthcare studies.

- Phase 3: Model Setup
  Four different LLMs were picked for testing: ChatGPT, Grok, LLaMA and Gemini. Each model has its strengths and is popular in the AI. Simple prompt templates will bew made for both summarising and QA tasks so each model got similar input. Their answers were collected for scoring later. Models will be accessed through APIs or web tools depending on availability.

- Phase 4: Evaluation Design
  To measure how good each models answers were, different evaluation methods will be used:

  - ROUGE and BLEU scores to check how well summaries match with the expected outputs.
  - Accuracy and F1-score for the QA tasks to see how often the models got the answers right.

  All model outputs will be saved and compared side by side. Some model will also be manually reviewed to get errors or replies especially if the model hallucinated and made up things.

- Phase 5: Analysis and Interpretation
  Finally, all the results will be put together and analysed. This will include checking:

  - which models gave the most correct answers.
  - which ones made mistakes or gave vague responses.
  - whether domain-specific data (like PubMedQA) will help improve results.
  - any problems like hallucinations or biases.

  This phase will help to understand which model is best and also where each one works well or needs improvement.

3.2 Timeline and Milestone Plan

The project was structured across 24 weeks with work distributed between Semester 1 (weeks 4–12) and Semester 2 (weeks 13–24) as Part A and Part B research project. Each phase was planned with academic milestones such as proposal submission, presentation and final report. The Gantt chart in Figure 4 shows the full timeline.

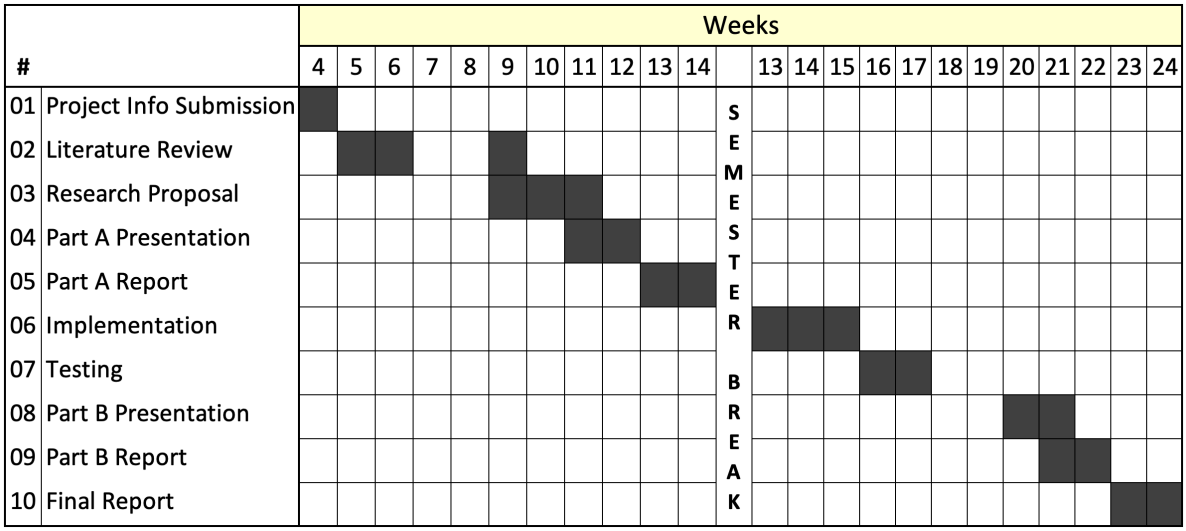

Figure 4: Gantt chart showing timeline and milestone plan for the Part A and Part B

In week 3, the project title was allocated and the supervisor was assigned. Project work began in week 4 and started with literature review and the development of a research proposal. From weeks 4 to 6, the main milestone was to read research papers and doing the literature review and planning the research questions. Weeks 9 to 10 were used to write the proposal and week 11 prepare for the Part A presentation. The Part A report was started in week 12. And currently we are in week 14 and in this week we have to submit the report. There is a semester break after this (shown as a blank space in the Gantt chart). Also there was a mid semester break of 2 weeks (week 7 and 8) and in next semester as well.

In Semester 2, starting from Week 13, project will be continued with setting up models, datasets (MedMCQA and PubMedQA) and creating prompts for testing from Weeks 13 to 16. Weeks 17 and 18 will be used to test the models and collect results using BLEU, ROUGE, accuracy and F1-score. In Weeks 19 and 20, the Part B presentation will be prepared. And in final weeks will be used to write and finish the final report.

4 Experimental Analysis

4.1 Initial Implementation Setup

Before using the datasets or conducting full-scale testing, an initial comparison was made to understand how different LLMs respond to a simple but realistic medical question. This phase involved no fine-tuning or dataset-based inputs and the focus was to observe how these models behave when presented with a healthcare query. The models selected for this early-stage test were:

- ChatGPT (general-purpose LLM)
- Grok (general-purpose but with conversational capabilities)
- ChatDoctor (domain-specific LLM designed for medical applications)

This comparison was to observe how general versus domain-specific models handle sensitive health-related information especially under minimal input.

4.2 Prompt Used and Model Responses

All three models were given the same input prompt to simulate a real-world scenario that might be asked by a patient refer Figure 5, 6, 7:

"I was having unexplained weightloss even though I used to eat clean. It was really hard for me to gain weight. I could not digest some foods. So, I did a colonoscopy in 2023. And the report said I have mild non-specific colitis. What chronic disorder am I suffering from? Give me the answer in one line."

Each model produced a different kind of response:

- ChatGPT and Grok provided direct and confident diagnoses which stated the user may have inflammatory bowel disease (IBD).
- ChatDoctor, on the other hand, responded cautiously. Instead of giving a diagnosis it acknowledged the presence of colitis but pointed out that the term 'non-specific' means the exact cause isn't clear. It then suggested follow-up with a healthcare provider which is a more appropriate, safe and human-like response in a medical context.

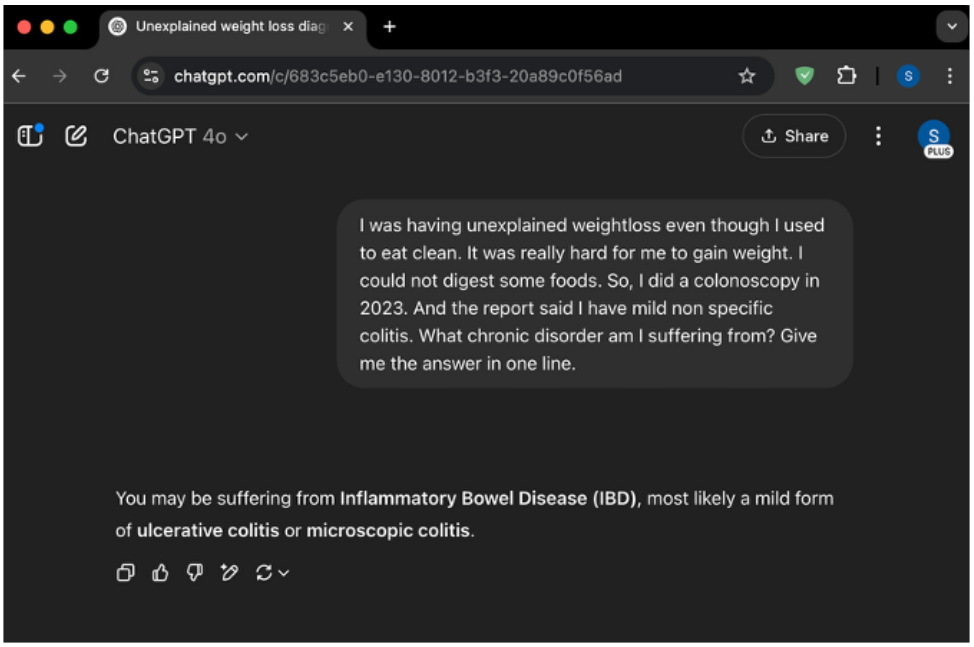

Figure 5: Response from ChatGPT to the given medical prompt

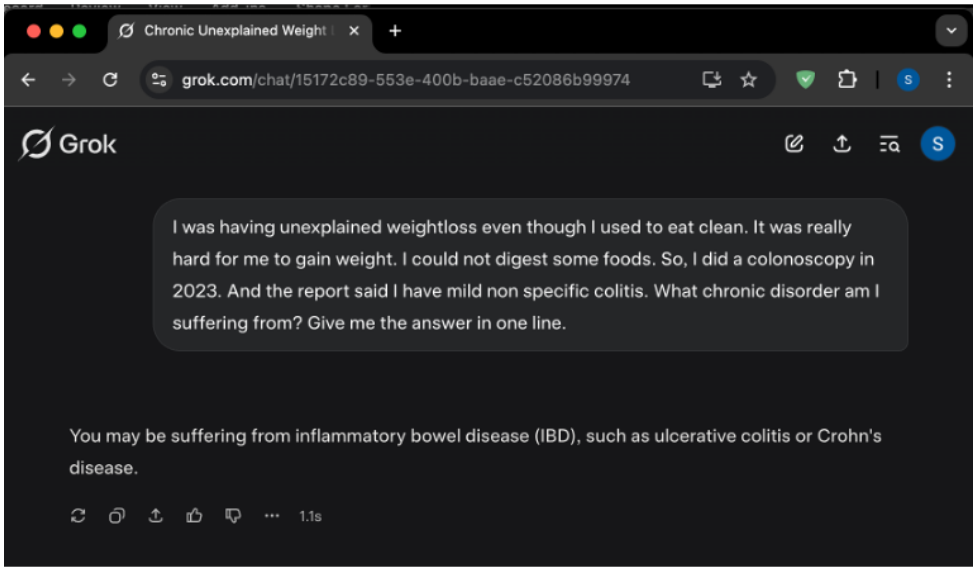

Figure 6: Response from Grok to the given medical prompt

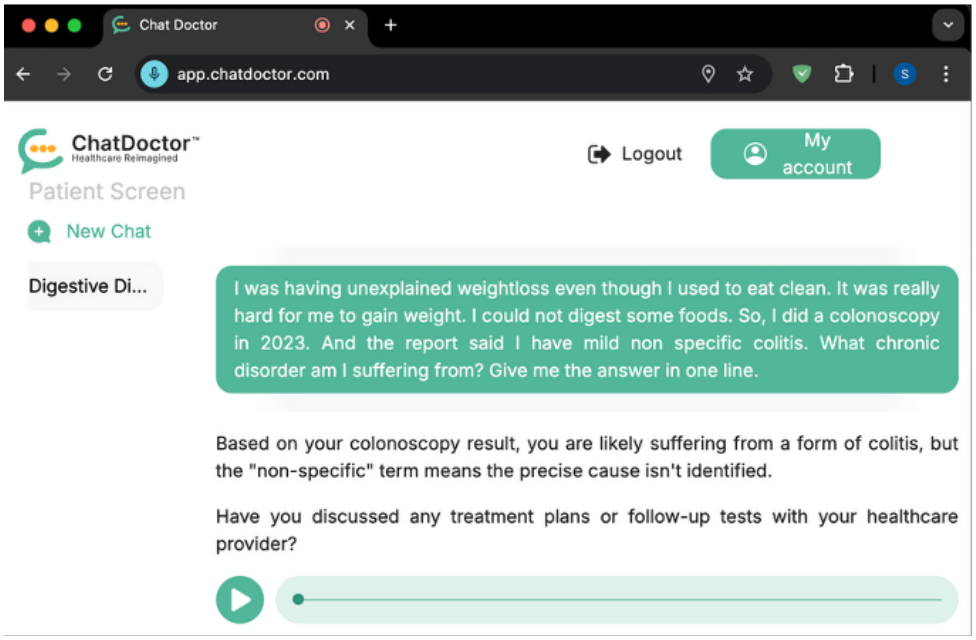

Figure 7: Response from ChatDoctor showing a more cautious and contextual reply

4.3 Observations and Insights

From this initial experiment, several important points were observed.

- General-purpose models (ChatGPT, Grok) tend to provide fast but overconfident answers. While technically correct in parts, these answers can be misleading or unsafe without proper follow-up especially in a clinical context.
- ChatDoctor showed greater responsibility. Instead of making assumptions, it asked follow-up questions and gave the uncertainty of diagnosis.
- This difference highlights a problem while generic LLMs are good at language generation but they lack the domain-aware safeguards needed in healthcare.
- These results support the importance of conducting structured evaluation using medical datasets which will be the next phase of this project.

All code, prompts, experiment logs and sample outputs will be shared in the following GitHub repository: https://github.com/subinsanthosh/llms-healthcare-project

5 Conclusion and Future Work

5.1 Conclusion

From the research till date, various LLMs fit for the project scope has been identified. The initial phase of the project focused on understanding the potential and limitations of LLMs in various healthcare applications and healthcare settings. A literature review was conducted from the week 4 to the 13th week to see the current capabilities of LLMs such as ChatGPT, LLaMA, Grok and Gemini in tasks like medical note summarisation and question answering. This phase also involved defining the research questions and identifying key use cases that align with the overall aim of the project. Based on insights gathered from the literature, open-access datasets such as MedMCQA and PubMedQA were shortlisted. These datasets are widely recognised and offer diverse challenges that reflect real-world medical tasks. Their selection ensures that future model evaluations will be both transparent and relevant to practical healthcare needs. During this initial phase different performance metrics such as BLEU, ROUGE, F1-score and accuracy which will be required to assess and analyse the model were identified. The Part A report and presentation is the combination of this research and planning phase. The work completed till date includes a well-documented literature review, the formulation of research objectives, the selection of datasets and the identification of comparing criteria. The Part A report and presentation collectively summarise these findings which will help for the upcoming implementation and testing phases in Semester 2.

5.2 Future Work

After the submission of the first report, the main focus of the project will shift from planning and research to implementation and evaluation. During the start of Semester 2 (Weeks 13 to 16) the main focus areas will be setting up the development environment and preparing the datasets for the model comparison. This includes designing prompt queries for interacting with various LLMs and loading the selected models like ChatGPT, Grok, LLaMA and Gemini with the prompts. All the models will be given healthcare-related tasks such as summarising clinical notes and answering medical questions. Efforts will be made to ensure reproducibility and consistency across all model evaluations.

Once the models are configured and tested, the next step is to measure the different model performances using appropriate evaluation metrics. Weeks 17 and 18 will focus on collecting and analysing model outputs using tools like BLEU and ROUGE for summarisation tasks. During this, the accuracy or F1-score for question answering tasks will also be evaluated

for all the models. The aim is to identify which model performs best for specific healthcare tasks and understand the limitations of each. This phase will also include documenting any inconsistencies or model behaviour that may raise any safety concerns. After this the models will be ranked based on their performances. Some of the main parameters include faster results and accurate results. The final project report will summarise the entire research from literature review to implementation and evaluation. This will also highlight the potential of LLMs in healthcare and provide suggestions for future research.

References

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

[2] J. W. Ayers, A. Poliak, M. Dredze, E. C. Leas, Z. Zhu, J. B. Kelley, D. J. Faix, A. M. Goodman, C. A. Longhurst, M. Hogarth, and D. M. Smith, "Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum", JAMA Internal Medicine, vol. 183, no. 6, pp. 589, 2023.

[3] M. Cascella, J. Montomoli, V. Bellini, and E. Bignami, "Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios", Journal of Medical Systems, vol. 47, no. 1, pp. 33, 2023.

[4] D. Duong and B. D. Solomon, "Analysis of large-language model versus human performance for genetics questions", European Journal of Human Genetics, vol. 32, no. 4, pp. 466–468, 2024.

[5] S. Yang, H. Zhao, Z. Senbin, G. Zhou, H. Xu, Y. Jia, and H. Zan, "Zhongjing: Enhancing the Chinese Medical Capabilities of Large Language Model through Expert Feedback and Real-world Multi-turn Dialogue", 2023.

[6] T. H. Kung, M. Cheatham, A. Medenilla, C. Sillos, L. De Leon, C. Elepaño, M. Madriaga, R. Aggabao, G. Diaz-Candido, J. Maningo, and V. Tseng, "Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models", PLOS Digital Health, vol. 2, no. 2, e0000198, 2023.

[7] Y. Li, Z. Li, K. Zhang, R. Dan, S. Jiang, and Y. Zhang, "ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge", Cureus, vol. 15, no. 6, e40895, 2023.

[8] S. Bedi, Y. Liu, L. Orr-Ewing, D. Dash, S. Koyejo, A. Callahan, J. A. Fries, M. Wornow, A. Swaminathan, L. S. Lehmann, H. J. Hong, M. Kashyap, A. R. Chaurasia, N. R. Shah, K. Singh, T. Tazbaz, A. Milstein, M. A. Pfeffer, and N. H. Shah, "Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review", JAMA, vol. 333, no. 4, pp. 319–328, 2025.

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

[10] T. Y. C. Tam, S. Sivarajkumar, S. Kapoor, et al., "A framework for human evaluation of large language models in healthcare derived from literature review", npj Digital Medicine, vol. 7, pp. 258, 2024.

[11] U. Mumtaz, A. Ahmed, and S. Mumtaz, "LLMs-Healthcare: Current applications and challenges of large language models in various medical specialties", Artificial Intelligence in Health, vol. 1, no. 2, pp. 16–28, 2024.

[12] H. Nori, N. King, S. McKinney, D. Carignan, and E. Horvitz, "Capabilities of GPT-4 on Medical Challenge Problems", 2023.

[13] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, P. Payne, M. Seneviratne, P. Gamble, C. Kelly, A. Babiker, N. Schärli, A. Chowdhery, P. Mansfield, D. Demner-Fushman, B. Agüera y Arcas, and V. Natarajan, "Large language models encode clinical knowledge", Nature, vol. 620, no. 7972, pp. 172–180, 2023.

[14] A. Szabo and G. Dolatkhah, "Comparative Evaluation of Large Language Models for Medical Education: Performance Analysis in Urinary System Histology", 2025.

[15] Z. Al Nazi and W. Peng, "Large Language Models in Healthcare and Medical Domain: A Review", Informatics, vol. 11, p. 57, 2024.

[16] R. Yang, T. F. Tan, W. Lu, A. J. Thirunavukarasu, D. S. W. Ting, and N. Liu, "Large language models in health care: Development, applications, and challenges", Health Care Science, vol. 2, no. 4, pp. 255–263, 2023.

[17] J. A. Omiye, H. Gui, S. J. Rezaei, J. Zou, and R. Daneshjou, "Large Language Models in Medicine: The Potentials and Pitfalls : A Narrative Review", Annals of Internal Medicine, vol. 177, no. 2, pp. 210–220, 2024.

[18] M. Al-Garadi, T. Mungle, A. Ahmed, A. Sarker, Z. Miao, and M. Matheny, "Large Language Models in Healthcare", 2025.

[19] P. Yu, H. Xu, X. Hu, and C. Deng, "Leveraging Generative AI and Large Language Models: A Comprehensive Roadmap for Healthcare Integration", Healthcare, vol. 11, p. 2776, 2023.

[20] R. AlSaad, A. Abd-Alrazaq, S. Boughorbel, A. Ahmed, M. A. Renault, R. Damseh, and J. Sheikh, "Multimodal Large Language Models in Health Care: Applications, Challenges, and Future Outlook", Journal of Medical Internet Research, vol. 26, e59505, 2024.

[21] K. Nassiri and M. A. Akhloufi, "Recent Advances in Large Language Models for Healthcare", BioMedInformatics, vol. 4, no. 2, pp. 1097–1143, 2024.

[22] J. Haltaufderheide and R. Ranisch, "The ethics of ChatGPT in medicine and healthcare: a systematic review on Large Language Models (LLMs)", NPJ Digital Medicine, vol. 7, no. 1, p. 183, 2024.

[23] L. Wang, Z. Wan, C. Ni, Q. Song, Y. Li, E. Clayton, B. Malin, and Z. Yin, "Applications and Concerns of ChatGPT and Other Conversational Large Language Models in Health Care: Systematic Review", Journal of Medical Internet Research, vol. 26, e22769, 2024.

[24] C. Peng, X. Yang, A. Chen, K. E. Smith, N. PourNejatian, A. B. Costa, C. Martin, M. G. Flores, Y. Zhang, T. Magoc, G. Lipori, D. A. Mitchell, N. S. Ospina, M. M. Ahmed, W. R. Hogan, E. A. Shenkman, Y. Guo, J. Bian, and Y. Wu, "A study of generative large language model for medical research and healthcare", NPJ Digital Medicine, vol. 6, no. 1, p. 210, 2023.

[25] M. Karabacak and K. Margetis, "Embracing Large Language Models for Medical Applications: Opportunities and Challenges", Cureus, vol. 15, no. 5, e39305, 2023.

[26] S. Tian, Q. Jin, P. T. Lai, Q. Zhu, X. Chen, Y. Yang, Q. Chen, W. Kim, D. Comeau, R. Dogan, A. Kapoor, X. Gao, and Z. Lu, "Opportunities and Challenges for ChatGPT and Large Language Models in Biomedicine and Health", 2023.

[27] H. Qin and Y. Tong, "Opportunities and Challenges for Large Language Models in Primary Health Care", Journal of Primary Care and Community Health, vol. 16, 215, 2025.

[28] C. R. Subramanian, D. A. Yang, and R. Khanna, "Enhancing Health Care Communication With Large Language Models—The Role, Challenges, and Future Directions", JAMA Network Open, vol. 7, no. 3, e240347, 2024.

[29] F. Busch, L. Hoffmann, C. Rueger, E. H. van Dijk, R. Kader, E. Ortiz-Prado, M. R. Makowski, L. Saba, M. Hadamitzky, J. N. Kather, D. Truhn, R. Cuocolo, L. C. Adams, and K. K. Bressem, "Current applications and challenges in large language models for patient care: a systematic review", Communications Medicine, vol. 5, no. 1, p. 26, 2025.

[30] U.S. National Library of Medicine, "MEDLINE: Description of the database", National Institutes of Health, 2025.

[31] Mayo Clinic, "Patient care and health information", Mayo Foundation for Medical Education and Research, 2025.

[32] Q. Jin, B. Dhingra, Z. Liu, W. Cohen, and X. Lu, "PubMedQA: A Dataset for Biomedical Research Question Answering", in MNLP-IJCNLP, pp. 2567–2577, Hong Kong, 2019.