RAG-LLM Healthcare Interaction System

Report submitted in partial fulfillment of the requirements
for the

B.Tech. in

Computer Science and Engineering Artificial Intelligence

By

Ojas (2021UCA1825)

Nikita Kanodia (2021UCA1803)

Under the supervision of

Prof. M.P.S. Bhatia

Computer Science and Engineering (CSE)

Netaji Subhas University of Technology, Delhi

Department of Computer Science and Engineering

NETAJI SUBHAS UNIVERSITY OF TECHNOLOGY
DELHI-110078

DECEMBER 2024

CERTIFICATE

This is to certify that the project titled **RAG-LLM Healthcare Interaction System** is a bonafide record of the work done by

Ojas (2021UCA1825)

Nikita Kanodia (2021UCA1803)

under my supervision and guidance in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology** in **Computer Science and Engineering Artificial Intelligence** of the **Netaji Subhas University of Technology, DELHI-110078**, during the year 2024-2025.

The original research work was carried out by the team under my guidance and supervision in the academic year 2024-2025. This work has not been submitted for any other diploma or degree from any university. Based on the declaration made by the group, we recommend the project report for evaluation

DATE:

Prof. M.P.S. Bhatia

Professor

Department of Computer Science and Engineering

Netaji Subhas University of Technology

DECLARATION

This is to certify that the work which is being hereby presented by us in this project titled "**RAG-LLM Healthcare Interaction System**" in partial fulfilment of the award of the Bachelor of Technology submitted at the Department of Computer Sciene and Engineering, Netaji Subhas University of Technology, Delhi, is a genuine account of our work carried out during the period from August 2024 to December 2024 under the guidance of Prof. M.P.S. Bhatia, Department of Computer Science and Engineering, Netaji Subhas University of Technology, Delhi.

The matter embodied in the project report to the best of our knowledge has not been submitted for the award of any other degree elsewhere.

DATE:

Ojas (2021UCA1825)

Nikita Kanodia (2021UCA1803)

ACKNOWLEDGEMENT

We would like to express our gratitude and appreciation to all those who made it possi ble to complete this project. Special thanks to our project supervisor (s) **Prof. M.P.S. Bhatia** whose help, stimulating suggestions, and encouragement helped us in writing this report. We also sincerely thank our colleagues for the time spent proofreading and correcting our mistakes.

We would also like to acknowledge with much appreciation the crucial role of the staff in Computer Science & Engineering, who gave us permission to use the lab and the systems and gave permission to use all necessary things related to the project.

Ojas
(2021UCA1825)

Nikita Kanodia
(2021UCA1803)

TABLE OF CONTENTS

Chapter 1

Introduction

1.1 General Introduction

Due to the advancement of technology in society, especially in the healthcare sector, it has become very important for the patients as well as providers to have easy access to medical information and to be able to manage it effectively. People find it difficult to comprehend their complicated medical records, including prescriptions, laboratory results, and treatment regimens, which results in confusion that may lead to wrong health decisions. In contrast, doctors and administrative employees engaging in computing this large patient data obtained from the Electronic Health Records (EHR) have to spend considerable time in data search and processing with high chances of making errors.

As for now, with the help of LLMs, such as LLaMA or Mistral and Retrieval-Augmented Generation (RAG), the AI systems can produce responses relevant to the context with the help of an external knowledge base. The described models can make a significant improvement for the communication process in the field of healthcare since they can offer immediate and correct solutions to patients' questions as well as help healthcare workers in finding the needed information. However, existing systems still have some limitations, like hallucinations, meaning the generation of false information, and some other problems, which concern the security and protection of the users' rights and the data in compliance with the rules of DISHA [8] and DPDPA [9]. To the end, this thesis suggests a design and implementation of a RAG-based LLM system suitable for healthcare context with a goal to improve the patient's and the provider's satisfaction, as well as protect the privacy and confidentiality of the medical data.

1.2 Motivation

The answer to the origin of this motivation comes from the fact that healthcare data are becoming complex and there is a need for a reliable and responsible AI-driven system to process these data. Some recent studies on RAG systems show that the external knowledge integration capability of RAG systems can enhance the performance of LLM by minimising hallucinations and maximising the response accuracy. But they are yet to be installed in real-life applications in a healthcare system or where patients' lives are at risk, and if the information given is inaccurate, it can lead to harm.

The patient's records are often incomprehensible to the particular patient, which may, in turn, entail wrong decisions about his/her treatment and wellbeing. Health care providers, on the other hand, require quicker and more efficient means of retrieving and analysing patient data, especially in emergency situations. By combining LLMs with a more robust RAG framework, these issues are not as much of a hurdle because they provide users with context-specific information. Furthermore, when human-in-the-loop strategies were incorporated into the system, it was possible to attain procedural accuracy and clinical validity, thus making the technology appropriate for sensitive healthcare applications.

1.3 Problem Statement

Modern healthcare organisations deal with massive volumes of data of various types, ranging from patient records to lab results. Patients struggle to understand what doctors tell them about themselves or what test results mean. They too have trouble making sense of the information presented to them. On the other hand, healthcare providers struggle to manage, retrieve, analyse the data quickly. There is a major missing link in terms of ease of use, time and context sensitivity of the current systems available for not only the patient but also the clinician. In the current system, they do lack both contextual awareness and accuracy, which are present in the current chatbots and digital healthcare solutions.

Even though models including LLaMA are accurate in using language to provide health information, they have problems with hallucinations and facts. These drawbacks have been considered to be tackled by the adoption of RAG systems since they provide additional knowledge to the generation process. However, some issues arise, like keeping the privacy of data, protecting patient data, and designing a system that can conform to current medical databases. In the scope of this thesis, the following are the challenges that this project is aimed to solve: The RAG-LLM system proposed in this project intends to come up with a reliable, scalable prototype that offers relevant, context-aware information to patients and care providers while satisfying legal requirements such as those presented by DISHA [8] and DPDPA [9].

Chapter 2

Literature Review

2.1 Prior Work

This section discusses the prior work done in the support of LLMs with RAGs for hospital based environment and how the concept evolved from this area and how can this help in making a solution that can help the heathcare interaction system simplify there process.

New trends in medical informatics have incorporated the use of the graph and the RAG frameworks as approaches to enhance information retrieval and question answering in the medical field. A paper proposed the Med-GraphRAG framework [1], which utilizes a hierarchical graph model to enhance the retrieval process by linking medical entities across three tiers: by the data furnished by the patient, basic medical knowledge, and medical encyclopedias or dictionaries respectively. The U-retrieve method which they used does a good job at handling a balance between global context and indexing efficiency, which in the end gives better results.

The latter domain was further explored in the next work, which developed the MIRAGE benchmark [2] for field-testing the performance of the RAG system in the context of medical question-answering. This showcases the potential of RAG when it comes to creating high-class retrieval pipelines, precisely for medical applications. So when working in a high-class retrievals pipelines for medical application, this showed the potential of RAG.

Another notable contribution, who developed a learning-to-rank approach [3] aimed at improving RAG-based search engines for Electronic Medical Records (EMR). Their system adapts to user search semantics by learning from user feedback, adjusting the ranking of retrieved documents based on relevance, which significantly improves retrieval accuracy in EMR systems.

The SelfRewardRAG method [4] introduced a self-evaluation layer into the retrieval process. This technique enables large language models (LLMs) to reflect on and refine their responses based on past outputs. By synthesizing information from multiple sources and applying chain-of-thought reasoning, SelfRewardRAG enhances the safety and accuracy of medical reasoning.

To address misalignment issues between user queries and retrieved content, Yang and other introduced Query-Based RAG (QB-RAG) [5]. This system generates a comprehensive set of pre-defined queries, helping to bridge the gap between user questions and the relevant content, thereby improving retrieval accuracy in handling complex medical queries.

In the realm of Electronic Health Records (EHRs), a paper demonstrated the application of generative AI combined with RAG for summarizing and extracting key clinical information [6]. Their system achieved an impressive high accuracy in identifying risk factors, showcasing RAG's potential for managing large volumes of structured and unstructured clinical data.

Further refining the interaction between retrieval and generation, the Iterative Retrieval-Generation (I-RetGen) [7] method integrates these processes iteratively, with each generation step guiding more precise retrieval results. This approach is especially useful for handling complex medical queries, where successive refinement leads to more accurate responses.

Lastly, FLARE (Active Retrieval) [7] introduces a dynamic aspect to the retrieval process by determining in real time when and what to retrieve during generation. This adaptability allows RAG systems to adjust to evolving user queries more effectively, making FLARE particularly valuable in dynamic medical environments.

2.2 Objectives

The primary goal of this thesis is to employ Retrieval-Augmented Generation (RAG) and Large Language Models (LLM) to build an AI-assisted healthcare communication system to improve the experience of the patient or healthcare provider relationship. The points for the objective are as follows:

- Improve Patient Access to Medical Information: Develop a chatbot that allows patients to interact with their medical data in simple, natural language, providing clear and accurate responses to queries about lab reports, prescriptions, and treatment plans.
- Streamline Data Retrieval for Healthcare Providers: Create a system that enables healthcare workers to retrieve patient data quickly and efficiently through natural language queries, reducing the time spent on manual searches and improving decision-making.
- Mitigate Hallucinations and Enhance Reliability: Integrate RAG systems to augment the LLM generation process with certified external knowledge, minimizing the risk of hallucinations and ensuring trustworthy, accurate responses.
- Ensure Data Privacy and Compliance: Implement robust encryption, access control mechanisms, and audit logs to secure patient data and ensure compliance with healthcare regulations like DISHA [8] and DPDPA [9].
- Incorporate Human-in-the-Loop Mechanisms: Introduce a human-in-the-loop mechanism for reviewing and validating ambiguous or critical queries to ensure the highest level of accuracy in the system's responses.

Chapter 3

Requirements

For this project, simulating the development and testing environment for the RAG-based LLM healthcare system requires a robust infrastructure capable of handling AI model training, retrieval systems, and large data processing. Below are the key requirements and platforms that will be employed:

- Programming Languages and Frameworks: Python, LangChain, FastAPI
- Models: LLaMA, Mistral, Nemotron
- Vector Database: Milvus or Pinecone (Paid)
- Cloud Infrastructure: AWS EC2 and S3 Buckets
- Security and Compliance: Implement OAuth 2.0 and IAM
- Testing Environment: Jupyter Notebook, Postman
- User Interface: Reactjs (web version)

Chapter 4

Conclusion

This project presents a promising solution to one of healthcare's most pressing challenges: promoting patient engagement in their care and the timeliness, relevance, and confidentiality of their health information. Thus, using the advanced RAG systems with additional large language models, we hope to provide a system that helps patients be more aware of their medical histories and help the doctors and nurses quickly access crucial information.

The integration of external knowledge sources through vector databases ensures that the responses generated are not only contextually relevant but also grounded in factual, certified information, minimising the risks of AI hallucinations. Furthermore, the project places a significant focus on data privacy and regulatory compliance, ensuring that all interactions with the system adhere to DISHA [8] and DPDPA [9] standards, making it suitable for real-world deployment in hospital environments.

Bibliography

[1] J. Wu, J. Zhu, and Y. Qi, "Medical graph rag: Towards safe medical large language model via graph retrieval-augmented generation," 2024. [Online]. Available: https://arxiv.org/abs/2408.04187

[2] G. Xiong, Q. Jin, Z. Lu, and A. Zhang, "Benchmarking retrieval-augmented generation for medicine," 2024. [Online]. Available: https://arxiv.org/abs/2402.13178

[3] C. Ye, "Exploring a learning-to-rank approach to enhance the retrieval augmented generation (rag)-based electronic medical records search engines," *Informatics and Health*, vol. 1, no. 2, pp. 93–99, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2949953424000146

[4] Z. Hammane, F.-E. Ben-Bouazza, and A. Fennan, "Selfrewardrag: Enhancing medical reasoning with retrieval-augmented generation and self-evaluation in large language models," in *2024 International Conference on Intelligent Systems and Computer Vision (ISCV)*, 2024, pp. 1–8.

[5] E. Yang, J. Amar, J. H. Lee, B. Kumar, and Y. Jia, "The geometry of queries: Query-based innovations in retrieval-augmented generation," 2024. [Online]. Available: https://arxiv.org/abs/2407.18044

[6] M. Alkhalaf, P. Yu, M. Yin, and C. Deng, "Applying generative ai with retrieval augmented generation to summarize and extract key clinical information from electronic health records," *Journal of Biomedical Informatics*, vol. 156, p. 104662, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1532046424000807

[7] S. Ghanbari Haez, M. Segala, P. Bellan, S. Magnolini, L. Sanna, M. Consolandi, and M. Dragoni, "A retrieval-augmented generation strategy to enhance medical chatbot reliability," in *Artificial Intelligence in*

Medicine, J. Finkelstein, R. Moskovitch, and E. Parimbelli, Eds. Cham: Springer Nature Switzerland, 2024, pp. 213–223.

[8] F. F. F. P. K. Ganapathy, MCh (NEURO), "Artificial intelligence and healthcare regulatory and legal concerns," *Telehealth and Medicine Today*, vol. 6, no. 2, Apr. 2021. [Online]. Available: https://telehealthandmedicinetoday.com/index.php/journal/article/view/252

[9] K. Sundara and N. Narendran, "The digital personal data protection act, 2023: analysing india's dynamic approach to data protection," *Computer Law Review International*, vol. 24, no. 5, pp. 129–141, 2023. [Online]. Available: https://doi.org/10.9785/cri-2023-240502