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 20242025. 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 take this opportunity to acknowledge the support of all those without whom the completion of this project in fruition would not be possible.

——— write more here———

TABLE OF CONTENTS

List of Tables

List of Figures

Chapter 1

Introduction

type Introduction here

1.1 Section 1 of Intro

Cite like this type Introduction here

1.2 Section 1 of Intro

Cite like this

Chapter 2

Motivation

Chapter 3

Literature Review

The use of Retrieval-Augmented Generation (RAG) together with Large Language Models (LLMs) within the healthcare field is one of the current trends, which has been the centre of attention in the recent years mainly because of the issues connected with the quality as well as context comprehension of the medical data. Haez et al. (2024) have suggested improved RAG strategy to increase the trust level in the medical chatbot by adding an initial interaction cycle in the RAG pipeline. This involves the LLM creating a mock document to use in requesting from a certified information source hence minimizing hallucinations in responses. Their work also shows that despite the current challenges, RAG-LLMs can improve the user trust specifically in maternal health domains by using the certified knowledge sources for the responses [1].

In the same vein, Al Ghadban et al., (2023) discuss the feasibility of using RAG models in healthcare education learning with frontline health workers in LMICs. One tool developed by them is known as "SMARThealth GPT" [2] that employ RAG to generate targeted, context-sensitive information to foster comprehension of the existing gaps in the delivery of community health services. It supports the RAG's capacity in individually catering LLMs for expanding educational ends, in boosting the health worker's ability on accurate guideline-based care. Furthermore, another study is about using generative AI with RAG to derive the critical clinical data from the EHRs. This way patient data summarization is performed and examples are shown on how a RAG system can reduce the burden of data management on clinicians while at the same time providing them with context relevant information [3].

Comparing the medical application of RAG also have a significant part to investigate its performance. Studies that compared RAG in requiring the healthcare domain present the advantage and drawbacks of using LLMs for searching for medical information. Therefore, this work lays a foundation for the development of future implementations in healthcare by raising the element of the need to include retrieval mechanisms that will enable the delivery of contextually relevant and precise information [4]. Research on the effectiveness of EMR search engines continues to indicate that learning to rank techniques have the potential of further boosting RAG systems' performance in dealing with the large volumes of medical information. This study shows that augmenting learning-to-rank approaches can enhance the process of document search and enhance patient treatment by offering better outcomes [5].

The other significant area appropriate for RAG-LLM development is dealing with the issue of semantic uncertainty and making the answers more accurate. Query-based innovations in RAG systems are described in a study but the approaches employed in the study are applied in reducing ambiguity and enhancing the retrieved documents relevance. This way, the given approach enhances the validity of LLM-produced responses in medical situations, which should bring the enhanced trust of the users in automatized healthcare systems [6]. To supplement the reliability of medical chatbots, SelfRewardRAG [7] provides LLMs with a self-evaluation function so as to enable them critically evaluate their generated responses in terms of accuracy and relevancy. This helps in minimizing the frequency of hallucinations and improves the quality of generated responses demonstrating that self-evaluation can effectively cause enhanced LLM performance in the medical context. There is also some emerging safety issues which have also been discussed in the recent papers as applying AI to generate medical advice. The use of graph-based RAG systems in one study therefore trains the system with rules and regulation that will make the LLM utterances conform to certified medical standard. The use of graph retrieval techniques improves on the safety and accuracy of the interactions and especially on patients' sensitive data. This emphasis on using verified information sources show the key idea of developing safe and rather reliable

artificial intelligence applications in healthcare. paper 568

Chapter 4

Problem Statement

It have been noticed that in the present day's health care delivery system the communication between patients and healthcare providers is of paramount important. One of the main issues, which arise dealing with medical information, including electronic health records (EHRs), laboratory results, and prescriptions, is that all the information can be overwhelming. This results to producing confusion and wrong decisions regarding their health. On the other hand, the health care providers are experiencing challenges on how to address the big issue of managing and searching large amount of data from multiple hospital information systems using repetitive and manual methods which are inconveniencing. Such existing and currently popular tools, and chatbots, for instance, provide simple solutions but cannot give a context-aware answer in real-time for both the patient and the health care personnel.

Recently, Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) systems present the solution path to these problems. Although the investigated LLMs are capable of providing natural language outputs, they are inclined to certain problems such as hallucinations. RAG systems therefore propose to improve the reliability of these models by incorporating external sources of knowledge for production of accurate and reliable results. The study reveals that RAG systems lead to noteworthy reduction in hallucination and increase in quality of the response as the system incorporates only certified medical knowledge to derive response and such complex health care sectors like maternal care are most vulnerable to benefit from the implementation of RAG systems.

Nevertheless, there are difficulties in the development of RAG-LLM systems that should support the calculation of the result in healthcare, making the system as reliable as possible, and at the same time, consuming minimal resources. Some of the key stakeholders' concerns include privacy of the data, accuracy of medical information in the system, and integration of the new system management with the existing hospital management systems. However, there is a need for human-in-the-loop approaches to handle the vagueness as well as to make sure that the last decisions are made by doctors. This paper aims at developing a RAG based LLM system in healthcare context with special reference to both patient and provider interfaces. The solution seeks to reduce the patient's involvement as much as possible with their records while at the same time making the retrieval of information smooth, fast and efficient for the healthcare providers; it makes communication in the health sector safe and reliable.

Chapter 5

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/healthcare provider relationship. The proposed system targets to solve major issues arising from information seeking and sharing in the healthcare domain, enabling patients and providers to effectively and efficiently acquire context-specific results. The points for the objective are as follows:

- Improve Patient Access to Medical Information
- Streamline Data Retrieval for Healthcare Providers
- Mitigate Hallucinations and Enhance Reliability
- Ensure Data Privacy and Compliance
- Incorporate Human-in-the-Loop Mechanisms

Bibliography

[1] 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.

[2] 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

[3] 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

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

[5] 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

[6] 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

[7] 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.

Appendices

Appendix A

Code Attachments

A.1 Lorem Ipsum

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum

```
def get_parameters(data, chunk_size=410):
  #Store the activity label to add later
  activity = data['Activity']
    '''
    Define a dictionary of functions. Sets of readings will be
    aggregated as per these functions
    '''
  func_dict = {
    'min': np.min,
    'max': np.max,
    'diff': lambda x: np.max(x) - np.min(x),
    'std': np.std,
    'iqr': stats.iqr,
    'rms': lambda x: np.sqrt(np.mean(np.square(x))),
    'mad': lambda x: x.mad(),
    'mediad': mediad
  }
  aggregations = {
    'X': func_dict,
    'Y': func_dict,
    'Z': func_dict
  }
  data_groups = []
    '''
    Transform the dataset into rolling windows of 410 readings each
    and store them in a Pandas data group.
    '''
```

```
26    for i in range(int(data.shape[0]/(chunk_size/2)) - 1):
27      temp = data.iloc[int(i*(chunk_size/2)):int((i+2)*(chunk_size/2))]
28      temp['k'] = i
29      data_groups.append(temp)
30    data_groups = pd.concat(data_groups).groupby('k', as_index=False)
31      #Run the aggregations on all data groups
32    stats_data = data_groups.agg(aggregations)
33    stats_data.columns = [''.join(col).strip() for col in stats_data.
       columns.values]
34    activity = activity.reset_index(drop=True)
35      #Add activity label
36    stats_data = pd.concat([stats_data, activity[:len(stats_data)]],
       axis=1)
37    del stats_data['k']
38    return stats_data
```