National University of Computer and Emerging Sciences

Project Final Report

Conversational AI for Healthcare (Lumbar Spine Diseases) Queries Using Retrieval-Augmented Generation (RAG)

Group Members

Muhammad Saib	21L-7708
M. Hannan Fareed	21L-5392
M.Ibrahim Raza	21L-1795

Supervised by

Mamoona Akbar

FAST School of Computing

National University of Computer and Emerging Sciences

Lahore, Pakistan

Abstract

This project aims to develop a conversational AI system designed to address healthcare queries, with a special focus on lumbar spine diseases such as Spinal Stenosis and Scoliosis. The system uses Retrieval-Augmented Generation (RAG), which combines retrieval-based models with generative models, to provide accurate, relevant, and contextually accurate answers. By integrating domain-specific medical documents into a large language model (LLM), the system provides accurate, context-aware responses. This system will benefit healthcare professionals, patients, and researchers by providing a reliable tool for medical advice and clarifications.

Introduction

Healthcare-related queries often require precise, accurate, and contextually relevant responses. However, misinformation and inaccurate medical advice are widespread across many platforms, which could lead to harmful consequences for patients. This project addresses the growing concern of misinformation in healthcare by developing a conversational AI system that ensures accurate, contextually relevant, and trusted responses to medical queries. Leveraging the Retrieval-Augmented Generation (RAG) approach, the system draws on verified sources like PubMed, clinical guidelines, and medical textbooks to provide reliable information. It focuses on prevalent lumbar spine diseases such as Spinal Stenosis and Scoliosis, which significantly impact the quality of life for many, especially older adults. Traditional methods of providing medical information, like printed materials and static websites, often fall short in offering up-to-date, personalized guidance. By using advanced AI models, this project aims to enhance the delivery of accurate healthcare information and reduce the risks associated with misinformation.

Problem Statement

The challenge addressed by this project is the development of a conversational AI system that can accurately respond to healthcare queries about lumbar spine diseases by integrating real-time, domain-specific medical knowledge. This system aims to reduce misinformation and provide users with reliable, contextually relevant answers.

Literature Review

This report presents a comprehensive literature review on the application of conversational AI in healthcare, particularly focusing on its use for healthcare queries through Retrieval-Augmented Generation (RAG). The studies analyzed various aspects of AI chatbots, their applications, challenges, and gaps in existing research.

1. Transforming Healthcare with Chatbots

Explores the use of AI-driven chatbots for diagnosing conditions and providing preliminary health information. It highlights the benefits of chatbots in improving healthcare processes and patient engagement. However, it identifies a gap in the integration of chatbots with healthcare infrastructure and data security measures.

2. **Achieving Health Equity through Conversational AI**

   Addresses the ethical considerations and inclusivity aspects of conversational AI in healthcare. The study emphasizes the importance of minimizing bias but lacks empirical evidence of the effectiveness of such methods in real-world healthcare settings.

3. **Conversational AI with Large Language Models to Increase Uptake of Clinical Guidelines**

   Investigates the potential of conversational AI, similar to ChatGPT, in implementing clinical guidelines. While it underscores the benefits, the study lacks validation through pilot studies or real-world testing.

4. **Roles, Users, Benefits, and Limitations of Chatbots in Healthcare**

   Reviews the roles of chatbots in healthcare, categorizing their functions into administrative, patient tracking, and therapeutic delivery. The study identifies the need for long-term studies on patient satisfaction and knowledge base maintenance.

5. **Era of Generalist Conversational AI to Support Health Systems**

   Focuses on the potential of generalist AI systems to improve public health communication. However, it lacks data on the practical application and success of such systems in healthcare environments.

6. **Systematic Review and Meta-Analysis of the Effectiveness of Chatbots in Healthcare**

   Evaluates the success of chatbots in mental health interventions, such as depression, anxiety, and substance abuse treatment. Despite positive results, it calls for larger-scale studies to validate long-term efficacy.

7. **AI-Powered Chatbots for Chronic Diseases**

   Discusses AI chatbots designed to assist with chronic disease management. While effective in guiding patients, the study highlights the need for standardized evaluation metrics to assess the effectiveness of such tools.

8. **Inclusive Healthcare and Chatbots**

   Emphasizes the importance of designing chatbots that meet the needs of elderly users and those with disabilities. It identifies a gap in the availability of such tools, suggesting the need for more inclusive design approaches.

9. AI Chatbots for Promoting Health Behavior Change

Reviews the use of AI chatbots in promoting healthy behavior, particularly in diet, exercise, and medication adherence. It calls for more detailed reporting on chatbot features and extended follow-up periods to measure behavior change.

10. Perioperative Application of Chatbots

Evaluates the use of chatbots in perioperative care, finding positive patient acceptance and effective communication without compromising outcomes. However, it stresses the need for standardized implementation guidelines.

11. AI-Based Conversational Agents for Mental Health

Examines how AI conversational agents can reduce psychological distress. While promising, the study notes the need for more research into the long-term effects and integration with existing mental health services.

12. Integration of AI-Powered Chatbots in Emergency Medicine

Explores the potential for AI chatbots in emergency care settings. It acknowledges their feasibility but lacks real-world implementation data, particularly in high-stress situations.

In conclusion, while AI chatbots have shown potential in various healthcare domains, significant gaps exist, including the need for empirical studies, real-world validation, and standardized evaluation metrics. The reviewed studies underline the importance of addressing ethical concerns, inclusivity, and long-term effectiveness to ensure that conversational AI systems can deliver reliable and impactful healthcare services.

Methodology

Data Collection

Medical documents related to lumbar spine diseases were collected from various sources, including research articles, clinical guidelines, and textbooks. These documents were processed to extract relevant information for the RAG system.

Document Processing Pipeline

A custom pipeline was developed to process the collected documents:

- **Document Loading**: Utilizing LangChain's `PyPDFLoader`, `TextLoader`, and `Docx2txtLoader` to load documents in various formats.
- **Text Cleaning**: Implementing functions to normalize medical text, correct OCR errors, and standardize units of measurement.

- **Text Splitting**: Using `RecursiveCharacterTextSplitter` to divide documents into manageable chunks, preserving context and section headers.
- **Embedding Generation**: Employing HuggingFace embeddings to convert text chunks into vector representations.
- **Vector Store Creation**: Storing embeddings in a FAISS vector store for efficient retrieval.

Conversational AI System

A chatbot interface was developed using Gradio, integrating the RAG system to handle user queries:

- **Retrieval Chain**: Using LangChain's `ConversationalRetrievalChain` to manage interactions and retrieve relevant information.
- **Model Integration**: Connecting to the Ollama model (`llama3`) for generating responses based on retrieved data.
- **User Interface**: Designing an intuitive interface with tabs for chat and system setup, including options to initialize the RAG system and load medical documents.

Implementation Details

1. System Architecture Overview

The system is built around a **Conversational Retrieval-Augmented Generation (RAG)** framework, where medical queries are answered by retrieving relevant documents from a pre-built vector store and generating responses using an AI language model. The architecture consists of several key components:

- **Document Processing Pipeline**: Handles the ingestion of medical documents (e.g., PDFs) and splits them into smaller chunks for efficient retrieval.
- **Embedding Generation**: Converts textual data into vector representations that capture the semantic meaning of the documents.
- **Vector Store (FAISS)**: A highly optimized data structure for performing similarity searches across large sets of embeddings, enabling fast retrieval of relevant documents.
- **Conversational Chain**: Facilitates the generation of responses to user queries based on retrieved documents, enhancing the interaction with a natural language model.
- **Gradio Interface**: Provides a web-based interface for users to interact with the chatbot and receive answers to their medical queries.

2. Document Processing

The initial step in the RAG system is to process the medical documents, typically in PDF format, to extract relevant information. This process involves the following steps:

- **Document Loading**: Using the `DirectoryLoader` class from the Langchain library, PDF documents are loaded from the designated directory (`./medical_docs`). The documents are then passed through the `PyPDFLoader` to extract text content.
- **Text Splitting**: Medical documents, especially lengthy ones, are split into smaller, manageable chunks using the `RecursiveCharacterTextSplitter`. This step is critical as it ensures that the information is

chunked into smaller segments, typically 1000 characters in size with 200 characters of overlap. This chunking allows for efficient retrieval of relevant information during query processing.

3. Embedding Generation and Vector Store Setup

The textual content from the medical documents is transformed into vector representations using **HuggingFaceEmbeddings**. These embeddings capture the semantic meaning of the text, enabling similarity searches to retrieve documents that are relevant to a user's query.

- **Embedding Model**: The model used for generating embeddings is `sentence-transformers/all-MiniLM-L6-v2`, which is a pre-trained transformer model optimized for sentence-level embeddings.
- **FAISS Vector Store**: The embeddings generated for each chunk are stored in a **FAISS (Facebook AI Similarity Search)** vector store. FAISS is used for efficient nearest neighbor searches, allowing the system to retrieve the top `k` most relevant documents based on the similarity of their embeddings to the query embedding.
- **Saved Vector Store**: The vector store can either be generated from scratch or loaded from a previously saved version. If the vector store is created, it is saved locally for future use, ensuring that the system does not need to reprocess the documents each time.

4. Conversational Model Integration

The conversational aspect of the system is powered by **ChatOllama**, a wrapper around the Ollama language model. This model generates responses based on the query and the context of the conversation. It is initialized with a specified model name (e.g., `llama3`), and the temperature parameter is set to 0.5 for balanced response generation.

- **ConversationalRetrievalChain**: This Langchain component is crucial in linking the retrieval and generation processes. It is responsible for taking the user's query, retrieving the top `k` documents from the vector store using their embeddings, and passing this context to the language model. The model then generates a response that is informed by the retrieved documents.
- **Response Augmentation**: In the response generation, the query is formatted with a prompt that instructs the system to behave as a medical specialist, ensuring that the answers are tailored to the context of the medical field. Additionally, the response is enriched with citations to the source documents used to generate the answer, providing transparency and reliability.

5. Gradio Web Interface

To allow user interaction, a **Gradio web interface** is used, which is composed of two primary tabs:

- **Chat Tab**: Users can enter their medical queries, and the chatbot responds with relevant information based on the documents stored in the system. The conversation history is maintained, and users can view both the query and the corresponding response.
- **Setup Tab**: This tab allows users to initialize the RAG system by specifying the path to the medical documents, the model name for Ollama, and the embedding model. Users can choose to use a saved vector store to avoid reprocessing documents each time the system is initialized.

Components:

- **Textbox**: For users to enter their medical queries.

- **Chatbot**: Displays the ongoing conversation, showing both user inputs and assistant responses.
- **Buttons**: The "Send" button submits the query, and the "Clear Chat" button resets the conversation history.
- **State**: Maintains the chat history across interactions.

The Gradio interface is designed to be intuitive and user-friendly, with custom CSS applied to ensure that the medical disclaimer and chatbot interface are clearly visible.

6. Error Handling and Logging

The system includes robust error handling to ensure that users are informed if the system is not initialized or if an error occurs during query processing. **Logging** is enabled using Python's `logging` module, providing detailed logs for debugging and tracking the flow of execution. The logs capture events such as:

- System initialization status
- Document loading and processing errors
- Query processing errors
- System exceptions

These logs are saved to a file and also displayed in the console for real-time monitoring.

7. Medical Disclaimer

To ensure compliance with medical guidelines, a **medical disclaimer** is displayed prominently in the interface, informing users that the responses provided by the chatbot are for general informational purposes only and should not replace professional medical advice.

Results & Discussion

System Performance

The RAG-based conversational AI system demonstrated:

- **Accuracy**: High relevance and correctness in responses, as evaluated against a set of test queries.
- **Efficiency**: Quick response times, with the system retrieving and generating answers within seconds.
- **User Satisfaction**: Positive feedback from users regarding the clarity and helpfulness of the information provided.

Limitations

Despite its strengths, the system has some limitations:

- **Scope of Knowledge**: Limited to the documents processed; may not cover all aspects of spinal health.
- **Dependency on Input Quality**: The accuracy of responses depends on the quality and comprehensiveness of the input documents.

Conclusion & Future Work

Conclusion

The project successfully developed a conversational AI system that leverages RAG to provide accurate and context-aware responses to healthcare queries about lumbar spine diseases. This approach enhances the reliability of AI-generated information, making it a valuable tool for both medical professionals and patients.

Future Work

Future enhancements include:

- **Expanding Knowledge Base**: Incorporating a broader range of medical documents to cover more aspects of spinal health.
- **Multilingual Support**: Developing capabilities to handle queries in multiple languages.
- **Integration with Medical Records**: Connecting the system to electronic health records for personalized responses.
- **Continuous Learning**: Implementing mechanisms for the system to learn from user interactions and improve over time.

References

- Chen, J., Lin, H., Han, X., & Sun, L. (2024). Benchmarking Large Language Models in Retrieval-Augmented Generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16), 17754-17762. https://doi.org/10.1609/aaai.v38i16.29728

- Nadarzynski, T., Knights, N., Husbands, D., Graham, C. A., Llewellyn, C. D., Buchanan, T., Montgomery, I., & Ridge, D. (2024). Achieving health equity through conversational AI: A roadmap for design and implementation of inclusive chatbots in healthcare. *PLOS Digital Health*, 3(5), e0000492. https://doi.org/10.1371/journal.pdig.0000492

- Lizée, A., Beaucoté, P.-A., Whitbeck, J., Doumeingts, M., Beaugnon, A., & Feldhaus, I. (2024). Conversational Medical AI: Ready for Practice. *arXiv:2411.12808 [cs.AI]*. https://doi.org/10.48550/arXiv.2411.12808

- Laymouna, M., Ma, Y., Lessard, D., Schuster, T., Engler, K., & Lebouché, B. (2024). Roles, users, benefits, and limitations of chatbots in health care: Rapid review. *J Med Internet Res*, 26, e56930. https://doi.org/10.2196/56930

- Sezgin, E., & Kocaballi, A. (2025). Era of generalist conversational artificial intelligence to support public health communications. *J Med Internet Res*, 27, e69007. https://doi.org/10.2196/69007

- Laranjo, L., Dunn, A. G., Tong, H. L., Kocaballi, A. B., Chen, J., Bashir, R., & Coiera, E. (2024). Systematic review and meta-analysis of the effectiveness of chatbots in healthcare. *npj Digital Medicine*, 7(1), 1–12. https://doi.org/10.1038/s41746-023-00856-1

- Kurniawan, M. H., Handiyani, H., Nuraini, T., Hariyati, R. T. S., & Sutrisno, S. (2024). A systematic review of artificial intelligence-powered (AI-powered) chatbot intervention for managing chronic illness. *Annals of Medicine*, 56(1), 2302980. https://doi.org/10.1080/07853890.2024.2302980

- Vandemeulebroucke, T., Dierckx de Casterlé, B., & Gastmans, C. (2024). A systematic review of chatbots in inclusive healthcare: insights from aging and disability studies. *Universal Access in the Information Society*, 23(1), 45-60. https://doi.org/10.1007/s10209-024-01118-x

- Laranjo, L., Dunn, A. G., Tong, H. L., Kocaballi, A. B., Chen, J., Bashir, R., & Coiera, E. (2023). Artificial intelligence–based chatbots for promoting health behavior change: Systematic review. *Journal of Medical Internet Research*, 25, e40789. https://doi.org/10.2196/40789

- Semigran, H. L., Linder, J. A., Gidengil, C., & Mehrotra, A. (2024). Perioperative application of chatbots: A systematic review and meta-analysis. *BMJ Health & Care Informatics*, 31(1), e100985. https://doi.org/10.1136/bmjhci-2023-100985

- Abd-Alrazaq, A. A., Alajlani, M., Alalwan, N., Bewick, B. M., Gardner, P., & Househ, M. (2024). Systematic review and meta-analysis of AI-based conversational agents for mental health. *npj Digital Medicine*, 7(1), 1–12. https://doi.org/10.1038/s41746-023-00979-5

- Taylor, R. A., Moore, C. L., & Venkatesh, A. K. (2024). Integration of AI-powered chatbots in emergency medicine: Potential and challenges. *Annals of Emergency Medicine*, 83(3), 245–258. https://doi.org/10.1016/j.annemergmed.2024.01.005