CSE 366 Course Project Report

RAG Chatbot - The VoG (Virtual omni-medicine Guide) Based on Assorted Medicine Dataset of Bangladesh

Md. Arfan Ahmed ID: 2020-1-60-139
Farhan Tanvir ID: 2020-1-60-132
Sudharam Sur ID: 2020-1-60-114

June 8, 2024

Abstract

In this paper, we propose a method to obtain strong explanations for question answering that correlate well with the answers. We introduced a Retrieval-Augmented Generation (RAG) based chatbot question-answering system for assorted medical dataset of Bangladesh. This theoretical foundation is built on advances in artificial intelligence (AI). The model uses some advanced AI Technologies like Llama3, LangChain, HuggingFace, ChromaDB, and Streamlit. These technologies make it perfect for enlightenment in the medical sector as a question-answering chatbot. Our Model provides the answer by using the RAG architecture consisting of Knowledge base, vector store, question-answering LLM, and user interface. A vector store using ChromaDB, facilitated by Hugging Face embeddings, to effectively retrieve data from the medicine dataset. By integrating these technologies, the chatbot responds to queries regarding medicine details such as brand names, generics, and manufacturers. We observe that the proposed method performs better with the given dataset. Future work will be focused on expanding dataset and improving resource efficiency.

Contents

1 Introduction

The way we access and use information has changed significantly as a result of the widespread adoption of advanced artificial intelligence capabilities. One such crucial industry is medicine, where an ever-expanding body of knowledge needs effective ways for application and retrieval. This work focuses on developing a Question and Answering chatbot based on assorted medicine dataset of Bangladesh, utilizing modern AI technologies to enhance healthcare information accessibility and accuracy.

The core of this project is a Retrieval-Augmented Generation model which is a sophisticated approach that combines the strengths of retrieval-based systems and generative models. The chatbot app is designed to provide accurate and contextually relevant answers on assorted medicine dataset of Bangladesh by utilizing Llama3, an advanced language model known for its superior natural language understanding and generation capabilities as well as LangChain, HuggingFace, ChromaDB, and Streamlit.

1.1 The Rise of Large Language Models and Conversational AI

The appearance of Large Language Models is the key milestone in the evolution of language and text technologies. These models, based on large text corpora, show impressive performance in understanding and producing human-like text. Not only does this development have the potential to enable new applications, such as question-answering chatbots, but it has also widened the scope of their implementation radically. LLMs are behind conversational AI-powered systems, which have the potential to transform the field of human-computer interaction by making it more natural and intuitive.

1.2 Retrieval-Augmented Generation (RAG)

One effective method for handling knowledge-intensive tasks is retrieval-augmented generation. RAG uses facts from predetermined external databases to improve the accuracy and reliability of generative AI models by integrating retrieval methods with these models. Large language models frequently encounter problems such as giving inaccurate information when they don't know the answer, giving answers that are too general or out of current, and producing results that are dependent on sources that are not trustworthy that are included in their training data. By providing the LLM with instructions to get accurate information from reliable and well-established knowledge bases, RAG resolves these issues. This allows the LLM to use the retrieved data for context in producing responses that are much more accurate, relevant, and useful across a range of applications. RAG applications can increase user trust by revealing sources for the information being retrieved, thus making transparent how the LLM is generating the responses [1].

1.3 Llama3

Llama3 is the most powerful openly available LLM currently available which is developed by Meta AI. Like other LLMs, Llama 3 is trained on a massive dataset of text and code. This allows it to learn complex relationships between words and concepts. When we provide input, Llama 3 analyzes it using this knowledge to generate text, translate languages, write different kinds of creative content, and answer the questions in an informative way.

1.4 Vector Embeddings

We have used Hugging Face embedding and Chroma DB as a vector database in our work. Hugging Face's embedding system is a framework designed to represent text or other types of data in numerical forms known as embeddings. Those embeddings serve to encapsulate quite effectively the semantic essence of the underlying data and therefore tend to be extremely versatile in processing data of all kinds. On the other hand, Chroma DB is an open-source vector database designed carefully to store and retrieve vector embeddings effectively. While traditional databases very often struggle with these processes of numerical data representations, Chroma DB steps forward to optimize this process by guaranteeing effective management of vector embeddings.

1.5 Literature Review

In [1], retrieval augmented generation is implemented to enhance the accuracy and transparency of large language models for natural language question-answering tasks. This approach includes a knowledge base, vector store, question-answering LLM, and user interface components. The system proposed in the paper is called 'MufassirQAS' and is based on a vector database that stores and retrieves vectors to improve the accuracy of LLMs. The vector store is created using open-access books related to Islam, such as 'Kuran yolu Turkce meal ve tefsir,' Kutub-i Sitte, and 'Islam Ilmihali.' The use of Large Language Models for processing and generating natural language texts, trained on intensive amounts of text data using deep neural networks.

In [2], The research paper proposes the Retrieval Augmented Generation (RAG) method for enhancing Question-Answering (QA) systems by addressing document processing in Natural Language Processing. RAG associates search techniques in vector store and text generation mechanisms created by Large Language Models, specifically using Generative Pre-trained Transformer 3.5. The paper also proposes LangChain and FAISS frameworks for the RAG technique, associating parametric and nonparametric models to provide good results.

1.6 Problem Statement

This work aims to develop a Retrieval-Augmented Generation (RAG) question-answering chatbot specifically designed for the assorted medicine dataset of Bangladesh. The primary objective is to create a system capable of accurately and efficiently answering queries related to this dataset which offers users quick access to relevant information regarding various medicines available in Bangladesh. This will solve the complexity of traversing and querying insights from this large diversified medicine dataset, presentable in a user-friendly interface to query and retrieve specific details of medicines: brand, type, slugs, dosage form, generic, strength, manufacturer, and more. Developing a focused chatbot app on this dataset will increase accessibility to medical information and further support informed decisions on healthcare products in Bangladesh.

1.7 Research Objectives

- Developing an effective Retrieval-Augmented Generation model for the assorted medicine dataset of Bangladesh.
- Providing a user-friendly chatbot interface to query and retrieve the desired information from the above-mentioned dataset.
- Training the RAG model to interpret and respond to a wide array of queries related to medicines, brands, types, slugs, dosage forms, generic, strength, manufacturer, and other relevant details.
- Evaluating chatbot competence in replying to open-ended questions which includes the chatbot's ability to provide detailed and comprehensive answers to questions that go beyond just simple fact retrieval.

2 Methodology

This research concept centers on the creation and application of a Retrieval-Augmented Generation (RAG) model to generate a chatbot that can respond to questions and is customized for the assorted medicine dataset of Bangladesh. The research employs a mixed-methods approach, integrating qualitative and quantitative methodologies to provide a thorough investigation and assessment of the chatbot's functionality.

This theoretical foundation is built on advances in artificial intelligence (AI), with a specific emphasis on the interaction between generative models and retrieval-based systems. The project makes use of Llama3, a sophisticated language model created by Meta AI that is renowned for its exceptional natural language creation and interpretation skills. To create a

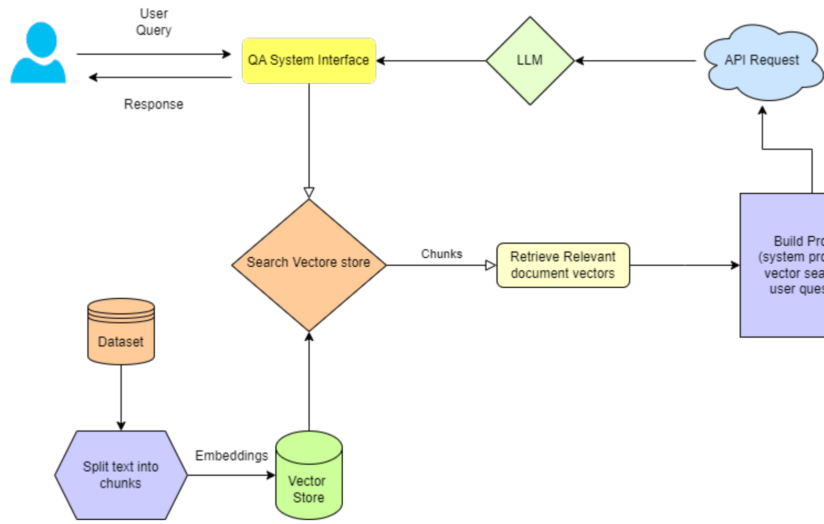

Figure 1: RAG based System

reliable and scalable system, the project also uses Streamlit, HuggingFace, ChromaDB, and LangChain.

Our Suggested system is shown in Figure 1 and is built using the RAG architecture. The components that the study consists of are: Knowledge base, vector store, question-answering LLM, and user interface.

2.1 Creating Vector Store

We used ChromaDB, an open-source vector database for storing and retrieving vector embeddings, to efficiently handle and use the wide range of assorted medicine data. The first step in the process is data preparation, which involves cleaning and standardizing the dataset of various medications, including information about brands, kinds, dosage forms, generic names, strengths, and manufacturers. The textual data is then transformed into numerical vector representations, or embeddings, using Hugging Face's embedding system. These embeddings effectively capture the semantic content of the original text. After that, these embeddings are kept in ChromaDB, a database designed specifically to manage high-dimensional vectors, making efficient storage and retrieval possible.

To enable vector search, the text in the knowledge base needs to be divided into fixed-length segments called chunks. It's important to assign a value to indicate how much overlap each chunk has with the previous one, known as the chunk overlap. To ensure that the language model understands sensitive content and maintains logical connections between parts of the document, we assign a higher value for chunk overlap.

2.2 Vector Search Mechanism

Using the same embedding technology that was used to create the vector store, vector search in ChromaDB works by converting user queries into vector embeddings. In response to the user's inquiry, the top-k comparable embeddings are chosen, and their associated data entries are obtained, yielding accurate and contextually relevant information.

2.3 Tuning Prompts for LLM

Once the context and question have been gathered, we must combine the system prompt, user question, and chunks to generate a prompt that can be sent to the LLM.

2.4 Data Collection

We used the "Assorted Medicine Dataset of Bangladesh" dataset [3]. This dataset, meticulously created by AHMED SHAHRIAR SAKIB, offers a solid basis for the knowledge base of our chatbot. The information was gathered by web scraping with Python libraries, guaranteeing an exhaustive and current compilation of data. The dataset includes comprehensive details about various medicines such as their prices, generics, indications, drug classes, dosage forms, and the pharmaceutical companies manufacturing them. The dataset contains 10 attributes. Those are brand id, brand name, type, slug, dosage form, generic, strength, manufacturer, package container, and package size. The Assorted Medicine dataset of Bangladesh consists of 21715 data.

3 Implementation

- **Accessing Llama3 via Groq API:** The Llama3 language model was accessed using the Groq API. Groq's robust infrastructure made it possible for us to effectively utilize Llama3's sophisticated natural language generating and interpretation capabilities. With this configuration, we were able to process intricate questions and produce precise, contextually appropriate answers.

- **Managing Vector Store with ChromaDB:** We used ChromaDB, an open-source vector database designed to handle high-dimensional vectors, to manage our vector store. ChromaDB was set up with a 1000 chunk size and a 100 chunk overlap. This configuration made sure that vector embeddings were stored and retrieved efficiently, which let us work with big datasets and get pertinent data fast and precisely.

- **Embedding System with Hugging Face:** We transformed our textual data into numerical vector representations using Hugging Face's embedding technology. These embeddings allowed for efficient processing and retrieval by capturing the semantic substance of the material. The strong architecture of Hugging Face made sure that our embeddings were accurate and dependable, which improved our chatbot's overall performance.

- **User Interface with Streamlit:** We created an interactive and user-friendly platform for the user interface by utilizing Streamlit, a robust framework. Users can ask questions with ease and read replies that have been enhanced with pertinent data from our vector store. Healthcare information was easily available and navigable because of the user-friendly layout, which offered a smooth experience.

4 Results

4.1 Findings

We present screenshots showcasing the interaction between users and our question-answering chatbot application. Users were prompted to ask various questions related to medicines available in Bangladesh, and the chatbot provided informative and contextually relevant answers based on the data from the assorted medicine dataset.

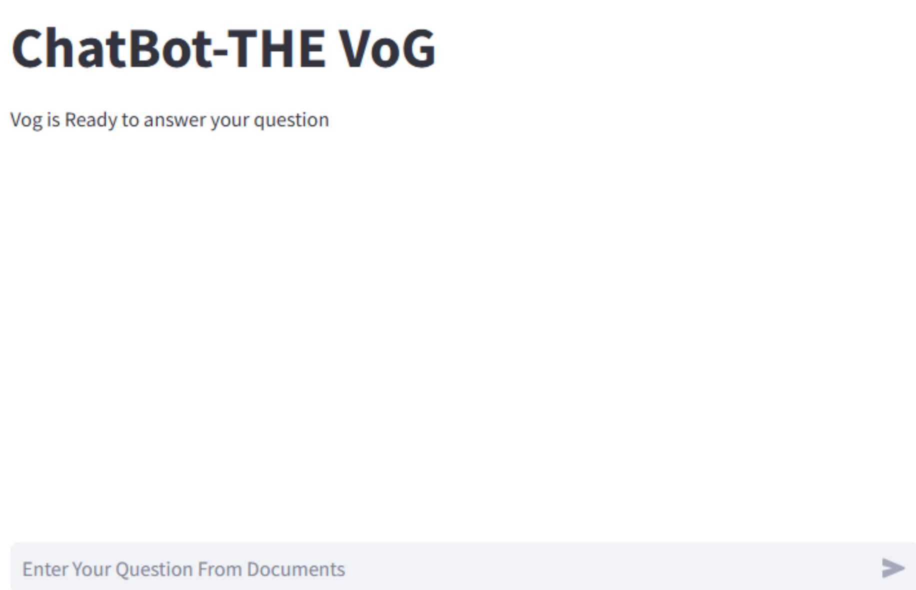

Figure 2:

Figure 3:

The Figure 3 demonstrates the chatbot's question-answering capabilities using the assorted medicine dataset of Bangladesh. It accurately responds to queries based on the knowledge it has been trained on.

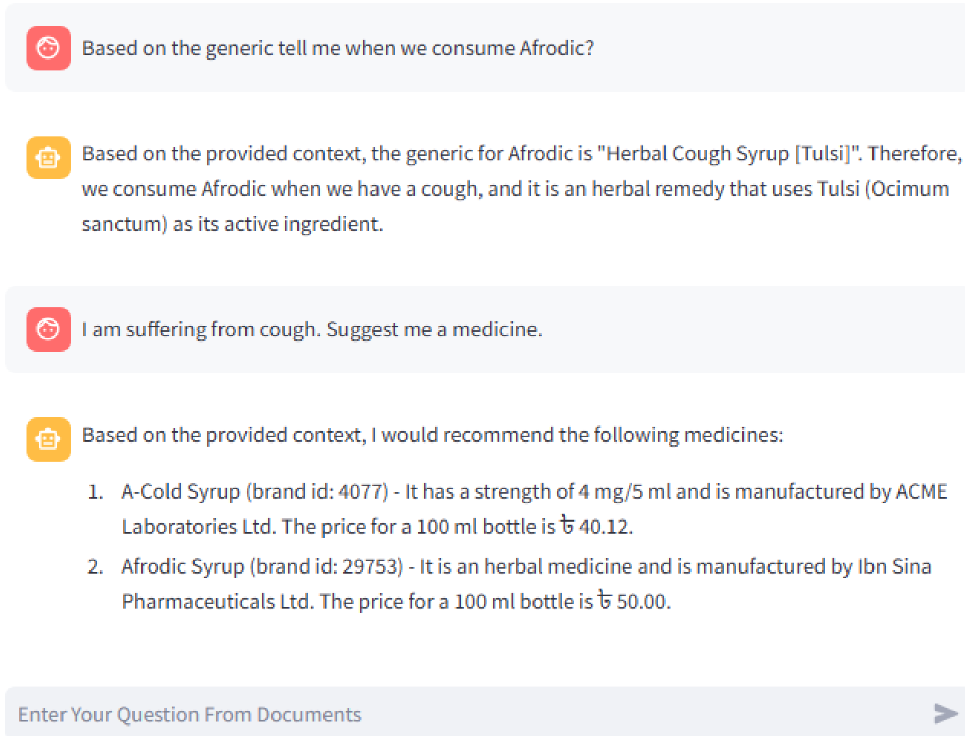

Figure 4:

Figure 4 showcases the chatbot's question-answering proficiency, highlighting its capacity to offer detailed and comprehensive responses that extend beyond the confines of the provided dataset. This ability is facilitated by the integration of Llama 3 intelligence which enables the chatbot to extrapolate information and provide correct answers to diverse queries.

Question	Response of chatbot-THE VoG	Expected Output
What is the slug of Carmina?	The slug of Carmina is carminasyrup.	The slug of Carmina is carminasyrup.
What is the strength of 5-Fluril?	According to the provided context, the strength of 5-Fluril is 25 mg/ml.	The strength of 5-Fluril is 25 mg/ml.
What is the generic of Adovas?	Based on the provided context, the generic of Adovas is Herbal cough syrup [Adhatoda Vasica].	The generic of Adovas is Herbal cough syrup [Adhatoda Vasica].
What are the type and dosage form of Apelin?	Based on the provided context, the type of Apelin is ”herbal” and the dosage form is ”Syrup”.	The type of Apelin is ”herbal” and the dosage form is ”Syrup”.
Who is the manufacturer of A-Cal?	The manufacturer of A-Cal is ACME Laboratories Ltd.	The manufacturer of A-Cal is ACME Laboratories Ltd.
Based on the generic, tell me when we consume Afrodic?	Based on the provided context, the generic for Afrodic is ”Herbal Cough Syrup [Tulsi]”. Therefore, we consume Afrodic when we have a cough, and it is an herbal remedy that uses Tulsi (Ocimum sanctum) as its active ingredient.	The generic for Afrodic is ”Herbal Cough Syrup [Tulsi]”. Therefore, we consume Afrodic when we have a cough.
Provide few medicines name of ACME Laboratories Ltd	Based on the provided context, the medicines manufactured by ACME Laboratories Ltd are: 1. Acme’s Chyabanprash (Oral Suspension) 2. Acmina (Syrup) 3. A-Care (Tablet)	1. Acme’s Chyabanprash (Oral Suspension) 2. Acmina (Syrup) 3. A-Care (Tablet)

Table 1: Some sample questions and answers given to the chatbot

4.2 Addressing Research Questions

The results of this project demonstrate that the developed Retrieval-Augmented Generation (RAG) model effectively meets the research objectives:

- **Developing a Good RAG Model:** The chatbot uses llama3, Hugging Face, and chromadb in developing a reliable Retrieval-Augmented Generation model customized for the Assorted Medicine Dataset of Bangladesh. The model's capability to retrieve and generate information from the dataset ensures that the model is working as intended, corresponding to the main research focus of developing a good RAG model for this dataset.

- **User Interface:** The interface design of the chatbot is user-friendly, and users can easily interact with it by querying and getting information. The interface is designed in such a way that one can interact with the chatbot by asking open questions regarding the different aspects of medicines. This satisfies the objective of providing a friendly and easy-to-use interface for querying the dataset.

- **Opens-Ended Questions and Competence Evaluation:** One of the key objectives has been an assessment of the chatbot's ability in replying to open questions, detailing comprehensive answers. From the given results, the chatbot can reply with profound detail, which is not only factual but also discusses quite complicated questions with a lot of information. That is, the chatbot can be said to be competent in giving extensive answers, which, in effect, has met the research objective of assessing its performance over open-ended questions.

5 Discussion

Chatbot The-Vog showcases the potential of advanced technology in improving access to medical information. This chatbot effectively answers queries related to various aspects of medicines such as brand names, dosages, generic information, and more. This indicates a significant step forward in the accessibility of pharmaceutical data.

5.1 Implications

- It will assist healthcare professionals, pharmacists, and patients in getting proper information about medicines quickly and precisely. It will also help in better patient education and support pharmacists in their daily operations.

- Its ability to provide instant and accurate responses reduces the time spent searching for information. This can lead to more efficient workflows in healthcare environments and potentially reduce errors associated with manual information retrieval.

- It will bridge the gap between patients and healthcare providers by offering easily understandable information on medication. This fosters better-informed healthcare decisions and discussions.

5.2 Limitations and Potential Impacts

Despite the promising outcomes, there are several limitations of this RAG chatbot which could impact the results and broader applicability.

- The chatbot is trained in the Assorted Medicine Dataset of Bangladesh only. So, it does work with a good success rate in this scenario, though it might not do well with other datasets or regions until it is not trained on more diversified data sources.
- The responses generated by this chatbot rely on how precise and comprehensive the underlying dataset is. Inaccuracies or gaps in the dataset could result in inaccurate or lacking responses, which could have serious implications in a healthcare setting.
- Llama3 is fine-tuned for English. As a result, our chatbot is able to process queries in English without any difficulties but may not respond appropriately if queries are raised in other languages. This kind of language restriction reduces the usability of the chatbot for users who are comfortable using any other language other than English. Additional models would be required to handle multiple languages, with further training on datasets that encompass those languages for accuracy and reliability.
- The performance of llama3, HuggingFace, and ChromaDB, all depend on the computational resources and infrastructure. In resource-constrained settings, deploying and maintaining such a system might be challenging, hence limiting its access and scalability.

6 Conclusion

In this paper, We introduce a Retrieval-Augmented Generation (RAG) chatbot using advanced AI technologies like Llama3, LangChain, HuggingFace, ChromaDB, and Streamlit for the assorted medicine dataset of Bangladesh. The dataset consists of 21715 data. In addition, the chatbot efficiently responds to questions regarding medicines for healthcare professionals and patients. The system acknowledges limitations such as dataset specificity, language constraints, and dependency on resources. In the Future, research will focus on expanding the dataset and improving resource efficiency. This RAG-based chatbot implements a noteworthy advancement in healthcare solutions, providing a potential foundation for future research in making medical information more accurate, and reliable for users.

7 References

References

[1] Ahmet Yusuf Alan, Enis Karaarslan, and Omer Aydin. A rag-based question answering system proposal for understanding islam: Mufassirqas llm. *arXiv preprint arXiv:2401.15378*, 2024.

[2] Kurnia Muludi, Kaira Milani Fitria, Joko Triloka, et al. Retrieval-augmented generation approach: Document question answering using large language model. *International Journal of Advanced Computer Science & Applications*, 15(3), 2024.

[3] Ahmed shahriar sakib. Assorted medicine dataset of bangladesh. 2022.