Table of Content

Link to Google Colab:

https://colab.research.google.com/drive/1bPK3zFUx5pI_sjbBSu0GtoTrSTCLQ8oq?usp=sharing

Introduction

Retrieval Augmented Generation (RAG) is an advanced approach that combines the strengths of information retrieval and generative models to produce more accurate and contextually relevant responses. By integrating LLMs with specific external data, RAG allows for more precise, reliable, and context-specific responses. The project begins with identifying a relevant domain and sourcing an appropriate dataset from publicly available repositories. The dataset forms the foundation of the RAG system, providing a rich knowledge base for subsequent analysis. Then, the selection of the chunking strategy and pre-trained embedding model are discussed, followed by the setup of an appropriate prompting template. The system's capability is evaluated through a set of predefined tests, ensuring it outperforms traditional LLMs in accuracy and efficiency. This project demonstrates the practical applications of RAG systems in enhancing LLM performance and handling specific, professional knowledge domains.

Knowledge-Domain Selection

In this project, I have chosen to use Full Text News data on HuggingFace (2023) as the Retrieval Augmented Generation (RAG) system input. The data contains 5,059 rows and 15 columns of English news data published from 10 May 2023 to 1 Aug 2023 from various sources; column ‘maintext’ contains the full-text data of each news and thus will be the main input data. Out of the 5,059 rows, 3,305 of them are unique entries; duplicated contexts are removed before moving on to further steps. Please refer to Appendix A for the data dictionary. The primary reason for employing news data is that it provides a rich, diverse, and current source of information, which enhances the system’s ability to generate accurate and up-to-date responses and is ideal for examining model performance.

Testing Questions Setup

To better assess the RAG model's ability to deliver precise responses, several testing questions have been formulated, each targeting a different evaluation perspective. The questions and their objectives are as follows:

- Question 1: Give me a comprehensive introduction of the shipping company Yellow Corp.

  This question is straightforward and general, with answers that remain relatively consistent over time. It aims to test if the RAG model, combined with the news data input, can provide more detailed answers compared to using the LLM model alone for general inquiries.
- Question 2: Who were the victim and perpetrator in the murder-suicide incident in Little Egg Harbor, New Jersey?

  This question is more specific. It is designed to test the RAG model's ability to extract detailed and accurate information from the news data input.
- Question 3: Why did 911 calls for severe allergic reactions nearly double in the summer? What measures can be taken to prevent serious allergic reactions?

  This is a more complex question. It aims to evaluate the RAG model's capacity to synthesize scattered information from the news data input, providing a comprehensive and detailed response compared to the LLM model alone.
- Question 4: Who is Emma Stone?

  The answer to this question is not provided in the given context. This question is designed to test whether the RAG model can recognize that the question falls outside the provided context and refrain from generating a response.

Chunking and Embedding

After loading the dataset and storing each row in the 'maintext' column into separate text files, the length of each text was found to be too long. Large text contexts can make locating relevant information harder and increase latency and computational requirements. Therefore, a semantic chunking method was employed before embedding. Semantic chunking was chosen over arbitrary or other methods because it preserves the meaning and context within each chunk without breaking paragraphs into excessively small pieces, like sentence-level chunking, ensuring more accurate and relevant embeddings for the following tasks. This method improves the model's performance and efficiency. After chunking, the original text data becomes 11,887 separate text and is stored in another folder as text files.

Next, a pre-trained sentence-transformers embedding model is used to map sentences to a multi-dimensional dense vector space. The model chosen is the 'bge-small-en-v1.5' from HuggingFace (Niklas Muennighoff et al., 2023), which maps sentences to a 384-dimensional vector space and has a model size of 33 million parameters. This relatively small embedding model is suitable for large datasets without being computationally demanding. Sentence-level embeddings, unlike word-level embeddings, capture the meaning of entire sentences, including structural elements, and show how semantic representation changes from single words to sentences (Fairhall, 2024). This enhances the efficiency and accuracy of processing and retrieving relevant information from the dataset by providing a more holistic representation of the data.

To implement this, the HuggingFace embedding model is initialized, and the LLM and embedding model are specified in LlamaIndex's settings. Documents are loaded using a SimpleDirectoryReader, and a ChromaVectorStore is set up with a PersistentClient database. A collection is created in the database, and the vector index is generated from the documents using the ChromaVectorStore to store the embedded sentences and serve as our query engine. This setup ensures efficient storage and retrieval of vectorized data.

Prompt Template Development

Following the vectorized data storage, a prompt template is set using the ChatPromptTemplate. This template includes a series of messages designed to instruct the model on responding to questions based on the provided context. The system message directs the model to say "I don't know" if the answer is not found in the context, ensuring accurate and context-dependent responses. This setup enhances the reliability of the generated answers. The prompt template details can be found in Figure 1 below.

```
from llama_index.core.llms import ChatMessage, MessageRole
from llama_index.core import ChatPromptTemplate

qa_prompt_str = (
    "Below is the context information.\n"
    "---------------------\n"
    "{context_str}\n"
    "---------------------\n"
    "Given the context information and not prior knowledge, "
    "answer the question: {query_str}\n"
)

# Text QA Prompt
chat_text_qa_msgs = [
    ChatMessage(
        role=MessageRole.SYSTEM,
        content=(
            "Please just say 'I don't know' if the answer is not provided in the given context."
        ),
    ),
    ChatMessage(role=MessageRole.USER, content=qa_prompt_str),
]

text_qa_template = ChatPromptTemplate(chat_text_qa_msgs)
```

Figure 1. Complete prompt template.

System Testing and Evaluation

To evaluate the system's performance and determine whether it outperforms the LLM model alone, their respective responses to each test question were compared. A response mode of 'tree summarize' was chosen to identify multiple answers from the text chunks, combine them into the context window, and then summarize them into a single response. This mode balances the benefits of both 'refine' and 'compact' modes.

For the first question, the LLM model provided a detailed background and industry overview of Yellow Corp, but it missed the recent news of the company's closure due to financial issues in 2023. Conversely, the RAG model captured this crucial information from the news text, highlighting Yellow Corp's significant role in the trucking industry and the impact of its closure on supply chains and employees. This demonstrates the RAG model's ability to integrate current, relevant information into general responses, enhancing the completeness and accuracy of its responses.

For the second question, the LLM model was unable to provide an up-to-date response, stating it lacked access to recent news and suggesting checking reliable sources for the most current information. In contrast, the RAG model accurately identified the victim and perpetrator, providing specific names and details about the incident. This shows the RAG model's capability to extract timely information from the given news data, making it more effective in responding to specific queries.

For the third question, the LLM model could not provide recent data, suggesting preventive measures like carrying an epinephrine auto-injector, reading food labels, and having an action plan. On the contrary, the RAG model explained the seasonal increase in 911 calls due to greater exposure to allergens and insect stings in summer, offering specific preventive measures such as staying vigilant, recognizing allergy symptoms, and having an Epipen ready. This highlights the RAG model's ability to accurately find and synthesize scattered information pieces in the input context.

For the fourth question, the LLM model provided a detailed background on Emma Stone, listing her notable roles and awards. On the other hand, the RAG model noted the absence of direct mentions of Emma Stone in the provided context but inferred her identity based on references to her popular movies. This demonstrates the RAG model's ability to make informed inferences from the given context, even when explicit information is missing.

Appendix B shows complete output responses from LLM alone and the RAG system.

Limitations and Discussions

Although the incorporation of the RAG system outshines the LLM model alone, there are limitations and areas for improvement. Firstly, the performance and effectiveness of RAG heavily depend on the organization and structure of the underlying data (Besen, 2024), but no measures were taken to assess the quality of the text data input in this project. Additionally, the project uses a small-size embedding model, which may sacrifice the ability to capture more nuanced meanings for shorter embedding times and lower computational complexity. Moreover, the system provides different responses each time the same question is re-queried, resulting in possibly inconsistent and unstable response quality.

Conclusions

In this project, the Retrieval Augmented Generation (RAG) system demonstrated significant improvements over the standalone LLM model by effectively integrating external data to provide accurate and contextually relevant responses. Through comprehensive testing, the RAG system outperformed the LLM model in capturing up-to-date and specific information. However, challenges such as dependency on data quality and potential response inconsistencies remain. Future enhancements could involve improving data assessment methods and exploring larger embedding models to capture nuanced meanings better, ensuring more consistent and stable outputs.

References

Besen, Sandi. (2024) 'The Limitations and Advantages of Retrieval-Augmented Generation (RAG)', *Towards Data Science*. Available at: https://towardsdatascience.com/the-limitations-and-advantages-of-retrieval-augmented-generation-rag-9ec9b4ae3729 (Accessed: 26 May 2024).

Fairhall, S.L., 2024. Sentence-level embeddings reveal dissociable word- and sentence-level cortical representation across coarse- and fine-grained levels of meaning. Brain and Language, 250, p.105389. Available at: https://doi.org/10.1016/j.bandl.2024.105389.

Muennighoff, N., et al. (2023) 'bge-small-en-v1.5', Hugging Face. Available at: https://huggingface.co/BAAI/bge-small-en-v1.5 (Accessed: 24 May 2024).

RealTimeData. (2023). News_August_2023 [Data set]. Hugging Face. Available at: https://huggingface.co/datasets/RealTimeData/News_August_2023

Appendix

A. Data Dictionary

Column Name	Column Description
authors	Names of individuals or organizations that wrote the articles.
date_download	Timestamp indicating when the article was retrieved.
date_modify	Timestamp of the last modification made to the article.
date_publish	Original publication date of the article.
description	Short text providing an overview of the article's content.
filename	File path or name where the article is stored locally.
image_url	Web address of any image linked to the article.
language	Language code (e.g., 'en' for English).
localpath	Local directory path where the article file is stored.
maintext	Full text of the news article.
source_domain	Website domain from which the article was retrieved.
title	Headline or title of the news article.
title_page	Title of the webpage hosting the article.
title_rss	Title of the article as it appears in the RSS feed.
url	Web address of the news article.

B. The Responses from LLM Model and RAG System

Question1	Give me a comprehensive introduction of the shipping company Yellow Corp.
LLM Model	Certainly, Yellow Corporation is one of the largest transportation companies in the world with its global network covering more than 200 countries. It was founded in 1853 and started as an ice cream delivery service in New York City before expanding into the trucking industry. Over time, it became a major player in the shipping and logistics space by acquiring various transportation-related businesses like Ryder, Con-way, and Cintas Corporation. Today, Yellow Corp is one of the top three largest freight carriers in North America with over 120,000 employees worldwide. It offers a wide range of services including air cargo, sea freight, trucking, and intermodal transport. With its extensive network, Yellow Corp provides efficient and reliable transportation solutions to businesses across various industries such as healthcare, technology, and retail.
RAG Model	Based on the context information provided, it appears that Yellow Corporation is an American trucking firm specializing in less-than-truckload service. The company was founded by two brothers who started their business with one truck and have since expanded to a fleet of over 2,000 trucks. They primarily transport goods for large corporations like Walmart and The Home Depot.

	However, Yellow Corporation has recently faced financial trouble and announced its closure. This has raised concerns about the impact it will have on supply chains across the country, as well as leaving many employees without jobs. Despite the uncertainty, it is clear that Yellow Corp had a significant presence in the trucking industry and played an important role in transporting goods to customers.
Question2	Who were the victim and perpetrator in the murder-suicide incident in Little Egg Harbor, New Jersey?
LLM Model	As an AI language model, I do not have access to up-to-date news or events. However, as of September 2021, there is no record of a recent murder-suicide incident that occurred in Little Egg Harbor, New Jersey. It's always important to check reliable sources for the most current information.
RAG Model	The woman found dead in her apartment was identified as Kimberly Hoffman, 49, and her attacker was her ex-husband, Carl Schulz Jr., 52.
Question3	Why did 911 calls for severe allergic reactions nearly double in summer? What measures can be taken to prevent serious allergic reactions?
LLM Model	I do not have access to recent data or statistics regarding the increase of 911 calls related to severe allergic reactions during the summer months. however, there are several measures that individuals can take to help prevent a serious allergic reaction, such as carrying an epinephrine auto-injector at all times, reading food labels carefully, and avoiding known allergens. it is also important for individuals with severe allergies to have an action plan in place in case of an emergency, which may involve seeking medical attention immediately and calling 911 if necessary.
RAG Model	In the summertime, people are more prone to insect stings and exposure to allergens such as peanuts, milk, and eggs at picnics and barbeques. As a result, BCEHS sees an increase of almost double the usual calls to 911 for severe allergic reactions. To prevent serious allergic reactions, it is important to stay vigilant and watch for the signs, including severe skin rash, swollen lips and eyes, swelling in the tongue or throat with difficulty swallowing, and trouble breathing. People with severe allergies should always have Epipen on hand and make sure they are not expired. If a person experiences anaphylaxis, it is important to call 911 immediately while remaining calm and following the advice of dispatch staff.
Question4	Who is Emma Stone?
LLM Model	Emma Stone is an American actress known for her roles in films such as "La La Land," " Easy Rider," "The Help," and "Crazy, Stupid, Love." She has won three Golden Globes and two Academy Awards, making her one of the most successful actresses in Hollywood.
RAG Model	Based on the context information provided in the given text files, it seems that there is no direct mention of Emma Stone as a person or character. However, based on her presence in popular media such as movies like "La La Land" and "The Help," she can be identified as an actress who has achieved great success in Hollywood.