Group: A11

Title: Enhancing Medical Question Answering with Retrieval Augmentation and Reranking Transformer Embeddings in Generative Language

Method:

Design:
The design of our Retrieval Augmented Generation (RAG) model was a careful and robust process, aimed at combining the power of large language models and high-performance retrieval techniques. This combination of state-of-the-art technology can be a powerful system designed for query and text generation in a medical day and Generator.

Retriever component:
The Retriever component was designed to take the vast amount of medical knowledge stored in our vector database and analyze it efficiently. The main objective of this phase was to identify and retrieve the most appropriate data for a given question or input. To achieve this, we used a state-of-the-art dense retrieval method that leverages the power of semantic similarity searching.

Architectures:

- **Vector database storage:** We realized the need for scalable and high-performance storage solutions to accommodate and index effectively. To meet this need, we used the FAISS (Facebook AI Similarity Search) vector store from the langchain.vectorstores module. FAISS is a highly optimized library for efficient similarity searching and clustering of dense vectors, making it the best choice for the retrieval component of our RAG model. By storing chunked and embedded data in this vector database we made sure to get the relevant information quickly in the generation process.

- **Text chunking and embedding:** We have implemented state-of-the-art text chunking and embedding techniques to facilitate high-quality text retrieval and indexing. Specifically, we used the 'intfloat/e5-large-v2' embedding model in conjunction with the RecursiveCharacterTextSplitter method from the LangChain library. In this way, pure text data was divided into manageable pieces, each 6,000 characters in size, with inconsistencies between adjacent texts on the basis of these slices and inserts, we enabled better contextualization and alignment during the process of the RAG model.

Generator Component:
The Generator component of our RAG model was responsible for aggregating the received data and generating coherent and contextual responses. The Mistral 7B was a state-of-the-art post-teaching speech model, known for its ability to understand complex instructions and work with exceptional natural language generation

The Mistral 7B model was optimized on medical textbooks, enabling it to gain a deeper understanding of the vocabulary, concepts, and linguistic structure of specific domains in medicine This specialized training ensured that the responses were not only grammatical meaning and coherence

To facilitate the integration in the generation process of retrieved data, we used a new system that combined the Retriever component with the Mistral 7B model This system called the Retrieval Augmented Generator (RAG) was provided the language model was able to focus on the input queries and retrieved-context simultaneously

Architectures:

1. **Quantization:** Similar to the Retriever part, we used the quantization techniques to reduce the memory footprint and computational overhead of the Mistral 7B model We used a '4-bit neural float' (nf4) quantization scheme, which allowed us to model using 4 bits load and activation can be represented, resulting in significant memory savings and improved computational efficiency

2. **Mixed Precision Training:** We used mixed precision training techniques to speed up the training and fine-tuning process. Mixed precision training using tensor cores and specialized hardware instructions enabled low-precision (e.g., 16-bit floating point) computations while maintaining high accuracy, resulting in significant performance gains without sacrificing accuracy

The simple integration of retriever generator components, using various optimization methods, resulted in a high-performance and scalable RAG model capable of providing accurate contextual queries in the medical domain of the This design not only took advantage of the latest developments in natural language processing and retrieval strategies Usage and efficiency were prioritized, ensuring a viable solution and it works for real-world application scenarios

Objective: The main objective of this study is to develop a state-of-the-art retrieval augmented generation (RAG) model designed for query and text generation tasks in the medical field. This will provide healthcare professionals, researchers, and students with easy access to reliable and up-to-date medical knowledge.

Basically, our goals were two fold: first, to build a robust RAG model that can efficiently derive relevant information from medical data sets, and second, to use empirical knowledge that will include state-of-the-art language models to create consistent and contextual responses. Achieving these goals will not only facilitate access to information but also facilitate the dissemination of knowledge, ultimately contributing to improvements in patient care, research development, and medical education

Study Selection: In order to meet our ambitious objectives, we recognized the critical importance of obtaining high-quality and comprehensive data from a reputable source in the medical field. After careful research, we selected the Medeasy website (www.medeasy.com) as our primary data source. Generally, it is a website where we can buy online medicine from our house, however, the website is filled with resources. it had all the details of medicine, price, and quantity to use also the side effects were listed in the description of the relevant medicine.
There were several main reasons for this decision:
Comprehensive collection: Medeasy boasts an extensive collection of medical information, covering a wide range of topics from disease definitions and treatment protocols to groundbreaking research findings and clinical guidelines
Trusted Reputation: Medeasy has established itself as a trusted and authoritative brand in the healthcare community, known for its commitment to accuracy, reliability, and integrity to the highest scientific and medical standards
Regular updates: The website is regularly updated by a team of experienced medical professionals and subject matter experts, ensuring that the information remains current and relevant in medical terms if is growing rapidly.
Access: Medeasy's user-friendly interface and comprehensive search functionality make it easy to navigate through any effort and retrieve the desired information, making it the right candidate for our data acquisition efforts
The first dataset obtained from Medeasy, hereafter referred to as 'Medeasy.csv', consisted of 4,831 rows and 7 columns, each row representing a specific medical issue or condition, with detailed description and metadata including the relevant

EDA:

Data Extraction: The scraped data was initially stored in a CSV file ('Medeasy.csv') with 4,831 rows and 7 columns, which had various columns such as descriptions, type, price of medicine, unit, company name and generic name of the medicine, URLs, and associated metadata. The ' 'description'' column that contained the original text data went through an extensive cleaning process to remove extraneous elements that could interfere with the performance of our RAG model This careful process involves the following steps. In the end, we have removed URLs, associated metadata, and other unnecessary columns from the processed_csv which were in the original CSV file.
After that, we get the data like this

```
train.head()
```

	Name	Price	unit	Type	Genric name	Company name	Description
0	Rivotril0.5 mg	Tk 8.28	/piece	Tablet	Clonazepam	Radiant Pharmaceuticals Ltd.	Medicine overviewEnglishIndications of Rivotri...
1	Montair10 mg	Tk 157.50	/10's Strip	Tablet	Montelukast	Incepta Pharmaceuticals Ltd.	Medicine overviewEnglishIndications of Montair...
2	Bislol5 mg	Tk 144.90	/14's Strip	Tablet	Bisoprolol Fumarate	Opsonin Pharma Ltd.	Medicine overviewEnglishIndications of Bislol ...
3	Rosuva5 mg	Tk 90.27	/10's Strip	Tablet	Rosuvastatin	Square Pharmaceuticals Ltd.	<img alt="Section" srcSet="/icons/note.svg 1x,...
4	Clopid75 mg	Tk 151.20	/14's Strip	Tablet	Clopidogrel Bisulphate	Drug International Ltd.	<img alt="Section" srcSet="/icons/note.svg 1x,...

- **HTML tag removal:** We used standard terminology to identify and systematically remove any HTML tags embedded in the text, ensuring that the data remained clean and uncluttered the right way
- **URL removal:** Similarly, we used regular terminology to identify URLs and removed them from the text, as they were considered irrelevant for the RAG project and could introduce noise or bias
- **Stop removing words:** To further enrich the textual content, we used the highly acclaimed NLTK (Natural Language Toolkit) library to identify and remove frequent grammatical words (e.g., 'the';, 'a', 'is') from the 'explanation' ' ' division. This step helped increase the signal-to-noise ratio by placing the data in more informative and relevant steps.
- **Irrelevant Pattern Removal:** Since we realized that some patterns such as emoticons or special characters could negatively affect the performance of our speech models, we used routine annotation-based expression filtering to remove those items this unnecessary from the text

After completing an extensive data preprocessing and transformation pipeline, the resulting cleaned and transformed dataset, henceforth referred to as 'processed_data.csv' contains 4,627 rows and 7 columns It is worth noting that this procedure resulted in a reduction of 204 rows from the original dataset, with the removal of non -string values, And because of strict maintenance procedures

By following this rigorous approach, we ensured a high-quality dataset from a reliable source, applied ethical and responsible web scraping practices, and conducted compliant data preprocessing and manipulation to optimize the dataset for restoration and enhanced generation. The development and subsequent evaluation of the model laid a solid foundation, preparing us to reach our goal of providing more accurate and relevant information in the medical field.