Retrieval Augmented Generation With MedQuAD Dataset

Team Name: Outliers
Team Members: Soumya Chowdary Daruru, Abishek Vanam, Susrutha Kanisetty

Stony Brook University
Computer Science

Introduction

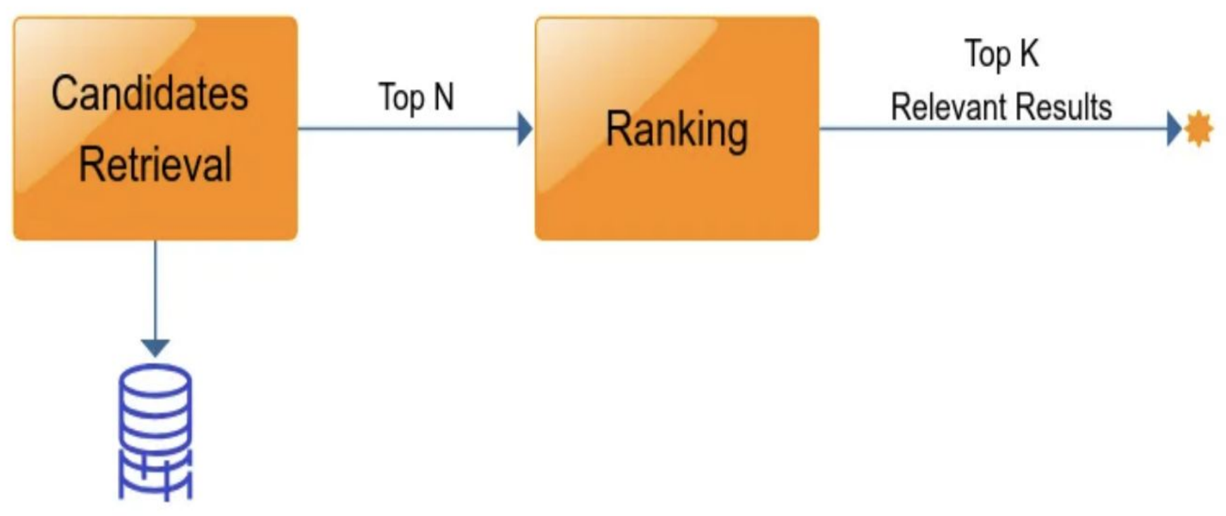
- In today's digital era navigating through vast amount of medical knowledge is cumbersome.
- Direct usage of LLMs is not reliable when it comes to querying specific data.
- It often produces misinformation and hallucinations
- We propose to use modern NLP techniques to implement RAG and extend it to medical domain using local data source for fast and reliable information.

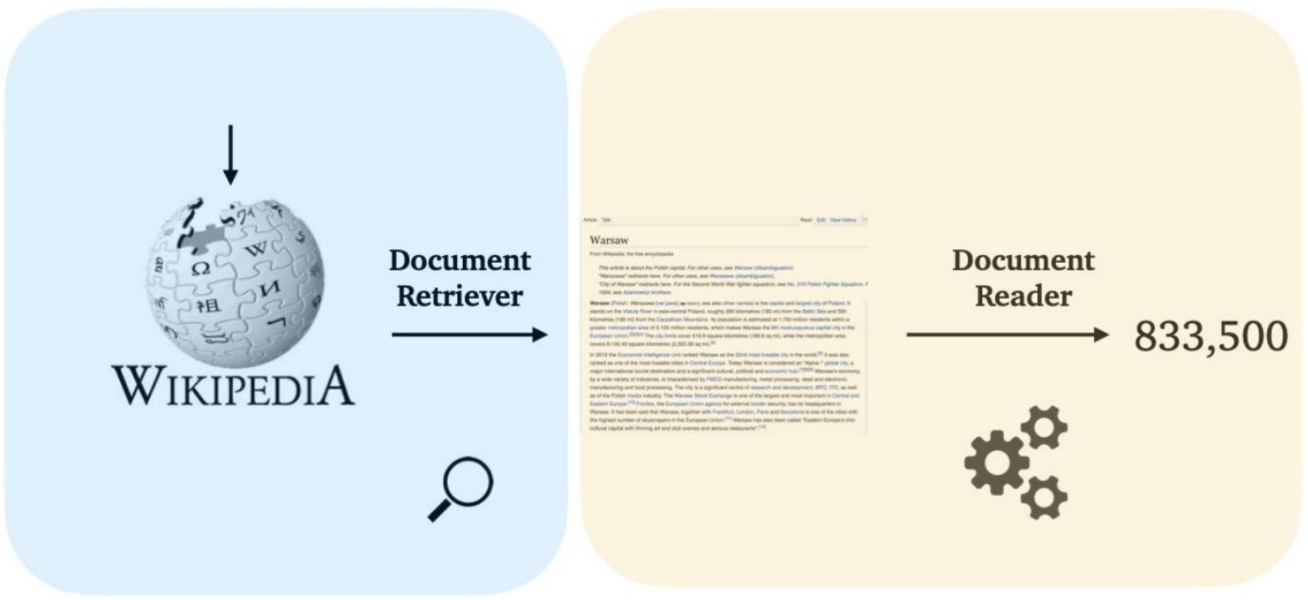
Background

RAG combines three core principles:

1. Retrieving information from various reliable sources.
2. Enhancing the prompts passed to LLMs with relevant contextual information and user queries, and
3. Generating responses using LLMs for accurate results.

This requires knowledge of information retrieval systems, the workings of transformers, LLMs, and effective prompt engineering.

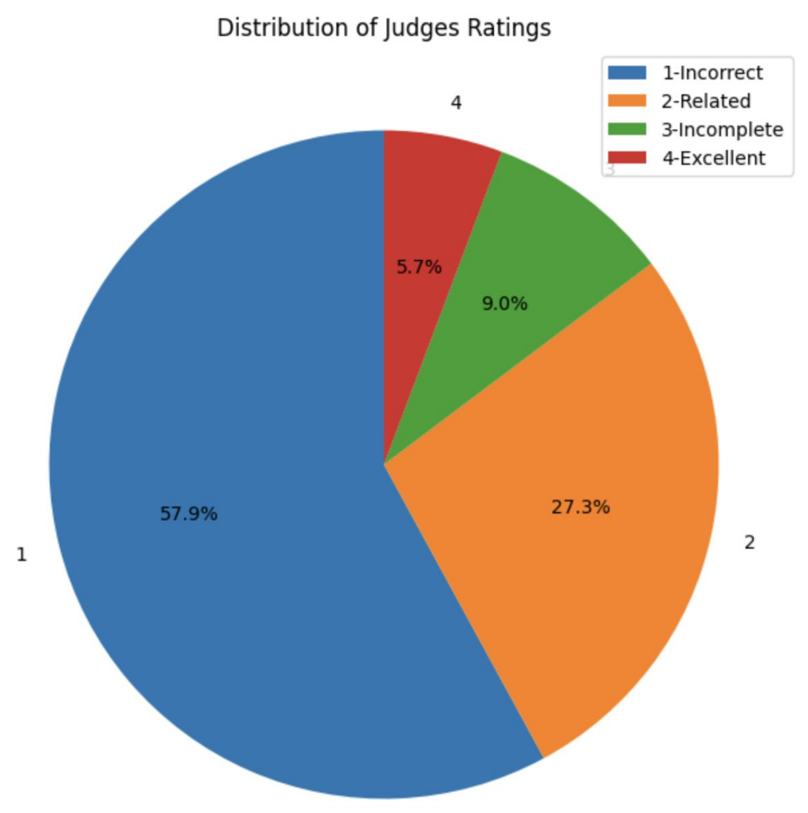




Data

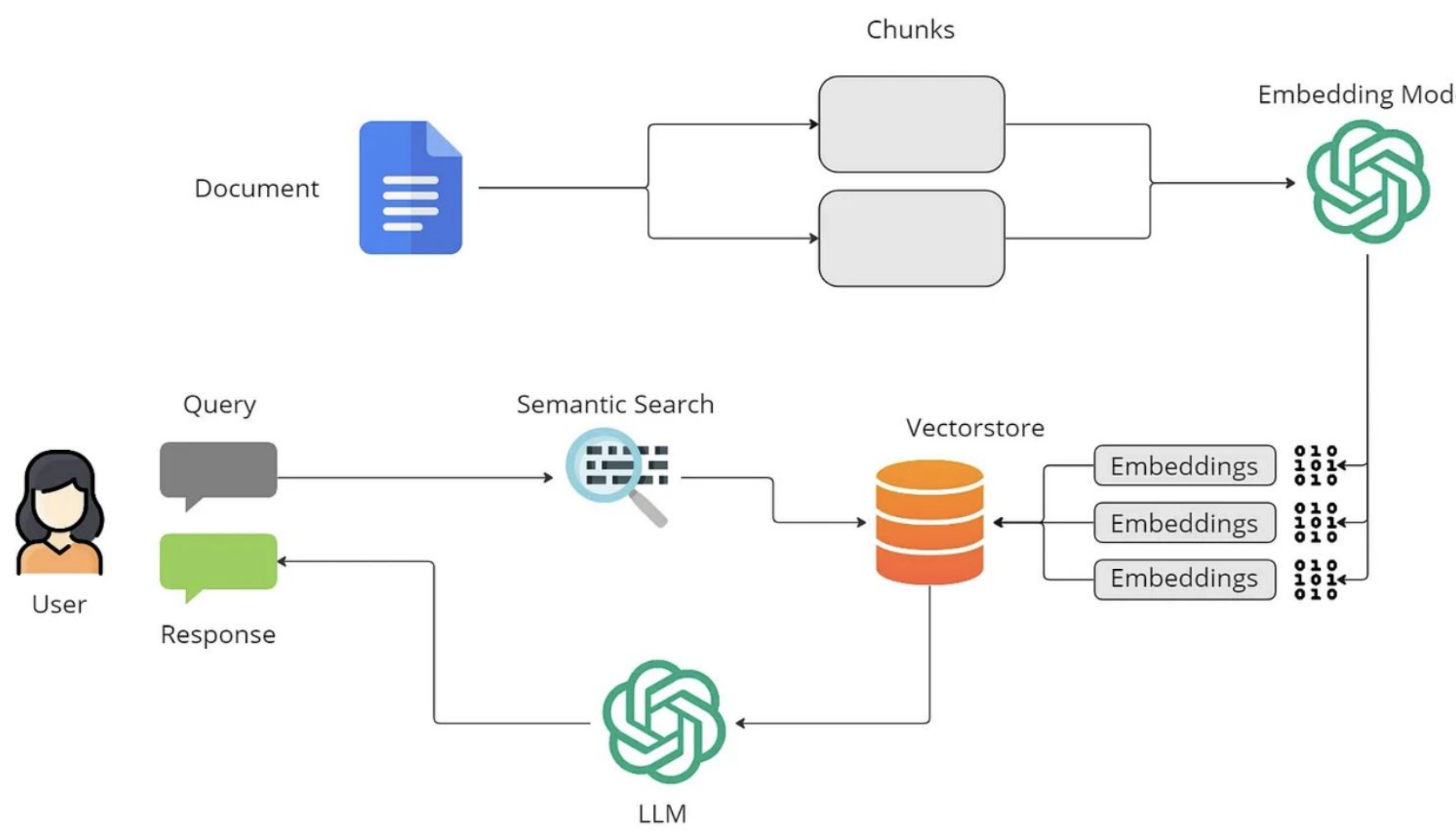
Used MedQuAD Dataset with 47,457 records as a source for IR and TREC-2017 LiveQA medical task Dataset with 2,479 judged answers with 4 ratings.

Statistic	Value
Total Word Count	6.1M
Total Token Count	8.2M
Avg.QA Sentence Length	151.6



Methods

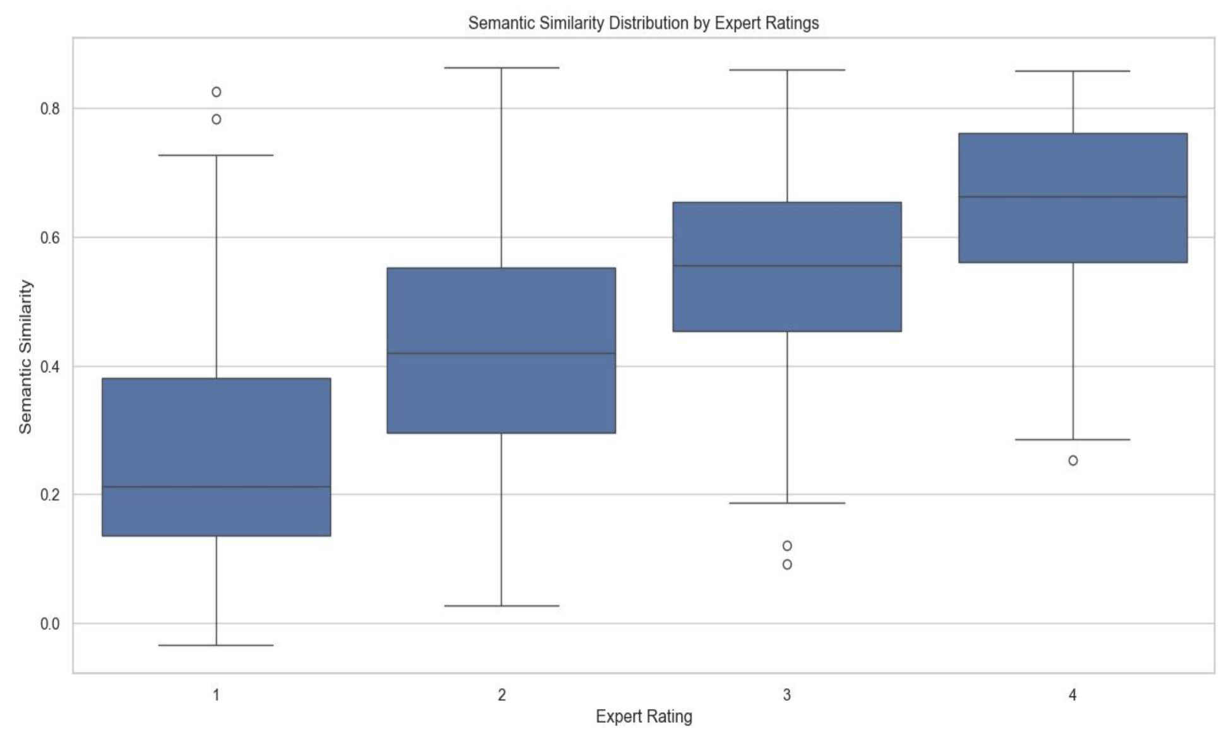
RAG pipeline:

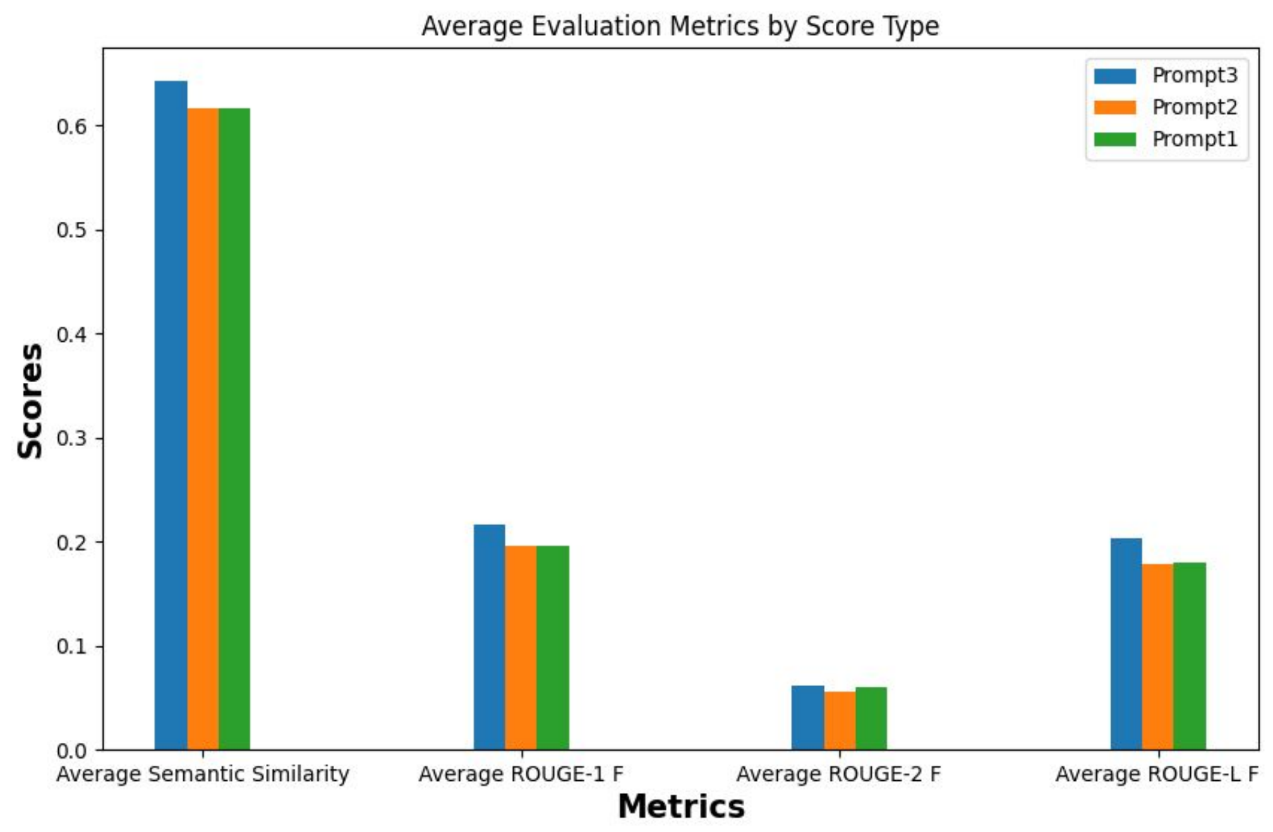


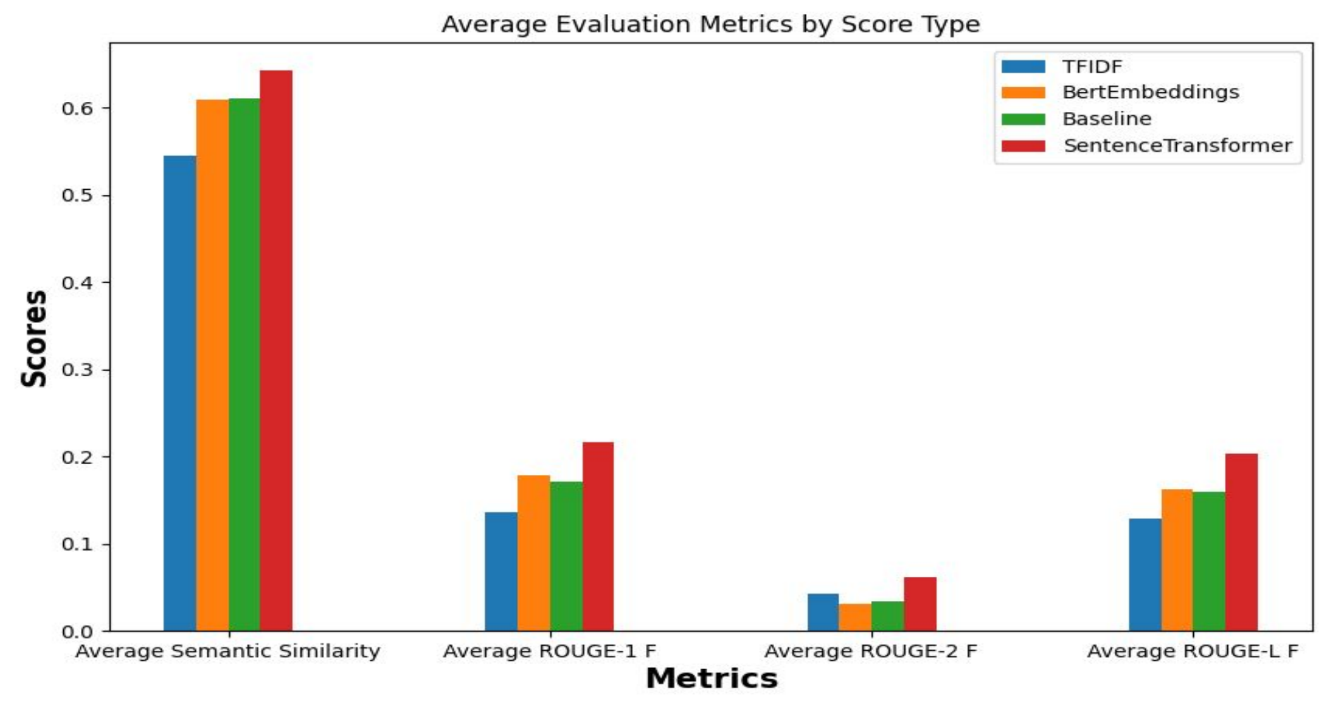
LLM Used: Gemma-2b

- Baseline 1: Passed Query directly to the LLM and tested the response
- Baseline 2: Used IR on MedQuAD dataset by embedding the data using TF-IDF and passed query+context to LLM.
- Enhancement 1: Used BERT Embeddings from k-1 layer for better contextual embeddings
- Enhancement 2: Used Sentence Transformers for embeddings
- Prompt Augmentation with specific instructions and examples to the LLM gave more relevant answers.

Results







- **Models Evaluated:** TFIDF, BertEmbeddings, Baseline (direct LLM use), SentenceTransformer.
- **Top Performer:** SentenceTransformer embedding model with Gemma-2b-it using augmented Prompt.
- The structure and specificity of prompts significantly influence the model's performance, with more detailed prompts leading to better outcomes.

***P** : User query: {query}. Use the following additional information retrived from local information storage to enhance the answers. Make sure the answers are explanatory.{context}. Answer:*

Conclusion

The implementation of RAG model with MedQuAD dataset, using SentenceTransformer embeddings and prompt engineering, has significantly improved medical question answering by addressing the limitations of general LLMs

References

1) https://medium.com/@prasadmahamulkar/introduction-to-retrieval-augmented-generation-rag-using-langchain-and-lamaindex-bd0047628e2a`
2) https://itnext.io/deep-learning-in-information-retrieval-part-i-introduction-and-sparse-retrieval-12de0423a0b9
3) https://mycourses.stonybrook.edu/d2l/lp/navbars/1135850/customlinks/external/50714