Aware RAG Project Group 1
Erdos Institute Data Science Bootcamp, April 2024
Executive Summary

Project team: Yangxiao Luo, Ness Mayker Chen, Merve Keskin
Mentor: Jason Morgan
Github: https://github.com/NessMayker/aware_nlp.git

Motivation and Goal

AI systems that generate responses use the most relevant data available to them. Nevertheless, there is a possibility of hallucinations leading to inaccurate outcomes. Hence, it is crucial to develop systems capable of delivering answers that are relevant to the context. Where is the source of this relevant data within the system? How well does it fit into the specific context?

Our objective for this project is to create a fast delivery system that incorporates external contextual reference data. The system needs to have the ability to efficiently search and rank millions of texts in response to user queries. The system also incorporates mechanisms for evaluating the qualitative performance of the retrieval outputs.

Data Organization

Challenge: create a fast and reliable retrieval augmented generation system that can accurately match a user query with documents that provide relevant context to then feed into a language model.

Our work analyzes a vast public dataset comprising 5 million Reddit submissions and comments across 21 subreddits. The dataset consists of written submissions and corresponding comments. In order to establish a reliable connection between our questions and answers in the data, we eliminated all submissions that lacked comments. Then, we reorganized our data into a new format where each submission is stored in one row, with a list of each posts' corresponding comment indicies are stored in a new column. This led us to create a submission-based data frame comprising 20613 observations.

Model

We use pre-trained sentence embedding for our retrieval. We compared the following embeddings "all-MiniLM-L6-v2", "nq-distilbert-base-v1", and "thenlper/gte-large". We found similar performance across all embeddings, so decided to use "all-MiniLM-L6-v2" as our embedding of choice due to its faster performance.

We combined the title, text, and subreddit category of each submission into one string. We used embeddings to transform the string into an encoded format. We went through this process for every submission and stored our vectors using LanceDB.

Then, we applied the KNN model to identify the reddit submissions that were the most relevant to a user's query. We feed the embedded query into LanceDB and convert the retrieved submissions into a response string comprised of the submission title, text, and top user comment. Together we feed the query and response into a prompt template that is targeted for use by the LLM.

Evaluation

We used Ragas to assess our LLM. Ragas considers both the retrieval and generation aspects of the model, including context recall, context relevancy, faithfulness, and answer relevancy.

By utilizing Ragas, we generated ground truths by analyzing the most frequently uploaded comments for each submission. We then compared the relevance of answers between the outcomes produced by our llama 3-8B models with no contexts and those with contexts provided by our RAG system. According to the model comparison, our RAG system demonstrates better performance in answer relevancy compared to a large language model without context.