BITS Pilani
Work Integrated Learning Programmes Division
M.Tech (AIML)
Conversational AI
AIMLCZG521
2023-24

CAI Assignment 1 - Problem Statement H

Instructions

1. Follow the instructions in each question carefully.
2. Only two files should be uploaded in canvas without zipping them. One is an ipynb file and the other one html output of the ipynb file. No other files should be uploaded.
3. Any assignment submitted using other python IDEs are not considered for grading.
4. Incorrect Assignment Set submitted will not be considered.

Dataset Link

1. Go to : *https://www.sec.gov/edgar/searchedgar/companysearch*
2. Under "**Company and Person Lookup**" Search by company name
    a. example: "Amazon.com"
3. In "**Search Result**" page, on right hand side under "Selected Filing" title
    a. Expand "**10-k (Annual Report)"** section
    b. Download 10-K filing for the year ending 2023.

Example for Amazon:
https://www.sec.gov/ix?doc=/Archives/edgar/data/1018724/000101872424000008/amzn-20231231.htm

You need Use 10-K filing for "Alphabet INC"

Problem Statement:

Implement an Advanced-RAG system using Vector Store of choice for Retrieval and Pre trained LLM (Any LLM of choice) Generation. You need to implement 2 services:

1. Ingestion Service: Ingest the data from files into a vector store.[Only text data]
2. Retrieval Service: Takes user query, Retrieve relevant data from vector store and generates answer to the user query.
3. You can use LLM and embedding models of your choice.
4. You have to use either Langchain or Llamaindex for implementation of the above services.

Tasks

Ingestion Service **[5 Marks]**

1. Extract text data from pdf files. Explain which library you have used for text extraction and why. [1 Mark]
2. Convert text into chunks. [2 Mark]
    a. Explain why you used a particular chunking strategy.
    b. Explain what chunk size / parameters you have used and why.
3. Vectorization of chunks into embedding vectors. Explain which embedding model you have used and its pros and cons. [1 Mark]
4. Ingest data to the vector store. [1 Mark]

Retrieval Service **[10 Marks]**

1. User provides a text query.
2. Query cleaning and rewriting [2 Mark]
    a. Rewrite user query for improved Retrieval
    b. Example: Simple rewriting, Fusion RAG, Query decomposition
3. Retrieve relevant chunks from the vector store. [3 Mark]
    a. Implement advanced retrieval technique for higher recall value.
    b. Simple retrieval would not get full marks.
4. Reranking [1 Mark]
    a. Implement reranking technique to select most relevant chunks
5. Response Generation [2 Mark]
    a. Showcase how hallucination can be minimized.
    b. Showcase how precision can be improved at this stage.
6. Showcase question-answer capability for 3 type of questions: [2 Mark]
    a. Simple: Easy and direct question whose answer is present on a single page.
    b. Medium: Question whose answer is present on multiple pages [2-3 pages of pdf].
    c. Complex: Multiple questions asked in a single query.