Problem statement - Assignment 2

1. Where would you rate yourself on (LLM, Deep Learning, AI, ML). A, B, C [A = can code independently; B = can code under supervision; C = have little or no understanding]

   Answer:

   AI: Comfortable building applied AI systems, and my strength lies in AI integration and product thinking.

   LLMs: Can use LLM APIs, Build Rag pipelines using LangChain, perform Prompt engineering

   ML: Can perform data preprocessing, train ML models, and evaluate models. I am also learning to use MLflow for monitoring the status of the deployed models.

   Deep Learning: Have knowledge on neural networks, ANNs, CNNs, RNNs, and published a paper on 'Prediction of Brain Diseases Using Machine Learning Models' - https://link.springer.com/chapter/10.1007/978-981-19-7753-4_74

   Overall I would rate myself B.

---

2. What are the key architectural components to create a chatbot based on LLM? Please explain the approach on a high-level

   Answer:

   The system starts with the User Interface layer, where users provide prompts and optionally upload documents like PDFs for additional context.

   The input is handled by a backend orchestration layer, which manages request flow, preprocessing, prompt construction, and communication between components.

   Next comes the RAG pipeline, which consists of a text splitter, embedding model, vector database, and retriever.

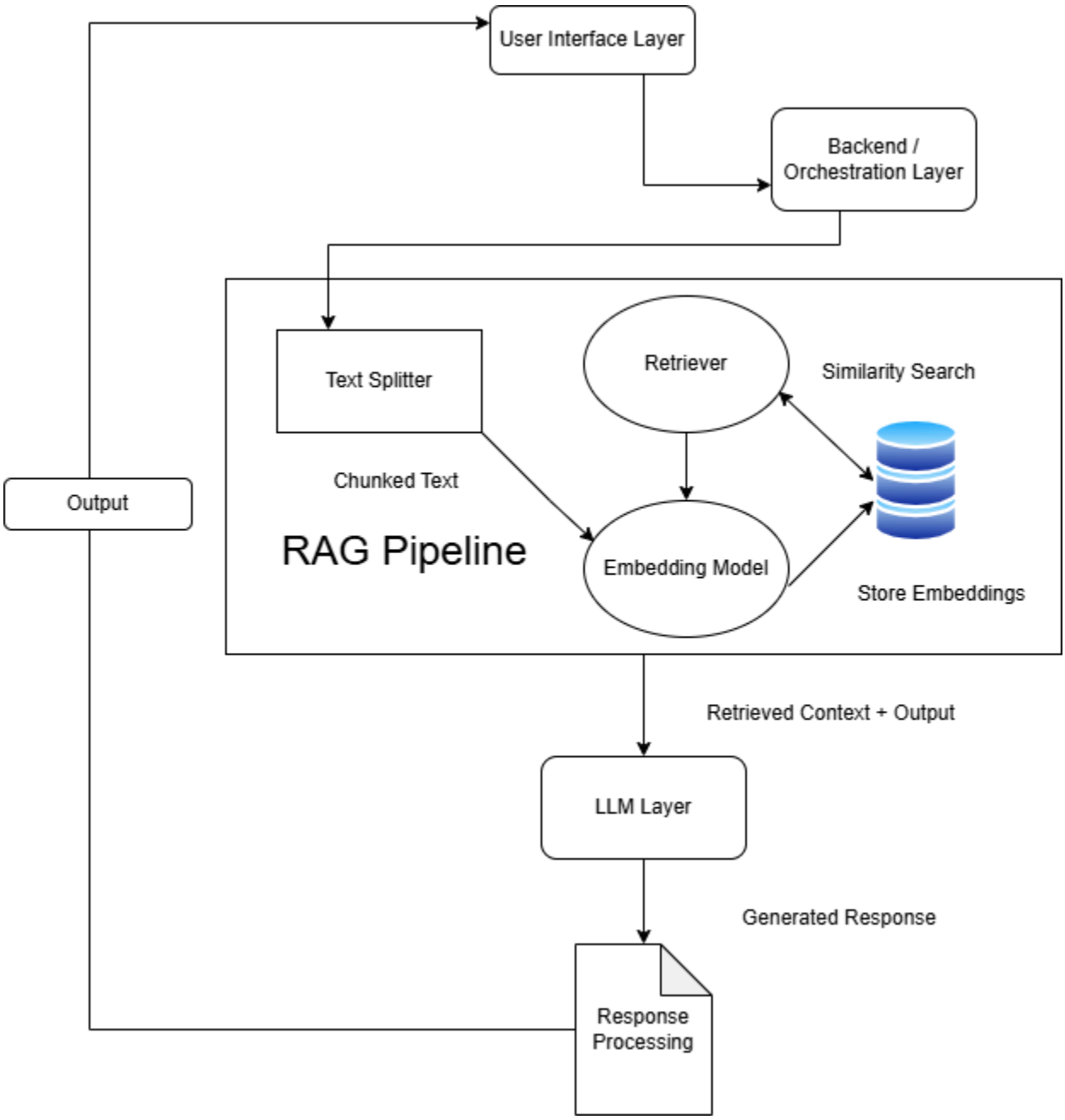
   The documents are split into chunks, converted into embeddings, and stored in a vector database.

   When a query is received, a similarity search retrieves the most relevant chunks to form contextual information.

   This retrieved context, along with the user query, is passed to the LLM layer, which generates a response grounded in the provided context.

Finally, a response or format processing layer handles formatting, post-processing, and presents the final output to the user.

Below is the Architecture



3. Please explain vector databases. If you were to select a vector database for a hypothetical problem (you may define the problem) which one will you choose, and why?

Answer:

Vector Databases store data in the form of embeddings, which are numerical representation of the data, which allows us to perform similarity search and find information that are semantically similar to the prompt or the question asked.

They are a key component for LLMs as they prevent hallucinations, by providing relevant context, based on the semantic search performed with the help of the user query, and the knowledge source provided.

Hypothetical Problem:

A SaaS company wants to build an AI-powered customer support chatbot that can answer user questions using:

- Product documentation
- Help articles
- Past support tickets
- Release notes

For a production-grade customer support chatbot with high traffic and large document volumes, I'd choose Pinecone because it's fully managed, without us having to manage any servers, has scalable vector database with low-latency semantic search, strong metadata filtering, and provides easy integration.