Explanation of the CRAG System Architecture

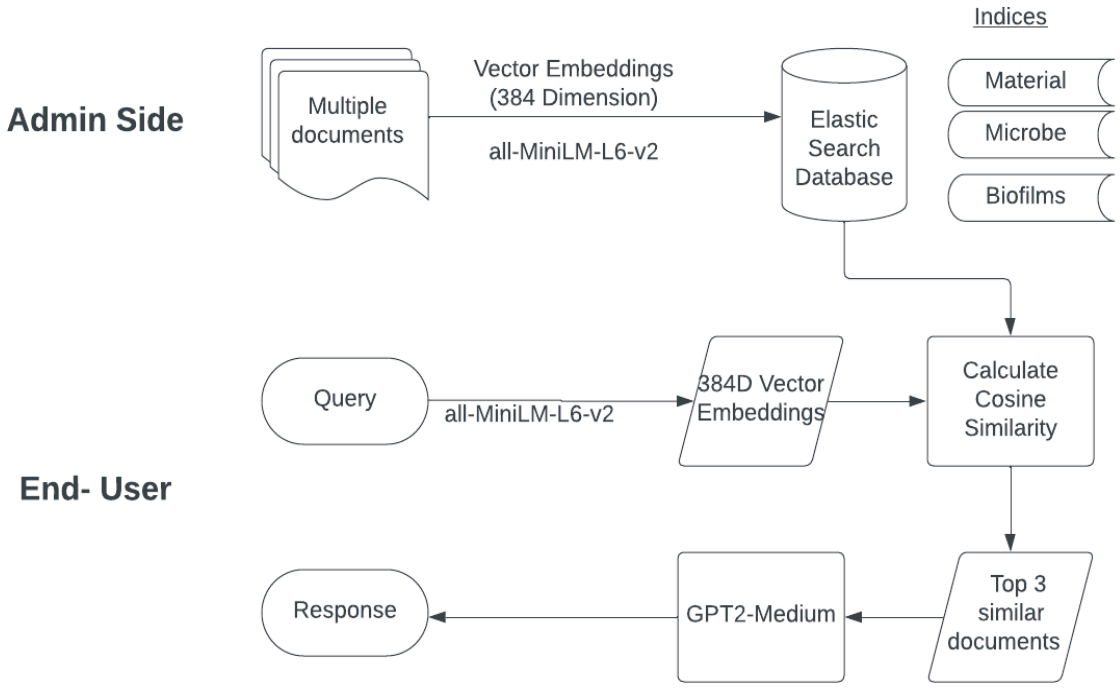

Fig: System Architecture of RAG System

The diagram illustrates the architecture of a Retrieval-Augmented Generation (RAG) system. This system comprises two main parts: the Admin Side and the End-User side.

Admin Side:

The admin side is responsible for preprocessing and storing documents that the RAG system will use to generate responses. Multiple documents are encoded into vector embeddings using the all-MiniLM-L6-v2 model, which reduces the text data into a 384-dimensional vector space. These vectors are then stored in an Elastic Search Database, which is organized into indices such as Material, Microbe, and Biofilms for efficient retrieval.

End-User Side: When an end-user inputs a query, the same all-MiniLM-L6-v2 model is used to convert the query text into a 384-dimensional vector embedding. This query vector is then compared against the document vectors in the Elastic Search Database using cosine similarity, a measure that quantifies how similar two vectors are. The system retrieves the top 3 most similar documents based on this similarity score.

Finally, the GPT-2-Medium model takes these documents as input and generates a coherent and contextually relevant response for the end-user.

In summary, the RAG system leverages vector embeddings for document retrieval and state-of-the-art language models for generating responses, aiming to provide accurate and contextually relevant information in response to user queries.