**University of North Texas**

**ADTA 5760-501**

**Natural Language Processing with Artificial Neural Networks**

**Financial Intelligence: A Generative AI-Powered Q&A System for Corporate Financial Analysis**

**Group Members** – Biniam Abebe, Srilekha Aduvala ,Nithin Marpu, Joshua Terrazas

**Part VI: Generative AI A&Q-Search System: System Design**

**High-Level Design:**

**Retrieval**

**User Documents Embeddings**

**LLM**

**Response**

**Query + Context**

**Vector DB**

**User Queries**

**Embeddings**

**Augmentation**

**Generation**

Relevant

Context

**Detailed Design:**

1. Using GCP vertex ai and Lang Chain open-source framework we build a Retrieval-Augmented Generation (RAG) pipeline that integrates large language models with knowledge bases for context-aware responses.
2. We Used GCSDirectoryLoader to load documents stored in Google Cloud Storage (GCS).
3. PDF files are parsed using PyPDFLoader from langchain\_community
4. RecursiveCharacterTextSplitter was used by us to split large documents into chunks
5. We used Google’s text-embedding-005 model for generating vector representations of text.
6. We Used PromptTemplate to structure input for the LLM, to ensure context and query are aligned for high-quality responses.
7. User Documents ->Embedding->Vector DB is all about pre-processing which is knowledge collection, chunking, and embedding. In the vector database, all the files will be organized and relevant context will be chosen based on user queries.
8. The queries received from the user were also fed into the vector database after embedding by the embedding model.
9. Based on the queries and context received from knowledge collection the LLM will generate the solution.