**ADTA 5770: Generative AI with LLM**

**Semester Project – Group Final Report**

Group: 8

**Group Members**:

* + Biniam Abebe
  + Srilekha Aduvala
  + Nithin Marpu
  + Joshua Terrazas

1. **Semester Project: Introduction**

The Financial Intelligence project aims to develop a Generative AI-powered Q&A system focused on corporate financial analysis. Our team selected corporate finance as our domain expertise field, specifically targeting financial statement analysis. This system leverages Google Cloud Platform's Vertex AI services to create an enterprise-level solution that enables users to conduct precise content searches and receive accurate answers about financial data contained in proprietary documents. The project responds to modern business organizations' need for fast, scalable solutions to process and interpret complex financial documented data, allowing financial professionals to access critical information efficiently.

1. **Semester Project: Knowledge Base**

Our knowledge base consists of 100 PDF documents specifically focused on corporate finance and financial statement analysis. These documents were systematically collected and curated to provide comprehensive coverage of financial reporting standards, analytical frameworks, and corporate financial practices. The collection process involved identifying authoritative sources in the financial domain, including Q4 financial statements from 100 public companies, ensuring the knowledge base contains real-world reporting examples. The contents encompass quarterly capital statements, financial analysis methodologies, reporting guidelines, and market research relevant to corporate finance. These documents were uploaded to our designated Google Cloud Storage bucket ("adta5770-docs-folder98") in the appropriate folder structure for processing.

1. **Semester Project: Business and Technical Requirements**

Business Requirements:

* Automate daily tasks including financial reporting and customer query responses
* Provide personalized financial advice and tips to clients
* Seamlessly integrate the AI system into existing workflow
* Enhance productivity by alleviating staff workload
* Improve customer satisfaction through faster, more accurate responses
* Modernize operations with current technology standards

Technical Requirements:

* Implement real-time APIs for customer interaction
* Ensure robust data security protocols for sensitive financial information
* Utilize Natural Language Processing capabilities for understanding complex financial queries
* Incorporate real-time data processing for up-to-date business insights
* Deploy on Google Cloud Platform (GCP) using Vertex AI
* Utilize Gemini 2.0 experimental as the large language model
* Implement vector embeddings and RAG architecture for accurate information retrieval

1. **Semester Project: System Analysis**

Our system analysis revealed strong potential for successful implementation of the Financial Intelligence Q&A system. The technical feasibility assessment confirmed that our selected technologies (GCP Vertex AI, Gemini 2.0, LangChain) provide the necessary capabilities to meet project requirements. Potential technical risks include model performance limitations with specialized financial terminology and integration complexity with existing systems.

The business feasibility analysis projected significant value from reducing analyst search time by 50%, improving customer satisfaction through faster responses, and enabling more personalized financial advice. Financial risks include potential development cost overruns and cloud computing expenses.

Operational feasibility appears strong, with high potential for adoption given the system's alignment with existing workflows. Operational risks include user adoption challenges and trust verification concerns among financial professionals. Our project timeline spans approximately 12 weeks, with phases including planning, design, setup, data preparation, development, testing, and deployment. The team of four members has been assigned specific responsibilities across each phase, with appropriate resource allocation to ensure comprehensive coverage of all system components.

1. **Semester Project: System Design: High-Level Design**

The high-level design of our Financial Intelligence Q&A system employs a modular architecture with five primary components:

1. Data Ingestion Layer - Handles collection, preprocessing, and storage of financial documents in GCP Cloud Storage
2. Embedding Generation Layer - Transforms processed documents into vector embeddings using Sentence Transformers
3. Query Processing Layer - Receives and analyzes natural language queries, converting them to vector embeddings
4. Retrieval Layer - Executes vector similarity search to identify relevant document chunks
5. Response Generation Layer - Implements Retrieval Augmented Generation using Gemini 2.0

This architecture follows the RAG (Retrieval Augmented Generation) pattern, whereby user queries are processed to retrieve relevant context from our document corpus before generating accurate responses grounded in that context. The workflow integrates our financial knowledge base with Gemini LLM to create a system that understands financial terminology and provides precise answers to domain-specific questions.

A diagram of a diagram

AI-generated content may be incorrect.

1. **Semester Project: System Design: Detailed Design**

The detailed design expands on the high-level architecture with specific module implementations:

1. Data Ingestion Component:

* Document Processing Module using PyPDF2 and LangChain document loaders
* Document Storage Module leveraging GCP Cloud Storage buckets
* Document Updating Module for maintaining current information

1. Embedding Generation Component:

* Text Chunking Module implementing intelligent document segmentation
* Embedding Model Module utilizing Sentence Transformers
* Vector Storage Module for efficient retrieval

1. Query Processing Component:

* Query Understanding Module analyzing natural language financial queries
* Query Embedding Module converting queries to vector representations
* Search Parameter Module optimizing retrieval parameters

1. Retrieval Component:

* Vector Search Module executing similarity searches
* Hybrid Retrieval Module combining vector and keyword-based approaches
* Context Preparation Module assembling coherent information

1. Response Generation Component:

* Prompt Engineering Module constructing effective queries
* LLM Integration Module interfacing with Gemini 2.0
* Response Formatting Module for financial-specific presentation
* User Feedback Module for continuous improvement

The system uses GCP vertex AI and LangChain framework to build the RAG pipeline, load documents from Google Cloud Storage, parsing PDFs, chunking text, and generating embeddings using Google's text-embedding-005 model. The detailed architecture ensures specialized handling of financial terminology and numerical data throughout the pipeline.

1. **Semester Project: System Development: Coding**

For the system development, we implemented a Retrieval-Augmented Generation (RAG) pipeline using GCP Vertex AI and the LangChain open-source framework. The development process followed these key phases:

1. Setting up the Google Cloud environment and service accounts
2. Creating storage buckets and folder structures for document management
3. Implementing document loading using GCSDirectoryLoader to access files in Cloud Storage
4. Developing PDF parsing functionality with PyPDFLoader
5. Creating text chunking logic with RecursiveCharacterTextSplitter
6. Implementing embedding generation using Google's text-embedding-005 model
7. Setting up vector database storage and indexing
8. Developing query processing and embedding functionality
9. Implementing context retrieval based on vector similarity
10. Integrating with Gemini 2.0 experimental for response generation

We utilized Gemini 2.0 experimental as our LLM model because of its strong performance in understanding and generating financial content. The model demonstrates excellent capability in handling complex financial terminology and numerical data, making it ideal for our corporate finance domain. We implemented Prompt Template structures to ensure optimal context formatting for high-quality responses from the LLM.

1. **Semester Project: Testing and Operationalization**

The Q&A Search system underwent comprehensive testing to ensure accuracy, response quality, and performance. Our testing approach included:

1. **Unit Testing**: Testing individual components of the pipeline (document loading, embedding generation, retrieval)
2. **Integration Testing**: Verifying the entire RAG pipeline functions correctly end-to-end
3. **Performance Testing**: Evaluating response times and resource utilization
4. **Accuracy Testing**: Assessing the quality and relevance of responses to financial queries

Our testing revealed that the system performs well with financial terminology and can accurately retrieve relevant information from our knowledge base. We optimized our prompt templates to ensure the LLM received appropriate context and instructions for generating high-quality financial analysis.

To interact with the system, users submit natural language queries related to corporate finance. The prompt structure includes the user's query along with retrieved context from the knowledge base. The system processes the query, retrieves relevant information from the vector database, and generates a response using Gemini 2.0 experimental. Responses include citations to source documents where appropriate, providing transparency and trustworthiness to the information provided.

1. **Semester Project: Q&A Search System: Prompts and Responses**

Our team tested the Financial Intelligence Q&A system with 10 diverse prompts covering various aspects of corporate financial analysis. Below is a summary of our testing approach and results.

We carefully designed our testing scenarios to evaluate the system's ability to retrieve relevant information and generate accurate, coherent responses to financial queries. The system demonstrated strong capabilities in analyzing corporate financial data, with the ability to extract specific financial metrics, identify trends, and provide contextual information.

For example, when querying "What is the Net loss ATT INC?", our system successfully retrieved and provided the answer that "AT&T reported a net loss attributable to the Company of $103 million in 2023," while also noting the system found references to adjustments for 2024 but without a specific total net loss figure.

Our testing also revealed the system's ability to handle different types of queries effectively:

* Direct fact-finding questions
* Questions requiring filtering by specific document
* Complex financial metrics inquiries
* Profitability analysis questions
* Risk assessment questions

The system utilized appropriate citations to source documents, making the responses transparent and verifiable. We structured responses to include both concise answers and supporting context from the knowledge base.

While most queries returned relevant information, some types of queries (particularly those about companies not in our knowledge base) appropriately returned "I don't know" responses, demonstrating the system's ability to recognize its limitations rather than providing speculative answers.

These tests validate that our RAG architecture successfully connects user queries with the most relevant information from our financial document corpus and produces high-quality, accurate responses that would be valuable to financial professionals.

1. **Semester Project: Group Contribution**

All team members contributed actively to the project, with responsibilities distributed based on individual strengths and expertise:

* Biniam Abebe: 100% - Led system design and setup, document collection and preprocessing, and integration efforts
* Srilekha Aduvala: 100% - Headed RAG pipeline implementation, data preprocessing, and functional testing
* Nithin Marpu: 100% - Developed query processing system, implemented embedding generation, and performed performance optimization
* Joshua Terrazas: 100% - Created UI components, developed response generation module, and led deployment activities

The team maintained excellent collaboration throughout the semester, with regular meetings both online and in-person. Clear communication channels ensured alignment on project goals and timely resolution of challenges. Each member participated actively in all phases of the project, from initial planning through final implementation and testing.