ADTA 5770: Generative AI with LLMs - Assignment 5

Student Name: Biniam Abebe
Domain Expertise Field: Finance/Investment

PART I: Generative AI Q&A-Search System: System Analysis (50 Points)

1. Introduction

This project aims to develop a generative AI-powered Q&A system specifically designed for corporate financial analysis. In today's rapidly evolving financial landscape, organizations face challenges in efficiently accessing, analyzing, and utilizing the vast amounts of financial information at their disposal. Our system, "Financial Intelligence," will leverage advanced generative AI technologies to enable company employees to perform content searches, ask questions, and receive accurate answers about the organization's proprietary financial documents.

The system will utilize Google Cloud Platform (GCP) Vertex AI services as the primary development environment, incorporating technologies such as vector embeddings, vector databases, and Retrieval Augmented Generation (RAG). By implementing this solution, financial organizations will be able to streamline workflows, enhance decision-making processes, and improve overall operational efficiency.

2. Problem Statement

In the financial industry, professionals face several critical challenges that our generative AI Q&A system aims to address:

Business Problems:

- Financial analysts spend excessive time searching through extensive documentation to find specific information, reducing their productivity and analytical capabilities.
- Customer service representatives struggle to quickly access accurate information to answer client queries, leading to longer response times and reduced customer satisfaction.
- The complexity of financial documents creates knowledge bottlenecks, where expertise is concentrated among a few individuals within the organization.
- Financial advice and reporting processes are often manual and time-consuming, limiting the organization's ability to scale services efficiently.

Technical Problems:

- Traditional search systems rely on keyword matching, which often fails to capture the semantic meaning behind financial queries.

- Financial documents contain specialized terminology and concepts that require contextual understanding for accurate information retrieval.
- Existing systems lack the ability to synthesize information across multiple documents to provide comprehensive answers to complex financial questions.
- The need for real-time, accurate financial information requires sophisticated data processing and retrieval capabilities beyond traditional database solutions.

Our Financial Intelligence Q&A system will solve these problems by creating an intuitive, AI-powered interface that understands natural language queries, retrieves relevant information from proprietary documents, and generates accurate, context-aware responses for financial professionals and clients alike.

3. System Requirements Analysis

Business Requirements

Our generative AI system must fulfill the following business requirements:

1. **Streamlined Workflow Enhancement**:
    - The system must reduce the time employees spend searching for financial information by at least 50%.
    - It should automate routine tasks such as generating financial reports and answering standard customer queries.
    - The solution must integrate seamlessly with existing workflows without disrupting current operations.
2. **Customer Experience Improvement**:
    - The system should provide personalized financial advice and tips based on client profiles and queries.
    - It must deliver consistent, accurate responses to customer inquiries across all interaction channels.
    - Response times for customer queries should be reduced to under 30 seconds.
3. **Operational Modernization**:
    - The solution should support multilingual capabilities to bridge language gaps between clients and employees.
    - It must allow financial advisors to review, refine, and approve AI-generated content before client delivery.

    - The system should maintain comprehensive audit trails of all interactions for compliance purposes.

4. **Data-Driven Decision Support**:
    - The system must provide analytical insights based on trends identified in financial documents.
    - It should support real-time data integration to ensure recommendations are based on current information.
    - The solution must enable the creation of customized financial reports tailored to specific business needs.

Technical Requirements

The technical implementation of our system will adhere to the following specifications:

1. **AI Platform and Model**:
    - Google Cloud Platform (GCP) Vertex AI will serve as the primary development environment.
    - Gemini 2.0 experimental will be the large language model driving the system's intelligence.
    - The system will utilize LangChain framework for orchestrating the AI workflow.
2. **Infrastructure and Integration**:
    - The solution must provide RESTful APIs for seamless integration with existing financial systems.
    - It should support secure authentication mechanisms compatible with financial industry standards.
    - The system must maintain high availability (99.9% uptime) to support critical financial operations.
3. **Data Processing Capabilities**:
    - The system will implement vector embeddings for efficient document representation and retrieval.
    - It must incorporate advanced Natural Language Processing (NLP) for understanding financial terminology.
    - The solution should utilize Retrieval Augmented Generation (RAG) to provide accurate, grounded responses.

4. **Security and Compliance**:
    - The system must implement end-to-end encryption for all data in transit and at rest.
    - It should support role-based access controls to protect sensitive financial information.
    - The solution must comply with relevant financial regulations such as SOX, GDPR, and CCPA.

Data Requirements

The effective operation of our Financial Intelligence system depends on the following data requirements:

1. **Document Corpus**:
    - The system requires a comprehensive collection of financial documents, including annual reports, financial analyses, market research, and regulatory guidelines.
    - All documents must be in PDF format for consistent processing.
    - The initial corpus will consist of 100 carefully selected financial documents relevant to the organization's operations.
2. **Data Quality and Structure**:
    - Documents must contain properly formatted text that can be accurately extracted for processing.
    - Financial tables, charts, and numerical data within documents should be properly structured for extraction.
    - Documents should include appropriate metadata (e.g., date, author, category) for efficient indexing and retrieval.
3. **Data Storage and Management**:
    - All documents will be stored in GCP Cloud Storage buckets with appropriate access controls.
    - Vector embeddings generated from documents will be stored in vector databases optimized for similarity search.
    - The system will implement data versioning to track changes in the document corpus over time.
4. **Data Processing Pipeline**:

- The system requires a robust ETL pipeline for document ingestion, processing, and indexing.
- It must support continuous updates to the knowledge base as new financial documents become available.
- The solution should implement data cleaning procedures to handle inconsistencies in document formatting.

4. Feasibility Analysis

Technical Feasibility Analysis

Can we complete the project successfully as required?

Yes, the project is technically feasible given the selected technologies and approaches. Google Cloud Platform's Vertex AI provides comprehensive tools for developing, deploying, and managing AI solutions. The Gemini 2.0 experimental model has demonstrated strong capabilities in understanding and generating financial content, and LangChain offers proven frameworks for implementing RAG systems.

Our team has the necessary expertise in Python, cloud computing, and machine learning to successfully implement the required components. Additionally, the RAG approach is well-documented and has been successfully used in similar applications across various industries.

Technical Risks:

1. **Model Performance Limitations**: Gemini 2.0, while powerful, may struggle with highly specialized financial terminology or complex numerical analyses. We may need to implement additional training or fine-tuning to achieve the desired performance.
2. **Integration Complexity**: Connecting the AI system with existing financial software and databases may present challenges, particularly if these systems use proprietary formats or protocols.
3. **Scalability Concerns**: As the document corpus grows, we may face challenges in maintaining performance and response times, requiring optimization of our vector storage and retrieval mechanisms.
4. **Technical Debt**: Rapid implementation to meet project deadlines could lead to suboptimal code structure that may require refactoring in later phases.
5. **Version Compatibility**: Updates to GCP services, Gemini API, or LangChain framework during the project lifecycle could introduce compatibility issues requiring additional development effort.

Business Feasibility Analysis

Will the project provide good business value after its completion?

The Financial Intelligence system is expected to deliver substantial business value by:

- Reducing the time financial analysts spend searching for information by 50%, potentially saving thousands of work hours annually
- Improving customer satisfaction through faster, more accurate responses to queries
- Enabling more personalized financial advice, potentially increasing client retention and acquisition
- Modernizing operations to maintain competitive advantage in the financial services sector

These benefits justify the investment in the system's development and ongoing maintenance.

Financial Risks:

1. **Development Cost Overruns**: Complex AI projects often encounter unforeseen challenges that can extend development timelines and increase costs beyond initial estimates.
2. **Cloud Computing Expenses**: GCP service costs could exceed projections, particularly if the system requires more computational resources than anticipated for processing large document volumes.
3. **ROI Timeline**: The full business value may take longer than expected to realize, as employees and customers adapt to the new system and workflows.
4. **Licensing Costs**: Future changes to the pricing models of third-party services or APIs could impact the ongoing operational costs of the system.
5. **Maintenance Budget**: Insufficient allocation for system maintenance and updates could limit the long-term viability and evolution of the solution.

Operational Feasibility Analysis

If we build the system, will it be used by the organization as expected?

The system has strong potential for adoption within the organization, given its alignment with existing workflows and focus on addressing clear pain points for both employees and customers. The financial services industry has shown increasing receptiveness to AI-powered tools, particularly those that enhance efficiency and decision-making capabilities.

Operational Risks:

1. **User Adoption Challenges**: Employees accustomed to traditional methods may resist adopting the new AI system, potentially limiting its effectiveness and ROI.

2. **Training Requirements**: Insufficient training resources could hamper users' ability to effectively leverage the system's capabilities.
3. **Change Management Issues**: Inadequate organizational change management could result in poor implementation and limited usage of the system.
4. **Trust and Verification Concerns**: Financial professionals may be hesitant to rely on AI-generated information without extensive verification, potentially limiting efficiency gains.
5. **Compliance and Regulatory Changes**: Future regulatory changes in the financial industry could impose new requirements on AI systems, necessitating modifications to ensure continued compliance.

5. Project Management

Project Timeline

Significant Tasks and Major Phases:

1. **Planning and Requirements Definition (Completed)**
   - Business and technical requirements gathering
   - Domain expertise field selection
   - Due Date: April 2, 2025
2. **System Analysis and Design (Current Phase)**
   - Detailed system analysis
   - High-level and detailed system design
   - Due Date: April 9, 2025
3. **System Setup (2 weeks)**
   - Cloud infrastructure setup
   - Development environment configuration
   - Due Date: April 23, 2025
4. **Data Preparation and Processing (3 weeks)**
   - Document collection and curation
   - Data preprocessing and cleaning
   - Embedding generation and storage

    - Due Date: May 14, 2025
5. **Core System Development (4 weeks)**
    - Implementation of RAG architecture
    - Query processing system development
    - Response generation and refinement
    - Due Date: June 11, 2025
6. **Integration and Testing (2 weeks)**
    - System integration
    - Functional testing
    - Performance optimization
    - Due Date: June 25, 2025
7. **Deployment and Final Presentation (1 week)**
    - System deployment
    - Documentation completion
    - Final project presentation
    - Due Date: July 2, 2025

Human Resources

Total Team Size: 4 members

- Biniam Abebe
- Srilekha Aduvala
- Nithin Marpu
- Joshua Terrazas

Resource Allocation by Phase:

1. **Planning and Requirements Definition**
    - All team members (4) participated in requirements gathering and domain selection
    - Biniam led business requirements documentation

- Srilekha led technical requirements documentation
- Nithin and Joshua researched domain expertise materials

2. **System Analysis and Design (Current Phase)**
   - All team members (4) are contributing to system analysis
   - Biniam is leading the business feasibility analysis
   - Srilekha is leading the technical feasibility analysis
   - Nithin is developing the high-level system design
   - Joshua is working on the detailed system design
3. **System Setup**
   - 3 team members will be assigned
   - Srilekha will lead GCP environment configuration
   - Nithin will set up storage buckets and authentication
   - Joshua will configure development environments
4. **Data Preparation and Processing**
   - All team members (4) will participate
   - Biniam will lead document collection and organization
   - Srilekha will develop data preprocessing scripts
   - Nithin will implement embedding generation
   - Joshua will configure vector storage solutions
5. **Core System Development**
   - All team members (4) will contribute
   - Biniam will develop the user interface components
   - Srilekha will implement the RAG architecture
   - Nithin will build the query processing system
   - Joshua will develop the response generation module
6. **Integration and Testing**
   - 3 team members will be assigned

- Biniam will lead integration efforts
- Srilekha will conduct functional testing
- Nithin will perform performance optimization

7. **Deployment and Final Presentation**
   - All team members (4) will participate
   - Joshua will lead deployment activities
   - Biniam will prepare documentation
   - Srilekha and Nithin will develop the final presentation

6. Conclusion

The Financial Intelligence generative AI Q&A system represents a significant opportunity to transform how financial organizations access, analyze, and utilize their proprietary information. While we anticipate challenges in areas such as model fine-tuning, system integration, and user adoption, our team is well-positioned to address these issues through careful planning, iterative development, and continuous stakeholder engagement.

The project's success will depend on balancing technical excellence with practical business considerations, ensuring that the final system not only demonstrates advanced AI capabilities but also delivers tangible value in real-world financial operations. With our comprehensive approach to system analysis, thoughtful design, and strategic resource allocation, we are confident that the Financial Intelligence system will meet both technical requirements and business objectives.

As we move forward with implementation, we will maintain flexibility to adapt to emerging challenges while staying focused on our core goal: creating an intuitive, powerful tool that empowers financial professionals to make better decisions, serve clients more effectively, and drive organizational success in an increasingly complex financial landscape.

PART II: Generative AI Q&A-Search System: High Level & Detailed Design (50 Points)

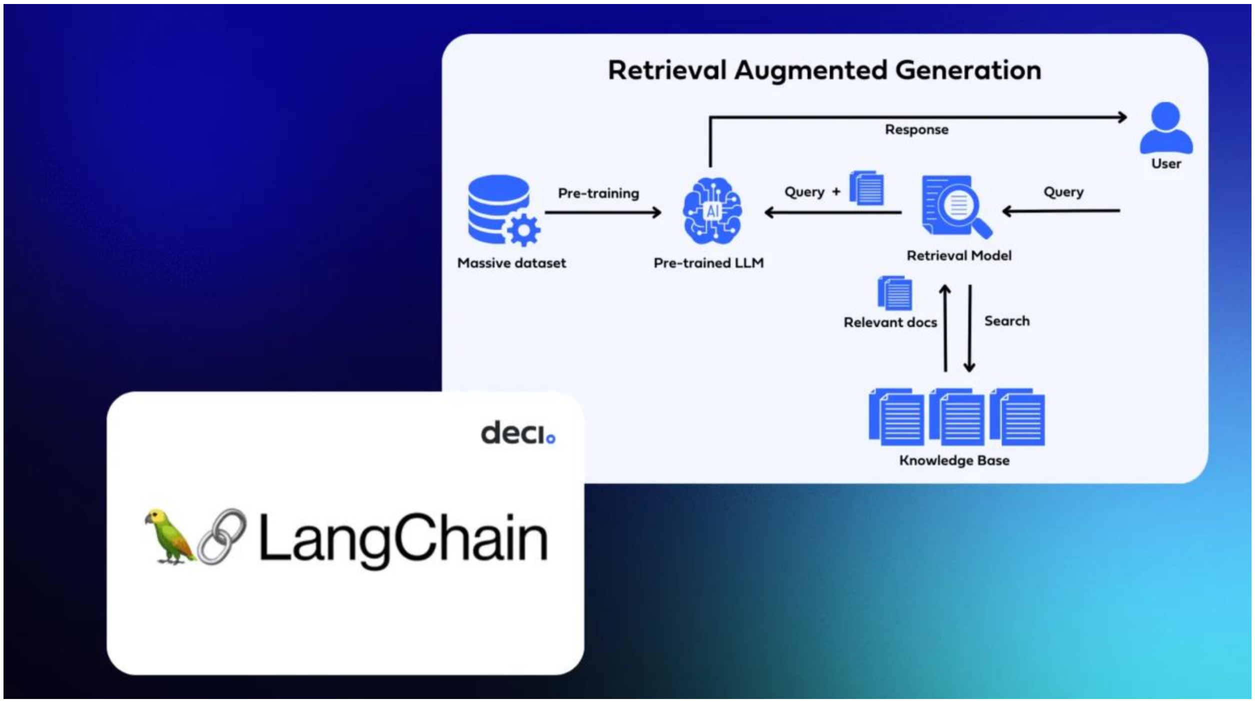
1 Q&A Search System: High-Level Design

The Financial Intelligence Q&A Search system employs a modular architecture designed to efficiently process financial documents, comprehend user queries, and produce accurate responses. The high-level architecture consists of five primary components:

1. **Data Ingestion Layer**
   - Handles the collection, preprocessing, and storage of financial documents
   - Converts PDF documents into processable text format

   - Extracts and preserves document structure, including tables and numerical data
   - Stores raw documents in GCP Cloud Storage

2. **Embedding Generation Layer**
   - Transforms processed documents into vector embeddings
   - Segments documents into appropriate chunks for effective retrieval
   - Utilizes Sentence Transformers for creating semantically meaningful embeddings
   - Stores generated embeddings in vector databases for efficient similarity search
3. **Query Processing Layer**
   - Receives and analyzes natural language queries from users
   - Converts queries into vector embeddings using the same model as documents
   - Applies query refinement techniques for handling financial terminology
   - Formulates appropriate search parameters based on query context
4. **Retrieval Layer**
   - Executes vector similarity search to identify relevant document chunks
   - Implements hybrid retrieval combining vector search with keyword matching
   - Ranks and filters search results based on relevance scores
   - Prepares retrieved context for the generation component
5. **Response Generation Layer**
   - Implements Retrieval Augmented Generation using Gemini 2.0
   - Combines retrieved context with the query to generate accurate responses
   - Applies financial domain-specific output formatting and verification
   - Provides citations linking responses to source documents



Workflow schema for integrating a corpus of docs with an LLM in RAG

https://medium.com/@leighphil4/basic-architecture-of-an-rag-system-cdd057e3a4d1

2 Detailed Design

2.1. Data Ingestion Component

Document Processing Module:

- Implements PDF extraction using PyPDF2 and LangChain document loaders
- Preserves document structure through custom parsers for financial tables and charts
- Maintains document metadata including source, date, author, and category
- Implements content cleaning to handle special characters and formatting

Document Storage Module:

- Utilizes GCP Cloud Storage buckets with hierarchical organization
- Implements document versioning to track changes over time
- Provides access control mechanisms for secure document management
- Maintains a document registry with metadata for efficient retrieval

Document Updating Module:

- Provides interfaces for adding new documents to the system
- Implements differential processing to update only changed content
- Supports batch processing for large document collections
- Includes validation checks to ensure document quality and compatibility

2.2. Embedding Generation Component

Text Chunking Module:

- Implements intelligent document segmentation based on content structure
- Maintains semantic coherence within chunks to preserve context
- Uses sliding window approaches with overlap for continuity
- Applies financial domain-specific chunking rules for numerical content

Embedding Model Module:

- Utilizes Sentence Transformers for generating document embeddings
- Implements batch processing for efficient embedding generation
- Includes specialized handling for financial terminology and numerical data
- Supports multiple embedding models for different types of content

Vector Storage Module:

- Implements vector database integration for storing and retrieving embeddings
- Utilizes indexing techniques for fast similarity search
- Provides metadata filtering capabilities for refined retrieval
- Includes backup and recovery mechanisms for embedding data

2.3. Query Processing Component

Query Understanding Module:

- Analyzes natural language queries to identify key financial concepts
- Implements query classification to determine intent (factual, analytical, advisory)
- Extracts relevant parameters such as time periods, financial entities, and metrics
- Applies query expansion for handling financial abbreviations and terminology

Query Embedding Module:

- Converts processed queries into vector embeddings
- Ensures consistency with document embedding approach
- Implements context-aware embedding for multi-part queries
- Provides query refinement based on user feedback

Search Parameter Module:

- Determines appropriate similarity thresholds based on query type
- Configures retrieval parameters including number of results and diversity
- Implements filtering based on document metadata and relevance
- Supports complex queries involving multiple financial concepts

2.4. Retrieval Component

Vector Search Module:

- Executes similarity search using optimized vector database queries
- Implements approximate nearest neighbor techniques for performance
- Provides configurable search parameters for precision vs. recall tradeoffs
- Includes relevance scoring based on semantic similarity

Hybrid Retrieval Module:

- Combines vector search with keyword-based retrieval
- Implements re-ranking algorithms to optimize retrieval quality
- Utilizes metadata filtering for contextually appropriate results
- Supports entity-based retrieval for financial instruments and organizations

Context Preparation Module:

- Assembles retrieved chunks into coherent context
- Applies deduplication to remove redundant information
- Formats retrieved content for optimal LLM consumption
- Preserves source attribution for citation generation

2.5. Response Generation Component

Prompt Engineering Module:

- Constructs effective prompts combining a query and the retrieved context
- Implements financial domain-specific prompt templates
- Provides instruction framing for different query types
- Includes guardrails to ensure responsible AI outputs

LLM Integration Module:

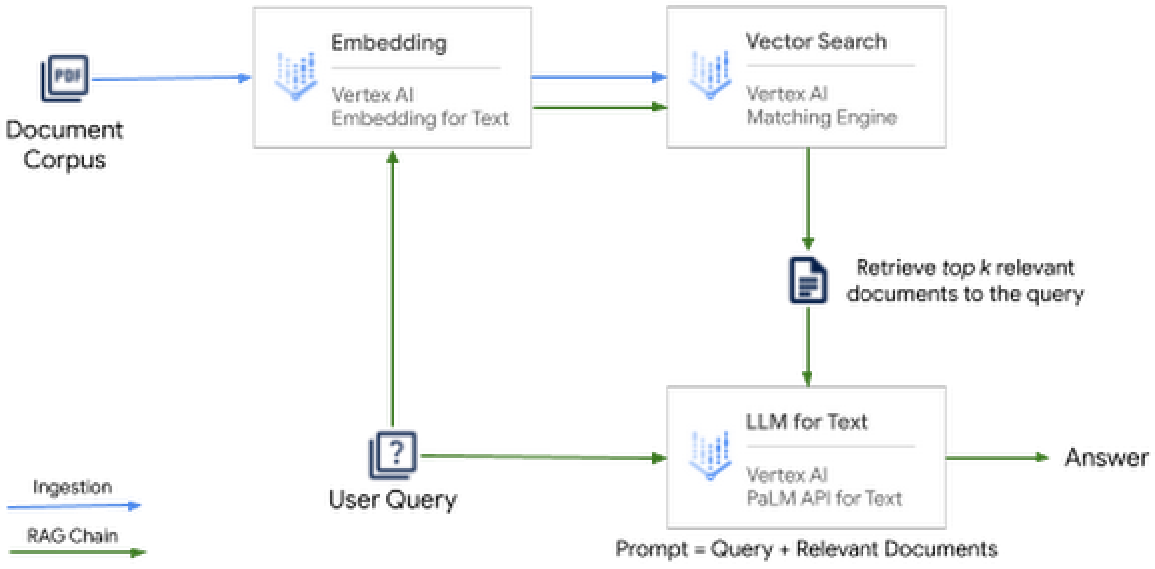
- Interfaces with Gemini 2.0 through Vertex AI APIs
- Manages token limits and context windows
- Implements fallback mechanisms for handling API limitations
- Provides streaming response capabilities for improved user experience

Response Formatting Module:

- Applies financial-specific formatting for numerical data and analyses
- Generates appropriate citations linking to source documents
- Implements confidence scoring for response reliability assessment
- Provides explanations for complex financial concepts when needed

User Feedback Module:

- Collects and processes user feedback on response quality
- Implements continuous improvement mechanisms
- Provides analytics on system performance and usage patterns
- Supports active learning for ongoing system refinement


Document Corpus
PDF
Embedding
Vertex AI Embedding for Text
Vector Search
Vertex AI Matching Engine
Retrieve top k relevant documents to the query
LLM for Text
Vertex AI PaLM API for Text
Answer
User Query
Prompt = Query + Relevant Documents
Ingestion
RAG Chain

High-Level Architecture of building QA system with RAG pattern

PART III: Teamwork Evaluation (10 Points)

1. **Group Number**: Group 8
2. **Group Members**:
   - Biniam Abebe
   - Srilekha Aduvala
   - Nithin Marpu
   - Joshua Terrazas
3. **Group Meetings**: Yes, our group has organized both online and in-person meetings to work on HW 5.
4. **Meeting Attendance**: All group members, including myself, attended the scheduled meetings.
5. **Participation**: Yes, all members made reasonable efforts to participate actively in the group work.
6. **Additional Comments**: Our team has maintained excellent collaboration throughout this assignment. We've established clear responsibilities based on individual strengths, with each member contributing significantly to different aspects of the system analysis and design. Regular communication through various channels has ensured everyone remains aligned with project goals and expectations.