**ADTA 5770: Generative AI with LLM**

**Semester Project – Group Final Report**

**Group 3:**

* Sonali Sabnam
* Young Yu
* Sonam Pohuja

# Semester Project: Introduction

**Objective:**

Having an efficient inventory and warehouse management system is a key driver for a successful business.

The team’s goal for the semester project is to build a Question-and-Answer search system that can help the users to find answers and serve as a source of information with respect to inventory/warehouse management.

**Domain:**

We are interested in building a Question-and-Answer system for an **Inventory/Warehouse management system** which comes under **supply chain management**. With the help of this system the user should be able to get answers on the existing inventory and the required items.

 This chatbot would be helpful to cater to the below mentioned items:

1. Process Improvement:

Inventory Management is critical to ensure smooth warehouse operations. Industries like food, pharmaceuticals, and retail require a rigorous tracking system and real-time monitoring of their inventory. They must comply with various regulations, maintain operational efficiencies and mitigate risks and a Question-and-Answer chatbot system would enhance process efficiency.

1. Traceability:

Many industries are heavily regulated with strict regulations on the origins of raw materials, handling of materials, and safety. Warehouse/Inventory management systems provide traceability on products from raw material to finished product, to meet these requirements, avoiding penalties and heavy fines. Traceability simplifies audits by providing readily available records of product history, ensuring compliance and transparency. Risk management is a core fundamental in business regardless of domain,

1. Navigation:

Effective navigation is a vital component of any Inventory/Warehouse management system. The User – be it a warehouse worker, manager, or even logistic personnel – need quick and accurate access to Inventory details, locations, and systemic functionalities.

The answer to this is a Q&A chatbot designed to answer system navigation queries in real-time, ensuring that the information is accessible with minimal effort.

 Some examples of Navigation Queries the chatbot should be able to answer are:

* *“Where can I find reports on current stock levels on XYZ?”*
* *“How can I locate where the highest quantity of ABC?”*
* *“What is the process for restocking low inventory?”*

Instant responses from the chatbot would reduce time spent manually searching for specific information.

1. Decision Making:

 Data driven decision making is key for the success of an enterprise. In today’s highly competitive world, quick decision making can get many benefits for the business and “first-mover’s advantage is one of them. Ease of access to actionable data can significantly boost the decision-making process and build confidence among the teams. Eventually it can be pivotal for the organization’s growth.

# Semester Project: Knowledge Base

* Our Knowledge base consists of 117 PDFs containing data from the Fishbowl inventory management system.
* **How the PDFs were Collected:** The team downloaded the data from the Fishbowl portal. One of the team mate had access to their data and API keys. The team used the same to scrape the PDFs from the Fishbowl portal and download it to their local machine.
* **Description of the Content:** The PDFs contain information about various modules of Fishbowl like
  + Kitting
  + Inventory Dashboard
  + Customer Portal

# Semester Project: Business and Technical Requirements

* **Business requirements:** Business goals or objectives that the organization tries to achieve with this generative AI system
  + **Enhance user experience:** The required information is readily available with the Q&A system making it a seamless experience for the users to access information.
  + **Workload reduction on Employees:** Employees can focus on actual tasks rather than looking for information which reduces the workload.
  + **Faster decision making with data:** As data is available seamlessly with accuracy, decisions can be made faster. This drives productivity and increases efficiency.
  + **Question and answer Capability:** LLM based system should be able to answer the questions with accuracy.
  + **Real time system:** The system will be providing real time data and information.
* **Technical Requirements:**
  + **GCP:** We will be using Google Cloud Platform to build this Q&A system.
    - **Bucket:** We will be using GCP buckets to keep the data (100 Pdfs) and use the same to build the Vector database.
  + **Vertex AI:** We plan to use Google’s Vertex AI platform for building the generative AI Question and Answer system. It is a fully managed service and easily pluggable in the Google Collab.
  + **LangChain:** LangChain is a framework having opensource components which enable application development for large language Models. It also provides seamless integration for third party tools. It provides platform for “development”, “productionization” and “deployment”, hence streamlining the entire application life cycle.

# Semester Project: System Analysis

* **Problem Statement**

We discuss the problem statement in terms of Business Problem and Technical Problem.

* **Business Problem:**
* **Operational Efficiency:** When employees spend their valuable time looking for information, it hinders the operational efficiency of the organization. With this Question-and-Answer system they can access critical information faster with the Q&A system and save time on requesting and receiving data from the respective departments.
* **Cost Savings:** Reducing dependency on human resources for data drives savings.
* **Competitive advantage and Innovation:** When decisions can be made quickly, operational efficiency increases, and the organization achieves a competitive advantage over its competitors.
* **Compliance and Risk Management:** The Q&A system can help to reduce legal risks by keeping all the up-to-date information readily available for audits and policy enforcement.
* **Technical Problem:**
* **Easily accessible system:** The Q&A system will solve the problem of information accessibility. This system will enable seamless access to the required information which otherwise would take a longer time and involve many other people and their time.
* **Eliminating dependency on Database systems:** This system would eliminate the need for logging into a database and run a query to retrieve data. This task otherwise would be done by a SQL Developer who has SQL skills.
* **System Requirements Analysis**
* **Business requirements:** Business goals or objectives that the organization tries to achieve with this generative AI system.
* **Enhance user experience:** The required information is readily available with the Q&A system making it a seamless experience for the users to access information.
* **Workload reduction on Employees:** Employees can focus on actual tasks rather than looking for information which reduces the workload.
* **Faster decision making with data:** As data is available seamlessly with accuracy, decisions can be made faster. This drives productivity and increases efficiency.
* **Question and answer Capability:** LLM based system should be able to answer the questions with accuracy.
* **Real time system:** The system will be providing real time data and information.
* **Technical requirements:** The technical requirements for the system are outlined below.
* **GCP**: We will be using Google Cloud Platform to build this Q&A system.
  + **Bucket:** We will be using GCP buckets to keep the data (100 Pdfs) and use the same to build the Vector database.
* **Vertex AI:** We plan to use Google’s Vertex AI platform for building the generative AI Question and Answer system. It is a fully managed service and easily pluggable in the Google Collab.
* **LangChain:** LangChain is a framework having opensource components which enable application development for large language Models. It also provides seamless integration for third party tools. It provides platform for “development”, “production” and “deployment”, hence streamlining the entire application life cycle.
* **Data requirements:** We downloaded the data from various sources including Fishbowl website. We have uploaded this data to the GCP bucket. We will be using the same bucket to build the Generative AI Question and Answer system.
  + **BUCKET Name = docs-genai-folder-1**
  + **Files are stored at docs-genai-folder-1 -> documents ->nlp-vlp-pdfs**
* **Feasibility Analysis:** The team performed a feasibility analysis to determine if the project can be successfully completed as required. We also did analysis to find out if there are any risks.
  + **Technical feasibility analysis:** The team performed a feasibility analysis to determine if the project can be successfully completed as required. We also did analysis to find out if there are any risks.
* Google **Vertex AI** provides managed LLMs like Gemini which are reliable and secure.
* **LangChain** offers a robust framework for building Retrieval-Augmented Generation (RAG) pipelines that bridge the gap between knowledge bases and generative models.
* The core system components, document loaders, vector stores, retrievers and LLM chains, are all available as pre-built models and can be customized for warehouse or inventory management systems. They all allow for rapid development with minimal coding.
* General-purpose LLMs may not perform well enough as they may not fully understand warehouse-specific terminology that is domain specific and would require fine-tuning.
* Vector store costs and complexity can rise as data volume grows (more data, would lead to more embedding, and each item embedded gets turned into a vector).
* There is a need to manage latency in real-time data retrieval and answer generation. There is usually a delay between being able to retrieve relevant information and generating a response to the question using the LLM. In retrieval, latency is high if the vector store is large as similar searches for relevance could take time. In generation, latency can be high when prompts have longer context, or the model size is larger).
  + **Business feasibility analysis:** This project will provide strong business value for warehouses and inventory-driven organizations after its completion. It would solve major operational problems using AI.
* It would reduce the human workload and time spent manually searching inventory systems and documentation.
* Initial costs would include GCP services, developer time, and team coordination, but in the long run, there would be fewer delays in processing inventory, reduced training costs (no need to memorize system paths), as well as streamlined audits due to better documentation access.
* Financial risks like Cloud costs and Vector storage costs would increase with data and complexity, but cost monitoring tools (Eg. Billing in GCP), and apply on usage quotas could help with this.
  + **Operation feasibility analysis:** The system, from an operational perspective, aligns well with existing workflows, and could probably be adopted with minimal resistance, as long as it is integrated carefully.
* User Adaptibility: The targeted users, in our case the Warehouse staff and managers, often need answers to routine queries but lack time to search multiple files and systems. A conversational interface that delivers accurate responses quickly would be welcomed by non-technical users. The risk here, is that if the system delivers irrelevant or slow responses, users may lose trust and avoid using it.
* Tech Support: A training module needs to be created to help users om how to phrase queries, understand responses and give feedback. There would also be a need to monitor usage and tune the system based on usage.
* Resistance to automation: Some team members may not fully understand or trust an AI system to find the right answers for their queries. Technical issues during deployment of this system may also disrupt access to other systems. To mitigate this, enough training, clear communication and emphasis on adding the use of the AI system as a regulatory practice during the initial rollout would help users trust the system.
* **Project Management:** We have made a project plan for this semester project. In the project plan we listed all the required tasks and assigned them to the respective person with timelines. It helped us to stay in track with respect to the activities with respect to the semester project.

A white sheet with black text

AI-generated content may be incorrect.

# Semester Project: System Design: High-Level Design

**A diagram of a diagram

AI-generated content may be incorrect.**

**Fig: High level system design**

# Semester Project: System Design: Detailed Design

A screenshot of a computer

AI-generated content may be incorrect.

**Fig: Detailed system design**

# Semester Project: System Development: Coding

**Q&A Search System Development Report:**

The system development was done as per the instructions provided by the instructor. We followed the provided ten phases to develop the components of the system.

There are three major components of the system which include:

* 1. **Knowledge Base:**
     + The knowledgebase consists of the PDFs that we downloaded from the fishbowl website.
     + These are stored in a GCP bucket
     + Indexes were built on these documents and saved in another folder in the same GCP bucket.
  2. **Retriever:** 
     + To implement search optimization, we used hybrid search strategies to get higher retrieval precision.
     + We used Vertex AI embeddings to convert text into vectors for similarity search.
     + Implemented LangChain retrieval-augmented generation (RAG) pipeline.
  3. **Text Generation**
     + The text generator takes the input as the ranked results from the Retriever.
     + The generator uses the context-rich prompt to generate a direct answer — converting a user's raw question into a fluent, human-readable response.
     + The test generator utilizes the prompt to get the answer. It converts the input text to a natural language response.

**Development Phases:**

* + **PHASE 1: Setting up the Development environment.**
    - Google Drive to access credentials and config files
    - Load .env file containing project-specific environment variables
    - Authenticate using GCP service account for Vertex AI access
  + **PHASE 2: Importing Necessary Libraries & Modules**
    - Install and import LangChain, Vertex AI, PyPDF, GCP SDKs, etc.
    - These libraries power document parsing, embedding, vector search, and question answering
  + **PHASE 3: Initialize the Vertex AI Platform**
    - Configure GCP project, region, and service credentials
    - Prepares system to access Vertex AI services like embeddings, LLMs, and vector indexing
  + **PHASE 4: Create Vector Search Index and Endpoint**
    - Checks if Matching Engine index and endpoint exist
    - If not, creates a new vector index for storing embeddings
    - Deploys the index to a public endpoint
  + **PHASE 5: Load PDF Documents from GCS**
    - Connects to Google Cloud Storage
    - Extracts text from all valid PDF files in the bucket
    - Skips folders and non-PDFs (e.g., .zip)
  + **PHASE 6: Build Vector Database with Embedded Documents**
    - This phase creates a vector store and adds embedded full documents (without chunking).
  + **PHASE 7: Test the Q&A Search System**
    - Performs a test query to verify retriever and vector store are functioning
    - Returns top-matching documents for a sample prompt
  + **PHASE 8: Set Up Prompt and Response Subsystem**
    - Defines prompt instructions to prevent hallucinations
    - Formats the LLM output with source references
  + **PHASE 9: Running Sample Queries and Generating Answers**
    - Uses RetrievalQA chain to run queries against document embeddings
    - Generates LLM answers using VertexAI LLMs
  + **PHASE 10: UN-DEPLOY Indexes and DELETE Indexes & End Points**
    - Get the list of end points that have been created
    - Un-deploy the indexes that have been deployed with these end points
    - Delete all the end points that have been created
    - Get the list of end points that have been created
    - Delete all the end points that have been created

# Semester Project: Testing and Operationalization

**Testing and Operationalization Report**

Once the system was successfully built, we performed various tests and tried to implement Operationalization for a smooth system performance.

* Content Validation: The content generated was validated by the team for it’s relevance to the question.
* Prompt-Response Evaluation: The response for the prompt was verified to check for it’s correctness.
* Test Scenarios: We asked the Q&A system to provide links. The system provided correct links.
* Error Handling Checks: We asked some incorrect questions which made no sense to identify if the system can handle them.
* End-to-End Testing: We tested the service account authentication, connectivity to Vertex AI and asked the question.

# Semester Project: Q&A Search System: Prompts and Responses

To be able to evaluate how our search system functions, we ran 10 queries/prompts through the model. We used content indexed from the data we scraped from Fishbowls website and saved as pdfs.

The system retrieved and each prompt and responded based on the similarity of the prompt to the documents embedded in our Vector store.

We have run the following Questions through the Q/A system to see the responses generated by it (Question 9 and 10 are meant to be irrelevant to our domain):

Q.1. "What are the different modules included with Fishbowl?"

**Response:** I cannot determine the answer to that. The provided context mentions that Fishbowl is organized into modules and gives examples such as the "My Fishbowl" module (in the General group), the "Sales Order" module (in the Sales group), and the "Accounting" module. However, it does not provide a complete list of all the modules included with Fishbowl. One document mentions an image that gives an overview of major modules, but the image content is not included in the text provided.

Q.2. “Where can I get Customer Support information?”

**Response:** Based on the provided context: The My Fishbowl module displays helpful information, such as support contract details. The Support tab within the My Fishbowl module displays the support tickets created when your company contacts Fishbowl, and a new request can be created by clicking the New Support Request button.

Q.3. "Where can I get the user manual?"

**Response:** Based on the provided context, the documentation for Fishbowl Advanced can be found at the following URL: `[https://help.fishbowlinventory.com/advanced/s/article/Documentation`](https://help.fishbowlinventory.com/advanced/s/article/Documentation%60). This page contains sections such as Getting started, General, Reporting, Sales, Purchasing, and others.

Q.4. "Can you give me a list of items with low inventory levels?"

**Response:** Based on the provided context, the \*\*Inventory Dashboard\*\* module includes graphs for \*\*short parts\*\*, which would represent items with low inventory levels.

Q. 5. "Can you provide the link for Inventory Dashboard?"

**Response:** The link for the Inventory Dashboard is: <https://help.fishbowlinventory.com/advanced/s/article/Inventory-Dashboard>

Q.6. "Can you provide information on pricing module?"

**Response:** The Pricing Rule module, found in the Sales group, enables automatic price adjustments for sales orders based on various factors like average cost, order quantity, customer, day, or product.  Key functions and features include: \*\*Creating Rules: \*\* You can create new pricing rules by clicking the "New" button, selecting applicable customers and products, and choosing an adjustment type and price/cost value. Pricing rules can also be imported. \*   \*\*Managing Rules: \*\* The module allows you to save, delete, duplicate, and inactivate pricing rules. \*   \*\*Adjusting Prices: \*\*Pricing rules can affect the sales price on a sales order and can also be used to adjust the actual "Price" field within the Product module. You can disable a rule for sales orders if it's only meant for adjusting product prices. \* \*\*Flexibility and Precedence: \*\* The module offers flexibility through combinations of types, tiers, and module options. Understanding pricing rule precedence is important to determine the final price on a sales order, as rules are applied in a specific order (e.g., customer rules often have high precedence). \*   \*\*Tabs: \*\* The module includes a General tab (for name, description, customers, products, adjustment types/values), an Advanced tab (for rounding options, dates, quantities, tiers), and a Details tab (for a summary and specific dollar amounts). \*   \*\*Customization: \*\* Some aspects of the Pricing Rule module can be customized via the Tools menu > Module Options.

Results: This response is one of the most detailed responses provided by the system. It outlines module functions, rule configurations and User Interface components. This shows the systems ability to understand structured documentation and provide a comprehensive answer.

 Q.7. "Where can I get details about my purchase order?"

**Response:** You can get details about your purchase order within the Purchase Order module.  Specifically, the context mentions the following tabs where you can find information: 1. The \*\*General tab\*\* displays important order information, including the vendor, order number, order status, addresses, order items, and more. 2. The \*\*Details tab\*\* displays important order details, including a link to the vendor, the last user to modify the order, shipping terms, payment terms, carrier, order dates, and more. 3. The \*\*Memo tab\*\* can be used to view or add information (memos) to the purchase order. 4. The \*\*Custom tab\*\* stores additional customized fields for the order.  You can use the \*\*Search pane\*\* on the left within the Purchase Order module to filter and find specific orders by Number, Vendor, or Status.

 Q.8."Provide details about Fishbowl Flow Chart"

**Response:** I cannot determine the answer to that. The provided context includes references to an article titled "Fishbowl Advanced - Flow Chart" but does not contain any actual content or details from that article.

 Q.9. "What is the meaning of life?"

**Response:** I cannot determine the answer to that.

 Q.10. “How do mangoes grow?”

**Response:** I cannot determine the answer to that.

# Semester Project: Group Contribution

* Sonali Sabnam: 100%
* Young Yu: 100%
* Sonam Pohuja: 100%