**ADTA 5900/5770.501: Generative AI with LLMs**

**Thuan L Nguyen, PhD**

**Assignment 5**

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

1. **Introduction** 
   1. Provide an overview of the project.

**Answer:**

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

With the vision of developing a Question-and-Answer system for making the data available for the users instead of depending on a database system and the technical expertise required to query data from it.

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.

**Company**: Fishbowl

1. **Problem statement** 
   1. Discuss in detail what problem (business, technical, …) the student is trying to solve with this project.

**Answer:  
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.

1. **System Requirements Analysis**
   1. **Business requirements** 
      1. Discuss in detail the business requirements of the Generative AI system.

**Answer:**

* **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

* 1. **Technical requirements** 
     1. **Discuss in detail the technical requirements of the generative AI system.**

**Answer:**

* **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.

* 1. **Data requirements** 
     1. **Discuss in detail the data requirements of the generative AI system.**

**Answer:**

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**

1. **Feasibility Analysis** 
   1. **Technical feasibility analysis:** 
      1. Can we complete the project successfully as required?
      2. Discuss any technical risks while working on the project.

**Answer:**

1. Yes, we can complete this project with available AI tools and cloud infrastructure. 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.

ii. Some technical risks we may face are:

* 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).
  1. **Business feasibility analysis:** 
     1. Will the project provide good business value after its completion?
     2. Discuss any financial risks while working on the project, e.g., running out of funding

**Answer:**

1. Yes, 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. There is less scope for operational downtime as well using real-time answers about inventory and warehouse-queries. This would also lead to faster and more confident decisions being made with on demand insights if necessary.
* 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.
  1. **Operation feasibility analysis** 
     1. If we build the system, will it be used by the organization as expected?
     2. Discuss any risks that may hinder the system’s deployment after its completion.

**Answer:**

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.

ii. With the introduction of any new process or system, comes risks that may hinder the systems deployment after completion. Some of them may be:

* **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.

1. **Project Management** 
   1. Discuss in detail the timeline of the project
      1. What kinds of significant tasks are to be done?
      2. What major phases need to be done until the completion?
2. Due date of each phase or task.

**Answer:**

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.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **SqNo** | **Task** | **Complexity** | **Done By** | **Timeline** | **Status** |
| 1 | Create a GCP Account | low | Sonam | 20-Jan | Done |
| 2 | Create a GCP project and bucket | low | Sonam | 25-Mar | Done |
| 3 | Data Acquisition | medium | Sonali/Young/Sonam | 14-Feb | Done |
| 4 | Loading the data to the shared folder | low | Young | 19-Feb | Done |
| 5 | Project proposal | medium | Sonali/Young/Sonam | 9-Feb | Done |
| 6 | System high level design | medium | Sonali/Young/Sonam | 31-Mar | Done |
| 7 | System low level design | medium | Sonali/Young/Sonam | 8-Apr | Done |
| 9 | Building the Q&A system (Coding) | high | Sonali/Young/Sonam | 22-Apr | in-progress |
| 10 | Mid-way presentation | medium | Sonali/Young/Sonam | 14-Apr |  |
| 11 | system development report | high | Sonali | 22-Apr |  |
| 12 | system testing report | high | Young | 25-Apr |  |
| 13 | Presentation | medium | Sonam | 28-Apr |  |

* 1. Discuss in detail the human resources needed for the project
     1. How many people, in total, are assigned to work on the project?

**Answer:** Our group has three people.

* + 1. How many people are assigned to work on each major phase of the project?

**Answer:** Everyone is assigned to a major work in this project. We are collaborating to successfully build a Question and Answer system using the required technology.

* + 1. For each phase, who does what?

**Answer:** As shown in the table above, we work on the assigned tasks as per the timelines and requirements.

1. **Conclusion** 
   1. Provide a short paragraph to express opinions about how the project will be done to conclude the report.
      1. For example: There are many challenges, but they will be successful with great effort.

**Answer:**

Even though there are challenges in building a domain-specific, AI-powered Q&A system, our project holds a strong technical and business value. With proper planning, modern cloud tools like LangChain and Vertex AI, and a collaborative team approach, the system could successfully improve inventory navigation, compliance, and decision-making within warehouse operations.

Our team is confident that with sustained effort and attention to detail, the project will be impactful and deployable.

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

**6.1 Q&A Search System: High Level Design**

**TO-DO**

* **Design (high-level) the semester project Q&A-Search system for the semester project**

**SUBMISSION REQUIREMENT: PART II #1:**

* **Submit the high-level design of the Q&A-Search System for the semester project**

**Answer:**

**A diagram of a diagram

AI-generated content may be incorrect.**

**6.2 Detailed Design**

**TO-DO**

* **Design (detailed) the Q&A-Search system for the semester project**

**SUBMISSION REQUIREMENT: PART II #2:**

* **Submit the detailed design of the Q&A-Search System for the semester project**

**Answer:**

A screenshot of a computer

AI-generated content may be incorrect.

1. **PART III: Teamwork Evaluation (10 Points)**

**SUBMISSION REQUIREMENTS PART VI:**

**Provide the information about your group activities by answering the following questions:**

1. **What group do you belong to? (Provide the group number)**

**Answer:** Group 3

1. **Who are the members of your group?**

**Answer:**

Sonali Sabnam

Sonam Pohuja

Young Yu

1. **Have the members organized meetings (ONLINE or IN-PERSON) to work on HW 5?**

**Answer:** Yes

1. **If YES to #3, which members, including the student himself/herself, showed up in the meeting?**

**Answer:**

Sonali Sabnam

Sonam Pohuja

Young Yu

1. **If YES to #3, do all the members make reasonable efforts to participate actively in the group work?**

**Answer:**

Yes, every team member contributes actively in all the discussions and meetings.

1. **If NO to #5, do you have any opinions to share about the group?**

**Answer:**

This is a proactive group with hardworking and dedicated team members.