Proof of Concept (POC): Retrieval-Augmented Generation (RAG) Using PDFs, Excel Files, and Q&A Files

## **Introduction**

This Proof of Concept (POC) demonstrates an innovative approach to building a Retrieval-Augmented Generation (RAG) system using multiple document formats, including PDF, Excel, and question-and-answer files. The goal is to extract, summarize, and efficiently retrieve relevant information using an open-source Unstructured library and multimodal Large Language Models (LLMs).

## **Objective**

The primary objective of this POC is to:

* Extract structured and unstructured data from PDFs, Excel files, and Q&A text files.
* Generate summaries for different data types (text, tables, images) using multimodal LLMs.
* Store and vectorize these summaries using a vector database (ChromaDB).
* Implement a retrieval mechanism that provides answers based on summarized vector data while ensuring traceability to the original document.

## **Technical Approach**

### **Step 1: Data Handling**

* The input consists of PDF, Excel, and Q&A files.
* The Unstructured library is used for handling these file types:
  + PDF Processing: partition\_pdf
  + Excel Processing: partition\_excel
  + Q&A Processing: partition\_text

### **Step 2: Data Extraction and Summarization**

* PDFs may contain a combination of text, tables, and images.
* Using the Unstructured library, we extract these elements separately:
  + Text Extraction: Extracted plain text is passed to the LLM for summarization.
  + Table Extraction: Tables are processed, and key insights are generated using the LLM.
  + Image Extraction: Images are analyzed using multimodal LLMs to generate relevant descriptions and summaries.
* A similar approach is applied to Excel files (tabular data processing) and Q&A files (text processing and summarization).

### **Step 3: Vectorization & Storage**

* The generated summaries from PDFs (text, tables, images), Excel files, and Q&A data are vectorized using ChromaDB.
* Doc\_ID Association:
  + The original data (raw, unprocessed content) is linked with its corresponding summary via a unique Doc\_ID.
  + This ensures that while the summary is used for retrieval, the final response can be traced back to the original document.

### **Step 4: Query Processing & Retrieval**

* When a user query is received, the system performs the following steps:
  + Searches the summary vector store for relevant content.
  + Retrieves the original document using the linked Doc\_ID.
  + Provides a final response using the original, high-fidelity data.

## **Multimodal LLMs Used**

The following multimodal LLMs are utilized in this system:

* Groq
* Gemini
* GPT

## **Key Benefits of This Approach**

* Efficient Summarization: Handles different file types, extracting meaningful insights from each.
* Improved Retrieval Accuracy: Vector-based search ensures faster and more relevant responses.
* Traceability & Reliability: Ensures responses are based on original documents, reducing the risk of hallucinations.
* Scalability: The system can be extended to handle more document formats and additional processing layers.

## **Conclusion**

This POC successfully integrates Unstructured library with multimodal LLMs and ChromaDB to create an effective RAG system. The approach ensures efficient summarization, retrieval, and traceability, making it highly suitable for real-world applications.

We believe this solution can be a game-changer for enterprises handling vast amounts of unstructured data. We look forward to discussing how we can implement this in a full-scale project for your organization.

Next Steps: We are open to further discussions to tailor this POC to your specific needs and integrate additional enhancements based on your requirements.