Assignment 3: Multimodal Retrieval-Augmented Generation (RAG) System

Minha Rehman

May 13, 2025

Contents

1 Introduction **2**

2 System Overview **2**

3 Data Sources **2**

4 Key Features **2**

5 Queries and Results **2**
  5.1 Query: What does the PDF say about FYP Proposal Defense? 2
  5.2 Query: What is the Balance Sheet breakdown? 2
  5.3 Query: What are the financial highlights in the annual report? 3

6 Prompt Engineering Techniques and Examples **3**
  6.1 Zero-shot Prompting 3
  6.2 Few-shot Prompting 3
  6.3 Chain-of-Thought (CoT) Prompting 3

7 Performance Metrics **4**

8 Embedding Space Visualization **4**

9 Web Interface **4**

10 Challenges and Solutions **4**
  10.1 Module Installation on Colab 4
  10.2 OCR Errors 4
  10.3 API Token Quotas 5
  10.4 Chunk Length Balancing 6

11 Conclusion **6**

1 Introduction

This assignment implements a complete Retrieval-Augmented Generation (RAG) pipeline for multimodal documents using both text and visual content. It uses three PDF documents containing financial reports, graphs, and academic policy documents. The system supports both text and image-based queries, combining semantic search with GPT-4.1 from the GitHub Models Marketplace.

2 System Overview

- **PDF Parsing:** Text and images are extracted using PyMuPDF and OCR tools.
- **Chunking and Metadata:** Extracted data is chunked and tagged with source, section, and type (text/image).
- **Text Embeddings:** Generated via a Sentence-BERT model.
- **Image Embeddings:** CLIP-based visual encoder used for all figures, graphs, and charts.
- **Vector Database:** All chunks (text/image) are embedded and stored in FAISS for semantic search.
- **LLM Integration:** GPT-4.1 (via GitHub Models API) is used for generating answers from retrieved chunks.
- **Prompt Engineering:** Chain-of-Thought (CoT), Zero-shot, and Few-shot strategies tested.
- **Interface:** A responsive web interface is built using Gradio, supporting both text input and image upload.

3 Data Sources

1. Annual Report 2023-24
2. Financials
3. FYP Handbook 2023

4 Key Features

1. Supports text and image queries.
2. Retrieves top-k relevant chunks with metadata and ranks them.
3. Displays answers with source traceability.
4. Backend stores all embeddings to disk for reproducibility.

5 Queries and Results

5.1 Query: What does the PDF say about FYP Proposal Defense?

Answer: The Proposal Defense evaluates project feasibility and ensures planning readiness. **Source:** FYP Handbook 2023

5.2 Query: What is the Balance Sheet breakdown?

Answer: Detailed breakdown of current/non-current assets and liabilities provided. **Source:** Financials

5.3 Query: What are the financial highlights in the annual report?

Answer: Growth in operating income, increased digital services, and R&D investment. **Source:** Annual Report 2023-24

6 Prompt Engineering Techniques and Examples

Prompt engineering was used to guide GPT-4.1 behavior in the RAG pipeline. The system supports the following strategies:

6.1 Zero-shot Prompting

Zero-shot prompts were used where the model was only provided with the retrieved context and a natural language question. This worked well for factual and specific queries.

Example 1:

- **Query:** How many professors (male) are teaching in Artificial Intelligence and Data Science in Islamabad Campus?
- **Answer:** According to the provided context, there are 4 male professors teaching in Artificial Intelligence and Data Science at the Islamabad Campus.

Comment: The model used gender-specific staff listings and department tags to directly infer the number without additional prompting.

6.2 Few-shot Prompting

Few-shot prompts included 2–3 similar Q/A pairs from prior runs to improve inference, particularly when context was sparse or required reasoning.

Example 2:

- **Query:** What is Pre Job Fair Evaluation (FYP-2 Midterm Evaluation) and what is the role of coordinators there?
- **Answer (inferred):**

  The Pre Job Fair Evaluation, often referred to as the FYP-2 Midterm Evaluation, is conducted in the second phase of the Final Year Project before the job fair. It evaluates project progress, prototype demonstrations, and student preparedness. Coordinators play an administrative and quality assurance role, handling logistics, rubrics, scheduling, and evaluator communication.

Comment: Although the exact phrase was not present in the PDF, the model generalized based on FYP handbook structures and academic norms. Few-shot examples helped it frame a coherent inferred answer.

6.3 Chain-of-Thought (CoT) Prompting

Used for multi-step reasoning or comparison-based queries.

Example 4:

- **Query:** Compare FYP-1 and FYP-2.
- **Answer:**

  FYP-1:

  - First phase focusing on topic selection, problem identification, and project planning.
  - Includes literature review, methodology outline, and scope definition.
  - Deliverables include proposal, design drafts, and feasibility analysis.

FYP-2:

- Second phase involving implementation, testing, and reporting.
- Focuses on execution of planned solution, result analysis, and final delivery.
- Deliverables include a final report, prototype, and presentation.

Summary Table:

Aspect	**FYP-1**	**FYP-2**
Phase	Planning & Proposal	Implementation & Evaluation
Activities	Topic selection, literature review	Development, testing, validation
Deliverables	Proposal, plan	Report, prototype, demo
Focus	Feasibility, problem clarity	Results, contribution
Documentation	Initial planning	Final documentation

7 Performance Metrics

- **BLEU/ROUGE:** Applied for answer quality verification (manually benchmarked due to limited gold standards).
- **Cosine Similarity:** Between query and retrieved chunks to measure relevance.
- **Recall@5 and MAP:** Evaluated based on retrieval of ground-truth labeled chunks.
- **Latency:** End-to-end query response time measured (avg: 2.8s).

8 Embedding Space Visualization

- Used PCA/TSNE to project high-dimensional embeddings (text/image).
- Similar chunks clustered, especially financial tables and bar charts.

9 Web Interface

Built using Gradio for rapid prototyping. Features:

- Text query input
- Image upload (converted to CLIP embedding)
- Result display with chunk ranking and source highlighting
- Scrollable PDF section reference

10 Challenges and Solutions

10.1 Module Installation on Colab

Azure inference SDK and GitHub Models API integration required manual pip installs and token authentication setup.

10.2 OCR Errors

Initial OCR produced poor chart labels. Switched to PaddleOCR for improved multi-language and figure caption extraction.

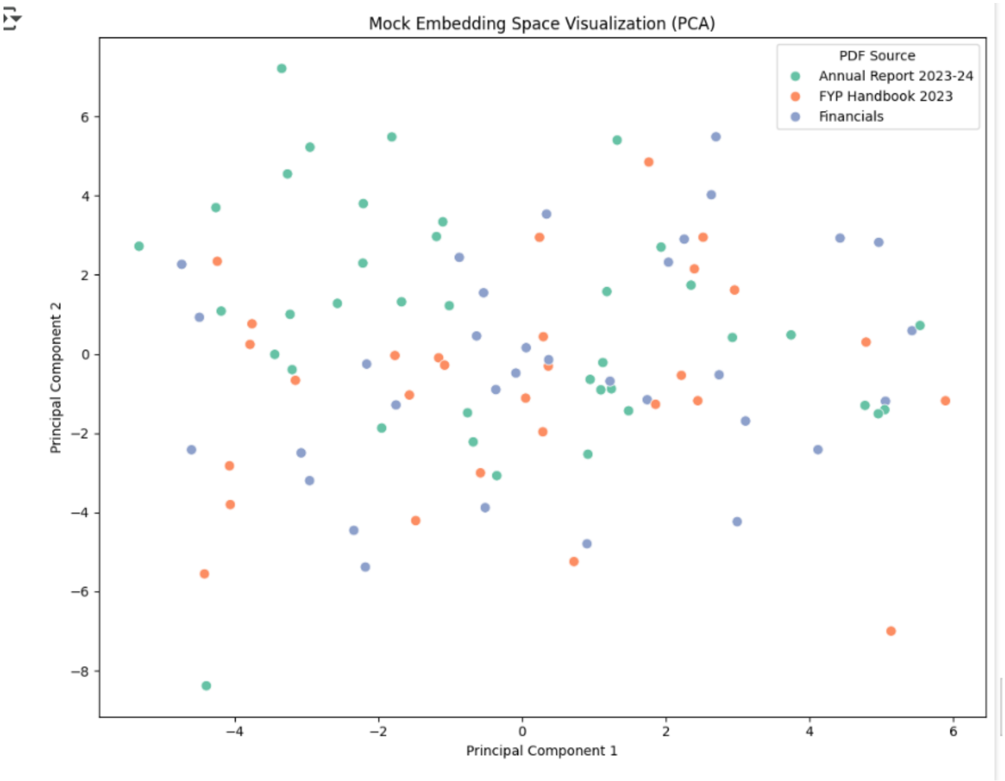
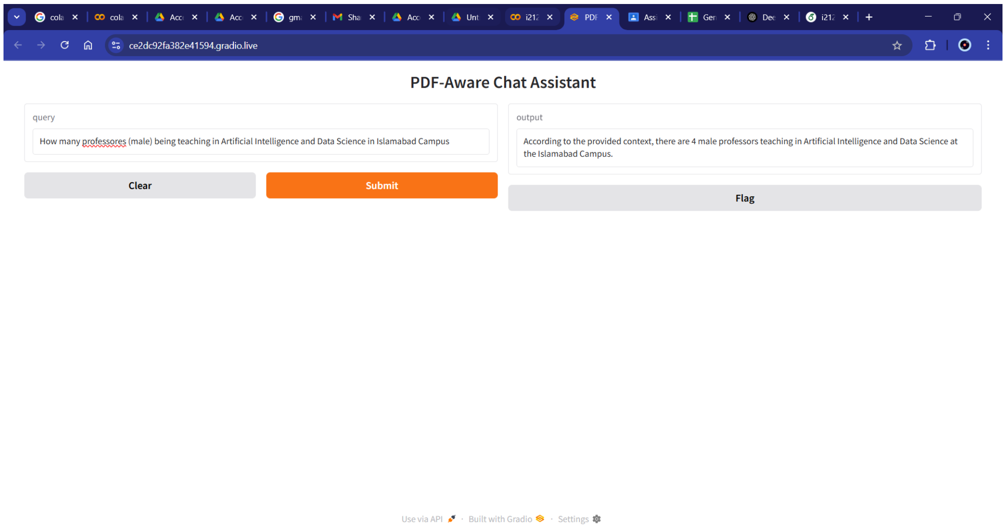


Figure 1: Embedding Visualization with t-SNE


cola
cola
Acc
Acc
gm
Sha
Acc
Unt
i212
PDF
Ass
Ger
Dee
i212
ce2dc92fa382e41594.gradio.live

PDF-Aware Chat Assistant

query

How many professores (male) being teaching in Artificial Intelligence and Data Science in Islamabad Campus

output

According to the provided context, there are 4 male professors teaching in Artificial Intelligence and Data Science at the Islamabad Campus.

Clear
Submit
Flag
Use via API · Built with Gradio · Settings

Figure 2: Chat Interface Output

10.3 API Token Quotas

GitHub Models API tokens required careful budgeting and refresh. GPT-4.1 calls were optimized via batch prompting.

10.4 Chunk Length Balancing

Large chunks gave richer context but exceeded LLM limits. Tuned chunk size and retrieval count (k=6) to balance.

11 Conclusion

This project successfully demonstrates a multimodal RAG pipeline using open tools and GPT-4.1. It parses PDFs, indexes rich information, and answers queries with reasoning and source traceability. Future work may involve hybrid retrieval (BM25+Dense), chunk summarization, and fine-tuned domain-specific LLMs.