Multi-Modal RAG System — Technical Report

This report provides an overview of the Multi-Modal RAG prototype, covering the ingestion pipeline, retrieval system, and Streamlit interface.

1. Overview

The system extracts text, tables, and images (with OCR) from PDF files and converts all extracted content into vector embeddings. These embeddings are stored in a FAISS index and later retrieved through a simple Streamlit UI.

2. Ingestion Pipeline

• Extract text using pdfplumber

• Extract tables and convert them into CSV-like text

• Extract images and run OCR using Tesseract

• Chunk long text into manageable parts

• Encode chunks using SentenceTransformer (all-MiniLM-L6-v2)

• Store embeddings inside a FAISS IndexFlatIP

• Save metadata (page, type, source, etc.) into metadata.json

3. Embedding Model

The model all-MiniLM-L6-v2 is used to generate 384-dimensional embeddings optimized for semantic similarity retrieval.

4. FAISS Indexing

Workflow:

• Normalize embeddings

• Build IndexFlatIP

• Save faiss.index and metadata.json

5. Retriever Module

• Loads FAISS index

• Loads metadata

• Encodes query

• Returns top-k similar chunks with metadata

6. Streamlit UI

The Streamlit interface allows users to:

• Enter questions

• Retrieve top relevant PDF chunks

• View metadata such as source and page number

7. Limitations

• LLM answer generation is not included (optional in assignment)

• Prototype focuses only on extraction + retrieval

8. Conclusion

This project demonstrates a working end-to-end multi-modal retrieval-augmented system:

PDF → Extraction → Embeddings → FAISS → Retrieval → UI.