Multi-Modal Document Intelligence System

Retrieval-Augmented Question Answering over Complex Financial Documents

Candidate: Samyak Mittal
Position: AI/ML Intern
Organization: Big AIR Lab

Problem Motivation

Research and policy institutions such as the IMF produce information-dense reports that combine narrative text, structured tables, charts, scanned figures, and footnotes. Many important insights are embedded outside plain text, making traditional text-only retrieval and QA systems insufficient.

The goal of this assignment was to design and implement a **multi-modal Retrieval-Augmented Generation (RAG) system** capable of answering questions grounded in heterogeneous document elements while maintaining faithfulness, traceability, and clarity.

—

System Overview

The system is designed as a modular pipeline, allowing independent evolution of ingestion, retrieval, and generation components.

High-level flow:

> Document → Multi-modal ingestion → Chunking → Embeddings → Vector index → Retrieval → Context-grounded answer generation

The architecture emphasizes:

- Modality-agnostic retrieval
- Explicit source attribution
- Research-friendly extensibility

—

Multi-Modal Ingestion

Text Processing

- Structured text extracted from PDFs
- Page numbers and section context preserved
- Footnotes handled as separate logical units

Table Processing

Tables are serialized into structured textual representations:

- Table title (if available)
- Header and row values
- Page-level metadata

This representation enables numerical reasoning while remaining LLM-compatible.

Image and Chart Processing

- Images extracted from PDFs
- OCR applied to scanned text and charts
- Extracted text enriched with contextual metadata

All modalities are normalized into a unified textual format with metadata.

—

Chunking Strategy

A hybrid **structural and semantic chunking** approach is used:

- Text chunks: 300–500 tokens with overlap
- Tables: one logical table per chunk
- Images: OCR text grouped by page and visual context

Each chunk stores modality type, page number, and source identifier to support accurate citation.

—

Embedding and Vector Index

All chunks, regardless of modality, are embedded into a **shared vector space**. This design enables cross-modal retrieval, such as retrieving table or image-derived content in response to text queries.

- Unified embedding model
- FAISS-based vector index
- Metadata-aware retrieval

—

Retrieval and Answer Generation

Retrieval

Top-K similarity search is performed over the vector index. Retrieved chunks are ranked and passed as contextual evidence.

Answer Generation

The language model is constrained to generate answers solely from retrieved context, reducing hallucinations.

Each answer includes:

- Page-level citation
- Section or table reference

—

Demo Application

A lightweight interactive QA interface was implemented using Streamlit.

- User-driven query input
- Real-time retrieval
- Citation-backed responses

The demo supports queries grounded in text, tables, and OCR-extracted image content.

—

Evaluation

An evaluation set consisting of benchmark queries across modalities was used.

Query Modality	Performance
Text-based policy questions	Accurate
Table-driven numerical queries	Accurate
Image/OCR-based queries	Accurate
Cross-modal retrieval	Successful

Key observations:

- Multi-modal ingestion improves answer completeness
- Table serialization is critical for numerical faithfulness
- Metadata-aware chunking reduces hallucination risk

—

Design Considerations

The system prioritizes:

- Faithfulness over speculative generation
- Modularity over monolithic design
- Clear engineering-research trade-offs

Heavy fine-tuning was intentionally avoided to focus on system-level robustness.

—

Future Work

Potential extensions aligned with applied research directions include:

- Cross-modal reranking with vision-language embeddings
- Hybrid lexical and dense retrieval
- Retrieval fine-tuning using contrastive learning
- Executive briefing and summary generation

—

Conclusion

This project demonstrates a practical yet research-aligned approach to multi-modal document intelligence. By emphasizing grounded retrieval, clear representation, and modular design, the system provides a strong foundation for real-world policy and financial document QA applications.