Candidate Assignment: Multi-Modal Document Intelligence (RAG-Based QA System)

Project Context

Modern LLM applications must handle real-world documents — not just plain text. Financial and policy documents, such as IMF Article IV reports, often include text, tables, charts, figures, scanned images, and footnotes. These diverse elements carry critical insights that a conventional text-only QA system might miss. Your challenge is to design and implement a Multi-Modal Retrieval-Augmented Generation (RAG) system that can accurately answer questions based on such complex, information-rich documents.

Objectives

- Develop a document ingestion pipeline that parses text, tables, and images (OCR included).
- Create a chunking and embedding strategy for multi-modal data.
- Build a vector-based retrieval system combining multiple modalities.
- Implement a chatbot/QA interface that generates context-grounded, citation-backed answers.

Expected Features

Feature	Description
Multi-modal ingestion	Handle text, tables, images (OCR), and chart metadata extraction
Vector index	Unified multi-modal embedding space
Smart chunking	Semantic and structural segmentation for LLM-friendly retrieval
QA chatbot	Interactive question answering grounded in document context
Source attribution	Include page or section-level citations
Evaluation suite	Benchmark queries across multiple modalities

Deliverables

1 Codebase – Well-structured and documented; modular components for ingestion, retrieval, and answer generation.

2 Demo Application – Streamlit/Gradio/FastAPI or CLI QA interface supporting context retrieval and response generation.

3 Technical Report (max 2 pages) – Summarize architecture, design choices, benchmarks, and key observations.

4 Video Demonstration (Mandatory) – 3–5 minute screen recording showing pipeline, demo, and insights.

Evaluation Criteria

Criteria	Weight	Description
Accuracy & Faithfulness	25%	Quality of generated answers and factual grounding.

Multi-modal Coverage	20%	Ability to process text, tables, and images.
System Design & Architecture	20%	Scalability, modularity, and clarity of pipeline.
Innovation & Tooling Choice	15%	Smart use of models, embeddings, and retrieval techniques.
Code Quality & Clarity	10%	Readability, modularity, and documentation.
Presentation & Report	10%	Clarity of documentation, UX, and communication.

Bonus (Excellence Track)

- Implement cross-modal reranking combining vision-text embeddings.
- Add retrieval fine-tuning using contrastive learning or hybrid search (RRF).
- Include summarization or briefing generation features.
- Develop an evaluation dashboard showing retrieval metrics and latency.

Duration Options

24-hour challenge: Deliver a minimal working prototype (MVP).
48-hour challenge: Deliver a complete ingestion → retrieval → generation pipeline with UI and report.

Outcome

This assignment mirrors real-world engineering challenges faced in building multi-modal document intelligence systems. We are looking for clarity of thought, technical rigor, and creativity — not just model usage. Your submission should reflect your ability to translate complex requirements into an elegant, functional prototype.