Technical Report: Multi-Modal RAG System

1. Introduction

Economic and policy reports contain heterogeneous information such as narrative text, numerical tables, and visual figures. Question answering over such documents requires accurate retrieval and strong safeguards against hallucination.

This project implements a **Multi-Modal Retrieval-Augmented Generation (RAG) system** that answers questions strictly from document content, provides **page-level citations**, and safely refuses unsupported queries. The system is evaluated on an IMF-style report on **Qatar's macroeconomic outlook**.

---

2. System Design

The system follows a modular **ingestion → retrieval → generation → verification** pipeline.

During ingestion, the PDF is parsed into three modalities:

- **Text** extracted per page.
- **Tables** extracted and preserved as structured text.
- **Images/Charts** processed using **OCR (Tesseract)** to extract visible textual elements such as legends and labels.

Extracted content is chunked (**400 characters, 50 overlap**) and embedded **using sentence-transformer (MiniLM-L6-v2)**. Dense semantic search is performed using **FAISS**, while **BM25** enables sparse keyword retrieval. Results from both methods are merged and reranked using a cross-encoder to improve precision.

Answer generation is performed under strict constraints. The system generates answers **only when explicitly supported by retrieved context**; otherwise, it returns *"Not found in document."* A verification step ensures that all generated statements are grounded in cited content.

---

3. Evaluation

Evaluation is qualitative, as required by the assignment, and focuses on correctness, grounding, and hallucination prevention.

3.1 Text-Based Queries

Query: *"Qatar has made limited progress in which diversification?"*
Answer: *Economic diversification (Page 8).*
This response is explicitly stated in the document and demonstrates correct retrieval of narrative analysis.

3.2 Table-Based Queries

Query: *"Write about GDP."*
Answer: The system correctly summarizes **real and nominal GDP growth projections** from macroeconomic tables (Page 61) without inventing values, demonstrating reliable handling of numerical data.

3.3 Image/OCR-Based Queries

Query: *"What type of inflation is represented in the chart legend?"*
Answer: *Wage Inflation (Page 10).*
This response is derived from **OCR-extracted text within a chart legend**, confirming multimodal support for image-based text.

3.4 Unsupported Queries

Query: *"What is Qatar's unemployment rate in 2024?"*
Answer: *Not found in document.*
The system correctly refuses to answer, demonstrating effective hallucination prevention.

---

4. Observations

- The system performs well on **explicitly stated textual and numerical information**.
- OCR is effective for extracting **literal text from figures**, while visual trend interpretation is intentionally avoided.
- Broad or abstract questions are conservatively rejected to maintain reliability.

---

5. Limitations

- The system does not perform visual reasoning over images or charts; OCR is limited to extracting explicit text such as legends and labels.
- Due to strict grounding and verification, some queries may return *"Not found in document"* even when related information exists but is not retrieved or ranked with sufficient confidence.
- Dense or complex tables are not always retrieved reliably, reflecting limitations in table-aware retrieval and ranking.
- The system prioritizes precision over recall, reducing hallucinations at the cost of answer coverage.

---

6. Conclusion

This project demonstrates a practical **multi-modal RAG system** for grounded document question answering. By combining hybrid retrieval, reranking, OCR integration, and strict verification, the system prioritizes **accuracy and safety**. The evaluation confirms correct handling of supported queries and appropriate refusal of unsupported ones, meeting the objectives of the assignment.