Week 5 – Unified RAG Pipeline Write-Up

Introduction

In Week 5, I extended my baseline RAG pipeline (from Week 4) with advanced retrieval techniques, multimodal support, and evaluation guardrails. The goal was to make retrieval more accurate, safer, and adaptable to both text and image-based project materials. This work is directly connected to my project documents (papers, datasets, technical notes, and charts), where improved retrieval quality and multimodal support are crucial for analyzing both written reports and accompanying visual data.

Track A: Reranking & Context Optimization

- **Techniques Added**:
    - Reciprocal Rank Fusion (BM25 + dense embeddings).
    - Maximal Marginal Relevance (MMR) for diversity.
    - Cross-Encoder reranker for fine-grained scoring.
    - TextRank compression to reduce long contexts.
- **Findings**:
    - Baseline recall improved with RRF and reranking.
    - Compression reduced average context length by ~30–40%, which lowers token cost without harming recall.
    - Reranker ensured top-ranked chunks were more semantically aligned with the query.
- **Connection to Project**:
    - Helps focus retrieval on precise technical sections in papers.
    - Makes outputs more concise for downstream summarization tasks.

Track B: Multimodal Retrieval

- **Implementation**:
    - Integrated CLIP/BLIP2 models to generate embeddings for both text chunks and images.
    - Built a joint FAISS index to allow text-only, image-only, and hybrid queries.
    - Demonstrated retrieval of charts (e.g., error trend plots) alongside text passages.
- **Findings**:
    - The system successfully returned relevant charts when queried with textual descriptions.
    - Hybrid queries (text + chart context) gave richer answers than text-only.
- **Connection to Project**:
    - Enables querying both research reports and experimental charts/datasets.
    - Critical for scientific workflows where insights are often visual (plots, tables, diagrams).

Track C: Evaluation & Guardrails

- **Evaluation**:
    - Built an eval set (`eval_queries.jsonl`) with factual, interpretive, and adversarial queries.
    - Computed metrics: Recall@4, context precision/recall, correctness, faithfulness, latency, and token cost.
- **Guardrails**:
    - Enforced citation requirement (answers must cite sources).
    - PII redaction for emails, phone numbers, and API keys.
    - Refusal template for unsafe/adversarial queries.
- **Findings**:

    - Guardrails worked: adversarial queries (“leak API key”) triggered safe refusal.
    - Recall was stable across variants; rerank+compression had the best balance of recall and latency.
- **Connection to Project**:
    - Ensures system outputs are trustworthy and do not leak sensitive data.
    - Provides a reproducibility framework with `ablation_results.csv`.

Conclusion

This unified pipeline integrates advanced retrieval, multimodal support, and guardrails into a single reproducible framework.

- **Strengths**: More accurate retrieval, multimodal flexibility, safer outputs.
- **Next Steps**:
    - Expand evaluation set to 15–20 queries.
    - Add more robust correctness/faithfulness scoring (LLM-based evaluators).
    - Explore scaling to larger embedding/reranker models.

Overall, the Week 5 pipeline significantly improves over the Week 4 baseline and provides a practical foundation for applying RAG to my project’s mix of technical papers, datasets, and charts.