A Multimodal Retrieval-Augmented Generation System for Banking Knowledge Access

In recent years, the emergence of Retrieval-Augmented Generation (RAG) has enabled many sectors to exploit their internal data more effectively, leading to increased productivity and better utilization of organizational knowledge. However, traditional RAG systems remain limited, as in domains such as finance and healthcare, information is not exclusively textual. Instead, valuable knowledge is distributed across multiple modalities including images, videos, and audio making unimodal retrieval approaches insufficient.

We propose a multimodal RAG system designed to enable intelligent access to banking data by grounding multiple modalities into a unified textual representation. The pipeline integrates text, PDFs, images, and video transcripts, with embeddings generated using the BAAI/bge-m3 model and stored in separate Qdrant vector collections per modality. A grounding mechanism ensures all modalities are consistently transformed into a primary textual space, facilitating semantic alignment.

To improve retrieval quality, we incorporate **RAG-Fusion**, which generates multiple reformulations of the user query, aggregates results, and reranks them for higher relevance within each modality. The retrieved candidates are then combined across modalities through a **late fusion strategy**, ensuring both intra-modal precision and cross-modal scalability.

The generation stage leverages **LLaMA 3.1**, which synthesizes coherent and context-aware responses tailored to financial queries. Preliminary evaluations on internal banking datasets indicate improvements in both accuracy and explainability compared to text-only RAG baselines.

This contribution highlights how combining multimodal RAG with advanced retrieval strategies can empower financial institutions with richer, context-sensitive AI assistants, while offering a scalable framework applicable to other high-stakes domains.

Keywords: Retrieval-Augmented Generation, Multimodal Retrieval, RAG-Fusion, Late Fusion, Grounding, Banking AI, Financial Knowledge Management