Hybrid Retrieval-Augmented Generation (RAG) System
Academic Evaluation Report

Abstract

This project implements a Hybrid Retrieval-Augmented Generation (RAG) system that combines dense semantic retrieval, sparse keyword-based retrieval (BM25), and Reciprocal Rank Fusion (RRF) to answer user queries from a large Wikipedia corpus. The system is evaluated using an automated framework over 100 generated questions.

System Architecture Overview

The architecture includes Data Processing, Retrieval Layer, Fusion Layer, and Generation Layer. Queries are processed in parallel using dense and sparse retrievers, fused using RRF, and passed to the generator model.

Methodology

Dense retrieval uses SentenceTransformers with FAISS indexing. Sparse retrieval is implemented using BM25. Results are combined using Reciprocal Rank Fusion with k=60. Final context is passed to Flan-T5-base for answer generation.

Evaluation Setup

The evaluation pipeline automatically runs all questions across Dense, Sparse, and Hybrid modes while recording retrieval accuracy and latency metrics.

Metrics and Justification

MRR is used to measure ranking quality at the document level. Recall@5 measures retrieval coverage. Average latency evaluates efficiency. These metrics together provide balanced evaluation.

Results Summary

Hybrid mode achieved the highest performance with improved MRR and Recall@5 compared to dense-only and sparse-only approaches.

Ablation Study

Dense-only, Sparse-only, and Hybrid modes were compared. Hybrid consistently outperformed individual retrievers.

Error Analysis

Common errors included ambiguous queries, overlapping Wikipedia topics, and partial chunk relevance.

User Interface

The Streamlit interface shows answer output, retrieval transparency, chunk scores, response time, and retrieval breakdown tabs.

Installation and Reproducibility

The system uses open-source libraries and can be executed locally with reproducible scripts.

Conclusion

The project demonstrates that hybrid retrieval improves robustness and retrieval quality while maintaining efficient response times.