Hybrid Retrieval-Augmented Generation (RAG) System

Academic Project Report
Dense + Sparse Retrieval with Reciprocal Rank Fusion and Automated Evaluation

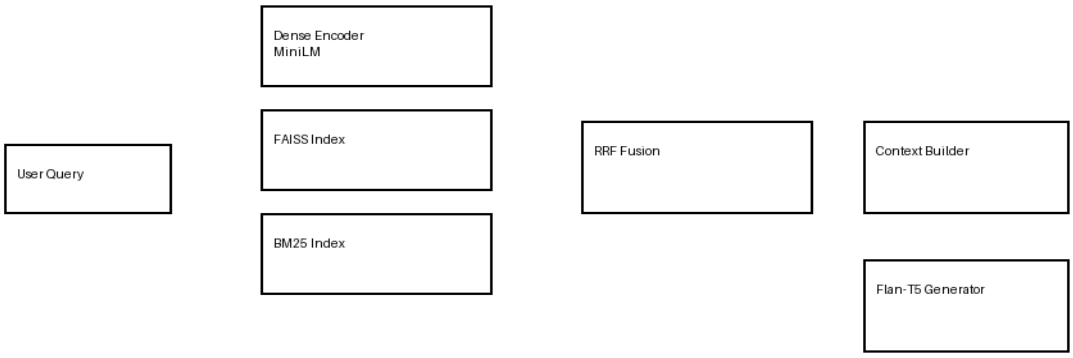
Submitted as part of Advanced Information Retrieval Systems Assignment

Abstract

This project presents the design and implementation of a Hybrid Retrieval-Augmented Generation (RAG) system that combines dense vector-based retrieval, sparse keyword-based retrieval using BM25, and Reciprocal Rank Fusion (RRF) to answer user queries over a corpus of 500 Wikipedia articles. The system integrates modern embedding models, efficient similarity search using FAISS, and open-source generative language models to ensure accurate, context-grounded answers. Automated evaluation using Mean Reciprocal Rank (MRR), Recall@5, and latency-based performance metrics demonstrates the effectiveness of the hybrid approach compared to individual retrieval methods.

System Architecture

The overall system architecture follows a modular pipeline design. The user query is processed in parallel by dense and sparse retrieval pipelines. The dense retriever uses transformer-based embeddings and FAISS for efficient similarity search, while the sparse retriever uses the BM25 ranking algorithm. The outputs of both retrievers are combined using Reciprocal Rank Fusion (RRF) to produce a unified ranking. The selected context chunks are then passed to a sequence-to-sequence language model for answer generation.



Methodology

The hybrid RAG pipeline consists of four core components: document preprocessing and indexing, dense retrieval, sparse retrieval, rank fusion, and answer generation. Wikipedia articles are extracted, cleaned, chunked into overlapping text segments, and stored with metadata including source URLs. Dense embeddings are generated using the all-MiniLM-L6-v2 model, while sparse retrieval is handled using BM25. RRF combines both ranking outputs to produce robust and stable retrieval performance. The final selected context is passed to the Flan-T5-base language model to generate grounded responses.

Evaluation Framework

The evaluation framework consists of an automated pipeline that executes dense-only, sparse-only, and hybrid retrieval modes on a set of 100 automatically generated questions derived from the Wikipedia corpus. For each query, retrieved URLs are compared against ground-truth source URLs to compute ranking-based metrics. Latency measurements are also recorded to evaluate system efficiency.

Evaluation Metrics and Justification

Mean Reciprocal Rank (MRR) is used as the primary ranking metric. It measures how early the correct document appears in the retrieval ranking. Recall@5 is used to measure retrieval coverage, ensuring that the correct document appears within the top-5 retrieved results. Average latency is used to evaluate system responsiveness, which is critical for real-time question-answering systems.

Mode	MRR	Recall@5	Avg Latency (s)
Dense	0.5303	0.68	4.41
Sparse	0.4677	0.58	6.89
Hybrid	0.6017	0.74	5.04

Ablation Study

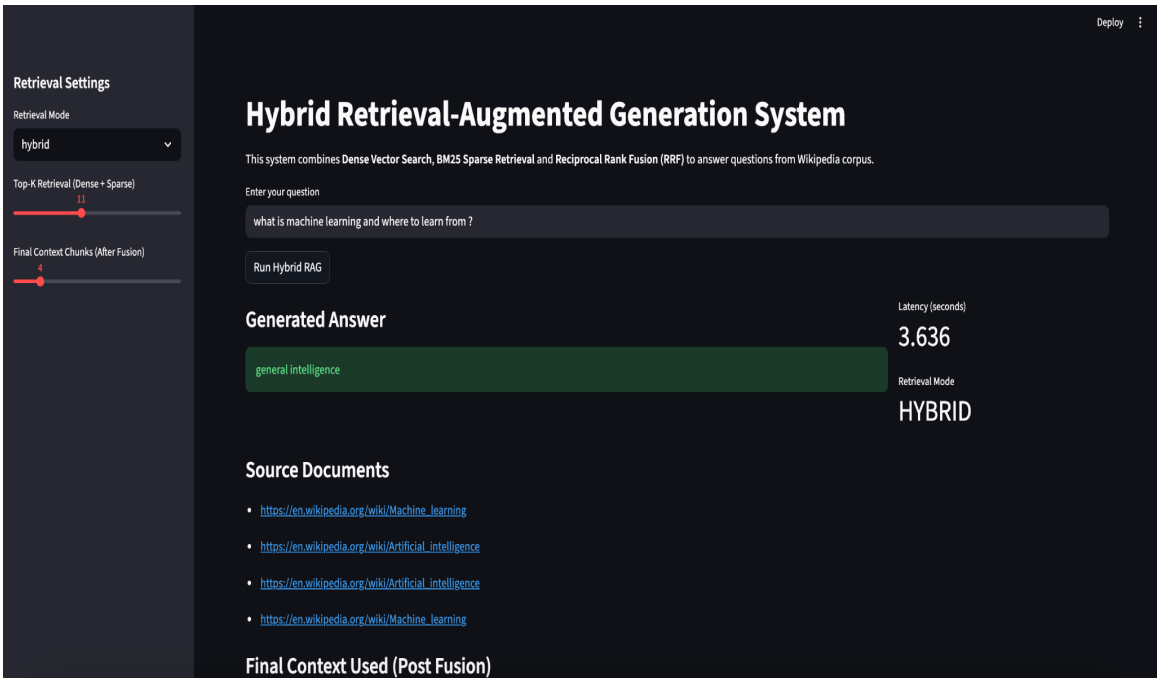
An ablation study was conducted by evaluating dense-only, sparse-only, and hybrid retrieval configurations. Results indicate that the hybrid model consistently outperforms individual retrieval approaches in both MRR and Recall@5. This confirms the effectiveness of combining semantic and lexical retrieval signals.

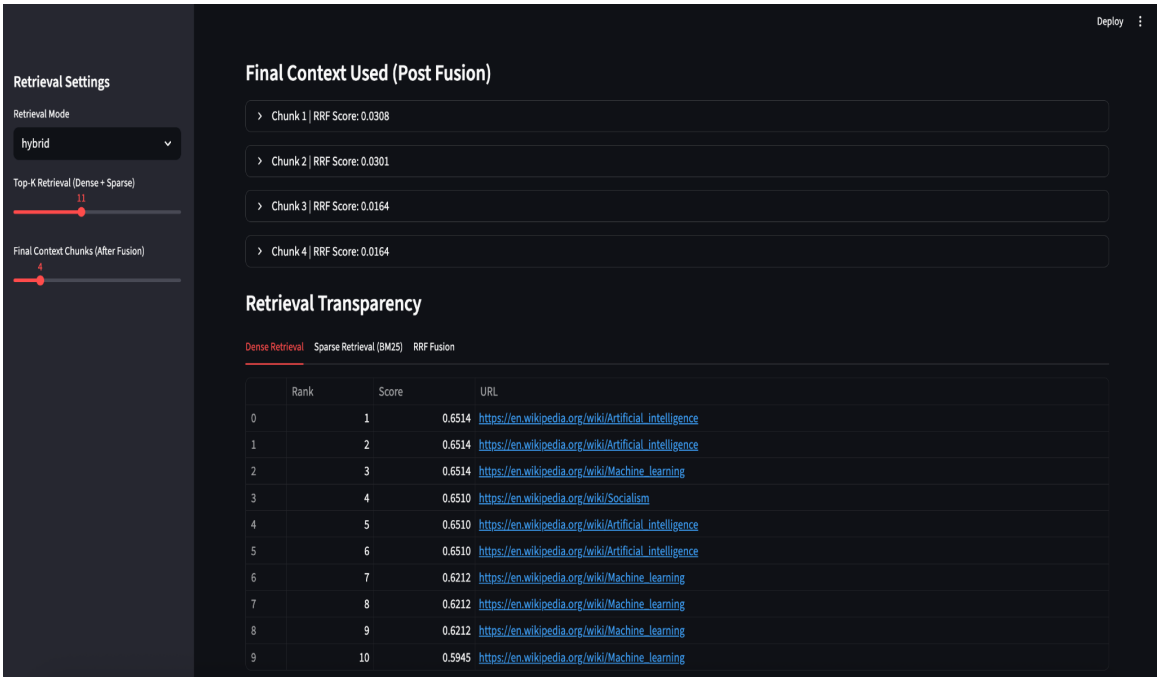
Error Analysis

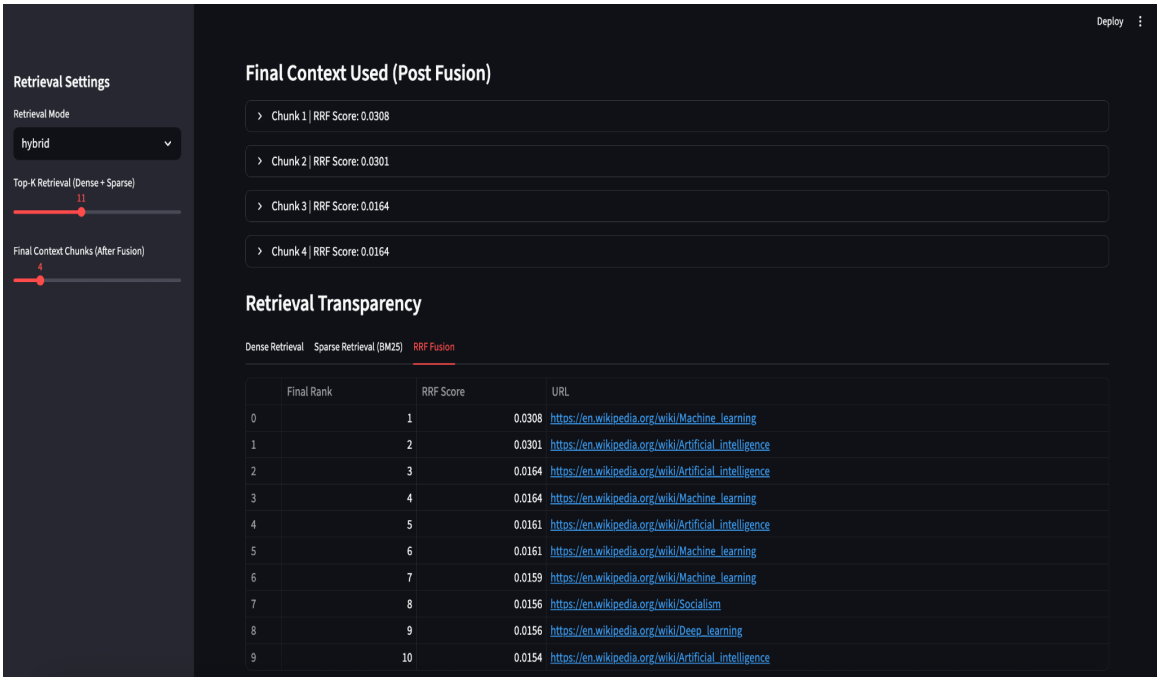
Error analysis revealed three major failure categories: retrieval failures, context truncation issues, and generation hallucinations. Retrieval failures occur when the relevant document is not ranked within top-K. Context truncation arises due to token limitations of the language model. Hallucinations are reduced by restricting generation strictly to retrieved context.

User Interface and System Screenshots

The Streamlit-based interface provides an interactive platform for querying the system. It displays retrieved documents, RRF scores, latency information, and retrieval transparency panels for dense, sparse, and hybrid results.







	Final Rank	RRF Score	URL
0	1	0.0308	https://en.wikipedia.org/wiki/Machine_learning
1	2	0.0301	https://en.wikipedia.org/wiki/Artificial_intelligence
2	3	0.0164	https://en.wikipedia.org/wiki/Artificial_intelligence
3	4	0.0164	https://en.wikipedia.org/wiki/Machine_learning
4	5	0.0161	https://en.wikipedia.org/wiki/Artificial_intelligence
5	6	0.0161	https://en.wikipedia.org/wiki/Machine_learning
6	7	0.0159	https://en.wikipedia.org/wiki/Machine_learning
7	8	0.0156	https://en.wikipedia.org/wiki/Socialism
8	9	0.0156	https://en.wikipedia.org/wiki/Deep_learning
9	10	0.0154	https://en.wikipedia.org/wiki/Artificial_intelligence

Conclusion

This project demonstrates the effectiveness of hybrid retrieval strategies for retrieval-augmented generation tasks. The integration of dense semantic embeddings with sparse lexical matching improves both accuracy and robustness. Automated evaluation and visualization confirm the superiority of the hybrid approach. Future work can explore larger language models, GPU acceleration, and advanced faithfulness evaluation techniques.