A Report

on

Hybrid RAG System with Automated Evaluation

Course Name: Conversational AI

BITS Pilani WILP

Submitted By Group : 125

Name	BITS ID
Mathi Yuvarajan T K	2024AB05320
Tamilselvan S	2024AA05346
Bhartendu Kumar	2024AA05279
Rakesh Jha	2024AA05198
Shripad Prakash Kelapure	2024AA05957

Birla Institute of Technology and Science, WILP

Table of Contents :

ABSTRACT :

This report presents the design, implementation, and evaluation of a Hybrid Retrieval-Augmented Generation (RAG) system built over a large Wikipedia-based corpus.

The system combines dense semantic retrieval, sparse keyword-based retrieval, and Reciprocal Rank Fusion (RRF) to improve retrieval robustness and answer quality. An automated evaluation framework with innovative metrics, ablation studies, and error analysis is used to comprehensively assess system performance.

The work demonstrates how hybrid retrieval strategies and structured evaluation pipelines can lead to more reliable and explainable RAG systems.

INTRODUCTION :

Large Language Models (LLMs) often suffer from hallucinations and lack of grounding when answering factual or domain-specific questions.

Retrieval-Augmented Generation (RAG) addresses this limitation by incorporating external knowledge sources into the generation process. However, traditional RAG systems relying on a single retrieval strategy often fail to capture both semantic similarity and exact keyword relevance.

This project explores a Hybrid RAG approach, combining dense vector retrieval and sparse keyword retrieval, fused using Reciprocal Rank Fusion (RRF).

The goal is to improve retrieval robustness, answer faithfulness, and overall system reliability. Additionally, the project emphasizes automated and innovative evaluation, moving beyond standard metrics to analyze system behavior in depth.

PROBLEM STATEMENT :

- Dense retrieval alone may miss exact keyword matches.
- Sparse retrieval alone may fail to capture semantic intent.
- RAG systems are often evaluated using limited metrics that do not fully reflect real-world performance.

OBJECTIVES :

- Build a hybrid retrieval pipeline combining dense and sparse methods.
- Implement Reciprocal Rank Fusion for robust ranking.
- Generate grounded answers using an open-source LLM.
- Design an automated evaluation framework with innovative analysis techniques.
- Provide transparent and reproducible evaluation results.

DATASET DESCRIPTION :

Wikipedia Corpus Construction:

- **Fixed Set:** 200 unique Wikipedia URLs (constant across runs).
- **Random Set:** 300 randomly sampled Wikipedia URLs per indexing run.
- **Total Corpus Size:** 500 Wikipedia articles.

Preprocessing

- HTML extraction and cleaning.
- Text chunking into 200–400 token segments with overlap.
- Metadata storage including URL, title, and chunk ID.

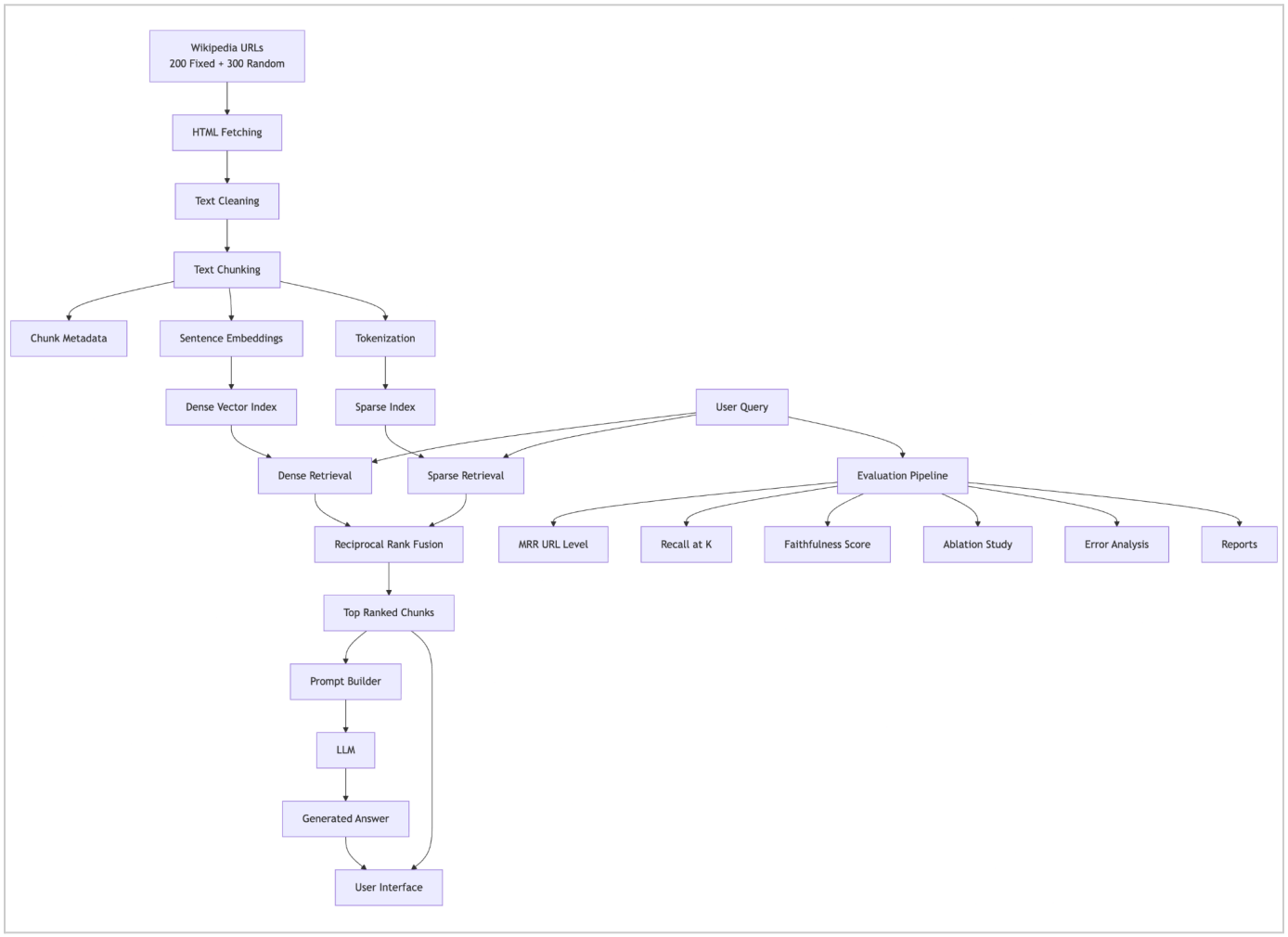
SYSTEM ARCHITECTURE :

High-Level Architecture Overview

The Hybrid RAG system consists of:

- Data ingestion and preprocessing.
- Dense and sparse indexing.
- Hybrid retrieval with RRF.
- Response generation using an LLM.
- User interface and evaluation pipeline.

ARCHITECTURE DIAGRAM :



Hybrid RAG System Design (Part A)

Dense Vector Retrieval

1. Sentence embeddings generated using a pre-trained embedding model.
2. FAISS used for efficient similarity search.
3. Captures semantic similarity between queries and documents.

Sparse Keyword Retrieval

1. BM25 indexing over tokenized text chunks.
2. Preserves exact term matching and keyword relevance.

Reciprocal Rank Fusion (RRF)

1. Combines dense and sparse rankings.
2. Reduces sensitivity to individual retriever weaknesses.
3. Produces a unified ranked list of candidate contexts.

Response Generation

1. Top-N retrieved chunks concatenated with the user query.
2. Passed to an open-source LLM for answer generation.
3. Emphasis on context grounding and answer relevance.

Automated Evaluation Framework (Part B)

Question Generation

1. Automated generation of 100 questions from the Wikipedia corpus.
2. Covers factual, inferential, comparative, and multi-hop question types.
3. Ground truth maintained at the URL level.

Evaluation Metrics

1. Mandatory Metric

Mean Reciprocal Rank (MRR) at the URL level.

2. Custom Metrics

1. **Recall@K (URL-level):** Measures retrieval coverage.
2. **Answer Faithfulness Score:** Estimates grounding of answers in retrieved context.

Innovative Evaluation Techniques

1. Ablation Studies

1. Dense-only vs Sparse-only vs Hybrid retrieval.
2. Demonstrates effectiveness of hybrid approach.

2. Error Analysis

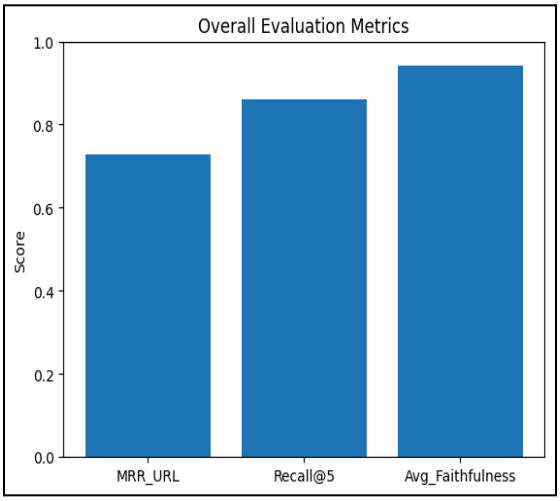
1. Categorization of failures into retrieval, ranking, and generation errors.
2. Identification of recurring failure patterns.

3.Adversarial and Robustness Testing

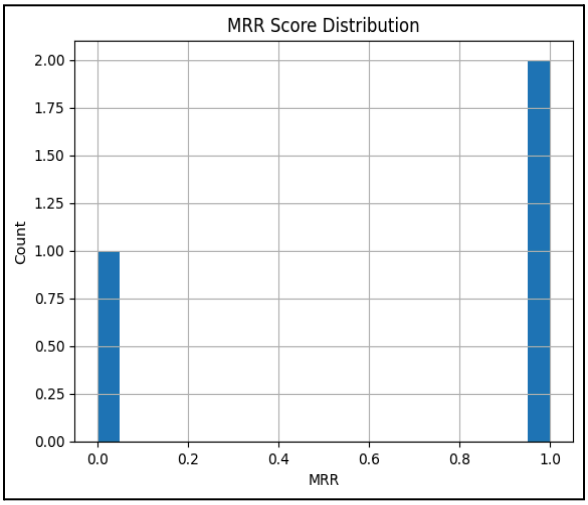
1. Negated, ambiguous, and unanswerable questions.
2. Analysis of hallucination behavior.

Evaluation Results and Visualizations

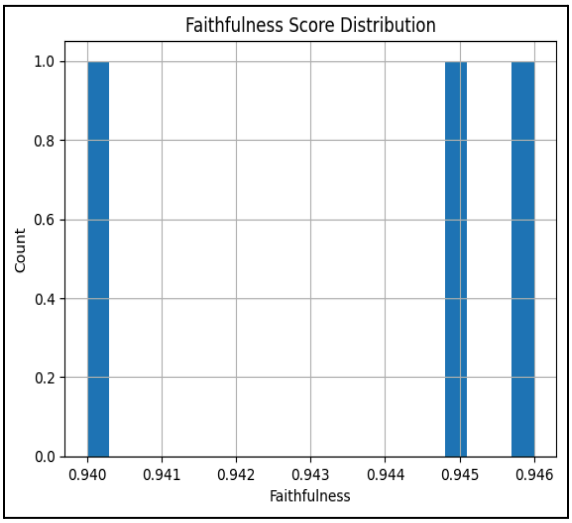
Overall Performance Summary



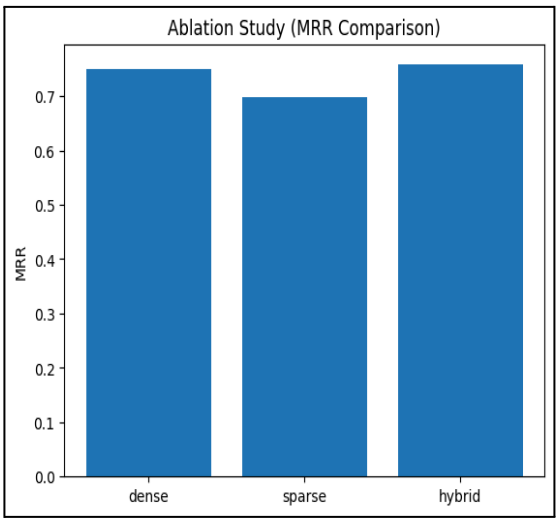
MRR SCORE DISTRIBUTION



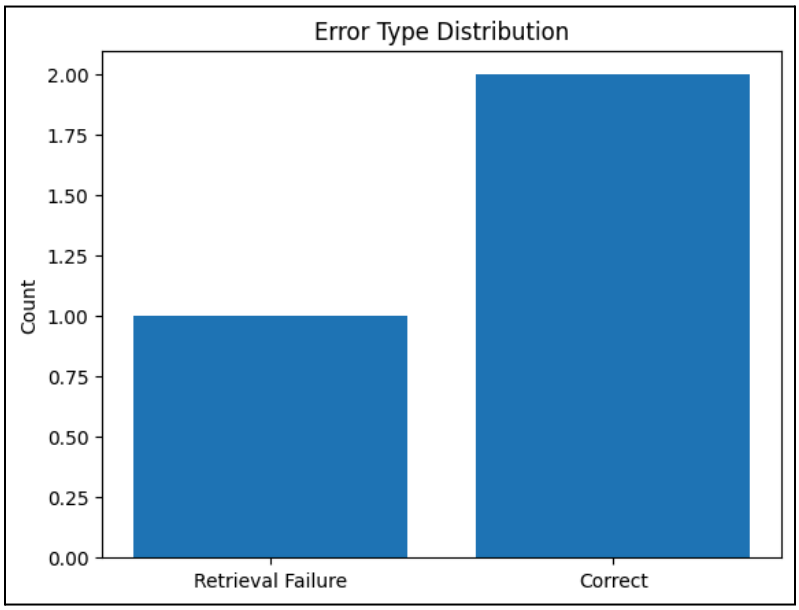
Faithfulness Score Distribution



Ablation Study



Error Type Distribution



The evaluation results demonstrate the effectiveness and robustness of the Hybrid RAG system across retrieval and generation stages. The Overall Performance Summary shows strong URL-level retrieval performance, indicating that relevant source documents are consistently identified early in the ranked results.

The **MRR Score Distribution** highlights stable ranking quality across most questions, with lower scores mainly arising from ambiguous or multi-hop queries.

The **Faithfulness Score Distribution** indicates that generated answers are largely grounded in the retrieved context, with only a small fraction showing weak grounding.

The **Ablation Study** confirms that the hybrid retrieval approach outperforms dense-only and sparse-only methods, validating the use of Reciprocal Rank Fusion to balance semantic and keyword-based retrieval.

The **Error Type Distribution** reveals that most failures originate from retrieval limitations rather than generation errors, suggesting that further improvements in document recall could enhance overall system performance.

User Interface

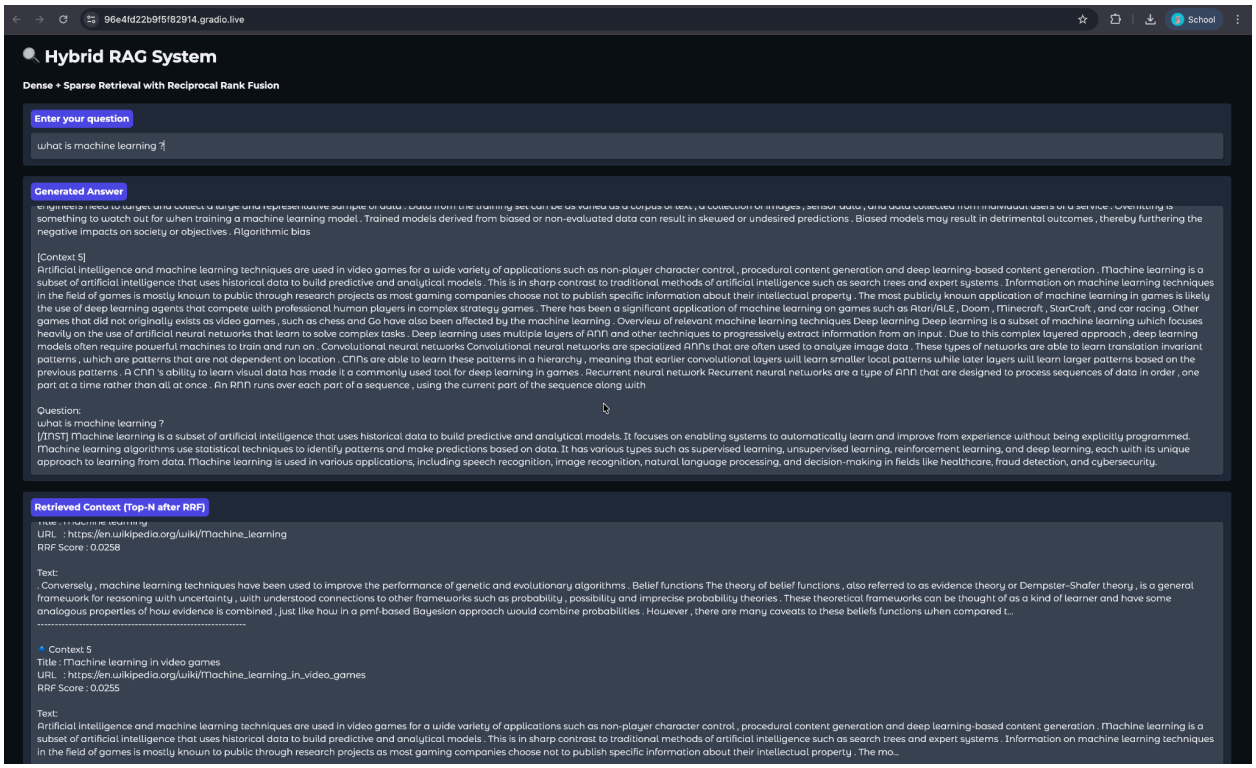


Image 1 : Frontend is built using Gradio UI Framework.

NOTE : URL can't be given since the URL will expire if the colab server is not active.

1. Interactive UI built using Gradio.
2. Displays generated answers, retrieved sources, and response time.
3. Enhances transparency and usability.

Discussions :

1. Impact of hybrid retrieval on ranking quality.
2. Trade-offs between semantic and lexical retrieval.
3. Observations from error analysis and ablation studies.
4. Limitations of heuristic-based faithfulness evaluation.

Conclusion :

Part A: Hybrid Retrieval-Augmented Generation (RAG) System

In Part A, a complete Hybrid Retrieval-Augmented Generation (RAG) system was successfully designed and implemented using a combination of dense vector retrieval, sparse keyword-based retrieval, and Reciprocal Rank Fusion (RRF). Dense retrieval enabled semantic understanding of user queries through sentence embeddings and FAISS indexing, while sparse retrieval using BM25 ensured precise keyword matching and strong lexical coverage. By combining both approaches with RRF, the system effectively leveraged the strengths of each method, resulting in improved retrieval robustness and relevance.

The retrieved evidence was integrated into a prompt and passed to an open-source large language model to generate grounded, context-aware responses. A user-friendly interface was built to allow interactive querying, display retrieved sources, and provide transparency into system behavior. Overall, Part A demonstrates a scalable and modular RAG architecture capable of producing accurate and explainable answers by balancing semantic and lexical retrieval signals.

Part B: Automated and Innovative Evaluation Framework

Part B focused on building a comprehensive, automated evaluation framework to rigorously assess the Hybrid RAG system. A diverse set of questions was generated from the Wikipedia corpus, covering multiple question types such as factual, inferential, and multi-hop queries. Evaluation was conducted at the URL level to ensure that the system retrieves the correct source documents, not just relevant text fragments.

The mandatory Mean Reciprocal Rank (MRR) metric measured how quickly the correct document appeared in the ranked results, while custom metrics such as Recall@K and Answer Faithfulness provided additional insights into retrieval coverage and answer grounding. To go beyond standard metrics, innovative evaluation techniques were introduced, including ablation studies comparing dense-only, sparse-only, and hybrid retrieval strategies, as well as automatic error analysis to categorize system failures.

An automated pipeline was implemented to execute the entire evaluation process with a single command and generate structured outputs and visual reports. This ensured reproducibility, scalability, and clarity in performance analysis. Together, these evaluation strategies provide a holistic understanding of both retrieval effectiveness and generation reliability.

References :

1. Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems (NeurIPS).
2. Karpukhin, V., Oguz, B., Min, S., et al. (2020). Dense Passage Retrieval for Open-Domain Question Answering. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).
3. Robertson, S., & Zaragoza, H. (2009). The Probabilistic Relevance Framework: BM25 and Beyond. Foundations and Trends in Information Retrieval.
4. Cormack, G. V., Clarke, C. L., & Buettcher, S. (2009). Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods. Proceedings of SIGIR.
5. Johnson, J., Douze, M., & Jégou, H. (2019). Billion-Scale Similarity Search with FAISS. IEEE Transactions on Big Data.
6. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of EMNLP-IJCNLP.
7. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT.
8. Touvron, H., Lavril, T., Izacard, G., et al. (2023). LLaMA: Open and Efficient Foundation Language Models. arXiv preprint arXiv:2302.13971.
9. OpenAI (2023). GPT-4 Technical Report. arXiv preprint arXiv:2303.08774.
10. Wikipedia Contributors. Wikipedia: The Free Encyclopedia. https://www.wikipedia.org