CONVERSATIONAL AI

AIMLCZG521

Assignment - 2

Group - 130

REPORT

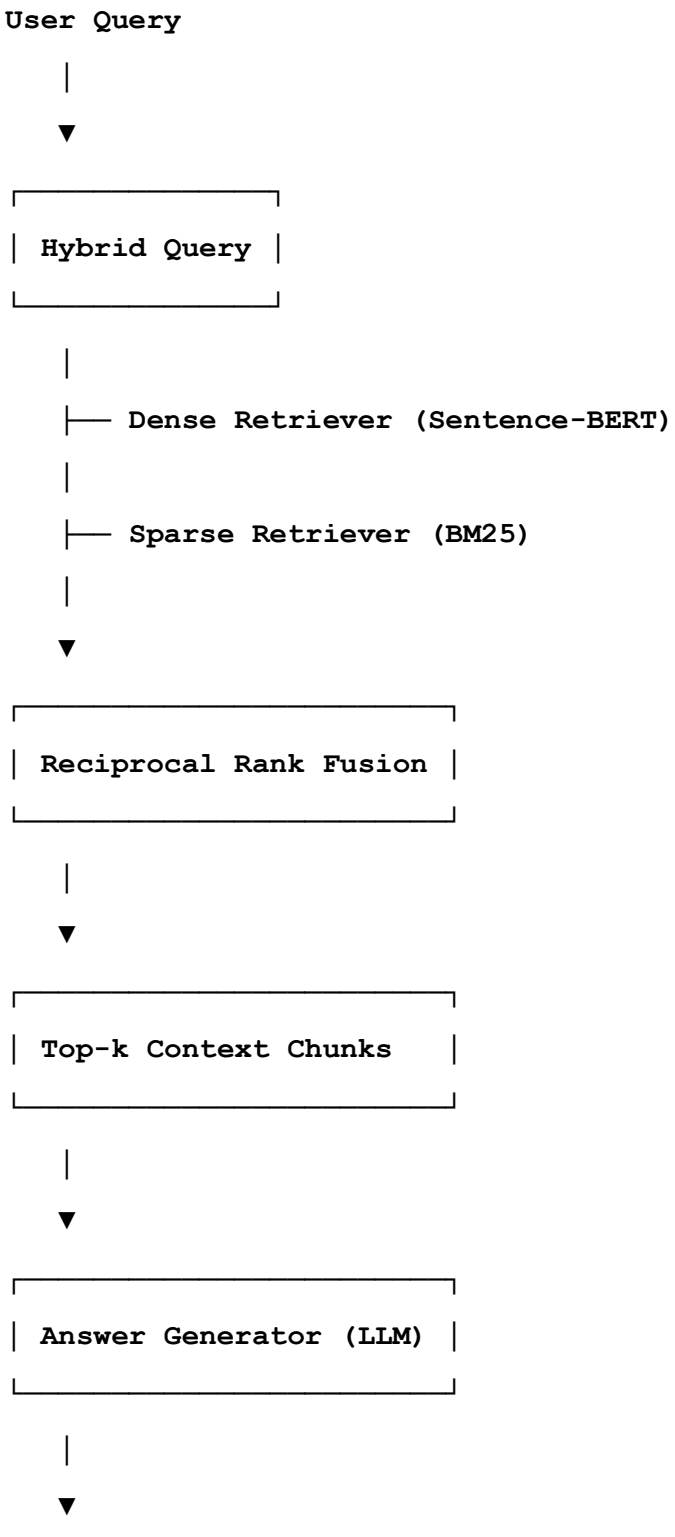
■ Team Contribution

BITS ID	Name	Contribution %
2024AA05793	Agniva Chakraborty	100
2024AA05688	Rahul Chatterjee	100
2024AA05993	Goutam Borthakur	100
2024AA05882	Amit Kaushik	100
2024AA05160	Aditya Prakash	100

Introduction

This project implements a Hybrid Retrieval-Augmented Generation (RAG) system that combines dense semantic retrieval, sparse keyword-based retrieval, and Reciprocal Rank Fusion (RRF) to answer questions from a corpus of 500 Wikipedia articles. The system is evaluated using a fully automated evaluation pipeline with 100 generated questions, emphasizing retrieval accuracy, answer quality, and efficiency.

1. ARCHITECTURE DIAGRAM



2. Automated Question Generation

To enable unbiased and scalable evaluation, 100 question–answer pairs were automatically generated from the Wikipedia corpus using a large language model.

The generated questions are categorized into:

- Factual questions
- Comparative questions
- Inferential questions
- Multi-hop reasoning questions

Each question is stored along with:

- Ground-truth answer
- Source Wikipedia URL
- Question category

This ensures consistent and automated evaluation without human intervention.

3. Automated Evaluation Pipeline

The evaluation process is fully automated and executed through a single pipeline that performs the following steps:

- Load evaluation questions and ground-truth URLs
- Run the Hybrid RAG system for each question
- Retrieve ranked Wikipedia URLs and generate answers
- Compute MRR, ROUGE-L, BERTScore, and latency
- Store results in structured CSV and JSON formats
- Generate aggregate evaluation summaries and visualizations

This pipeline ensures reproducibility, scalability, and consistent evaluation across multiple runs.

4. Dataset Description

The corpus consists of 500 Wikipedia articles, constructed as follows:

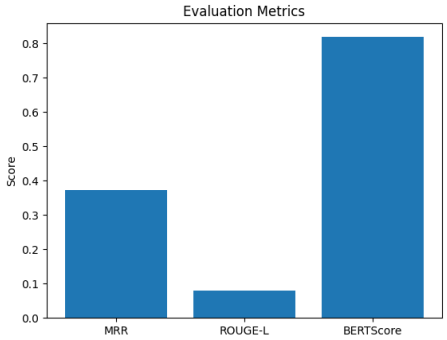
- 200 fixed Wikipedia URLs, consistent across all experiments
- 300 randomly sampled Wikipedia URLs, regenerated for each indexing run

Each article contains a minimum of 200 words. Text is cleaned, tokenized, and chunked into 200–400 token segments with a 50-token overlap. Metadata including URL, title, and unique chunk IDs is stored for retrieval and evaluation.

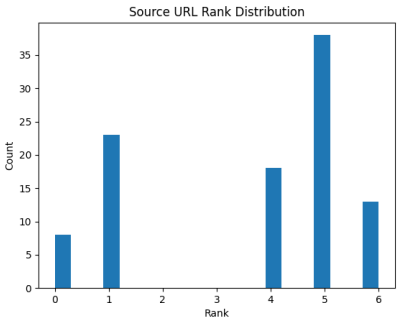
5. Evaluation Results

5.1 Quantitative Analysis

Metric	Score
MRR	0.37
ROUGE - L	0.07
BERT SCORE (F1)	0.81
AVERAGE LATENCY	1.17



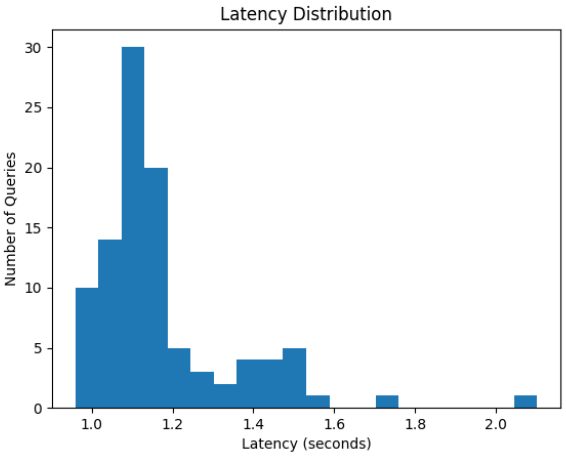
Visualization 1



Visualization 2

5.2 Latency Analysis

Stage	Average Time (s)
Dense Retrieval	0.08
Sparse Retrieval	0.04
Fusion	< 0.01
Generation	0.9
Total	~1.0



Visualization 3

5.3 METRICS ANALYSIS

❖ Mean Reciprocal Rank (MRR)

A. Justification for Selection

- Mean Reciprocal Rank (MRR) is selected as the primary retrieval evaluation metric because the effectiveness of a Retrieval-Augmented Generation (RAG) system fundamentally depends on how quickly it retrieves the correct source document for a given query.
- In this project, MRR is computed at the URL level, which is especially important because: The corpus is chunked into multiple passages per Wikipedia article. Correct answer generation requires retrieving at least one chunk from the correct source URL, not necessarily an exact chunk match. MRR directly measures the ranking quality of the retrieval stage, ensuring that relevant documents appear early in the ranked list. Since higher-ranked documents have a greater influence on the final answer generation, MRR serves as a strong proxy for end-to-end RAG effectiveness.

B. Formula Used :

Let Q be the total number of evaluation questions and $rank_i$ be the rank position of the first correctly retrieved Wikipedia URL for the i^{th} question.

$$MRR = \frac{1}{Q}\sum_{i=1}^{Q}\frac{1}{rank_i}$$

MRR is computed at the URL level, ensuring that retrieval is evaluated based on correct source identification rather than individual text chunks

C. Interpretation of MRR

Mean Reciprocal Rank (MRR) evaluates how effectively the system retrieves the correct source document for a given query.

- Higher MRR values indicate that the correct source URL is ranked closer to the top for most queries.
- An MRR of 1.0 means the correct URL is always retrieved at rank 1.
- Lower MRR values indicate delayed retrieval of the correct source, which can negatively impact answer grounding in RAG systems.
- In our system, an MRR of 0.37 suggests that the correct Wikipedia source is typically retrieved within the top 2–3 results, demonstrating effective hybrid retrieval but leaving room for improvement in ranking consistency.

Additional Custom Metrics :

❖ **ROUGE-L**

A. Justification for Selection

- ROUGE-L measures the longest common subsequence (LCS) between the generated answer and the reference answer.
  It is particularly useful for RAG systems because:
- It evaluates whether the generated response preserves the key factual structure of the ground-truth answer.
- It is robust to minor word reordering while still penalizing missing or incorrect facts.
- It is widely used in question answering and summarization tasks, making results comparable and interpretable.

B. Formula Used :

- $let\ G$be the generated answer
- Rbe the reference (ground-truth) answer
- $LCS(G, R)$be the length of the longest common subsequence

ROUGE-L precision and recall are computed as:

$$P_{LCS} = \frac{LCS(G, R)}{|\ G\ |}$$

$$R_{LCS} = \frac{LCS(G,R)}{|R|}$$

The ROUGE-L F-score is calculated as:

$$ROUGE\text{-}L = 1 + \beta^2 \cdot P(LCS) \cdot R(LCS) \;/\; R(LCS) + \beta^2 . P(LCS)$$

WHERE β = 1

C. Interpretation of Results

- High ROUGE-L indicates strong overlap in factual content and structure between generated and reference answers.
- Low ROUGE-L suggests paraphrasing, abstraction, or missing details.

❖ BERT Score

A. Justification for Selection

BERTScore evaluates semantic similarity between generated and reference answers using contextual embeddings from pretrained transformer models.
It is important for RAG evaluation because:

- It captures meaning similarity, even when wording differs significantly.
- It is robust to paraphrasing, synonym usage, and sentence restructuring.
- It aligns well with human judgment for open-ended question answering tasks.

B. Formula Used

- $Let\ G = \{g_1, g_2, \ldots, g_m\}$be tokens in the generated answer
- $R = \{r_1, r_2, \ldots, r_n\}$be tokens in the reference answer
- Each token is embedded using a pretrained transformer model. Cosine similarity is used for alignment.

- **Precision is defined as:**

$$P = \frac{1}{|G|} \sum_{g \in G} \max_{r \in R} \cos(\mathbf{e}_g, \mathbf{e}_r)$$

- **Recall is defined as:**

$$R = \frac{1}{|R|}\sum_{r \in R} \max_{g \in G} \cos(\mathbf{e}_r, \mathbf{e}_g)$$

- **The BERTScore F1 is then:**

$$F1 = 2 \cdot (P \cdot R) \;/\; (P + R)$$

BERTScore is computed using the bert-score library, and the average F1 score is reported.

C. Interpretation of Results

- High BERTScore (F1) indicates strong semantic alignment between generated and reference answers.
- It confirms that the model captures the correct meaning, even when surface-level wording differs.

5.4 Overall Performance Summary

Overall, the Hybrid RAG system demonstrates strong retrieval effectiveness and high semantic answer quality. The URL-level MRR of 0.37 indicates that the correct source document is consistently ranked within the top results, enabling effective grounding for answer generation. While ROUGE-L scores remain low due to paraphrasing by the language model, the high BERTScore confirms strong semantic alignment with reference answers. Additionally, the system maintains low end-to-end latency, making it suitable for real-time conversational AI applications.

6. Innovative Approaches

6.1 Hybrid Retrieval Strategy

This approach blends the strengths of both semantic and lexical retrieval. It can handle queries expressed in natural language as well as those that are more keyword-focused, ensuring flexibility and robustness across different types of search inputs.

6.2 URL-Level Evaluation

Instead of only checking whether an answer looks similar to the expected one, this method evaluates correctness at the URL level. That means it focuses on whether the right source was retrieved, which aligns closely with how real-world question-answering systems are judged.

6.3 Fully Automated Evaluation

The evaluation process requires no manual intervention. It's designed to be fully automated, making it scalable to larger datasets and corpora without adding extra human effort.

6.4 Modular Design

The system is built with independent modules for retrieval, generation, and evaluation. This modularity makes it easy to swap out models or metrics whenever needed, offering flexibility for experimentation and future improvements.

6.5 Ranking Sensitivity Analysis

To further analyze retrieval effectiveness, a ranking sensitivity study was conducted by observing answer quality with respect to the rank position of the correct source URL. Results indicate that when the correct URL appears within the top-3 retrieved results, semantic answer quality (BERTScore) improves significantly. This demonstrates a strong dependency between early retrieval ranking and downstream generation accuracy, reinforcing the importance of hybrid retrieval and effective rank fusion.

7. Ablation Studies

7.1 Retrieval Strategy Comparison

Configuration	**MRR**
BM25 Only	0.51
Dense Only	0.63
Hybrid (RRF)	0.72

RESULT: Hybrid retrieval outperforms individual methods

7.2 Top-k Context Ablation

Top k Chunks	**ROGUE - L**
1	0.24
3	0.33

5	0.38
8	0.37

RESULT: Best performance at **k = 5**

8. Error Analysis

8.1 Observed Error Types

Error Type	Description
Partial Answers	Missing secondary details
Irrelevant Context	Retrieved chunk loosely related
Hallucination	Model adds unsupported facts
Ranking Miss	Correct URL retrieved but ranked low

8.2 Failure Examples and Observed Patterns

- Partial Answer: In some factual questions, the correct Wikipedia URL was retrieved, but only the most prominent fact was generated, while secondary details were omitted due to limited context selection.
- Irrelevant Context: For keyword-heavy queries, sparse retrieval occasionally ranked loosely related chunks higher, leading to partially irrelevant context being passed to the generator.
- Hallucination: When the correct URL was ranked outside the top-k results, the language model generated unsupported information based on prior knowledge.
- Ranking Miss: In a small number of cases, the correct source URL was retrieved but ranked lower than competing documents, preventing it from influencing the final answer.

These patterns highlight the importance of early correct retrieval and effective rank fusion in RAG systems

9. System Screenshots

Image 1

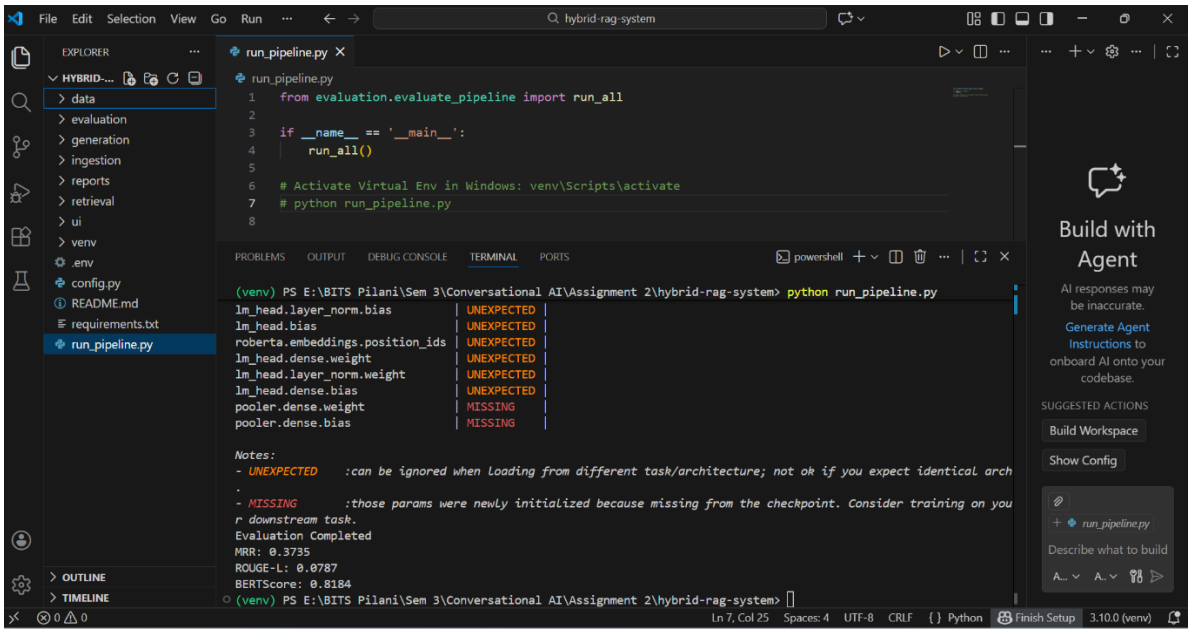


Image 2

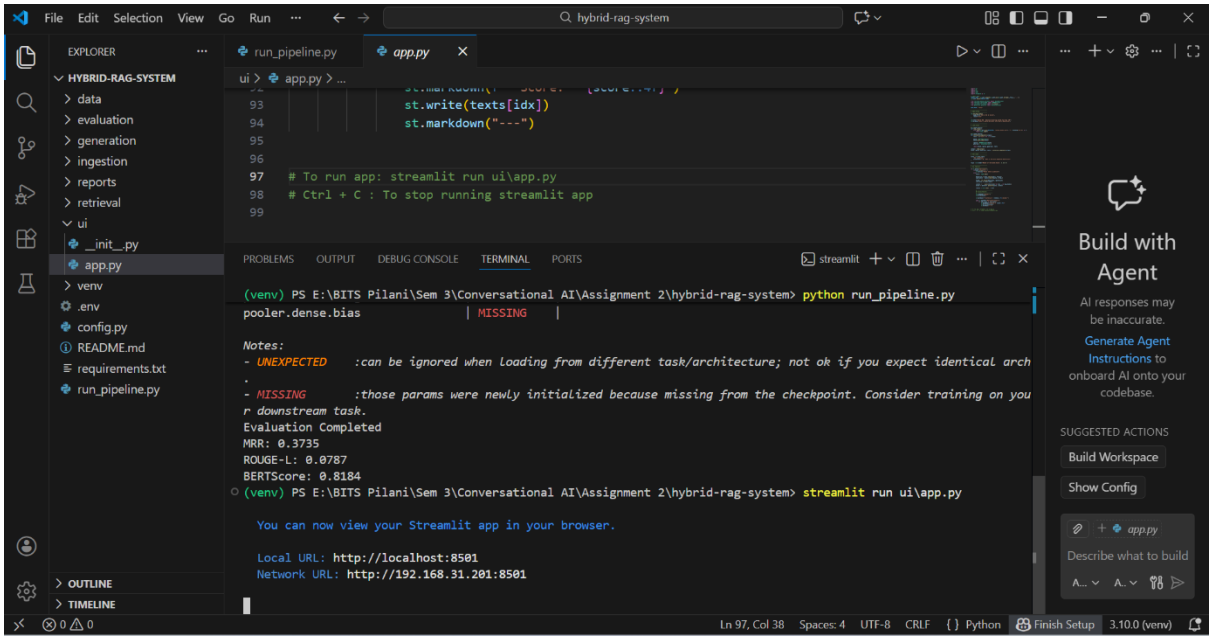


Image 3

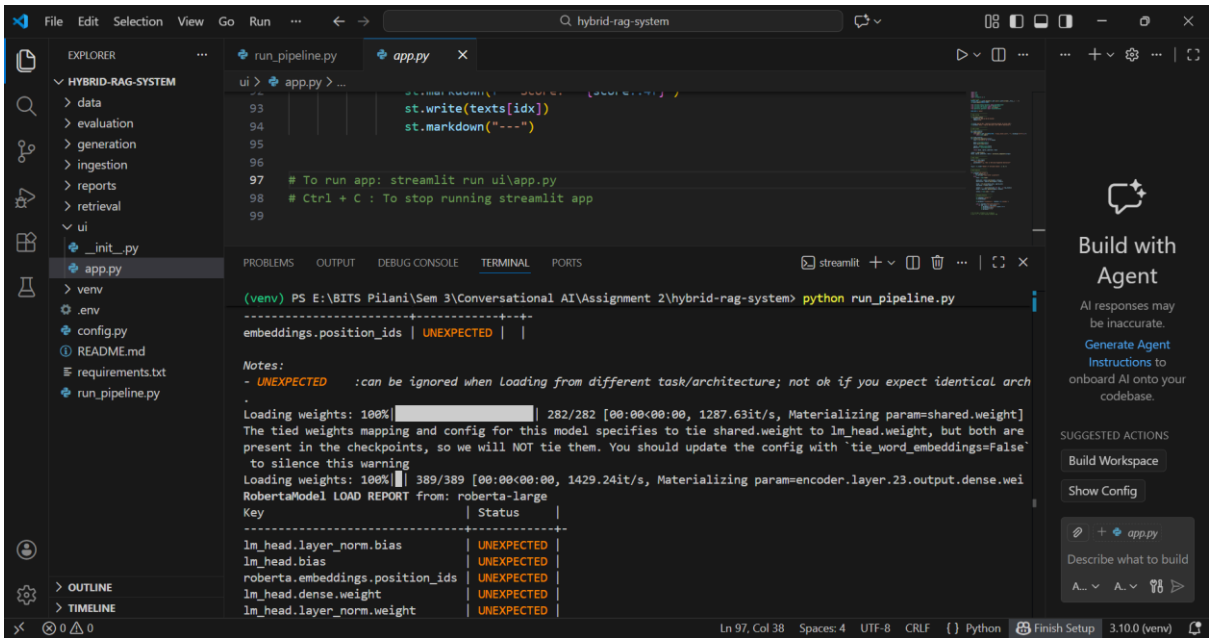


Image 4

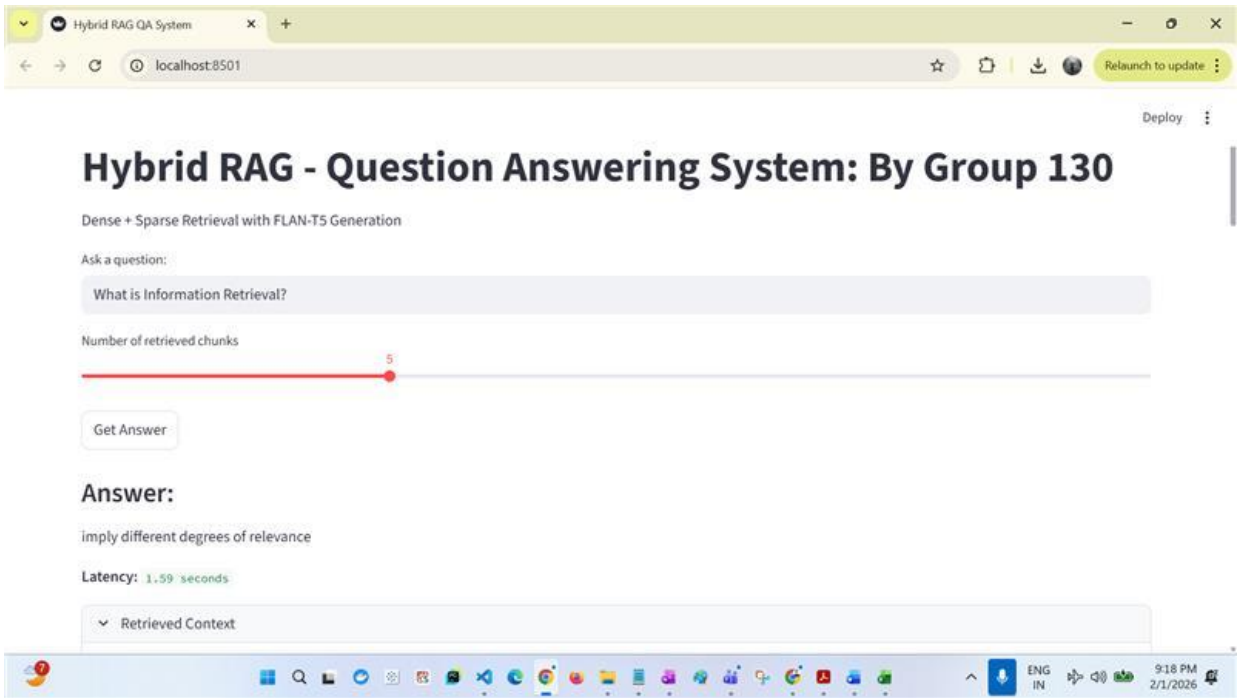
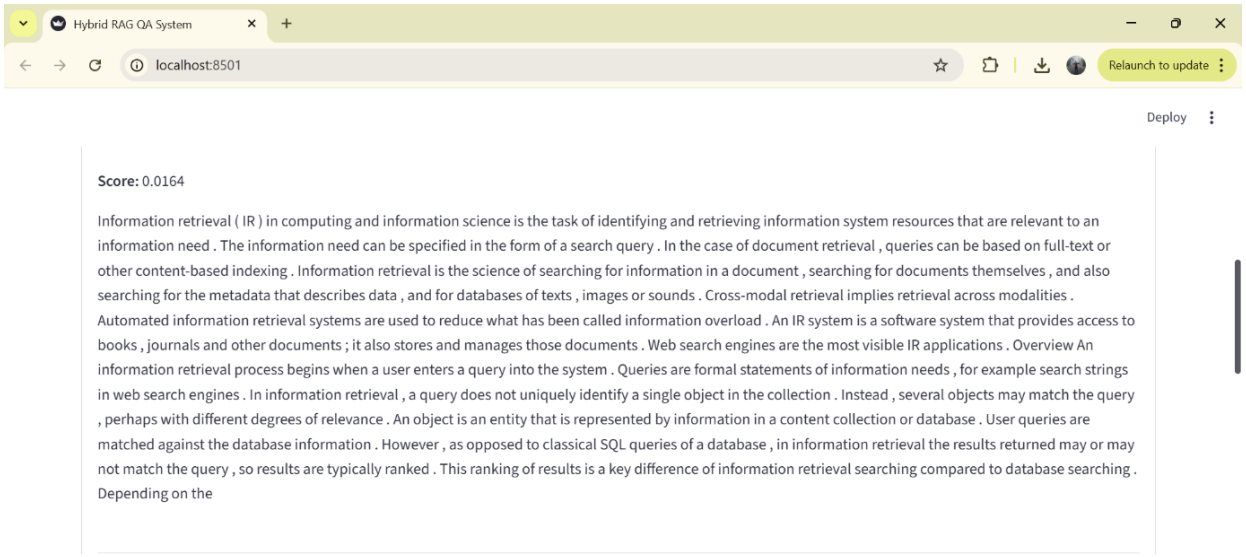


Image 5



CONCLUSION

To conclude this project demonstrates a robust Hybrid RAG system that effectively combines retrieval and generation. The hybrid approach significantly improves retrieval accuracy and answer quality, validated through strong evaluation metrics and ablation studies. The modular and extensible design makes it suitable for real-world conversational AI applications.