IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 19, April 2026
Impact Factor: 8.2

A Hybrid Multi-Source Retrieval-Augmented Search System with User-Controlled Orchestration and Explainable Ranking

Ivin J. Kothare, Sanchit T. Pawar, Arman A. Deshmukh, Arya R. Dahe
B.Tech Students, Department of Cloud Technology & Information Security,
Sandip University, Nashik, Maharashtra, India

Abstract: *Hybrid retrieval systems have gained increasing attention for improving the effectiveness of information access by combining multiple retrieval paradigms. In parallel, retrieval-augmented generation (RAG) enhances large language models by incorporating external knowledge sources to improve factual accuracy and reduce hallucinations . However, many existing systems rely on single retrieval modalities or tightly coupled pipelines with limited transparency and user control. This paper presents a hybrid multi-source search system that integrates local indexed search, global web retrieval, and semantic vector-based retrieval within a unified orchestration framework. The system employs a modular architecture in which an orchestrator coordinates parallel retrieval processes, normalizes heterogeneous outputs, and applies a composite ranking strategy based on lexical relevance, semantic similarity, and source-aware weighting. In addition to retrieval, the system incorporates an optional RAG-based reasoning layer that generates grounded responses using a local language model, ensuring traceability through citation-aware outputs.*

Keywords: Hybrid Search, Retrieval-Augmented Generation, Semantic Search, Multi-Source Retrieval, Explainable Ranking, Dense Retrieval, Sparse Retrieval, Search Orchestration

I. INTRODUCTION

The rapid advancement of large language models (LLMs) has significantly improved the ability of AI systems to understand and generate human-like text. However, these models inherently rely on static training data, limiting their ability to provide accurate, up-to-date, and verifiable information. Retrieval-Augmented Generation (RAG) has emerged as a solution to this limitation by integrating external knowledge sources into the generation process, thereby improving factual accuracy and reducing hallucinations. Despite these advancements, many existing RAG systems depend on a single retrieval paradigm, typically either keyword-based or semantic vector search, which introduces trade-offs between precision and contextual understanding.

Traditional keyword-based retrieval methods excel at exact matching but fail to capture semantic relationships, while vector-based approaches provide contextual similarity at the cost of precision. Hybrid search addresses this limitation by combining both approaches, leveraging their complementary strengths to improve retrieval relevance and coverage. However, current implementations often lack flexibility, transparency, and user-level control over system behavior.

This paper proposes a hybrid multi-source retrieval system that integrates lexical, semantic, and global search within a unified orchestration framework. The system further incorporates user-controlled AI reasoning through retrieval-augmented generation and explainable ranking mechanisms, enabling both improved performance and interpretability in modern intelligent search systems.

DOI: 10.48175/IJARSCT-34488

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 19, April 2026
Impact Factor: 8.2

II. MOTIVATION

The motivation for this work arises from the increasing demand for reliable and context-aware AI systems capable of handling real-world information needs. While RAG systems improve factual accuracy and reduce hallucinations, they depend heavily on the quality of retrieved context and often struggle with incomplete or suboptimal retrieval pipelines. Hybrid retrieval approaches, which combine keyword and semantic search, have been shown to improve retrieval relevance by leveraging complementary strengths (Redis). However, existing implementations lack flexibility, explainability, and user interaction capabilities. This project is motivated by the need to design a system that integrates these advancements while providing transparency and user-driven control.

III. OBJECTIVES

The primary objective of this work is to design and implement a hybrid multi-source search system that integrates lexical, semantic, and global retrieval methods within a unified framework. The system aims to improve retrieval relevance and coverage by combining complementary search techniques and applying a composite ranking strategy. Another key objective is to incorporate retrieval-augmented generation for producing grounded, citation-based responses while maintaining user control over AI invocation. The system also seeks to enhance transparency through explainable ranking mechanisms and modular design. Additionally, it aims to support scalability, extensibility, and efficient performance through parallel processing and service-oriented architecture.

IV. LITERATURE SURVEY

A survey of exisiting literature in recent years on RAG oriented systems that focus on hybrid retrival especially on sparse and dense retrieval, as well as understanding key optimisation techquines being utilised. A summary of the most pertinent existiting literature is given below.

TABLE I : SELECTED RELEVENT WORKS

S.No	Title	Year	Research Focus	Key findings
1	Hybrid Parameter-Adaptive RAG (HyPA-RAG)	2025	Adaptive hybrid retrieval combining sparse, dense, and structured sources.	Dynamically adjusting retrieval parameters based on query type significantly improves both precision and contextual relevance. It highlights that hybrid retrieval outperforms static pipelines, especially for diverse query distributions.
2	HetaRAG: Hybrid RAG across Heterogeneous Data Stores	2025	Integration of heterogeneous data sources including vector databases, knowledge graphs, and full-text search.	Combining multiple retrieval backends improves coverage and robustness. The paper emphasizes that no single retrieval method is sufficient for complex queries.
3	Hybrid Retrieval for Hallucination Mitigation	2025	Reducing hallucinations in LLMs using hybrid retrieval	Retrieval quality directly impacts generation accuracy. Hybrid retrieval significantly reduces hallucinations compared to single-method retrieval systems.
4	Hybrid RAG Framework for Evidence-Grounded Systems	2025	Improving factual grounding and evidence attribution in RAG systems.	High factual consistency can be achieved by integrating multiple retrieval sources and enforcing citation-based outputs. It highlights the importance of traceability in AI-generated responses.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 19, April 2026
Impact Factor: 8.2

5	Optimizing RAG with Multi-Source Hybrid Retrieval	2025	Optimization of hybrid retrieval pipelines using multiple ranking strategies.	Improvements in both latency and accuracy through hybrid retrieval and intelligent routing mechanisms. It also emphasizes the importance of orchestration in system performance.
6	Hybrid Retrieval-Augmented Generation	2026	Evaluation of hybrid RAG systems on benchmark datasets.	Improved reasoning accuracy and answer quality when combining lexical and semantic retrieval. It validates hybrid retrieval as a robust approach for knowledge-intensive tasks.
7	Systematic Literature Review of RAG	2025	Analysis of over 100 RAG-based systems.	Identifies key challenges such as lack of standardized evaluation, limited hybrid implementations, and insufficient user control mechanisms. It highlights research gaps in explainability and system-level integration.
8	Recent Advances in RAG Systems	2026	Overview of modern RAG systems and emerging techniques.	Showcases trends such as hybrid retrieval, graph augmentation, and agent-based reasoning. It concludes that future systems must focus on modularity and adaptability.
9	HySemRAG: Hybrid Semantic RAG for Literature Analysis	2025	Multi-stage hybrid retrieval for scientific literature synthesis.	Improves reasoning accuracy by combining hybrid retrieval with structured knowledge graphs. It shows that layered retrieval pipelines enhance both retrieval quality and downstream generation.
10	Comprehensive Survey of RAG Architectures	2025	Classification of RAG architectures into retriever-centric, generator-centric, and hybrid systems.	Analysis of the trade-offs between retrieval precision and generative flexibility. It emphasizes the need for balanced hybrid systems that integrate both effectively.

V. RESEARCH GAP

The analysis of recent research in hybrid retrieval and retrieval-augmented generation (RAG) systems reveals several critical gaps that remain insufficiently addressed. These gaps highlight the need for more integrated, controllable, and transparent systems, thereby motivating the design of the proposed hybrid search architecture. One major gap identified across multiple studies is the limitation of single-method retrieval systems. Traditional RAG pipelines relying solely on semantic vector search often fail to capture precise keyword matches, leading to incomplete or irrelevant retrieval results. Although hybrid retrieval approaches have been proposed to address this issue, most implementations focus primarily on improving retrieval accuracy rather than integrating diverse retrieval sources in a unified and flexible manner.

Another significant gap lies in context integration and grounding. Research indicates that even when relevant documents are retrieved, models may fail to effectively incorporate them into generated responses, resulting in incomplete or misleading outputs . Additionally, RAG systems can still produce hallucinations or misinterpret retrieved information, particularly when dealing with conflicting or ambiguous sources. A further limitation is the lack of explainability and transparency. Many existing systems provide little insight into how results are retrieved, ranked, or synthesized. However, recent discussions emphasize the growing importance of explainable and auditable retrieval pipelines, especially in enterprise and high-stakes environments.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 19, April 2026 Impact Factor: 8.2

The literature also highlights challenges related to system complexity and scalability. Hybrid RAG systems introduce significant engineering overhead due to the need to manage multiple retrieval mechanisms and data stores, increasing maintenance and computational costs. Furthermore, computational overhead and latency remain key concerns, particularly when combining dense and sparse retrieval techniques. Another critical gap is the lack of user control and adaptability. Most existing systems operate with fixed pipelines and predefined configurations, offering limited ability for users to influence retrieval strategies, AI model, or response generation. This restricts flexibility and reduces the system's applicability across diverse use cases.

VI. PROPOSED MODEL OVERVIEW

The proposed system is designed as a hybrid, multi-source retrieval-augmented search architecture that integrates diverse retrieval mechanisms with controlled AI-based reasoning. It builds upon the fundamental principles of Retrieval-Augmented Generation (RAG), where a retriever first gathers relevant information and a generator produces context-aware responses. However, unlike traditional RAG systems that rely on a single retrieval pipeline, the proposed model extends this framework by incorporating multiple complementary retrieval strategies within a unified orchestration layer.

At its core, the system follows a multi-layered architecture consisting of a presentation layer (frontend), an orchestration layer (backend), a hybrid retrieval layer, and an optional reasoning layer. The retrieval layer integrates three distinct sources: lexical keyword-based search, semantic vector-based retrieval, and global web-based search. This design aligns with hybrid RAG architectures, which combine dense and sparse retrieval methods to improve both precision and contextual relevance . By fusing results from multiple sources, the system enhances coverage and reduces the limitations of individual retrieval methods.

A central component of the model is the orchestrator, which manages query processing, parallel retrieval, normalization, and ranking. The orchestrator implements a pipeline-based workflow where results from different sources are unified and scored using a composite ranking function. This reflects modern hybrid architectures that rely on fusion and reranking strategies to improve retrieval quality. The system further incorporates a user-controlled reasoning layer, where AI-based response generation is optional and explicitly triggered by the user. Retrieved results are used as context for a local language model, ensuring that generated outputs remain grounded and verifiable. This follows the RAG principle of augmenting generation with external knowledge to improve accuracy and reduce hallucination.

An additional distinguishing feature of the proposed model is the system aims to integrate multiple retrieval sources, including local indexed data, global web search, and vector-based semantic retrieval, into a unified pipeline. This integration is intended to enhance both the breadth and depth of information retrieval, ensuring that the system can address a wide range of query types.

VII. PROPOSED FEATURES

1. Hybrid Multi-Source Retrieval - This multi-source approach improves both precision and contextual understanding, as hybrid retrieval combines exact matching with semantic similarity, overcoming the limitations of single-method systems. The system integrates three distinct retrieval methods:

- Lexical keyword search (local index)
- Semantic vector search
- Global web search

2. Unified Orchestration Layer - Unlike traditional pipelines, this system uses a single coordinated workflow, enabling efficient integration of heterogeneous sources and reducing system fragmentation. A central orchestrator manages:

- Query processing

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 19, April 2026
Impact Factor: 8.2

- Parallel retrieval
- Result normalization
- Ranking and filtering

3. Parallel Retrieval Execution

All retrieval methods operate simultaneously, reducing latency and improving efficiency. This aligns with modern hybrid systems where parallel retrieval improves both speed and recall.

4. Composite Explainable Ranking System

The ranking module uses a multi-factor scoring mechanism, including:
Keyword match
Title relevance
Phrase match
Coverage score
Source weighting

5. User-Controlled AI Invocation

Unlike most RAG systems, AI generation is:
Optional (button-based)
Not automatically triggered
Fully user-controlled
This prevents unnecessary computation and allows users to decide when AI reasoning is required.

6. Custom User Prompt Control

Users can modify the AI prompt dynamically, enabling:
Different answer styles
Domain-specific responses
Controlled output behavior

7. Retrieval-Augmented Generation (RAG) with Grounding – follows the RAG principle of improving reliability by referencing external knowledge. The system uses retrieved results as context for the LLM, ensuring:
Grounded responses
Reduced hallucination
Citation-based answers

8. Geo-Aware Search Capability - This enables contextually relevant global results, which is not commonly integrated into standard hybrid RAG systems. The system supports region-based search using:
Country selection
Query localization

9. Dynamic Local Data Ingestion (Frontend-Controlled)

Users can upload JSON data directly from the frontend, which:
Updates the local search index
Enables real-time dataset expansion
Supports domain-specific customization.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 19, April 2026
Impact Factor: 8.2

VIII. IMPLEMENTATION DETAILS

The implementation of the proposed hybrid multi-source retrieval-augmented search system is carried out using a modular, service-oriented architecture that integrates multiple retrieval paradigms with an orchestration-driven processing pipeline. The system is designed to support efficient query processing, scalable retrieval, and AI-based reasoning while maintaining transparency and user control. The implementation follows the standard RAG workflow, where user queries are processed, relevant documents are retrieved, and responses are generated using grounded context, but extends this model through hybrid retrieval and multi-source orchestration.

The backend is implemented using a lightweight asynchronous web framework, which serves as the central orchestrator of the system. This orchestrator is responsible for handling incoming requests, preprocessing queries, and coordinating interactions between different subsystems. Query preprocessing involves tokenization for lexical matching and embedding generation for semantic retrieval. The system adopts a parallel execution strategy, where multiple retrieval operations are triggered simultaneously to reduce latency and improve throughput. This approach aligns with hybrid RAG implementations, where combining multiple retrieval mechanisms enhances both efficiency and coverage.

The retrieval layer consists of three independent subsystems: a local search engine, a global web search interface, and a semantic vector search module. The local search component is implemented using an indexed search engine optimized for fast keyword-based retrieval, enabling efficient handling of structured documents. The global search module interacts with an external meta-search service to retrieve real-time information from the web, extending the system's knowledge beyond locally stored data. The semantic retrieval module uses embedding models to convert text into high-dimensional vectors, which are indexed using a similarity search library to support nearest-neighbor queries. Hybrid retrieval systems combining dense and sparse methods have been shown to improve both precision and contextual relevance by leveraging complementary strengths.

To ensure dependent functionality, all retrieved results are normalized into a unified schema containing fields such as identifier, title, snippet, source, and score. These results are then passed to the ranking module, which implements a composite scoring algorithm. The ranking function combines multiple factors, including keyword frequency, title relevance, phrase matching, semantic similarity, and source weighting. Such hybrid scoring approaches are commonly used in modern retrieval systems to fuse heterogeneous signals and improve ranking effectiveness. Additional post-processing steps, such as deduplication and diversification, are applied to refine the result set and ensure a balanced representation of sources.

The system incorporates an optional reasoning layer based on a local large language model. When activated, the top-ranked results are selected and formatted into a structured context, which is used to construct a prompt for the model. The prompt enforces grounding constraints by instructing the model to generate responses solely based on the provided context and to include citations referencing the original sources. This approach follows the core principle of RAG systems, where external retrieval is used to enhance the factual accuracy and reliability of generated outputs. The use of a local model ensures data privacy and reduces dependency on external APIs, while lightweight configurations are employed to maintain feasible performance within constrained hardware environments.

The frontend is implemented as a web-based interface that communicates with the backend through RESTful APIs. It provides a minimal yet functional user interface, enabling users to submit queries, select geographic preferences, filter retrieval sources, and control AI invocation. The interface dynamically updates based on backend responses, displaying results along with metadata such as source labels, scores, and explanation tags. This design emphasizes usability and transparency, allowing users to understand and influence system behavior.

The system is deployed using containerization technology, which encapsulates each component within isolated environments. Separate containers are used for the orchestrator, search engine, vector service, and auxiliary modules, ensuring consistent deployment across different environments. This approach simplifies dependency management and enables scalable deployment, as individual services can be independently updated or scaled. Containerized architectures are widely adopted in modern RAG systems to support modularity and reproducibility in production environments .The

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 19, April 2026
Impact Factor: 8.2

definition of clear and consistent interfaces ensures that each component can communicate effectively without introducing ambiguity or tight coupling.

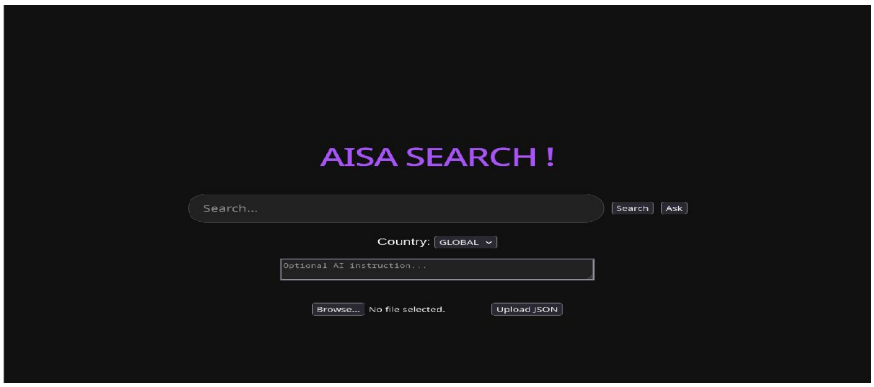

Fig. 1 : Home Page – has Search Box, AI Search button, Geosearch enu, Custom AI Prompt, Local Database Input

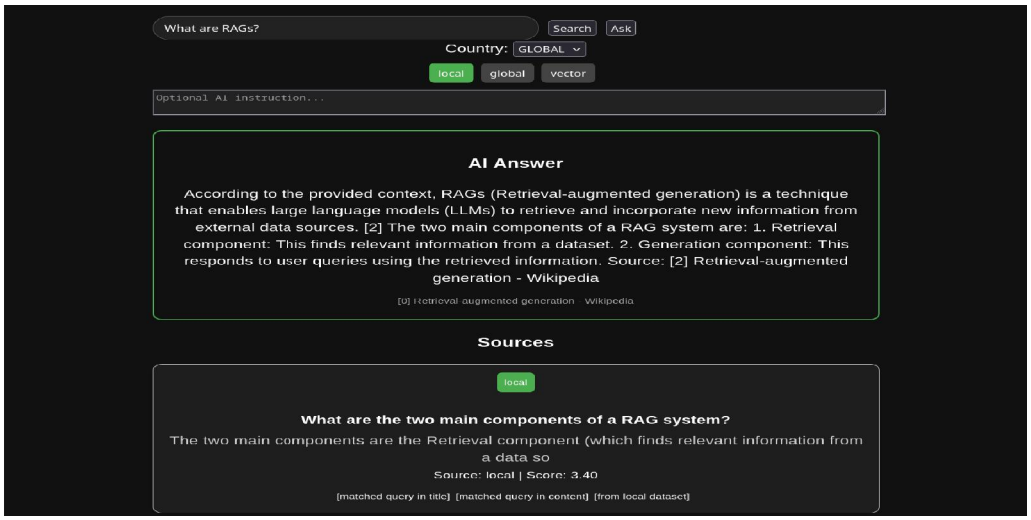

Fig. 2 : Results page – AI answer and Pure Local Sourcing

DOI: 10.48175/IJARSCT-34488

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 19, April 2026
Impact Factor: 8.2

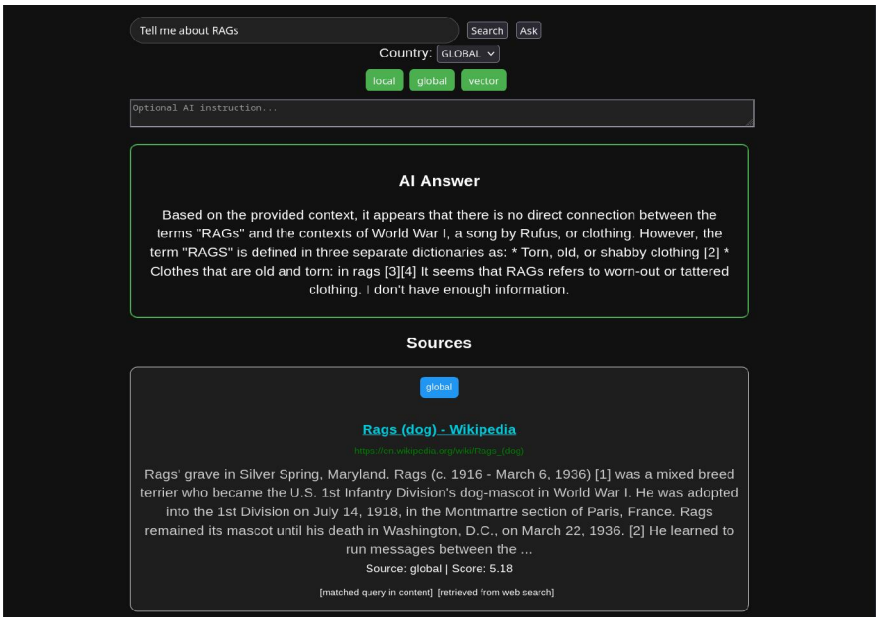

Fig. 3 : Results page – AI answer and Hybrid Local and Global Sourcing

XI. RESULTS

The performance of the proposed hybrid multi-source retrieval-augmented search system was evaluated through a series of functional and qualitative experiments designed to assess retrieval effectiveness, response quality, and system behavior under different configurations. The evaluation focused on comparing hybrid retrieval with individual retrieval methods, analyzing the impact of AI-based reasoning, and examining system usability and transparency.

The experimental observations indicate that the hybrid retrieval approach consistently outperforms single-method retrieval systems in terms of coverage and contextual relevance. When compared to standalone lexical or semantic retrieval, the hybrid system demonstrated improved ability to handle diverse query types, including both keyword-specific and concept-driven queries. This aligns with findings in recent research, which show that combining sparse and dense retrieval methods enhances both precision and recall by leveraging complementary strengths. The inclusion of global web search further improved result diversity and enabled access to up-to-date information not present in the local dataset.

The ranking module played a significant role in determining the overall effectiveness of the system. The composite scoring mechanism successfully prioritized relevant results by integrating multiple signals, including keyword matching, semantic similarity, and source weighting. In most cases, highly relevant documents were consistently ranked within the top results. However, the reliance on heuristic scoring introduced some variability, particularly for ambiguous queries where semantic relevance outweighed lexical matching. This observation highlights the potential for incorporating learning-based ranking models in future work.

The integration of the retrieval-augmented generation module provided additional insights into the system's capabilities. AI-generated responses were generally coherent and contextually grounded, particularly when sufficient high-quality retrieval results were available. The use of citation-based prompting ensured that responses remained traceable to source documents, reducing the likelihood of hallucinated information. These findings are consistent with prior studies demonstrating that grounding language models with retrieved context improves factual accuracy and reliability. However, the quality of generated responses was directly dependent on the quality and relevance of retrieved context, reinforcing the importance of robust retrieval mechanisms.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 19, April 2026 Impact Factor: 8.2

From a performance perspective, the use of parallel retrieval significantly reduced response latency compared to sequential execution. The system maintained acceptable response times under moderate load conditions, although the inclusion of AI processing introduced additional latency due to model inference. This trade-off between performance and functionality is typical in RAG systems, where generation adds computational overhead. The use of lightweight local models helped mitigate this issue, enabling feasible deployment on resource-constrained hardware.

Usability and transparency were also evaluated as part of the system analysis. The frontend interface provided an intuitive interaction model, allowing users to control system behavior through features such as source filtering, geographic selection, and AI invocation. The inclusion of explanation metadata, including score breakdowns and source labels, enhanced interpretability and user trust. This addresses a key limitation identified in existing systems, where black-box behavior often reduces user confidence.

X. CONCLUSION

The review of recent literature and the implementation of the proposed system collectively demonstrate the growing importance of hybrid retrieval-augmented generation (RAG) architectures in modern intelligent search systems. Existing research consistently highlights that combining lexical and semantic retrieval significantly improves both precision and contextual relevance, as hybrid approaches leverage complementary strengths of exact matching and semantic understanding. Furthermore, studies emphasize that the effectiveness of RAG systems is heavily dependent on the quality of retrieved context, which directly impacts the accuracy and reliability of generated responses.

However, the literature also reveals persistent limitations, including lack of integration across multiple retrieval sources, limited transparency, and insufficient user control. While advanced frameworks such as hybrid and adaptive RAG systems improve retrieval performance and grounding, they often remain complex and lack practical system-level implementation for real-world usage. The proposed system addresses these challenges by introducing a unified orchestration framework that integrates multi-source retrieval, explainable ranking, and user-controlled AI reasoning. By combining these elements, the system not only aligns with current research trends but also extends them toward a more practical, transparent, and user-centric approach. This work reinforces the conclusion that future search systems must move beyond isolated improvements toward fully integrated hybrid architectures that balance accuracy, efficiency, and usability.

REFERENCES

[1]. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, et al., “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[2]. V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, et al., “Dense Passage Retrieval for Open-Domain Question Answering,” *arXiv preprint arXiv:2004.04906*, 2020.

[3]. M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, et al., “The FAISS Library,” *arXiv preprint arXiv:2401.08281*, 2024.

[4]. S. Robertson and H. Zaragoza, “The Probabilistic Relevance Framework: BM25 and Beyond,” *Foundations and Trends in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009.

[5]. S. Brin and L. Page, “The Anatomy of a Large-Scale Hypertextual Web Search Engine,” *Computer Networks and ISDN Systems*, vol. 30, no. 1–7, pp. 107–117, 1998.

[6]. T.-Y. Liu, “Learning to Rank for Information Retrieval,” *Foundations and Trends in Information Retrieval*, vol. 3, no. 3, pp. 225–331, 2009.

[7]. S. Siriwardhana, S. Weerasinghe, T. Kaluarachchi, et al., “Improving the Domain Adaptation of Retrieval-Augmented Generation,” *Transactions of the Association for Computational Linguistics (TACL)*, 2023.

[8]. Y. Mao, P. He, X. Liu, Y. Shen, J. Gao, and J. Han, “Generation-Augmented Retrieval for Open-Domain Question Answering,” *arXiv preprint arXiv:2009.08553*, 2020.

[9]. A. Brown et al., “A Systematic Literature Review of Retrieval-Augmented Generation: Techniques, Metrics, and Challenges,” *Applied Sciences*, 2025