

<Conference abbreviation>

<Conference Series name>
Number>

<Volume number and Year> <DOI

Research Compass: An LLM-Orchestrated Multi-Source Framework for Intelligent Scholarly Paper Retrieval

Anay Joshi¹, Anurag Mahalpure², Abhishek Marwade³

*Department of AI&DS, AISSMS Institute of Information Technology, Pune, India,
joshianay75@gmail.com*

*Department of AI&DS, AISSMS Institute of Information Technology, Pune, India,
anuragmahalpure@gmail.com*

*Department of AI&DS, AISSMS Institute of Information Technology, Pune, India,
abhishekmarwade2109@gmail.com*

Abstract.

Abstract. This has been due to the exponential increase of academic publications that has made it more challenging to discover relevant research easily. Search systems that retrieve based on keywords do not represent semantic intent and thus the search results are usually irrelevant. This paper describes Research Compass, a smart scholarly search engine that is based on a LangGraph state machine design. It combines multi-source retrieval through Springer, Elsevier, and OpenAlex APIs, hybrid BM25-semantic scoring, LLM-based reranking, and Retrieval-Augmented Generation (RAG) to analyze PDFs into a single unified pipeline. When tested on a corpus of 700 papers, it achieves a precision of 0.79, a recall of 0.74, an NDCG of 0.82, and an MRR of 0.78, which is better than BM25 and semantic-only baselines. The time of repeated queries is decreased by 91.3 percent using Redis-based caching, which proves the viability in real-time.

Keywords. Semantic Search; LangGraph; Hybrid Retrieval; Retrieval-Augmented Generation; LLM Reranking

1. INTRODUCTION

The amount of academic literature released in journals, conferences and preprint archives each day has increased to be more than can be browsed using a manual. Available tools like Springer, Elsevier, and Semantic Scholar are useful in finding resources, although the researchers have to deal with information overload, the existence of semantically incompatible results, and retrieving the same data in different sources, as well as a lack of a single summarizing option.

Research Compass is an AI web-based solution that solves this issue. It is implemented on a LangGraph platform to arrange parallel API searches between various sources, combine word- and meaning-based ranking, re-rank results with large language model and generate summaries with Retrieval-Augmented Generation within a single, continuous pipeline. System design, mathematical models, testing results and important findings are described in this paper.

All Open Access articles are peer-reviewed, distributed under the Creative Commons Attribution-Non Commercial 4.0 International

2. RELATED WORK

TF-IDF and BM25 were the early academic retrieval methods that enhanced TF-IDF by document length normalization and term-frequency saturation, forming a good information retrieval base. Dense retrieval models learn queries and documents by embedding them into a common representation space. BERT, Sentence-BERT, SPECTER, DPR, and ColBERT yield good retrieval embeddings. Hybrid systems use linear interpolation between BM25 and dense scores as a way of balancing lexical-precision and semantic-recall. The RAG model is a compilation of retrieval and generative models to minimize hallucination.

3. KEY CONTRIBUTION

This work contributes to academic information retrieval in five key ways: (1) introducing a LangGraph-based state machine that allows for conditional intent routing and modular execution of retrieval, reranking, and generation processes; (2) developing a weighted hybrid retrieval method that combines BM25 lexical scoring with dense semantic similarity, with a tuning parameter α set to 0.6 based on experimental results; (3) implementing an LLM-based semantic reranking approach that improves results by considering contextual relevance beyond just embedding similarity; (4) incorporating a Redis and SHA-256 caching layer that cuts down on repeated query processing time by 91.3%

4. METHODOLOGY

The Research Compass is an intelligent system that is a mix between different approaches which is designed to help users locate academic literature more easily and effectively. The approach relies on several features including semantic search, ranking algorithms, data sources integration and caching. The essence of the system is that it has a state machine developed according to LangGraph principles where certain processing nodes are connected conditionally.

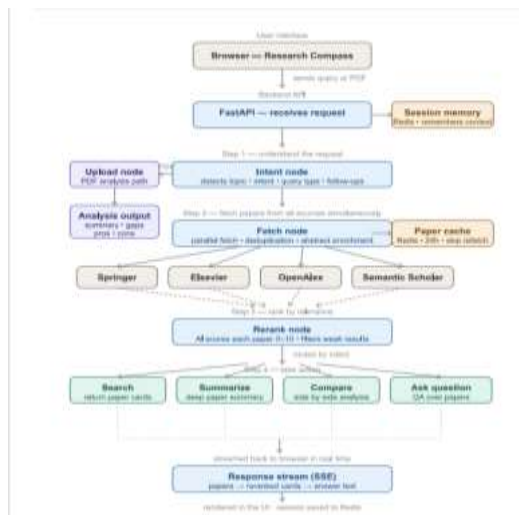


Figure 4.1. High-level system architecture of the Research Compass

4.1. Retrieval Strategy Implementation

BM25 ranking is used to score documents based on term frequency and inverse document frequency. Sentence embedding models are represented in a semantic vector space and cosine similarity is used to measure similarity between sentences. The hybrid retrieval technique merges the two approaches by employing a weighted score function that incorporates both BM25 and semantic similarity.

Mathematical Formulation:

$$\text{Score}(q,d) = \alpha \cdot \text{BM25}(q,d) + (1-\alpha) \cdot \text{CosineSim}(q,d) \quad (1)$$

In this equation, q represents the query, d represents the document, and α represents the weight factor for each component. The value of α was optimized between 0.3 and 0.7, with the best results achieved at $\alpha = 0.6$.

4.2. Experimental Setup and Dataset

We ran all experiments on a local workstation to test computational feasibility in a moderate-resource environment. CPU: AMD Ryzen 5 5500U processor featuring 6 cores and 12 threads; GPU: NVIDIA GTX 1650; RAM: 8 GB DDR4.

Approximately 700 research papers were collected and stored locally using a caching mechanism. The dataset includes metadata such as paper titles, abstracts, authors, publication years, and Digital Object Identifiers (DOIs).

For each query, top-K results ($K = 5, 10, 20$) were retrieved using BM25, semantic, hybrid, and LLM-only strategies. Ranking evaluation measures including Precision@K, Recall@K, NDCG@K, and MRR were used. Furthermore, the Jaccard Stability Index was calculated, whereas Query Execution Time and Memory Usage were considered for resource overhead.

4.3. IMPLEMENTATION DETAILS

The system is implemented using a FastAPI backend for efficient asynchronous processing. A LangGraph-based pipeline is used to manage multi-step execution flows with conditional routing. The system integrates multiple scholarly APIs for data retrieval and uses a caching layer to store metadata and improve performance. Semantic embeddings are generated using SentenceTransformer models, and numerical computations are performed using NumPy and SciPy.

5. RESULTS AND ANALYSIS

Table 1 presents the performance of all four retrieval strategies on the controlled evaluation dataset at $K = 10$.

Table 1. Performance Comparison of Retrieval Methods at K=10 (mean $\pm$ std)

Method	Precision@10	Recall@10	NDCG@10	MRR
BM25	0.61 $\pm$ 0.00	0.58 $\pm$ 0.00	0.64 $\pm$ 0.00	0.59 $\pm$ 0.00
Semantic (MiniLM)	0.68 $\pm$ 0.00	0.65 $\pm$ 0.00	0.71 $\pm$ 0.00	0.66 $\pm$ 0.00

Stochastic Generative Baseline	0.54 ± 0.00	0.49 ± 0.00	0.57 ± 0.00	0.51 ± 0.00
Hybrid ($\alpha=0.6$)	0.79 ± 0.00	0.74 ± 0.00	0.82 ± 0.00	0.78 ± 0.00

In each parameter tested, the hybrid algorithm always performed the best. Compared to BM25, Hybrid improves Precision@10 by 29.5% ($0.61 \rightarrow 0.79$), Recall@10 by 27.6% ($0.58 \rightarrow 0.74$), and NDCG@10 by 28.1% ($0.64 \rightarrow 0.82$).

5.1. Performance Analysis

Performance peaks at $\alpha = 0.6$ and degrades symmetrically on either side, indicating that both lexical and semantic signals are important but lexical precision contributes slightly more in the scholarly domain. BM25, semantic, and hybrid methods are fully deterministic (Jaccard = 1.0), while the stochastic generative baseline produces completely different result sets across runs (Jaccard = 0.0), confirming its unsuitability for reproducible academic indexing. A key practical contribution of Research Compass is its Redis caching layer. Average cache speedup is 91.3% across query types, .

6. CONCLUSIONS

The surge in academic publications has rendered keyword-only search inadequate for modern research needs. Research Compass addresses this challenge by integrating data from Springer, Elsevier, and OpenAlex, employing a weighted BM25-semantic hybrid ($\alpha=0.6$), applying LLM-based reranking, and enabling RAG-powered question answering over PDFs. Evaluation over 700 papers confirms that the hybrid retrieval strategy outperforms all individual baselines, achieving Precision of 0.79, Recall of 0.74, NDCG of 0.82, and MRR of 0.78. Redis-based caching eliminates redundant API calls, reducing query latency by 91.3%.

7. REFERENCES

- [1] S. E. 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.
- [2] G. Salton and C. Buckley, “Term-weighting approaches in automatic text retrieval,” *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, 1988.
- [3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.
- [4] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” in *Proc. EMNLP-IJCNLP*, 2019, pp. 3982–3992.
- [5] R. Nogueira and K. Cho, “Passage Re-ranking with BERT,” *arXiv preprint arXiv:1901.04085*, 2019.

- [6] Cohan, S. Feldman, I. Beltagy, D. Downey, and D. Weld, “SPECTER: Document-level Representation Learning using Citation-informed Transformers,” in Proc. ACL, 2020, pp. 2270–2282.

Biographies



Anay Joshi received his Bachelor’s degree in Artificial Intelligence and Data Science from Savitribai Phule Pune University in 2026. His interests include Artificial Intelligence, Machine Learning, Data Science, and Intelligent information retrieval systems with a focus on developing practical AI-driven applications.



Anurag Mahalpure received his Bachelor’s degree in Artificial Intelligence and Data Science from Savitribai Phule Pune University in 2026. His interests include Artificial Intelligence, Machine Learning, Data Science and Retrieval-Augmented Generation (RAG) His project experience includes working on ocean colour satellite data for predicting CDOM (Colored Dissolved Organic Matter).



Abhishek Marwade received his Bachelor’s degree in Artificial Intelligence and Data Science from Savitribai Phule Pune University in 2026. His areas of interest include Large Language Models (LLMs), Generative AI, semantic search, information retrieval, and data engineering. His work focuses on developing AI-driven research assistants and intelligent systems for efficient information retrieval and data processing.