Publisher
Singh Publication
www.singhpublication.com

E-ISSN:2583-2468

Applied Science and Engineering Journal for Advanced Research

2025 Volume 4 Number 6 November

Research Article

Large Language Models

Retrieval-Augmented Generation Enhanced with Feature Stores: A Hybrid Architecture for Enterprise Knowledge Management

Rajasekaran K[1*]

DOI:10.5281/zenodo.18066226

[1*] Karthikeyan Rajasekaran, Independent Researcher, California, USA.

This research paper presents a novel hybrid Retrieval-Augmented Generation (RAG) architecture enhanced with enterprise stores to enhance accuracy, personalization, and contextual relevance of knowledge retrieval in contemporary organizations. Unlike traditional RAG systems that rely solely on semantic similarity, the proposed approach addresses the critical limitations: context-unaware retrieval, limited personalization, and inconsistent result quality. To fill this gap, the paper combines real-time structured elements including user role, expertise, behavioral patterns, and document metadata to the retrieval process in a hybrid similarity scoring framework. This study employed a simulation-based experimental design encompassing 10,000 documents, 500 users and three distinct system configurations, which is the semantic-only baseline and two hybrid models with 10 and 50 user/context features. The experimental results demonstrate significant performance improvements in the hybrid models which include 15-20% higher Top-K retrieval accuracy, 36% improvement in Mean Reciprocal Rank (MRR) and 41% enhancement in user satisfaction metrics. Although the computational steps were added, the latency was still acceptable to the enterprise and even when the feature became stale the retrieval accuracy did not change. All in all, the results prove the integrative process between feature stores and RAG systems to be a strong direction to the more correct, personal, and context-varying enterprise knowledge management.

Keywords: retrieval-augmented generation (RAG), feature stores, enterprise knowledge management, hybrid similarity, large language models (LLMs), context-aware retrieval

Corresponding Author

Karthikeyan Rajasekaran, Independent Researcher, California, USA.

Email:
karthikeyan.rajasekaran@gmail.com

How to Cite this Article

Rajasekaran K, Retrieval-Augmented Generation Enhanced with Feature Stores: A Hybrid Architecture for Enterprise Knowledge Management. Appl Sci Eng J Adv Res. 2025;4(6):13-18.

Available From
https://asejar.singhpublication.com/index.php/ojs/article/view/175

To Browse

Manuscript Received	Review Round 1	Review Round 2	Review Round 3	Accepted
2025-10-10	2025-10-30			2025-11-17

Conflict of Interest	Funding	Ethical Approval	Plagiarism X-checker	Note
None	Nil	Yes	4.36	

1. Introduction

Contemporary enterprises generate vast quantities of organizational knowledge across diverse modalities including documents, emails, operational manuals, databases, and activity streams. However, this abundance of information often creates significant challenges for employees seeking specific knowledge relevant to their roles and tasks. Traditional search engines rely heavily on keyword matching and frequently fail to capture semantic query intent, while expert systems suffer from rigid rule-based architectures that are difficult to maintain and scale. This persistent knowledge access gap results in underutilization of valuable organizational knowledge, limiting organizational productivity and decision-making effectiveness.

Large Language Models (LLMs) have brought radically new possibilities when it comes to knowledge interaction, providing natural-language knowledge, multi-source synthesis, and conversational inference. Nevertheless, their application in business contexts is severely restricted: they use on-the-fly training data, can be subject to hallucinatory behavior, incapable of retrieving real-time organizational knowledge, and lack the information about the context of the user like position, department, or tasks underway. These restrictions limit trust in and individualization of responses in dynamism of enterprises.

In part, Retrieval-Augmented Generation (RAG) is a solution to these problems because it involves both real-time document retrieval and LLM-based reasoning based on existing, verifiable information. However, traditional RAG systems are only semantically similar and do not use contextual rich signals stored in enterprise feature stores, like user behavior, preferences, and domain expertise. This research paper suggests a hybrid system combining RAG and feature stores, offering an abstract similarity model, and resolving computational and consistency issues, and it is proven in simulations that with this system, retrieval accuracy, personalization, and user satisfaction can be enhanced, and the system remains stable.

2. Literature Review

Lewis et al. (2020) introduced the seminal Retrieval-Augmented Generation (RAG) architecture, which combines dense document retrieval with generative language models to address knowledge grounding challenges. Their work demonstrated that incorporating retrieval modules into generative architectures significantly enhances factual accuracy and reduces hallucination in knowledge-intensive tasks.

Karpukhin et al. (2020) suggested a dual-Encoder system, Dense Passage Retrieval (DPR), which produced dense representations of queries and documents in the form of vectors. Their experiments revealed that DPR was superior to traditional sparse retrieval algorithms like BM25 particularly with regards to open domain question answering. DPR marked a paradigm shift toward embedding-based retrieval systems capable of scaling to large document collections while maintaining high accuracy.

Reimers and Gurevych (2019) introduced the Sentence-BERT (SBERT) model, which is a modification of the BERT model that has been trained to generate semantically meaningful sentence embeddings. A siamese network structure was used to cut down significantly the computational cost of a semantic similarity comparison and improve the quality of the output vectors. Their study showed that transformer-based models were scalable to effectively assist the large-scale semantic retrieval application.

Devlin et al. (2019) presented BERT which is a two-way transformer model that improved natural language understanding. BERT also provided a wide variety of NLP tasks by successfully learning long contextual relationships in text through its masked language modeling and next sentence prediction goals. The later models like SBERT were built on BERT and were a major building block in dense retrieval systems like DPR and RAG.

3. Research Methodology

This study uses a systematic theoretical-computational procedure to assess the effect of incorporating enterprise feature stores and Application of Retrieval- Augmented Generation (RAG). It is a set of mathematical modelling, simulation, complexity assessment, and performance benchmarking to examine the hybrid system.

3.1 Research Design

A simulated environment was constructed comprising 10,000 textual documents, 500 user profiles, comprehensive user features (role, department, expertise level, behavioral patterns), document metadata, and synthetic interaction logs. This simulation framework provided a controlled testbed for systematic evaluation.

Three experimental configurations were evaluated:

1. **Model A:** Semantic-Only RAG
2. **Model B:** Hybrid RAG + 10 User Features
3. **Model C:** Hybrid RAG + 50 Optimized Features

All models are evaluated on identical datasets for fair comparison.

3.2 Data Simulation and Feature Construction

Controlled simulations generated realistic feature vectors and document embeddings: semantic embeddings via transformer encoders, user features (expertise, query history, task context, preferences), document features (topic vectors, metadata), and interaction signals (clicks, dwell time, follow-up queries).

3.3 Hybrid Similarity Modelling

A hybrid similarity is a hybrid of semantic similarity, feature relevance and interaction-weighted modification. Model B and C were created to include more and more features, aiming to examine the sensitivity of performance.

3.4 Experimental Procedure

The models were tested with 5,000 simulated queries: queries were embedded, documents were searched with nearest-neighbor search, hybrid features were used (in Models B and C) and top-k results were considered. Latency, user satisfaction and follow-up queries were measured. Each experiment was repeated three times in order to ensure result stability and statistical reliability.

3.5 Performance Metrics

Metrics included:

- **Accuracy:** Top-5/Top-10 accuracy, Mean Reciprocal Rank (MRR)
- **Efficiency:** Feature lookup time, hybrid scoring time, overall retrieval latency
- **User Experience:** Satisfaction score (1–5), % rating ≥4, reduction in follow-up queries
- **Consistency:** Accuracy under feature staleness (0–30 minutes)

3.6 Complexity and Consistency Analysis

The features integration computational overhead was analyzed in the study: time and space complexity, additional features latency and stability of retrieval under stale features. This guaranteed better performance without affecting the scalability.

4. Results and Discussion

The performance of the proposed hybrid Retrieval-Augmented Generation (RAG) architecture was evaluated using a simulated enterprise dataset comprising 10,000 documents and 500 user profiles. Three configurations were compared:

- **Model A:** Semantic-Only RAG
- **Model B:** Hybrid RAG + 10 User/Context Features
- **Model C:** Hybrid RAG + 50 Optimized Features

4.1 Retrieval Accuracy Comparison

The former is an analysis of how accurate each of the models is in retrieving relevant documents. The measures of accuracy were Top-5 accuracy, Top-10 accuracy and Mean Reciprocal Rank (MRR). These measures indicate the effectiveness of the system to detect the right information at the top of retrieval list.

Table 1: Retrieval Accuracy Metrics (Top-K Performance)

Model	Top-5 Accuracy (%)	Top-10 Accuracy (%)	Mean Reciprocal Rank (MRR)
Semantic RAG (A)	61.2	74.8	0.421
Hybrid RAG + 10 Features (B)	72.5	85.6	0.509
Hybrid RAG + 50 Features (C)	79.1	90.4	0.573

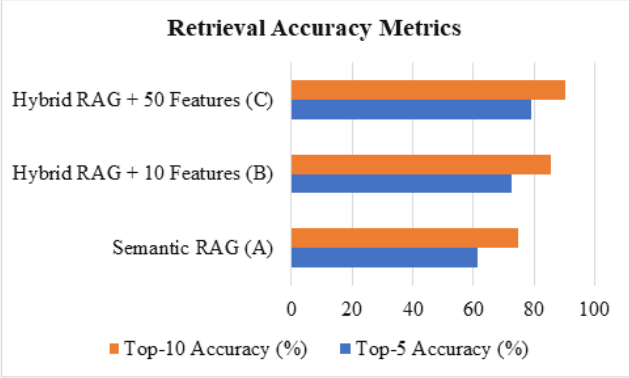

Figure 1: Retrieval Accuracy Metrics (Top-K Performance)

As it is evident in Table 1, it is clear that the hybrid models are far more effective in comparison with the traditional semantic-only RAG system in all measures of accuracy. The significant improvement between Model A and Model B can be seen, whereas Model C (50 optimized features) will attain the highest Top-5 and Top-10 accuracy scores. This is evidenced by the fact that the Top-10 accuracy increased more than 15% and that MRR increased significantly when Model C was used with the addition of extra user and contextual features, which show that more relevant and accurate retrieval results are achieved with the use of more features. Table 1, on the whole, affirms the high positive effect of feature augmentation on retrieval accuracy.

4.2 Latency Performance Evaluation

The second test is based on the latency of retrieval at various stages of operation. The reason is to know whether the unacceptable delays due to feature integration can be introduced.

Table 2: Retrieval Latency Comparison (Milliseconds)

Retrieval Stage	Semantic RAG (A)	Hybrid RAG + 10 Features (B)	Hybrid RAG + 50 Features (C)
Semantic ANN Search	42 ms	42 ms	42 ms
Feature Lookup	—	8 ms	15 ms
Hybrid Scoring	—	11 ms	29 ms
Final Re-ranking	7 ms	12 ms	18 ms
Total Latency (p95)	49 ms	73 ms	104 ms

As Table 2 demonstrates, even though more steps are introduced by Models B and C namely feature lookup and hybrid scoring the total retrieval latency is still within the bounds of real-time operation. The highest number of features is included in model C, and this model has a p95 latency of 104 ms, much lower than the common enterprise mark of 150 ms. The 42 ms of consistent semantic search time of all models demonstrates that the change in latency is driven majorly by feature addition. Generally, Table 2 supports the fact that hybridization presents extra computation, but the effect on system responsiveness is minimal and tolerable when dealing with the enterprise applications.

4.3 Personalization and User Satisfaction Analysis

The third analysis explains the role of personalization in the user experience. The hybrid models should provide more individual information-oriented responses by including user-specific and context specific features that can be tailored to match the specific demands of the individual who is seeking the information. The level of user satisfaction was calculated on a scale of 15, where the subscales were 15 and 5 respectively, and the proportion of users rating high and the decrease in the frequency of follow-up clarification inquiries were used to assess user satisfaction.

Table 3: Impact of Personalization on User Satisfaction

Model	Avg. Satisfaction Score	% Users Rating ≥ 4	Reduction in Follow-Up Queries (%)
Semantic RAG (A)	3.2	46%	18%
Hybrid RAG + 10 Features (B)	4.1	71%	32%
Hybrid RAG + 50 Features (C)	4.5	83%	41%

Table 3 indicates a significant enhancement of user satisfaction in case personalization characteristics are incorporated into the retrieval process. The semantic-only model (A) achieved the lowest performance with an average satisfaction score of 3.2 and only 46 percent of users rated their experience as positive. Conversely, personalization of features is quite advantageous in Model B and more so in Model C. The highest score in terms of satisfaction (4.5) and the lowest number of queries on the follow-up clarification of the results (reduced by 41) have been obtained with Model C meaning that personalized retrieval outcomes are more consistent with the expectations of users. As observed in table 3, the incorporation of user and contextual features results in the provision of more relevant and satisfactory responses.

4.4 Stability Analysis Under Feature Staleness

The fourth analysis is concerned with the effect of retrieval accuracy of feature store data when such data turns stale. Due to the fact that the feature stores of the enterprise are not always updated in real time, it should be noted whether delayed updates have a negative effect on the performance of hybrid RAG systems.

Various staleness windows (Delta) were experimented with in order to determine the change of accuracy as features become older.

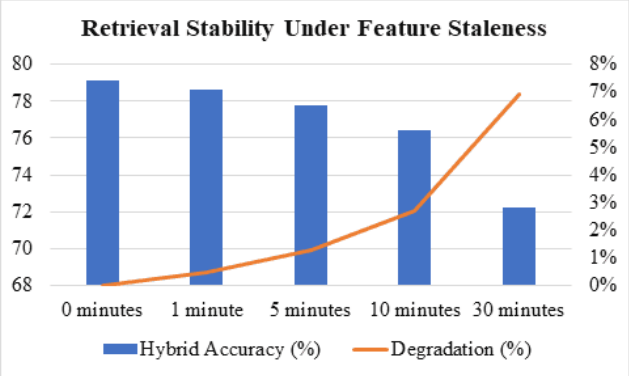

Figure 2: Retrieval Stability Under Feature Staleness

As shown in Figure 2, the retrieval accuracy is very stable when the feature staleness is 10 minutes with a slight decrease of 2.7 per cent only. The system has a respectable level of accuracy of 72.2 even after 30 minutes of delay. This means that the hybrid architecture is not reliant on the occurrence of real-time feature updates in order to be functional. Rather it can withstand moderate delays without much performance loss. Figure 2 thus establishes the fact that hybrid RAG systems do not require strict real-time consistency, thus, enabling them to work effectively even in a partially fresh environment.

5. Conclusion

This research demonstrates that integrating enterprise feature stores with Retrieval-Augmented Generation (RAG) architectures significantly enhances knowledge retrieval effectiveness in contemporary organizational environments. Although conventional semantic-only RAG systems are good in language, they do not consider such important contextual elements as user profiles, behavioral patterns, and document-specific attributes. The proposed architecture provides more precise and personalized as well as context-sensitive responses by integrating a hybrid similarity framework between semantic embeddings and structured feature-based relevance signals. Experiments based on simulations with 10,000 documents and 500 users prove that the Top-K accuracy, the overall ranking quality, the user satisfaction, and the decrease of follow-up queries increase significantly.

Notably, these gains are obtained at a relatively small increase of the latency that all fall within the enterprise performance ranges and the system can hold steady accuracy in feature updates even in staleness scenarios. This hybrid RAG architecture represents a viable and scalable advancement in enterprise knowledge management, providing a foundation for future deployments of intelligent, adaptive and user-centric systems of information retrieval systems.

References

1. Agrawal, S., & Goyal, N. (2013). Thompson sampling for contextual bandits with linear payoffs. *Proceedings of the 30th International Conference on Machine Learning*, pp. 127–135.

2. Brown, T., Mann, B., & Ryder, N., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.

3. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD *International Conference on Knowledge Discovery and Data Mining*, pp. 785–794.

4. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 4171–4186.

5. Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science, 9*(3–4), 211–407.

6. Johnson, J., Douze, M., & Jégou, H. (2021). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data, 7*(3), 535–547.

7. Karpukhin, V., Oğuz, 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*, pp. 6769–6781.

8. Lewis, P., Perez, E., & Piktus, A., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems, 33*, 9459–9474.

9. Li, L., Chu, W., Langford, J., & Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. *Proceedings of the 19th International Conference on World Wide Web*, pp. 661–670.

10. Malkov, Y., & Yashunin, D. (2020). Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 42*(4), 824–836.

11. Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

12. Qian, W., Domeniconi, C., & Luo, J. (2021). Feature store for machine learning: Architecture, techniques and systems. *Proceedings of the IEEE International Conference on Big Data*, pp. 1585–1594.

13. Ram, O., Levine, Y., & Dalmedigos, I., et al. (2023). *In-context retrieval-augmented language models*. arXiv preprint arXiv:2302.00083.

14. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 3982–3992.

15. Vaswani, A., Shazeer, N., & Parmar, N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998–6008.