International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048
Volume 10, Issue 5, pp 183-195, May-2026
https://doi.org/10.47001/IRJIET/2026.105024

Agentic RAG with Hybrid Retrieval and RBAC for Secure and Explainable Enterprise Knowledge Management

[1]Yamagani Niharika, [2]Voruganti Hasitha, [3]Manas Kumar Rath

[1,2]Department of Computer Science and Engineering, Mahatma Gandhi Institute of Technology, Hyderabad, India
[3]Assistant Professor, Department of Computer Science and Engineering, Mahatma Gandhi Institute of Technology, Hyderabad, India
E-mail: [1]yniharika_cse2305k3@mgit.ac.in, [2]vhasitha_cse2305k1@mgit.ac.in, [3]manaskumarrath_cse@mgit.ac.in

***Abstract* - Enterprise organizations generate large volumes of unstructured data such as policy documents, reports, spreadsheets, presentations, and internal records. Retrieving relevant information from such heterogeneous data remains challenging due to fragmented storage, semantic ambiguity, and strict access-control requirements. Traditional keyword-based search systems often fail to capture contextual meaning, while purely semantic retrieval approaches may overlook exact keyword relevance and enterprise security constraints. To address these limitations, this paper proposes an Agentic Retrieval-Augmented Generation (RAG)-based Enterprise Knowledge Retrieval System designed to provide accurate, secure, and context-aware information retrieval across organizational data repositories.**

The proposed system integrates hybrid retrieval using dense vector search (Qdrant) and sparse keyword search (Elasticsearch BM25), combined through Reciprocal Rank Fusion (RRF) and cross-encoder reranking to improve retrieval quality and contextual relevance. In addition, the framework incorporates a Plan–Act–Verify agentic reasoning loop with self-reflection and confidence-based verification to reduce hallucinations and improve answer reliability. Role-Based Access Control (RBAC) is enforced directly within the retrieval pipeline to ensure secure access to enterprise information. The system further supports ingestion of heterogeneous document formats including PDF, DOCX, XLSX, and PPTX files through a scalable processing pipeline. Experimental evaluation demonstrates improved retrieval precision, contextual relevance, and response reliability, making the proposed system suitable for enterprise-scale knowledge management applications.

Keywords: Retrieval-Augmented Generation (RAG), Enterprise Knowledge Retrieval, Hybrid Search, Agentic AI, RBAC, Semantic Search, Large Language Models.

I. INTRODUCTION

In recent years, enterprise organizations have experienced rapid growth in the volume of unstructured digital data generated through reports, policy documents, spreadsheets, presentations, and internal communications. Managing and retrieving relevant information from such heterogeneous data sources has become increasingly challenging. Employees often spend significant time searching for information across fragmented systems, leading to reduced productivity, delayed decision-making, and inefficient knowledge utilization.

Traditional enterprise search systems primarily rely on keyword-based retrieval methods, which often fail to capture semantic meaning and contextual relationships within documents. Although semantic vector search techniques improve contextual understanding, they may overlook exact keyword relevance and domain-specific terminology. In addition, enterprise environments require strict access-control mechanisms to ensure that sensitive information is accessible only to authorized users. Existing retrieval systems frequently lack integrated security-aware retrieval and reliable answer verification mechanisms.

The emergence of Retrieval-Augmented Generation (RAG) systems has introduced a powerful approach for combining information retrieval with Large Language Models (LLMs) to generate context-aware responses. However, deploying RAG systems in enterprise environments introduces several challenges including retrieval accuracy, hallucination reduction, response reliability, low-latency processing, and secure access control. Furthermore, enterprise documents exist in multiple formats such as PDF, DOCX, XLSX, and PPTX, requiring scalable ingestion and preprocessing pipelines.

To address these limitations, this paper proposes an Agentic Retrieval-Augmented Generation (RAG)-based Enterprise Knowledge Retrieval System designed to provide secure, accurate, and context-aware information retrieval over

ISSN (online): 2581-3048
Volume 10, Issue 5, pp 183-195, May-2026
https://doi.org/10.47001/IRJIET/2026.105024

enterprise data repositories. The proposed framework integrates hybrid retrieval using dense vector search and keyword-based retrieval combined through Reciprocal Rank Fusion (RRF) and cross-encoder reranking to improve retrieval quality. In addition, the system incorporates a Plan–Act–Verify agentic reasoning loop with self-reflection and confidence-based verification to reduce hallucinations and improve answer reliability.

The proposed system also enforces Role-Based Access Control (RBAC) directly within the retrieval pipeline to ensure secure information access. Redis-based caching is utilized to improve response efficiency and reduce retrieval latency. By integrating hybrid retrieval, agentic reasoning, and secure access control, the proposed system aims to provide a scalable and reliable enterprise knowledge management solution.

II. LITERATURE SURVEY

The rapid growth of enterprise data and the increasing adoption of Large Language Models (LLMs) have significantly accelerated research in Retrieval-Augmented Generation (RAG)-based knowledge retrieval systems. Enterprise organizations generate large volumes of unstructured data including policy documents, reports, spreadsheets, presentations, emails, and internal records. Efficient retrieval of relevant information from such heterogeneous repositories has become a major challenge due to fragmented storage systems, semantic ambiguity, and strict organizational security requirements. Consequently, researchers have focused on developing intelligent retrieval systems capable of combining semantic understanding with accurate information retrieval.

Traditional enterprise search systems primarily rely on keyword-based retrieval techniques such as BM25 and lexical matching algorithms. These systems are effective for exact term matching and structured document retrieval; however, they often fail to capture contextual meaning and semantic relationships between queries and documents. In enterprise environments, users frequently express queries in natural language, making it difficult for conventional retrieval systems to identify relevant information accurately. In addition, keyword-only retrieval approaches may overlook contextually related documents that do not contain exact query terms.

Recent advancements in transformer-based language models and dense vector embeddings have significantly improved semantic retrieval capabilities. Roychowdhury et al. proposed ERATTA, an enterprise-focused RAG framework for table-to-answer generation using Large Language Models. Their work demonstrated the importance of retrieval grounding and contextual reasoning in enterprise question-answering systems. Similarly, Balakrishnan and Purwar evaluated the efficacy of open-source LLMs within enterprise-specific RAG systems and highlighted the scalability and flexibility advantages of hybrid retrieval architectures.

Several studies have investigated retrieval optimization techniques aimed at improving contextual relevance and reducing hallucinations in generated responses. Sawarkar et al. proposed Blended RAG, which integrates semantic vector search with hybrid query-based retrievers to improve retrieval consistency and answer quality. Zhang et al. introduced Prolog-RAG, combining symbolic reasoning with retrieval augmentation to improve logical coherence in generated outputs. Furthermore, AboulEla et al. explored retrieval-grounded verification approaches to minimize hallucinations in enterprise LLM systems and improve answer reliability.

Research has also focused on enterprise-specific deployment challenges including security, explainability, scalability, and adaptive reasoning. Hamayat et al. developed SEEBot, a secure and economical enterprise chatbot integrating open-source LLMs with RAG-based retrieval mechanisms. Thanh et al. proposed an intelligent enterprise assistant capable of connecting LLMs with structured organizational data for enterprise decision support. Likewise, Tiwari et al. introduced an enterprise automation framework integrating NLP, RAG, and advanced security mechanisms to support scalable enterprise workflows.

Another significant area of research involves agentic reasoning and adaptive retrieval strategies. Hariharan et al. proposed an Agentic RAG framework integrating hybrid vector-graph retrieval and multi-agent orchestration for intelligent reasoning tasks. Jayavardhana and Hadinata further improved retrieval quality through fine-tuned BERT cross-encoders and GPT-based adaptive reranking mechanisms. These approaches demonstrated improved contextual reasoning and retrieval precision for complex enterprise queries.

Despite these advancements, existing enterprise RAG systems still face several limitations. Many systems suffer from inconsistent retrieval quality, hallucination generation, inadequate role-based security integration, and limited support for heterogeneous enterprise documents. Furthermore, most current approaches focus primarily on retrieval optimization or response generation independently without integrating secure retrieval, confidence-based verification, adaptive query refinement, and agentic reasoning within a unified framework.

Therefore, there is a growing need for scalable enterprise knowledge retrieval systems capable of combining hybrid retrieval, intelligent reasoning, secure access control, and

ISSN (online): 2581-3048
Volume 10, Issue 5, pp 183-195, May-2026
https://doi.org/10.47001/IRJIET/2026.105024

hallucination reduction mechanisms within a single architecture. This paper proposes an Agentic Retrieval-Augmented Generation (RAG)-based Enterprise Knowledge Retrieval System integrating hybrid retrieval, Reciprocal Rank Fusion (RRF), cross-encoder reranking, Role-Based Access Control (RBAC), and self-reflective agentic reasoning to improve retrieval accuracy, reliability, security, and scalability in enterprise environments.

III. RELATED WORK

Enterprise knowledge retrieval and Retrieval-Augmented Generation (RAG) systems have attracted substantial attention from both academia and industry due to the increasing demand for intelligent organizational knowledge management solutions. The rapid growth of enterprise-scale data repositories and the emergence of Large Language Models (LLMs) have motivated researchers to develop advanced retrieval architectures capable of generating context-aware and reliable responses from enterprise documents.

Several enterprise search and document management systems currently available in the market provide functionalities such as keyword search, document indexing, metadata filtering, and query-based retrieval. These systems are widely used across organizations for managing reports, policies, and internal records. Although such systems assist users in locating documents efficiently, they primarily rely on lexical matching methods and often fail to capture semantic meaning and contextual relationships between queries and documents. As enterprise data becomes increasingly heterogeneous and unstructured, these limitations reduce retrieval quality and negatively affect user productivity.

Traditional retrieval systems based on BM25 and sparse keyword matching approaches are highly effective for exact term retrieval and structured data indexing. However, these approaches perform poorly when users express queries in natural language or when relevant documents do not contain exact keyword matches. To overcome these limitations, researchers have explored semantic retrieval approaches utilizing transformer-based embedding models and vector databases for contextual similarity search.

Roychowdhury et al. proposed ERATTA, an enterprise-focused RAG framework utilizing Large Language Models for table-to-answer generation and contextual enterprise question answering. Their work emphasized the importance of retrieval grounding and evidence-supported response generation. Similarly, Balakrishnan and Purwar conducted comparative studies on open-source LLMs within enterprise RAG environments and highlighted the scalability advantages of hybrid retrieval architectures integrating both semantic and keyword-based retrieval.

Several studies have focused on improving retrieval quality and minimizing hallucinations in generated responses. Sawarkar et al. introduced Blended RAG, combining semantic retrieval with hybrid query-based retrievers to improve contextual relevance and ranking consistency. Zhang et al. proposed Prolog-RAG, integrating symbolic reasoning with retrieval augmentation to improve logical consistency and explainability in generated outputs. AboulEla et al. further investigated retrieval-grounded verification strategies to reduce hallucinations and improve the trustworthiness of enterprise LLM systems.

Another important research direction involves enterprise-specific security and deployment considerations. Hamayat et al. developed SEEBot, a secure enterprise chatbot integrating open-source LLMs with RAG-based retrieval while emphasizing economical deployment and secure organizational communication. Thanh et al. proposed an intelligent enterprise assistant capable of integrating structured business data with retrieval-augmented generation systems to support enterprise-level decision-making. In addition, Tiwari et al. introduced an enterprise automation framework integrating NLP, RAG, and advanced security mechanisms for scalable organizational workflows.

Recent works have also explored agentic reasoning frameworks and adaptive retrieval mechanisms. Hariharan et al. proposed an Agentic RAG architecture integrating hybrid vector-graph retrieval and multi-agent orchestration to improve reasoning capabilities for complex enterprise tasks. Jayavardhana and Hadinata enhanced retrieval performance through fine-tuned BERT cross-encoders and GPT-based adaptive reranking mechanisms, demonstrating improvements in contextual relevance and retrieval precision.

Although significant advancements have been achieved, many existing enterprise RAG systems still face limitations including retrieval inconsistency, hallucination generation, insufficient role-based security integration, and limited support for heterogeneous enterprise documents. Most existing solutions focus either on retrieval optimization or response generation independently without integrating hybrid retrieval, confidence-based verification, secure access control, and adaptive reasoning into a unified enterprise architecture.

Therefore, there remains a strong need for intelligent enterprise knowledge retrieval systems capable of providing secure, scalable, and reliable retrieval over heterogeneous organizational data repositories. The proposed work addresses these limitations by integrating hybrid retrieval, Reciprocal Rank Fusion (RRF), cross-encoder reranking, RBAC-based filtering, and self-reflective agentic reasoning within a unified Enterprise RAG framework.

ISSN (online): 2581-3048
Volume 10, Issue 5, pp 183-195, May-2026
https://doi.org/10.47001/IRJIET/2026.105024

IV. PROPOSED SYSTEM

The proposed system is an Agentic Retrieval-Augmented Generation (RAG)-based Enterprise Knowledge Retrieval System designed to provide secure, accurate, and context-aware retrieval of enterprise information from heterogeneous organizational documents. The primary objective of the system is to improve enterprise knowledge management by integrating hybrid retrieval, intelligent reasoning, and secure access control within a unified architecture.

The proposed framework is designed to address the limitations of conventional enterprise search systems, which often rely solely on keyword-based retrieval and fail to capture semantic meaning and contextual relationships between queries and documents. Unlike traditional retrieval systems, the proposed solution combines dense semantic retrieval, sparse keyword search, adaptive reranking, and agentic reasoning mechanisms to improve retrieval precision and response reliability.

Enterprise documents within the proposed system are processed through a scalable ingestion pipeline supporting multiple heterogeneous document formats including PDF, DOCX, XLSX, PPTX, and TXT files. Extracted content is segmented into semantic chunks and converted into dense vector embeddings using transformer-based embedding models. These embeddings are stored in the Qdrant vector database for semantic retrieval, while the raw textual content is indexed in Elasticsearch for keyword-based search. Metadata associated with each document chunk, including file information and user-access permissions, is preserved throughout the pipeline.

One of the major components of the proposed system is the hybrid retrieval mechanism. The system performs semantic vector retrieval and sparse keyword retrieval in parallel to capture both contextual relevance and exact term matching. Results from both retrieval approaches are combined using Reciprocal Rank Fusion (RRF), which improves ranking consistency by prioritizing documents appearing in multiple retrieval outputs. Furthermore, a cross-encoder reranking model is applied to the top retrieved results to refine contextual relevance and improve final retrieval quality.

Another important feature of the proposed framework is the integration of an agentic reasoning mechanism based on the Plan–Act–Verify paradigm. The system initially analyzes the user query and determines whether query rewriting or refinement is required. Hybrid retrieval is then executed using the refined query, and the retrieved context is passed to the Large Language Model for response generation. A self-reflection and confidence-scoring mechanism subsequently evaluates the generated response and verifies its grounding against retrieved evidence. If the confidence score falls below a predefined threshold, the retrieval and reasoning process is iteratively refined to improve answer reliability and reduce hallucinations.

The proposed system also incorporates Role-Based Access Control (RBAC) directly within the retrieval pipeline. Each document chunk is associated with predefined access permissions corresponding to organizational roles such as admin, HR, finance, and general users. During retrieval, RBAC filtering is enforced within both vector search and keyword search operations to ensure that users can access only authorized enterprise information. This approach significantly improves enterprise data security and prevents unauthorized document exposure.

To improve efficiency and reduce latency, the system integrates Redis-based caching for storing previously processed query responses and retrieval results. Frequently accessed queries can therefore be served with reduced computational overhead and faster response generation. The system additionally maintains document metadata and user interaction records within PostgreSQL to support scalable enterprise deployment.

The proposed framework further includes a full-stack application architecture consisting of a FastAPI backend and a React frontend. The application supports document upload, ingestion monitoring, natural-language querying, source attribution, query history management, and role-based authentication. Figure 1 illustrates the overall architecture of the proposed Enterprise RAG system including document ingestion, hybrid retrieval, agentic reasoning, reranking, RBAC filtering, and response generation components.

Finally, the integration of hybrid retrieval, adaptive reasoning, confidence-based verification, and secure retrieval mechanisms enables the proposed system to provide scalable and reliable enterprise knowledge retrieval. The proposed architecture therefore offers a practical solution for enterprise-scale organizational knowledge management and intelligent information access.

V. SYSTEM ARCHITECTURE

The proposed Enterprise Retrieval-Augmented Generation (RAG) system is designed using a modular and scalable architecture to support secure, efficient, and context-aware enterprise knowledge retrieval. The architecture integrates document ingestion, hybrid retrieval, agentic reasoning, reranking, and response generation within a unified framework. The layered design improves maintainability, scalability, and efficient interaction between various system

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 10, Issue 5, pp 183-195, May-2026
https://doi.org/10.47001/IRJIET/2026.105024

components. The overall architecture of the proposed system is illustrated in Figure 1.

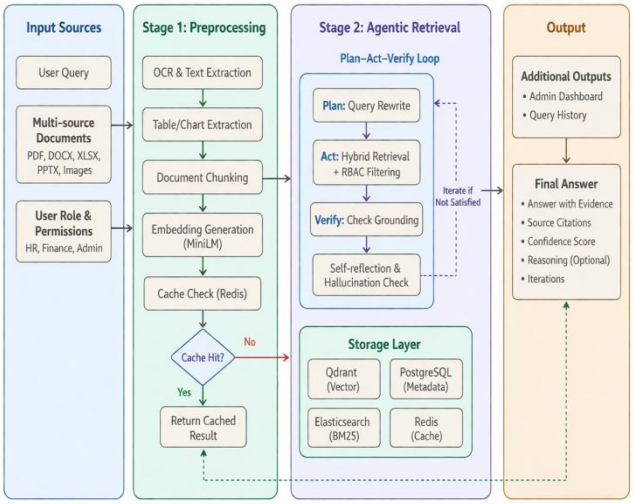


Figure 1: Proposed Enterprise RAG System Architecture

Figure 1 illustrates the architecture of the proposed Enterprise RAG system including document ingestion, hybrid retrieval, reranking, agentic reasoning, RBAC filtering, and response generation modules.

The architecture of the proposed system consists of four major layers: Presentation Layer, Processing and Retrieval Layer, Data Management Layer, and Platform Services Layer. Each layer performs specific responsibilities and communicates with other layers to ensure efficient and secure system operation.

A. Presentation Layer

The Presentation Layer represents the user interface through which users interact with the system. This layer is implemented using the React framework and provides interfaces for document upload, query submission, query history management, and response visualization.

The Presentation Layer includes modules such as Login Interface, Document Management Interface, Query Interface, Admin Dashboard, and Analytics View. The Query Interface allows users to submit natural-language enterprise queries and receive context-aware responses along with source attribution and confidence scores. The Document Management Interface enables administrators to upload enterprise documents and monitor document ingestion status.

All user queries and document-upload requests are captured within this layer and transferred to the Processing and Retrieval Layer for further processing.

B. Processing and Retrieval Layer

The Processing and Retrieval Layer represents the core intelligence of the proposed Enterprise RAG framework. This layer performs document preprocessing, embedding generation, hybrid retrieval, reranking, agentic reasoning, and response generation.

The Document Processing Module extracts textual information from heterogeneous enterprise documents including PDF, DOCX, XLSX, PPTX, and TXT files. OCR fallback mechanisms are utilized for scanned documents. Extracted content is segmented into semantic chunks and transformed into dense vector embeddings using the all-MiniLM-L6-v2 embedding model.

The Hybrid Retrieval Module performs semantic vector retrieval using Qdrant and sparse keyword retrieval using Elasticsearch BM25. Results from both retrieval methods are combined using Reciprocal Rank Fusion (RRF) to improve retrieval consistency and contextual relevance.

Another important component within this layer is the Cross-Encoder Reranking Module, which refines the ranking of retrieved results by jointly evaluating query–document pairs. The reranked context is then forwarded to the Large Language Model for response generation.

The Agentic Reasoning Module implements the Plan–Act–Verify paradigm. The system first analyzes the user query and performs query rewriting if required. Retrieved context is subsequently utilized by the Large Language Model to generate responses. A self-reflection mechanism then evaluates answer grounding and assigns a confidence score. If the generated response fails to satisfy the predefined confidence threshold, the retrieval and reasoning process is iteratively refined to improve reliability and reduce hallucinations.

Role-Based Access Control (RBAC) filtering is also enforced within this layer to ensure that users can retrieve only authorized enterprise information.

C. Data Management Layer

The Data Management Layer is responsible for data storage, indexing, and metadata management. This layer includes Qdrant Vector Database, Elasticsearch, PostgreSQL, and Redis Cache components.

Dense vector embeddings generated from enterprise document chunks are stored within Qdrant to support semantic similarity retrieval. Raw textual content and metadata are indexed in Elasticsearch to support keyword-based retrieval operations. PostgreSQL is utilized for storing document

ISSN (online): 2581-3048
Volume 10, Issue 5, pp 183-195, May-2026
https://doi.org/10.47001/IRJIET/2026.105024

metadata, user information, query logs, and RBAC permissions.

Redis caching is integrated to reduce redundant computations and improve retrieval latency by storing previously processed query responses and retrieval results.

The Data Management Layer ensures efficient storage, scalable indexing, and reliable access to enterprise knowledge repositories.

D. Platform Services Layer

The Platform Services Layer interacts with external services, APIs, authentication mechanisms, and system-level functionalities. This layer supports secure enterprise deployment and system integration.

Authentication and authorization services are implemented using JWT-based authentication and bcrypt password hashing. The Platform Services Layer also manages API communication between frontend and backend services through FastAPI endpoints.

Additional services including logging, monitoring, background task execution, and notification handling are managed within this layer. The incorporation of platform-level services improves scalability, reliability, and enterprise deployment capability.

E. Interaction Between Layers

Communication between the architectural layers is performed in a structured and modular manner. User requests generated within the Presentation Layer are transferred to the Processing and Retrieval Layer for document retrieval and reasoning. The Processing and Retrieval Layer interacts with the Data Management Layer to retrieve embeddings, metadata, and cached responses. Platform-level services support authentication, API communication, and backend operations.

The modular interaction between layers ensures scalability, maintainability, and efficient enterprise-level deployment without disrupting the functionality of other components.

F. Advantages of the Architecture

The proposed layered architecture provides several advantages including modularity, scalability, maintainability, and secure enterprise deployment. The separation of responsibilities between layers simplifies system maintenance and future upgrades.

The integration of hybrid retrieval, agentic reasoning, reranking, and RBAC filtering significantly improves retrieval quality, response reliability, and enterprise security. Furthermore, the use of Redis caching and scalable databases enables efficient handling of large-scale enterprise document repositories.

The proposed architecture therefore provides a practical and scalable solution for enterprise knowledge retrieval and intelligent organizational information access.

VI. IMPLEMENTATION

The implementation of the proposed Enterprise Retrieval-Augmented Generation (RAG) system focuses on developing a scalable, secure, and efficient enterprise knowledge retrieval platform capable of handling heterogeneous organizational documents and context-aware query processing. The system is implemented using modern AI frameworks, vector databases, web technologies, and scalable backend services following a modular architecture design.

The frontend of the application is developed using the React framework to provide an interactive and responsive user interface for enterprise users. The interface enables document upload, query submission, query history visualization, role-based authentication, and response monitoring. The backend services are implemented using FastAPI, which provides efficient REST API communication and supports scalable asynchronous processing.

The proposed system integrates multiple enterprise AI and retrieval technologies including Qdrant vector database, Elasticsearch, PostgreSQL, Redis cache, and Large Language Models (LLMs). Figure 2 illustrates the implementation workflow of the proposed Enterprise RAG system including document ingestion, embedding generation, hybrid retrieval, reranking, agentic reasoning, and response generation.

Figure 2 illustrates the end-to-end implementation pipeline including document preprocessing, embedding generation, hybrid retrieval, reranking, agentic reasoning, RBAC filtering, and response generation.

The implementation process follows a modular architecture in which the system is divided into multiple functional modules. Each module performs a specific role within the enterprise retrieval pipeline while communicating efficiently with other modules through defined interfaces.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 10, Issue 5, pp 183-195, May-2026

https://doi.org/10.47001/IRJIET/2026.105024

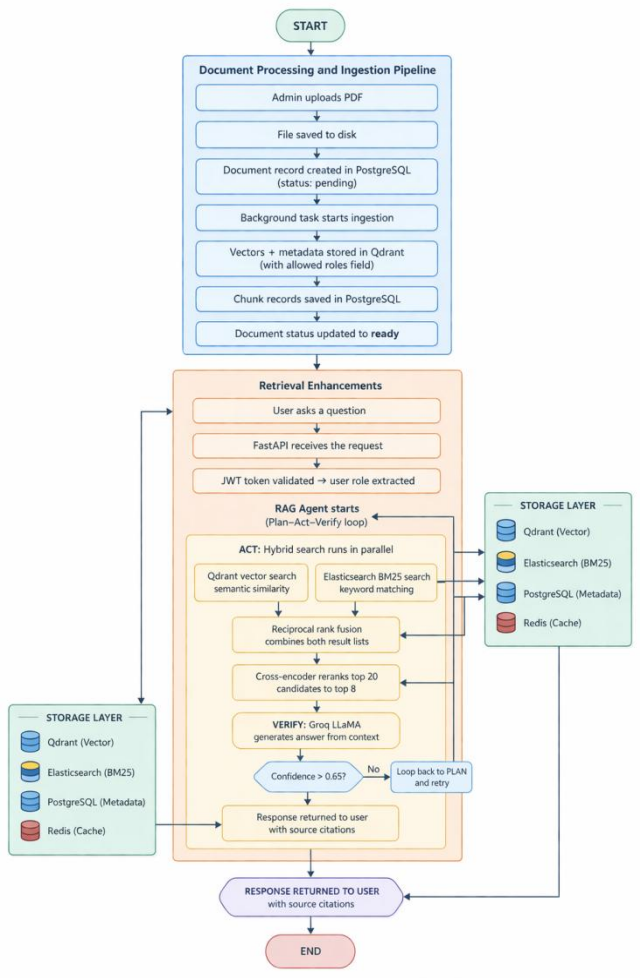


Figure 2: Implementation Workflow of the Enterprise RAG System

The major modules implemented in the proposed system include:

- Authentication and User Management Module
- Document Upload and Ingestion Module
- Embedding Generation Module
- Hybrid Retrieval Module
- Cross-Encoder Reranking Module
- Agentic Reasoning and Verification Module
- RBAC Security Module
- Query Processing and Response Generation Module
- Redis Caching Module
- Logging and Analytics Module

A. Authentication and User Management Module

The Authentication Module is responsible for secure user access and authorization. JWT-based authentication and bcrypt password hashing are implemented to ensure secure login and credential management. Users are assigned organizational roles such as admin, HR, finance, and general user, which are utilized for Role-Based Access Control (RBAC) during document retrieval.

The module also maintains user sessions, access logs, and authentication metadata to support enterprise-level monitoring and auditing.

B. Document Upload and Ingestion Module

The Document Upload and Ingestion Module handles ingestion of heterogeneous enterprise documents including PDF, DOCX, XLSX, PPTX, and TXT files. Uploaded documents are validated and processed through format-specific extraction pipelines.

PDF files are processed using OCR fallback mechanisms for scanned documents, while structured parsers are utilized for DOCX, XLSX, and PPTX files. Extracted textual content is segmented into semantic chunks using overlapping token windows to preserve contextual continuity.

Document metadata including file type, page number, uploader information, and RBAC permissions are preserved throughout the ingestion process.

C. Embedding Generation Module

The Embedding Generation Module converts semantic document chunks into dense vector representations using the all-MiniLM-L6-v2 transformer embedding model. Each chunk is represented as a 384-dimensional embedding vector capable of capturing semantic relationships between enterprise documents and user queries.

Generated embeddings are stored in the Qdrant vector database to support semantic similarity retrieval. The embedding generation process is optimized for efficient indexing and scalable enterprise deployment.

D. Hybrid Retrieval Module

The Hybrid Retrieval Module performs parallel semantic vector retrieval and sparse keyword retrieval to improve retrieval quality and contextual relevance. Semantic retrieval is performed using Qdrant cosine-similarity search, while keyword retrieval is implemented using Elasticsearch BM25 indexing.

Results from both retrieval approaches are combined using Reciprocal Rank Fusion (RRF), which improves ranking consistency by prioritizing documents appearing in multiple retrieval outputs.

This hybrid retrieval strategy significantly improves retrieval precision compared to single-method retrieval systems.

ISSN (online): 2581-3048
Volume 10, Issue 5, pp 183-195, May-2026
https://doi.org/10.47001/IRJIET/2026.105024

E. Cross-Encoder Reranking Module

The Cross-Encoder Reranking Module refines the ranking of retrieved document chunks by jointly evaluating query–document pairs using the ms-marco-MiniLM-L-6-v2 cross-encoder model.

The reranking process improves contextual alignment between enterprise queries and retrieved content, enabling the system to select the most relevant document chunks for response generation.

F. Agentic Reasoning and Verification Module

The proposed system integrates an agentic reasoning mechanism based on the Plan–Act–Verify paradigm. The module initially analyzes the user query and determines whether query rewriting or refinement is required.

Retrieved context is passed to the Large Language Model (LLaMA 3.1 via Groq) for answer generation. A self-reflection mechanism then evaluates the generated response against retrieved evidence and assigns a confidence score.

If the confidence score falls below the predefined threshold, the system iteratively refines retrieval and reasoning operations to reduce hallucinations and improve answer reliability.

G. RBAC Security Module

The RBAC Security Module ensures secure enterprise retrieval by enforcing role-based filtering directly within both vector search and keyword retrieval operations.

Each document chunk is tagged with predefined access permissions corresponding to organizational roles. During retrieval, only authorized document chunks are returned to the user, preventing unauthorized enterprise data exposure.

This security-aware retrieval approach significantly improves enterprise data protection compared to traditional application-level filtering mechanisms.

H. Query Processing and Response Generation Module

The Query Processing Module coordinates communication between retrieval, reranking, and reasoning components. User queries are processed through hybrid retrieval, reranking, and LLM-based generation pipelines to produce grounded enterprise responses.

Generated responses include contextual source attribution and confidence scores to improve explainability and user trust.

I. Redis Caching and Performance Optimization

Redis caching is integrated to reduce redundant retrieval and generation computations. Frequently accessed query responses and retrieval outputs are cached using role-aware cache keys, significantly improving response latency and reducing computational overhead.

The system also incorporates asynchronous API processing and optimized retrieval pipelines to support scalable enterprise deployment.

J. Integration and Testing

Module integration is performed systematically to ensure reliable communication between all architectural components. Unit testing and integration testing are conducted across document ingestion, retrieval, reranking, authentication, and response-generation modules.

The system is evaluated under various enterprise query scenarios including ambiguous queries, restricted-access queries, and multi-document reasoning tasks to validate retrieval accuracy, security enforcement, and response reliability.

K. Performance Considerations

System performance is improved through efficient vector indexing, Redis caching, asynchronous backend processing, and optimized reranking pipelines. The hybrid retrieval architecture enables accurate enterprise retrieval while maintaining low response latency.

The modular implementation design additionally supports scalability, maintainability, and future extension toward multimodal enterprise RAG applications.

VII. RESULTS

The proposed Enterprise Retrieval-Augmented Generation (RAG) system was implemented and evaluated to analyze its effectiveness in secure enterprise knowledge retrieval, contextual response generation, and intelligent reasoning. The system performance was evaluated with respect to retrieval quality, response reliability, latency, RBAC-based security enforcement, and overall system efficiency.

Initially, the retrieval performance of the proposed framework was analyzed by evaluating the effectiveness of the hybrid retrieval pipeline integrating semantic vector search and sparse keyword retrieval. Experimental observations demonstrated that combining Qdrant vector retrieval with Elastic search BM25 significantly improved contextual

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 10, Issue 5, pp 183-195, May-2026
https://doi.org/10.47001/IRJIET/2026.105024

relevance and retrieval precision compared to individual retrieval methods. Reciprocal Rank Fusion (RRF) further improved ranking consistency by prioritizing documents appearing in both semantic and keyword retrieval outputs.

Table 1 presents the quantitative performance evaluation of the proposed Enterprise RAG system.

Table 1: Performance Evaluation of Proposed Enterprise RAG System

Metric	Value
Precision@5	0.87
Recall@5	0.82
Answer Faithfulness	0.88
Avg. Confidence Score	0.84
Latency (seconds)	2.3

The evaluation results indicate that the proposed framework achieves high retrieval precision and strong grounding in generated responses while maintaining acceptable response latency for enterprise-scale deployment.

The retrieval performance of different retrieval strategies was further compared using Precision@5 and Recall@5 metrics. Experimental results demonstrated that the hybrid retrieval approach outperformed standalone vector retrieval and keyword retrieval systems.

Table 2: Retrieval Method Comparison

Method	Precision@5	Recall@5
Vector Search Only	0.78	0.74
BM25 (Keyword Search)	0.72	0.69
Hybrid Retrieval (Proposed)	0.87	0.82

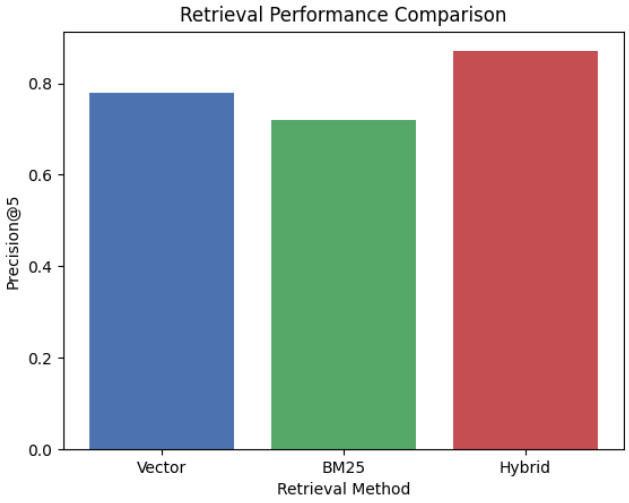


Figure 3: Retrieval Performance Comparison

Figure 3 illustrates the comparison between vector retrieval, keyword retrieval, and hybrid retrieval approaches. The hybrid retrieval mechanism achieved the highest retrieval precision and contextual relevance due to the integration of semantic and lexical retrieval techniques.

The effectiveness of the agentic reasoning mechanism was also evaluated. The Plan–Act–Verify framework enabled the system to iteratively refine retrieval and generation processes whenever low-confidence responses were detected. The self-reflection mechanism significantly reduced hallucinations and improved answer reliability.

Table 3: Agentic Reasoning Performance Analysis

Metric	Value
Single Iteration Success Rate	88%
Multi-Iteration Queries	12%
Avg. Iterations per Query	1.2
Confidence ≥ 0.65 Accuracy	90%

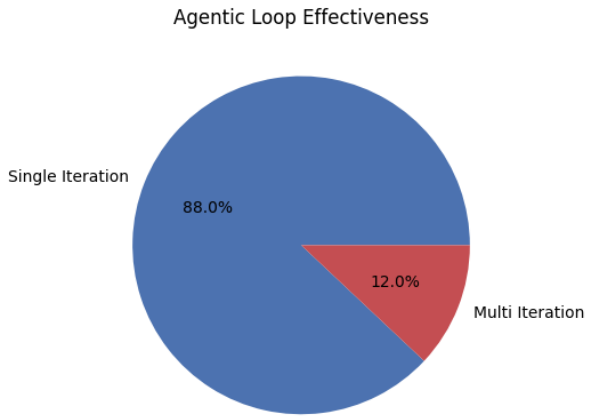


Figure 4: Agentic Loop Effectiveness

Figure 4 demonstrates that the majority of enterprise queries were successfully resolved within a single reasoning iteration, while only a small percentage required query refinement and re-retrieval.

The system latency and computational efficiency were further analyzed by evaluating the execution time of individual pipeline components including retrieval, reranking, and Large Language Model response generation.

Table 4: System Latency Analysis

Component	Time (seconds)
Retrieval	0.8
Reranking	0.6
LLM Generation	0.9
Total	2.3

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 10, Issue 5, pp 183-195, May-2026

https://doi.org/10.47001/IRJIET/2026.105024

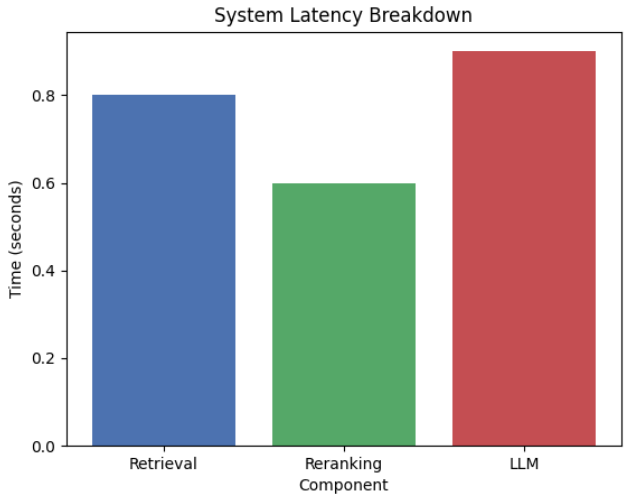


Figure 5: System Latency Breakdown

Experimental results showed that retrieval and reranking operations maintained low latency, while Large Language Model generation contributed the highest computational overhead. However, Redis-based caching significantly reduced repeated query processing time and improved overall response efficiency.

The contribution of individual architectural components was further investigated through an ablation study. Different configurations of the proposed framework were evaluated by selectively removing reranking and agentic reasoning modules.

Table 5: Ablation Study of Proposed System

Configuration	Precision@5	Faithfulness
Without Reranking	0.79	0.76
Without Agent Loop	0.81	0.78
Full System (Proposed)	0.87	0.88

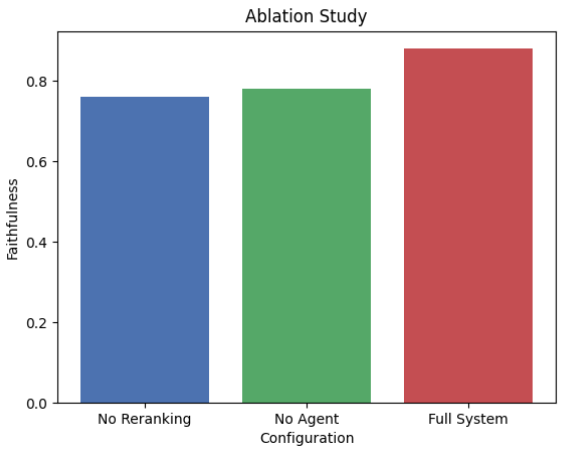


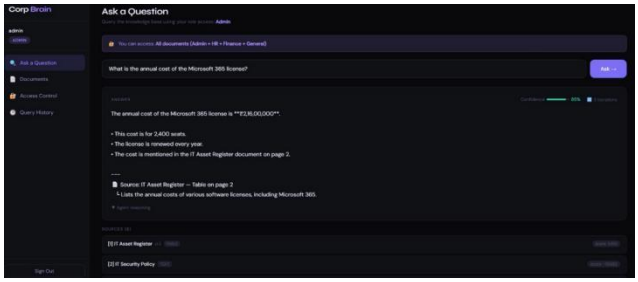
Figure 6: Ablation Analysis of Proposed System

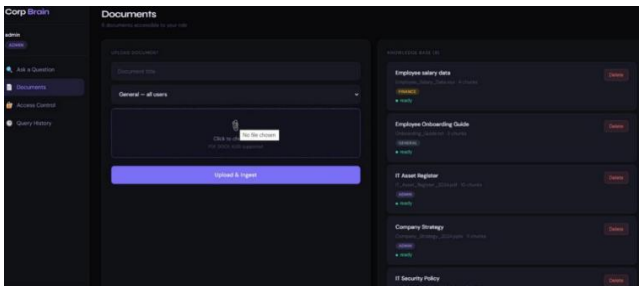
The ablation analysis demonstrated that both cross-encoder reranking and agentic reasoning significantly improved retrieval precision and answer faithfulness. The complete system achieved the highest performance among all evaluated configurations.

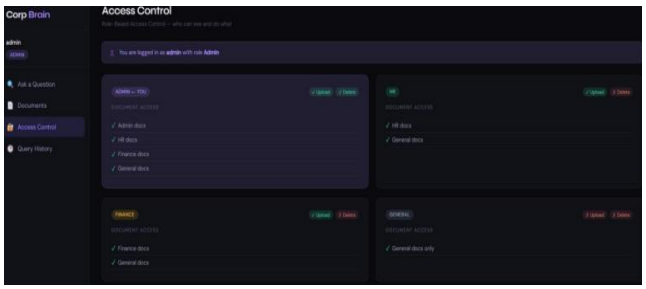
The RBAC security mechanism was also evaluated to verify secure enterprise information access. Experimental testing confirmed that users were restricted to retrieving only authorized enterprise documents according to predefined organizational roles. Unauthorized retrieval attempts were successfully blocked at the retrieval level within both vector search and keyword search operations.

In addition to quantitative evaluation, the proposed system was tested across various enterprise query scenarios including direct factual queries, ambiguous queries, restricted-access queries, and multi-document reasoning tasks. The system successfully generated context-aware and source-grounded responses while maintaining high reliability and security.

The functionality of the proposed framework is illustrated through screenshots of the implemented application interface as shown in Figure 7. The interface includes modules for document upload, enterprise querying, retrieval monitoring, and response visualization.







International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 10, Issue 5, pp 183-195, May-2026
https://doi.org/10.47001/IRJIET/2026.105024

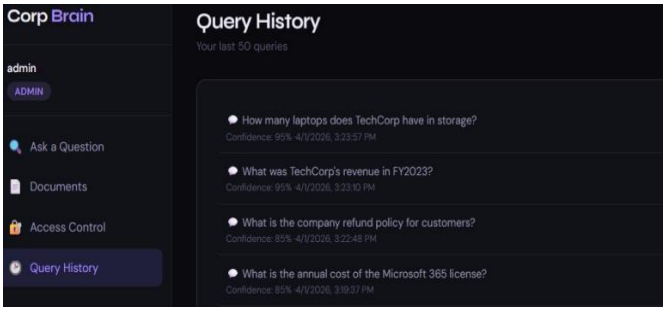


Figure 7: Application Interface Screens

Figure 7 illustrates the implemented frontend interfaces including login page, document management dashboard, enterprise query interface, and response visualization module.

Overall, the experimental results demonstrate that the proposed Enterprise RAG framework effectively integrates hybrid retrieval, agentic reasoning, reranking, RBAC filtering, and scalable document processing into a unified enterprise knowledge retrieval system. The proposed architecture therefore provides a practical and reliable solution for enterprise-scale intelligent information access and organizational knowledge management.

VIII. DISCUSSIONS

The implementation and evaluation of the proposed Enterprise Retrieval-Augmented Generation (RAG) framework provide several important insights regarding enterprise knowledge retrieval, contextual reasoning, and secure information access. The proposed system demonstrates clear improvements over traditional enterprise search systems by integrating hybrid retrieval, reranking, agentic reasoning, and RBAC-based secure retrieval within a unified architecture.

One of the most significant observations from the experimental evaluation is the effectiveness of the hybrid retrieval strategy combining semantic vector retrieval and sparse keyword retrieval. Traditional keyword-based retrieval systems are effective for exact term matching but often fail to capture contextual relationships within enterprise documents. Conversely, semantic vector retrieval improves contextual understanding but may overlook exact enterprise terminology and keywords. The proposed hybrid retrieval framework successfully combines both approaches, thereby improving retrieval precision and contextual relevance across diverse enterprise query scenarios.

Another important contribution of the proposed framework is the integration of Reciprocal Rank Fusion (RRF) and cross-encoder reranking mechanisms. The reranking module significantly improved contextual alignment between user queries and retrieved document chunks by jointly evaluating query–document pairs. Experimental analysis demonstrated that reranking was particularly effective for ambiguous and multi-document enterprise queries where contextual relevance plays a major role in retrieval quality.

The Plan–Act–Verify agentic reasoning framework also proved effective in improving answer reliability and reducing hallucinations. The self-reflection and confidence-based verification mechanism enabled the system to evaluate generated responses against retrieved enterprise evidence before returning answers to users. Queries with weak or insufficient grounding were iteratively refined through query rewriting and re-retrieval processes, thereby improving response trustworthiness and reducing incorrect answer generation.

Another major strength of the proposed framework is the incorporation of Role-Based Access Control (RBAC) directly within the retrieval pipeline. Unlike conventional systems where security filtering is applied only after retrieval, the proposed system enforces RBAC filtering during both vector retrieval and keyword retrieval operations. This significantly improves enterprise data security by preventing unauthorized document exposure during retrieval itself.

The proposed system also demonstrated strong flexibility and scalability in handling heterogeneous enterprise documents. The ingestion pipeline successfully processed multiple document formats including PDF, DOCX, XLSX, PPTX, and TXT files. OCR fallback mechanisms additionally enabled extraction of textual content from scanned enterprise documents. These capabilities improve the applicability of the framework within real-world enterprise environments where organizational knowledge exists across multiple heterogeneous sources.

Redis-based caching and optimized retrieval pipelines further improved overall system efficiency and response latency. Frequently repeated enterprise queries were processed faster using cached retrieval outputs and generated responses. The modular architecture additionally simplified communication between frontend interfaces, backend services, vector databases, and Large Language Models.

Table 6 presents a comparison between the proposed Enterprise RAG framework and existing enterprise retrieval systems.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 10, Issue 5, pp 183-195, May-2026

https://doi.org/10.47001/IRJIET/2026.105024

Table 6: Comparison with Existing Enterprise Retrieval Systems

System	Retrieval Type	Agentic Loop	RBAC	Reranking	Reliability
Basic RAG	Vector only	No	No	No	Medium
Hybrid RAG	Vector + Keyword	No	Partial	No	High
Agentic RAG (Literature)	Hybrid	Yes	Limited	Partial	High
Proposed System	Hybrid + RRF	Yes	Yes	Yes	Very High

The comparison demonstrates that conventional enterprise retrieval systems primarily focus on static keyword search and document indexing, whereas the proposed framework integrates hybrid retrieval, adaptive reasoning, reranking, hallucination reduction, and RBAC-aware secure retrieval within a single architecture.

Despite the strong performance of the proposed framework, several limitations still remain. The self-reflection mechanism depends on confidence-estimation heuristics and may occasionally require multiple retrieval iterations for highly ambiguous enterprise queries. In addition, Large Language Model inference contributes computational overhead during large-scale enterprise deployment involving high query throughput.

Another limitation involves the dependence on embedding quality and chunking strategies for retrieval effectiveness. Improper chunk segmentation may affect contextual retrieval quality in certain enterprise scenarios. Furthermore, although the current framework supports heterogeneous textual enterprise documents, multimodal enterprise data such as images, diagrams, and handwritten content remain challenging for current retrieval pipelines.

Overall, the proposed Enterprise RAG framework demonstrates strong potential as a scalable and intelligent enterprise knowledge retrieval system. The integration of hybrid retrieval, reranking, agentic reasoning, RBAC security, and confidence-based verification provides substantial improvements over conventional enterprise search systems and establishes a practical foundation for future enterprise AI applications.

IX. CONCLUSION

In this paper, an Agentic Retrieval-Augmented Generation (RAG)-based Enterprise Knowledge Retrieval System was proposed to address the challenges associated with enterprise-scale information retrieval, contextual reasoning, and secure organizational knowledge access. The proposed framework integrates hybrid retrieval, cross-encoder reranking, agentic reasoning, and Role-Based Access Control (RBAC) within a unified architecture to provide accurate, reliable, and secure enterprise knowledge retrieval.

The proposed system combines semantic vector retrieval using Qdrant and sparse keyword retrieval using Elasticsearch BM25 to improve contextual relevance and retrieval precision. Reciprocal Rank Fusion (RRF) and cross-encoder reranking mechanisms further improve retrieval consistency and contextual alignment between enterprise queries and retrieved document chunks. In addition, the integration of a Plan–Act–Verify agentic reasoning framework with self-reflection and confidence-based verification significantly reduces hallucinations and improves answer reliability.

Experimental evaluation demonstrated that the proposed framework effectively handles heterogeneous enterprise documents including PDF, DOCX, XLSX, PPTX, and TXT files through a scalable ingestion pipeline. The RBAC-aware retrieval mechanism successfully enforced secure enterprise information access by restricting retrieval operations according to organizational user roles. Redis-based caching additionally improved system efficiency and reduced response latency for repeated enterprise queries.

The modular architecture of the proposed framework improves scalability, maintainability, and enterprise deployment capability. The integration of hybrid retrieval, adaptive reasoning, reranking, and secure retrieval mechanisms enables the system to provide context-aware and source-grounded responses suitable for real-world enterprise applications.

Although the proposed framework demonstrates strong performance, several limitations remain. Large Language Model inference contributes computational overhead during high-throughput enterprise deployment scenarios, and retrieval effectiveness may vary depending on document chunking quality and embedding representation. In addition, the current framework primarily focuses on textual enterprise documents and provides limited support for multimodal enterprise data such as diagrams, images, and handwritten content.

Future enhancements may include integration of multimodal Retrieval-Augmented Generation techniques, adaptive chunking strategies, personalized retrieval mechanisms based on user behavior, and lightweight optimized LLM inference pipelines for large-scale enterprise deployment. Additional research may also focus on integrating

ISSN (online): 2581-3048
Volume 10, Issue 5, pp 183-195, May-2026
https://doi.org/10.47001/IRJIET/2026.105024

enterprise databases, workflow systems, and real-time organizational analytics into the proposed framework.

Overall, the proposed Enterprise RAG framework demonstrates strong potential as a scalable, secure, and intelligent enterprise knowledge retrieval system capable of improving organizational information access, response reliability, and enterprise knowledge management efficiency.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to Mr. Manas Kumar Rath, Assistant Professor, Department of Computer Science and Engineering, Mahatma Gandhi Institute of Technology, Hyderabad, India, for his valuable guidance, continuous support, and encouragement throughout the development of this research work. His technical insights and suggestions greatly contributed to the successful design and implementation of the proposed system.

The authors also extend their heartfelt thanks to Dr. Meera Alphy, Assistant Professor, Department of Computer Science and Engineering, Mahatma Gandhi Institute of Technology, Hyderabad, India, for her constant support, guidance, and valuable feedback during the course of this work.

Finally, the authors express their gratitude to the Department of Computer Science and Engineering, Mahatma Gandhi Institute of Technology, for providing the necessary facilities and academic environment to successfully carry out this research work.

REFERENCES

[1] S. Roychowdhury, M. Krema, A. Mahammad, B. Moore, A. Mukherjee, and P. Prakashchandra, "ERATTA: Extreme RAG for Enterprise – Table to Answers with Large Language Models," 2023.

[2] G. Balakrishnan and A. Purwar, "Evaluating the Efficacy of Open-Source LLMs in Enterprise-Specific RAG Systems: A Comparative Study of Performance and Scalability," 2023.

[3] V. S. Devarajulu, M. Usha Rani, N. V. Muthu Lakshmi, and L. U. K. Reddy, "RAG-Based Code Intelligence for Real-Time Fault Diagnosis in Enterprise Systems," 2023.

[4] F. Hamayat, L. Ejaz, M. Danish, A. Nazir, P. Ahadian, and R. F. Ahmad, "SEEBot: Leveraging Open-Source LLMs and RAG for Secure and Economical Enterprise Chatbots," 2023.

[5] R. Hu, S. Liu, P. Qi, J. Liu, and F. Li, "ICCA-RAG: Intelligent Customs Clearance Assistant Using Retrieval-Augmented Generation (RAG)," 2023.

[6] S. Bag, A. Gupta, C. Jain, and R. Kaushik, "RAG Beyond Text: Enhancing Image Retrieval in RAG Systems," 2023.

[7] D. D. Thanh, D. H. Nguyen, and B. N. T. Nguyen, "Towards Intelligent Enterprise Assistants: Leveraging RAG to Connect LLMs with Structured Business Data," 2023.

[8] B. Zhang, K. Peng, K. Meng, J. Li, and S. Zheng, "Prolog-RAG: A Symbolic Reasoning Approach to Retrieval-Augmented Generation," 2023.

[9] K. Sawarkar, A. Mangal, and S. R. Solanki, "Blended RAG: Improving RAG Accuracy with Semantic Search and Hybrid Query-Based Retrievers," 2023.

[10] S. AboulEla, P. Zabihitari, N. Ibrahim, M. Afshar, and R. Kashef, "Exploring RAG Solutions to Reduce Hallucinations in LLMs," 2023.

[11] G. Gill, R. Gupta, D. Lusson, A. Chandrashekar, and D. Nguyen, "From Search to Reasoning: A Five-Level RAG Capability Framework for Enterprise Data," 2023.

[12] A.Tiwari, A. R. Tiwari, A. S. Khan, and C. Yadav, "DSPY: A Next-Generation AI Chatbot for Enterprise Automation Using NLP, RAG, and Advanced Security," 2023.

[13] Y. Guo, L. Yan, J. Niu, D. Gao, and X. Yuan, "Integrating RAG and KAG Frameworks to Build Accurate Enterprise Question Answering Systems," 2023.

[14] M. Hariharan, S. Barma, S. Arvapalli, and E. Sheela, "Agentic RAG for Software Testing with Hybrid Vector-Graph and Multi-Agent Orchestration," 2023.

Citation of this Article:

Yamagani Niharika, Voruganti Hasitha, & Manas Kumar Rath. (2026). Agentic RAG with Hybrid Retrieval and RBAC for Secure and Explainable Enterprise Knowledge Management. *International Research Journal of Innovations in Engineering and Technology - IRJIET*, 10(5), 183-195. Article DOI https://doi.org/10.47001/IRJIET/2026.105024

****

