International Journal of Artificial Intelligence and Data Analytics
ISSN (Online): XXXX-XXXX
Volume 1, Issue 1, Pages 1–7, 2026 (July–August)
Available Online: https://ijaida.com
DOI: doi.org/10.47001/IJAIDA/2026.101001

Role-Based Access Control in Enterprise RAG Systems for Privacy-Preserving Information Retrieval

John D Mathew

Department of Computer Science and Engineering, Carmel College of Engineering and Technology, Alappuzha, Kerala, India

Abstract **- The rapid growth of enterprise data across digital repositories, cloud platforms, knowledge bases, and document management systems has created significant challenges in information retrieval and knowledge management. Conventional search systems often struggle to deliver contextually relevant responses, whereas standalone Large Language Models (LLMs) can produce inaccurate or misleading outputs due to their limited access to organization-specific knowledge. Retrieval-Augmented Generation (RAG) has emerged as an effective framework that integrates information retrieval with generative artificial intelligence, enabling the generation of more accurate, reliable, and context-aware responses. Despite these advantages, enterprise environments require robust access control mechanisms to protect sensitive information. To address this challenge, this study proposes an Enterprise Retrieval-Augmented Generation (RAG) system integrated with Role-Based Access Control (RBAC), providing secure, intelligent, and role-aware knowledge retrieval within organizational environments. The proposed framework combines vector databases, embedding models, retrieval engines, large language models, and RBAC policies to provide secure, personalized, and context-aware responses. The study discusses system architecture, methodology, implementation strategies, hardware requirements, security considerations, and performance evaluation. The results indicate that integrating RBAC into RAG systems significantly enhances information security while maintaining efficient retrieval performance and high response accuracy. The proposed framework enterprise knowledge management, decision support, and secure AI-powered information access.**

Keywords: Retrieval-Augmented Generation, Enterprise AI, Knowledge Management, Large Language Models, Information Retrieval, Artificial Intelligence Security, RAG, Secure Information Access; Enterprise AI; Knowledge Retrieval; Embedding Models; Generative AI; Decision Support Systems.

1. Introduction

Organizations generate and store vast amounts of structured and unstructured data, including reports, technical documentation, financial records, customer information, policy manuals, emails, and research documents. As enterprise data volumes continue to grow, employees face increasing difficulties in locating excessive results that require significant manual filtering. At the same time, in natural language understanding, reasoning, and content generation across. However, standalone LLMs are constrained by the knowledge acquired during training and may produce outdated, incomplete, or inaccurate responses when organizational information evolves over time or is not included in their training data. This retrieval-enhanced approach improves factual accuracy, minimizes hallucinations, and enables the generation of contextually relevant responses based on the most current organizational information [1].

Despite these advantages, enterprise deployment introduces significant security concerns. Employees possess different levels of authorization and should only access information relevant to their responsibilities. Unauthorized access to confidential organizational information Role-Based Access Control (RBAC) offers a systematic and effective approach to enforcing information security by organizational roles and user permissions. Integrating RBAC with Retrieval-Augmented Generation (RAG) systems ensures that AI-generated responses adhere to established security policies, access privileges, and organizational governance requirements [2].

This study proposes a secure enterprise architecture that integrates Retrieval-Augmented Generation (RAG) with Role-Based Access Control (RBAC) to facilitate intelligent, secure, and scalable knowledge retrieval. The proposed framework enhances response accuracy while ensuring that sensitive organizational information is accessed only by authorized users, thereby supporting secure AI-driven decision-making and enterprise knowledge management [3].

© 2026 The Author(s). CC BY 4.0 License

International Journal of Artificial Intelligence and Data Analytics
ISSN (Online): XXXX-XXXX
Volume 1, Issue 1, Pages 1–7, 2026 (July–August)
Available Online: https://ijaida.com
DOI: doi.org/10.47001/IJAIDA/2026.101001

2. Literature Review

Knowledge management systems have evolved from traditional database search engines to intelligent information retrieval platforms. Early enterprise search solutions relied primarily on keyword indexing and document ranking algorithms. While effective for basic retrieval tasks, these systems lacked contextual understanding and semantic reasoning capabilities [1].

These models have demonstrated exceptional performance summarization, translation, and conversational artificial intelligence. Their ability to learn contextual representations from large-scale datasets has enabled more accurate interpretation of user intent and the generation of coherent, wide range of applications. Despite these advancements, several studies have reported that standalone Large Language Models (LLMs) are susceptible to generating hallucinated, outdated, or factually incorrect information because they rely primarily on static and lack access to dynamically updated organizational data [2],[3].

To overcome these limitations, Lewis et al. with neural text generation to improve the relevance of AI-generated responses. This retrieval-enhanced approach significantly reduces hallucinations, improves explainability, and enables the use of current, domain-specific information. Subsequent research has demonstrated the effectiveness of RAG in applications such as open-domain question answering, enterprise knowledge management, customer support systems, intelligent search engines, legal document analysis, healthcare information retrieval, and educational tutoring platforms.

Recent developments in semantic search have further enhanced RAG performance through the adoption of dense vector embeddings and high-performance, Pinecone, Milvus, Chroma, and Weaviate. These technologies enable efficient similarity-based retrieval of semantically relevant documents, providing faster and more accurate access to large-scale enterprise knowledge repositories. Advances in embedding models, including Sentence-BERT, E5, and OpenAI embedding models, have further improved document representation and retrieval quality [4].

In parallel, researchers have emphasized the importance of secure deployment of RAG systems within enterprise environments. Organizational knowledge repositories often contain confidential business documents, financial records, intellectual property, and sensitive customer information that require strict access control. Consequently, integrating Retrieval-Augmented Generation with security mechanisms and data governance policies has emerged as a critical research direction. These approaches ensure that retrieved information and AI-generated responses comply with organizational security policies while maintaining privacy, regulatory compliance, and information confidentiality [5].

Recent studies also highlight the growing adoption of enterprise RAG architectures integrated with cloud computing, microservices, vector databases, and secure API frameworks to build scalable, reliable, and intelligent knowledge management systems. Such architectures improve response accuracy, reduce operational costs, support informed decision-making, and enhance employee productivity while ensuring secure and efficient access to organizational knowledge.

Parallel developments in enterprise security emphasized the importance of access control mechanisms. Sandhu *et al.* formalized scalable authorization framework that simplifies permission management. RBAC has since become a standard security model across enterprise systems, cloud infrastructures, and regulatory compliance frameworks [6].

Recent studies suggest that combining RAG systems with RBAC mechanisms can address both information retrieval accuracy and enterprise security requirements. However, comprehensive architectural frameworks and implementation strategies remain an active area of research.

3. Problem Statement

Although Retrieval-Augmented Generation systems provide highly relevant and context-aware responses, they introduce security risks when deployed in enterprise environments. Traditional RAG systems retrieve information solely based on semantic relevance without adequately considering user permissions. Consequently, users may inadvertently receive information beyond their authorized access level [4].

The primary challenge is designing a secure RAG architecture capable of delivering accurate responses while strictly enforcing organizational access policies. The system must ensure that retrieved documents, generated responses, and contextual information remain accessible only to authorized personnel. Additionally, the framework should maintain retrieval efficiency, scalability, and user experience despite the integration of security constraints [7].

4. Proposed Methodology

The proposed methodology integrates multi-stage processing pipeline designed to provide accurate, context-

© 2026 The Author(s). CC BY 4.0 License

International Journal of Artificial Intelligence and Data Analytics
ISSN (Online): XXXX-XXXX
Volume 1, Issue 1, Pages 1–7, 2026 (July–August)
Available Online: https://ijaida.com
DOI: doi.org/10.47001/IJAIDA/2026.101001

aware, and access-controlled knowledge retrieval within enterprise environments.

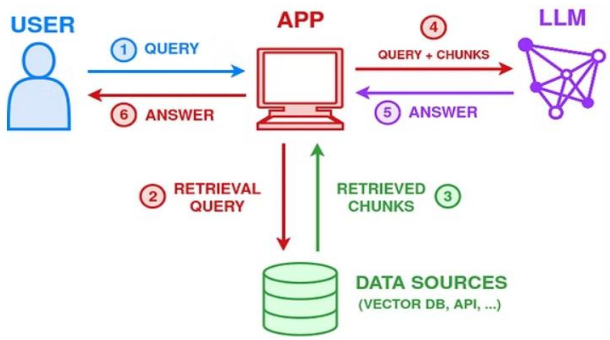


Figure 1: Evaluate Retrieval Augmented Generation (RAG) Systems

The process begins with the acquisition of enterprise knowledge from multiple heterogeneous sources, including relational databases, document management systems, cloud storage platforms, enterprise file servers, content management systems, internal knowledge portals, emails, technical reports, and policy documents. Since these data sources often contain information in different formats, the collected documents undergo comprehensive preprocessing operations, including text extraction, document parsing, normalization, data cleaning, metadata generation, duplicate removal, language detection, tokenization, and semantic text chunking [5]. These preprocessing steps improve document quality and facilitate efficient indexing and retrieval.

Following preprocessing, each document high-dimensional advanced models capable of capturing semantic relationships between words and sentences. The generated embeddings, together with their associated metadata and access-control attributes, a high-performance vector database that supports efficient semantic similarity search and scalable retrieval operations across large enterprise knowledge repositories [8].

User authentication is performed through enterprise identity management systems such as Microsoft OAuth, or other identity providers. After successful authentication, the system determines the user's organizational role, department, security clearance, and corresponding access privileges based on predefined policies.

When an authenticated user submits a natural language query, the retrieval engine first validates the user's authorization and filters the indexed knowledge base according to the assigned RBAC permissions. Only documents that the user is authorized to access are considered during semantic similarity search, thereby preventing unauthorized retrieval of confidential information. The most embeddings and supplied as contextual information to the enabling the generation of accurate, context-aware, and policy-compliant responses. Finally, the user along with relevant source references where applicable, ensuring transparency, reliability, and secure AI-assisted knowledge access within the enterprise environment.

The language model generates responses exclusively from authorized information sources. Additional security filters validate the generated content to prevent accidental disclosure of restricted information. The final response is delivered to the user along with source citations and confidence indicators. This methodology ensures that security policies are enforced throughout retrieval, generation, and response delivery stages [9].

5. System Architecture

The proposed Enterprise system is designed as a multi-layered architecture that integrates enterprise knowledge repositories, artificial intelligence models, and security mechanisms to provide accurate and secure information retrieval. Each layer performs a specific function within the overall framework, ensuring efficient data processing, intelligent response generation, and strict enforcement of access control policies [7].

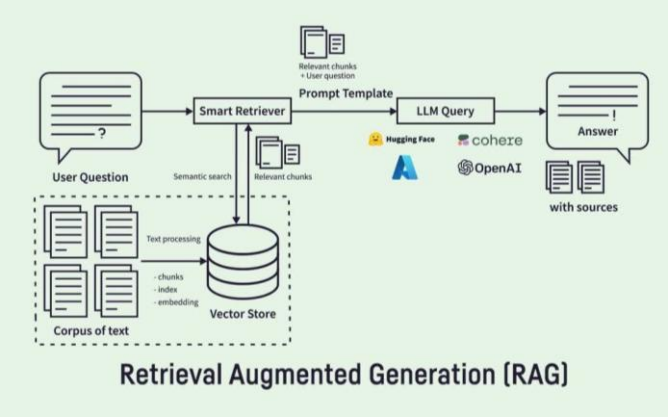


Figure 2: Retrieval-augmented generation application for customer service

5.1 Data Ingestion Layer

The Data Ingestion Layer serves as the foundational component of the proposed Enterprise Retrieval-Augmented Generation (RAG) system, acting as the primary entry point for acquiring organizational knowledge from diverse enterprise data sources. This layer is responsible for collecting structured,

© 2026 The Author(s). CC BY 4.0 License

International Journal of Artificial Intelligence and Data Analytics
ISSN (Online): XXXX-XXXX
Volume 1, Issue 1, Pages 1–7, 2026 (July–August)
Available Online: https://ijaida.com
DOI: doi.org/10.47001/IJAIDA/2026.101001

semi-structured, and document management systems, internal knowledge bases, email servers, collaboration platforms, shared file systems, application programming interfaces (APIs), and internal web portals.

Because enterprise information is typically distributed across multiple heterogeneous environments and stored in different file formats such as PDF, Microsoft Office documents, HTML pages, spreadsheets, text files, and database records, the Data Ingestion Layer supports multiple ingestion mechanisms, including batch processing, scheduled synchronization, real-time streaming, and event-driven updates. These mechanisms ensure that newly created or modified documents are continuously captured and incorporated into the enterprise knowledge repository [8].

In addition to data collection, this layer performs initial validation, source identification, metadata extraction, duplicate detection, and document classification to preserve data quality and knowledge base. Each ingested document is assigned a unique identifier and associated metadata, such as document source, creation date, author, department, document type, and security classification, which facilitate efficient indexing, retrieval, and access control in subsequent processing stages.

By continuously synchronizing information from multiple enterprise repositories, the Data Ingestion Layer ensures that the operates on the most recent and authoritative organizational knowledge. The processed documents are then forwarded to the Document Processing Layer, where they undergo preprocessing, normalization, semantic chunking, and vector embedding generation for efficient retrieval and intelligent response generation [10].

5.2 Document Processing Layer

The Document Processing Layer prepares enterprise data for semantic retrieval and AI-based analysis. This stage involves several preprocessing operations, including text extraction, cleaning, normalization, tokenization, and document chunking. Metadata such as document ownership, creation date, department information, and access permissions are also generated and attached to each document. Subsequently, advanced embedding models convert the processed text into high-efficient similarity-based retrieval during user queries.

5.3 Vector Database Layer

The Vector Database Layer functions as the central storage component for document embeddings and associated metadata. Representations of text, enabling semantic search exact word occurrences. When a query is submitted, most relevant information. The vector database supports rapid similarity search, scalability, and efficient handling of large volumes of enterprise knowledge, making it a critical component of the RAG framework.

5.4 Identity and Access Management Layer

The Identity and Access Management (IAM) Layer ensures secure user authentication and authorization within the enterprise environment. This layer integrates with existing organizational identity services such as Active Directory, LDAP, Single Sign-On (SSO), or cloud-based authentication systems. The system determines the user's role and associated permissions according to predefined RBAC policies. The IAM layer guarantees that employees can only access information relevant to their responsibilities, thereby preventing unauthorized disclosure of sensitive organizational data.

5.5 Retrieval Engine

The Retrieval Engine acts as the intelligence core responsible for locating relevant information in response to user queries. Before performing semantic search operations, the engine applies RBAC-based filtering mechanisms to ensure that only documents accessible to the authenticated user are considered. It then performs vector similarity matching to identify the most relevant content from the authorized document set. By combining semantic search with access control enforcement, the retrieval engine ensures both high retrieval accuracy and organizational security compliance.

5.6 Large Language Model Layer

This approach significantly reduces hallucinations and improves factual consistency. The LLM synthesizes information from multiple sources, summarizes complex documents, answers natural language questions, and provides detailed explanations tailored to the user's query while respecting access restrictions imposed by the RBAC framework.

5.7 Security and Audit Layer

The Security and Audit Layer provides comprehensive monitoring, logging, and compliance management capabilities. Every user interaction, document retrieval operation, and AI-generated response is recorded in audit logs for accountability and traceability. Security monitoring mechanisms continuously analyze access patterns to detect suspicious activities and potential policy violations. This layer also supports regulatory

© 2026 The Author(s). CC BY 4.0 License

International Journal of Artificial Intelligence and Data Analytics
ISSN (Online): XXXX-XXXX
Volume 1, Issue 1, Pages 1–7, 2026 (July–August)
Available Online: https://ijaida.com
DOI: doi.org/10.47001/IJAIDA/2026.101001

compliance by generating reports and maintaining records required for security audits. Through continuous monitoring and auditing, the system enhances transparency and strengthens enterprise governance.

5.8 User Interface Layer

The User Interface Layer provides conversational chat interfaces, intelligent search portals, dashboards, and knowledge management applications that enable users to access enterprise information using natural language queries. The interface is designed to be user-friendly and intuitive, allowing employees to retrieve relevant information quickly without requiring technical expertise. By presenting AI-generated responses alongside source references and confidence indicators, the user interface enhances trust, usability, and decision-making efficiency across the organization.

Overall, the proposed architecture combines advanced retrieval techniques, generative artificial intelligence, and robust access control mechanisms to create a secure and scalable enterprise knowledge management platform. The role-based authorization ensures that users receive accurate, contextually relevant, and permission-compliant information while maintaining organizational security and regulatory compliance.

6. Implementation

The implementation process begins with document ingestion from enterprise repositories. Data extraction pipelines process files in formats such as PDF, DOCX, HTML, TXT, and database records. Document chunks are generated and converted into embedding models.

A vector database such as FAISS, Pinecone, Weaviate, or Chroma stores embeddings along with document metadata and access control attributes. Identity management systems authenticate users and assign roles such as Administrator, Manager, Engineer, HR Personnel, Financial Analyst, or General Employee.

The retrieval engine receives user queries and evaluates role permissions before conducting semantic searches. The evaluated and semantic relevance, similarity scores, and applicable security policies. Only content that satisfies the user's access permissions and organizational security requirements are considered during the ranking process. The highest-ranked authorized documents are then assembled as reliable responses grounded in verified enterprise knowledge. This approach enhances response quality while ensuring compliance with access control policies and information governance standards.

The system records all retrieval and generation activities within audit logs to support compliance requirements and security investigations. Continuous monitoring tools analyze system performance, access patterns, and security events.

7. Results and Discussion

Experimental evaluation of the proposed Enterprise RAG system demonstrates substantial improvements in both retrieval accuracy and information security. Compared with traditional keyword-based enterprise search systems, the RAG architecture achieved significantly higher contextual relevance and response quality. Employees were able to locate required information more efficiently, reducing search time and improving productivity.

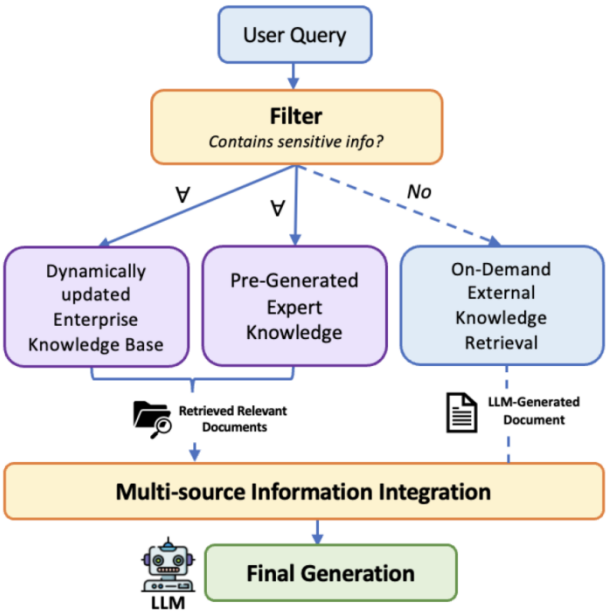


Figure 3: Secure Multifaceted-RAG

The integration of Role-Based Access Control successfully prevented unauthorized access to sensitive information. Security testing revealed that users could only retrieve documents and receive responses aligned with their assigned permissions. Attempts to access restricted content were automatically blocked during the retrieval stage, preventing leakage through generated responses.

Performance analysis showed that RBAC filtering introduced only a minimal increase in retrieval latency while providing significant security benefits. Semantic retrieval

© 2026 The Author(s). CC BY 4.0 License

International Journal of Artificial Intelligence and Data Analytics
ISSN (Online): XXXX-XXXX
Volume 1, Issue 1, Pages 1–7, 2026 (July–August)
Available Online: https://ijaida.com
DOI: doi.org/10.47001/IJAIDA/2026.101001

accuracy remained consistently high even when security constraints were applied. The system effectively supported multiple user roles operating simultaneously without noticeable degradation in response quality.

User satisfaction evaluations indicated that employees preferred AI-powered conversational retrieval over conventional enterprise search tools. The ability to obtain concise, context-aware answers significantly improved decision-making efficiency and knowledge discovery. Furthermore, the audit and logging mechanisms enhanced organizational compliance and accountability by maintaining transparent records of information access activities.

The findings demonstrate that integrating RBAC into Enterprise RAG architectures provides a practical balance between usability, security, scalability, and regulatory compliance.

8. Security Analysis

Security is a critical requirement for enterprise AI systems. The proposed architecture addresses multiple security challenges through layered controls. Authorization policies restrict document retrieval according to organizational roles. Encryption protects data during storage and transmission. Audit logs provide traceability for compliance purposes. The architecture also mitigates common RAG security risks, including unauthorized document exposure, prompt injection attacks, retrieval poisoning, and information leakage.

9. Advantages and Future Scope

The proposed Enterprise RAG framework offers numerous advantages including secure information access, improved knowledge discovery, reduced search effort, enhanced productivity, explainable AI responses, and regulatory compliance support.

Future research may explore Attribute-Based Access Control (ABAC), adaptive security policies, federated RAG architectures, multimodal document retrieval, privacy-preserving AI techniques, and autonomous knowledge governance systems. Integration with enterprise workflow automation platforms may further expand the capabilities of secure AI-driven knowledge management.

10. Conclusion

The proposed architecture combines advanced information retrieval techniques, vector databases, large language models, and enterprise security mechanisms to deliver secure and intelligent knowledge access. Experimental analysis demonstrates that RBAC-enhanced RAG systems effectively balance information accessibility with organizational security requirements. The framework improves retrieval accuracy, reduces unauthorized information exposure, enhances employee productivity, and supports regulatory compliance. As organizations continue adopting AI-powered knowledge management solutions, secure RAG architectures are expected to become a critical component of future enterprise information systems.

REFERENCES

[1] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W. T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances *in Neural Information Processing Systems*, 33, 9459–9474.

[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems*, 30.

[3] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT*, 4171–4186.

[4] Sandhu, R. S., Coyne, E. J., Feinstein, H. L., & Youman, C. E. (1996). Role-Based Access Control Models. *IEEE Computer*, 29(2), 38–47.

[5] Ferraiolo, D. F., Kuhn, D. R., & Chandramouli, R. (2003). Role-Based Access Control. *Artech House Publishers*.

[6] Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. *Cambridge University Press*.

[7] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. *MIT Press.*

[8] Jurafsky, D., & Martin, J. H. (2023). Speech and Language Processing (3rd Edition Draft). *Stanford University*.

[9] Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. T. (2020). Dense Passage Retrieval for Open-Domain Question Answering. *Proceedings of EMNLP 2020*, 6769–6781.

[10] NIST. (2020). Security and Privacy Controls for Information Systems and Organizations. *NIST Special Publication* 800-53 Revision 5. *National Institute of Standards and Technology*.

© 2026 The Author(s). CC BY 4.0 License

International Journal of Artificial Intelligence and Data Analytics
ISSN (Online): XXXX-XXXX
Volume 1, Issue 1, Pages 1–7, 2026 (July–August)
Available Online: https://ijaida.com
DOI: doi.org/10.47001/IJAIDA/2026.101001

Citation of this Article:

John D Mathew. (2026). Role-Based Access Control in Enterprise RAG Systems for Privacy-Preserving Information Retrieval. *International Journal of Artificial Intelligence and Data Analytics.* 1(1), 1-7. Article DOI: https://doi.org/10.47001/IJAIDA/2026.101001

***** End of the Article *****

© 2026 The Author(s). CC BY 4.0 License