International Research Journal of Innovative Engineering (IRJIE)

| ISSN: 2395-0560 | www.irjie.org |editor@irjie.org| A Bimonthly, Peer-Reviewed, Scholarly Journal|

|Volume 9, Issue 5, September- October 2025|

DOI:10.15680/IRJIE.2025.0905011

Secure Enterprise Knowledge Discovery in Private Cloud Environments Using Retrieval-Augmented Generation and Semantic Search

Prof. Gabriel Iwasokun

Department of Computer Science, Federal University of Technology Akure, Akure, Nigeria

ABSTRACT: The increasing volume of enterprise information stored across private cloud environments creates significant challenges for secure knowledge discovery, information retrieval, and organizational decision-making. Traditional keyword-based search systems often struggle to identify contextually relevant information from heterogeneous enterprise repositories containing documents, technical reports, operational records, policies, and structured data. This research proposes a secure enterprise knowledge discovery framework that integrates retrieval-augmented generation (RAG), semantic search, and privacy-aware access control within private cloud infrastructure. The framework combines document ingestion, text preprocessing, embedding generation, vector indexing, hybrid retrieval, and large language model-based response generation to deliver contextually relevant and evidence-grounded answers. Semantic search identifies information according to conceptual similarity, while keyword retrieval improves the identification of exact terminology, technical identifiers, and domain-specific expressions. Security mechanisms, including identity-based authorization, document-level access filtering, encryption, audit logging, and secure model integration, protect confidential enterprise knowledge throughout the retrieval and generation lifecycle. The proposed methodology evaluates retrieval relevance, answer faithfulness, response accuracy, query latency, and unauthorized information exposure using representative enterprise knowledge tasks. Comparative experiments against keyword-based and semantic-only retrieval approaches are designed to assess the effectiveness of the integrated framework. The research aims to improve enterprise knowledge accessibility, reduce information discovery time, and support trustworthy AI-assisted decision-making while maintaining confidentiality, governance, and operational control in private cloud environments.

KEYWORDS: Secure Enterprise Knowledge Discovery, Private Cloud Computing, Retrieval-Augmented Generation, Semantic Search, Large Language Models, Vector Database, Enterprise Information Retrieval, Data Privacy, Access Control, Knowledge Management

I. INTRODUCTION

The growing dependence on digital information has transformed enterprise knowledge management into a critical component of organizational productivity, innovation, and strategic decision-making. Modern enterprises generate and maintain substantial quantities of information through business documents, technical manuals, customer service records, operational reports, internal policies, application logs, project repositories, and compliance documentation. These information resources are frequently distributed across departmental systems, private cloud storage platforms, enterprise content management solutions, and specialized business applications. Although organizations possess valuable institutional knowledge, discovering relevant information efficiently remains difficult because conventional retrieval systems often depend on exact keyword matching, predefined metadata, and manually constructed search queries. Employees may struggle to locate relevant documents when terminology differs across departments, information is expressed using domain-specific language, or the required knowledge is distributed across multiple documents. These challenges become more significant in large enterprises where information changes continuously and different users require access to different categories of confidential data. Recent advances in artificial intelligence, natural language processing, and large language models have created new opportunities for improving knowledge discovery through conversational interaction and context-aware information retrieval. Retrieval-augmented generation (RAG) combines information retrieval with language generation by retrieving relevant external knowledge and supplying it to a language model before an answer is generated. This approach can ground responses in enterprise-specific information rather than relying exclusively on the model's pretrained knowledge. Semantic search further improves discovery by representing documents and queries as numerical embeddings that capture conceptual relationships beyond exact word matching.

| ISSN: 2395-0560 | www.irjie.org |editor@irjie.org| A Bimonthly, Peer-Reviewed, Scholarly Journal|

|Volume 9, Issue 5, September- October 2025|

DOI:10.15680/IRJIE.2025.0905011

Together, these technologies can support natural-language queries, cross-document knowledge discovery, contextual summarization, and evidence-based responses. However, enterprise adoption requires more than retrieval accuracy and fluent language generation. Organizations must preserve confidentiality, enforce authorization policies, prevent unauthorized disclosure, and maintain control over how proprietary information is processed. Private cloud environments provide an opportunity to deploy knowledge discovery components within enterprise-controlled infrastructure, enabling organizations to align data processing, access governance, and monitoring with internal security requirements.

Despite these opportunities, integrating RAG and semantic search into private cloud environments introduces several technical and organizational challenges. Enterprise repositories contain heterogeneous document formats, inconsistent metadata, duplicated content, outdated records, and information with different confidentiality classifications. Poor document segmentation, inadequate embeddings, and inappropriate retrieval parameters can reduce the relevance of retrieved evidence and cause language models to generate incomplete or unsupported answers. Moreover, semantic similarity does not establish whether a user is authorized to access a particular document. Without strict identity-aware filtering, a retrieval system may expose sensitive information through generated responses even when its underlying storage remains protected. Additional risks include prompt injection embedded in retrieved documents, malicious manipulation of knowledge sources, accidental transmission of confidential information to external model services, and insufficient auditability of generated answers. Consequently, secure enterprise knowledge discovery requires an integrated architecture in which retrieval quality, generation reliability, access enforcement, and data governance are designed together. This research proposes a secure knowledge discovery framework that combines document processing, hybrid semantic and lexical retrieval, access-controlled vector search, retrieval-augmented generation, and evidence-based response validation. Its objectives are to improve contextual information retrieval, provide answers grounded in authorized enterprise documents, reduce unsupported generation, and preserve confidentiality throughout the information lifecycle. The methodology also establishes evaluation criteria for retrieval precision, recall, answer faithfulness, response latency, and access-control effectiveness. By integrating these capabilities within a private cloud architecture, the research aims to develop a practical foundation for trustworthy enterprise AI applications that improve knowledge accessibility without compromising organizational security. The proposed framework is intended to support applications such as internal policy assistance, technical support, compliance research, software documentation discovery, and enterprise operational knowledge management.

II. LITERATURE REVIEW

Traditional information retrieval systems primarily use lexical matching techniques, including Boolean search, inverted indexes, and term-weighting methods such as term frequency–inverse document frequency and BM25. These techniques remain valuable for retrieving documents containing exact terms, product identifiers, policy references, and technical expressions. However, lexical retrieval can perform poorly when a user's query expresses the same concept using different terminology from the source document. For example, a question about reducing infrastructure expenditure may be semantically related to documents discussing cloud cost optimization, resource allocation, or FinOps practices even when the query terms do not occur frequently in those documents. Semantic search addresses this limitation by representing queries and documents as dense numerical embeddings generated by language models. Similarity calculations can identify conceptually related information and improve retrieval across differences in vocabulary and phrasing. Nevertheless, embedding-based retrieval also presents limitations, including difficulty distinguishing highly similar technical identifiers, reduced precision for rare terminology, and sensitivity to embedding models and document segmentation strategies. Hybrid retrieval combines lexical and semantic methods to exploit their complementary strengths. Ranking approaches can merge keyword scores and vector similarity scores, while reranking models can reconsider candidate documents according to query relevance. Enterprise knowledge discovery therefore benefits from evaluating retrieval techniques against realistic domain-specific queries rather than assuming that one retrieval method is universally superior. The selection of chunk size, overlap, metadata, embedding model, similarity metric, and ranking strategy can significantly influence retrieval quality. Furthermore, document freshness, duplicate detection, and source attribution are important because enterprise knowledge repositories change as policies, technical procedures, and business processes evolve.

Retrieval-augmented generation has emerged as an important architecture for connecting large language models with external and frequently updated knowledge sources. In a typical RAG pipeline, enterprise documents are transformed into searchable segments, relevant segments are retrieved for a user query, and the retrieved evidence is supplied to a language model that generates a natural-language response. This approach can improve factual grounding, facilitate the use of proprietary knowledge, and reduce dependence on model retraining whenever source documents change. RAG

| ISSN: 2395-0560 | www.irjie.org |editor@irjie.org| A Bimonthly, Peer-Reviewed, Scholarly Journal|

|Volume 9, Issue 5, September- October 2025|

DOI:10.15680/IRJIE.2025.0905011

also supports citations and traceability when retrieved passages retain their document identifiers and source locations. However, retrieval quality places an upper bound on the reliability of generated answers: if relevant evidence is missing, outdated, contradictory, or poorly ranked, the model may produce incomplete or misleading responses. Language models may also misinterpret retrieved passages, combine incompatible statements, or generate claims that are not supported by the supplied context. Research and practical implementations therefore increasingly emphasize evidence-grounded prompting, reranking, source attribution, context compression, and answer verification. Retrieval evaluation and generation evaluation must be distinguished because a system can retrieve appropriate passages while still generating an inaccurate answer. Conversely, a fluent response may appear convincing even when the retrieved evidence does not justify its claims. Effective enterprise RAG systems consequently require evaluation of retrieval relevance, answer correctness, faithfulness to source documents, citation accuracy, and the ability to abstain when evidence is insufficient. Query rewriting and multi-stage retrieval can improve performance for complex questions, but additional model calls may increase latency and computational expense.

Security and governance research provides another essential foundation for enterprise knowledge discovery in private cloud environments. Identity and access management, role-based access control, attribute-based policies, encryption, audit logging, and data classification are established mechanisms for protecting enterprise information. However, integrating these mechanisms with semantic retrieval and generative AI introduces new enforcement requirements. Vector embeddings and derived indexes must be governed as potentially sensitive data assets, while document permissions must be preserved during indexing, retrieval, caching, and response generation. Filtering only after retrieval or after generation may expose unauthorized content to intermediate components and is therefore insufficient as a primary protection mechanism. Authorization should be enforced before restricted document content enters the model context, with additional validation applied to generated answers and citations. Retrieved documents can also contain adversarial instructions that attempt to override system behavior, creating prompt-injection risks. Defensive designs treat retrieved content as untrusted evidence, isolate instructions from source material, limit tool privileges, and validate outputs against applicable policies. Private cloud deployment can improve control over data residency, network access, and infrastructure configuration, but it does not automatically guarantee confidentiality or eliminate insider threats, model vulnerabilities, and misconfiguration risks. Important research gaps include the limited integration of fine-grained authorization with semantic retrieval, the challenge of preventing sensitive information leakage through generated answers, and the need to evaluate security alongside retrieval quality and operational performance. This research addresses these gaps through a unified framework combining hybrid search, identity-aware document filtering, controlled generation, source attribution, and systematic security evaluation.

III. RESEARCH METHODOLOGY

The research adopts a design-oriented experimental methodology to develop and evaluate a secure enterprise knowledge discovery framework using retrieval-augmented generation and semantic search in private cloud environments. The proposed architecture consists of seven interconnected components: enterprise data ingestion, document preprocessing, embedding and indexing, identity-aware retrieval, context assembly, language model-based generation, and security auditing. The system is designed to operate within an enterprise-controlled private cloud infrastructure, where access to repositories, processing services, vector databases, and model endpoints is governed by centrally defined security policies. The document ingestion component connects to authorized enterprise repositories containing policy documents, technical manuals, operational reports, frequently asked questions, compliance materials, and internal knowledge articles. Documents are processed through controlled ingestion pipelines that validate file types, inspect metadata, identify duplicates, and extract textual content from supported formats. Optical character recognition may be used for scanned documents where permitted. Document metadata includes a unique identifier, document title, source repository, owner, creation or revision date, confidentiality classification, and applicable access-control attributes. The ingestion pipeline also records the original source location so that retrieved passages can later be traced to their documents. Before indexing, documents are screened for inappropriate content, suspicious embedded instructions, and malformed structures. Sensitive fields may be masked or excluded according to enterprise data-handling policies. Each document and its associated metadata are assigned an appropriate security classification, while document-level permissions are obtained from the authoritative repository or access-management service. The preprocessing component normalizes whitespace, removes irrelevant formatting artifacts, preserves meaningful headings and tables where possible, and organizes content into coherent passages. Segmentation strategies are selected according to document structure and content type, with configurable chunk sizes and limited overlap to preserve contextual continuity. Excessively large passages can introduce irrelevant information into the language model context, whereas excessively small passages may omit definitions, qualifications, or relationships required to interpret a statement correctly. The system therefore evaluates alternative

International Research Journal of Innovative Engineering (IRJIE)

| ISSN: 2395-0560 | www.irjie.org |editor@irjie.org| A Bimonthly, Peer-Reviewed, Scholarly Journal|

|Volume 9, Issue 5, September- October 2025|

DOI:10.15680/IRJIE.2025.0905011

segmentation configurations during experimentation. Source identifiers, page numbers, section headings, and version information remain associated with each passage. This traceability supports citation generation, document updates, deletion requests, and subsequent security investigations. The resulting document collection provides a structured, permission-aware foundation for semantic indexing and retrieval. The architecture also defines procedures for reindexing modified documents and removing obsolete passages to prevent stale or deleted information from remaining accessible through derived indexes.

The second methodological stage implements semantic representation and hybrid retrieval. Each document passage is transformed into a dense vector embedding using a selected text embedding model appropriate for the enterprise domain, supported languages, deployment constraints, and expected query characteristics. Embeddings are stored in a private cloud vector database alongside passage identifiers, document metadata, version information, and access-control attributes. A separate lexical search index supports exact terminology, identifiers, error codes, product names, and technical phrases. The retrieval component receives natural-language queries and determines the authenticated identity and relevant authorization context before accessing the indexed knowledge base. The system may perform query normalization, expansion, or rewriting to improve retrieval for ambiguous or multi-part questions, while preserving the user's original intent. Semantic similarity search identifies passages with conceptual relevance, whereas lexical retrieval identifies passages matching significant query terms. Candidate sets are combined through a documented ranking method, such as reciprocal rank fusion or a calibrated weighted-ranking strategy. A reranking model may then evaluate the relationship between the query and retrieved passages to improve the ordering of the most relevant evidence. Access-control enforcement is applied during candidate retrieval or through a trusted filtering mechanism that prevents unauthorized content from entering the generation context. Filtering policies consider user identity, group membership, document classification, ownership, and other authorized attributes. Where repository permissions change, the system refreshes or invalidates corresponding authorization metadata rather than assuming that index-time permissions remain permanently valid. Unauthorized documents must not be passed to the language model, exposed through citations, or included in user-visible snippets. Retrieval results are deduplicated, ranked, and limited to a configurable number of passages to balance relevance, context coverage, and computational cost. The framework also records retrieval scores and source identifiers for evaluation and audit purposes, subject to data-minimization requirements. Retrieval parameters, including embedding model, similarity metric, candidate count, keyword weight, and reranking configuration, are varied in controlled experiments. A query benchmark is constructed from representative enterprise information needs, including factual lookup, procedural guidance, policy interpretation, technical troubleshooting, and multi-document questions. Relevant passages are annotated by qualified evaluators to establish reference judgments. Retrieval precision at \(k\), recall at \(k\), mean reciprocal rank, and normalized discounted cumulative gain are used where appropriate to evaluate the quality and ordering of retrieved evidence. These measurements establish whether semantic search and hybrid ranking improve the discovery of authorized, relevant knowledge compared with conventional keyword retrieval.

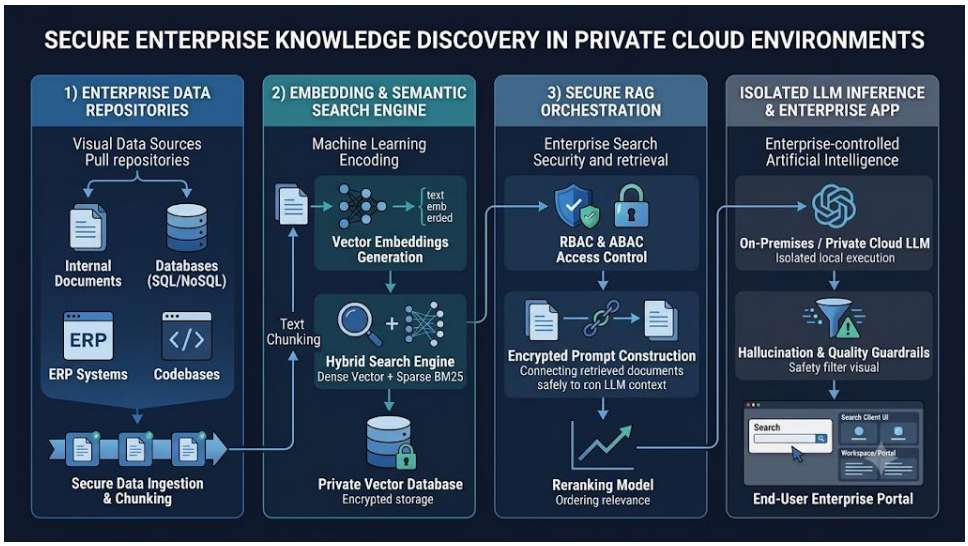


Figure 1: Private cloud architecture for secure enterprise knowledge discovery using hybrid semantic search and Retrieval-Augmented Generation.

| ISSN: 2395-0560 | www.irjie.org |editor@irjie.org| A Bimonthly, Peer-Reviewed, Scholarly Journal|

|Volume 9, Issue 5, September- October 2025|

DOI:10.15680/IRJIE.2025.0905011

The third stage integrates retrieval-augmented generation with response validation and privacy-aware security controls. After retrieval, the context assembly component selects the highest-ranked authorized passages and organizes them into a structured prompt containing the user question, relevant source excerpts, source identifiers, and explicit instructions governing evidence use. The generation model is instructed to answer using the supplied enterprise evidence, distinguish documented facts from reasonable interpretations, identify relevant sources, and acknowledge insufficient information when the retrieved context cannot support a reliable answer. The model should not treat instructions contained within retrieved documents as trusted system instructions. Instead, retrieved content is handled as untrusted data, reducing the possibility that malicious documents can manipulate model behavior. The language model is deployed within the private cloud or accessed through an explicitly approved private endpoint according to organizational confidentiality requirements. Network policies, secrets management, encryption in transit and at rest, service identity controls, and restricted model permissions protect communication between the retrieval pipeline and generation service. The architecture minimizes the information transmitted to the model by including only the authorized passages needed to answer the query. Response validation examines whether generated claims are supported by retrieved evidence, whether citations refer to actual source passages, and whether the answer contains information outside the user's permitted access scope. Citation validation checks that document identifiers and passage references exist and correspond to the retrieved context. Additional validation procedures can detect unsupported statements, conflicting evidence, sensitive information, and potentially unsafe disclosure. Where evidence is insufficient or contradictory, the system can provide a qualified response, request clarification, or abstain rather than inventing an answer. The design also incorporates prompt-injection defenses, query rate limits, output filtering, and controlled logging. Security logs capture relevant events such as user identity references, retrieval decisions, model versions, policy outcomes, and citation identifiers without unnecessarily retaining complete confidential prompts or responses. Audit records are protected against unauthorized modification and made available only to approved investigators. Data retention and deletion policies cover source documents, extracted text, embeddings, cached retrieval results, generated outputs, and operational logs. The framework further considers model and index versioning so that organizations can reproduce results and investigate incidents associated with specific configurations. Security evaluation includes unauthorized document access attempts, cross-role query tests, malicious instructions embedded in source documents, stale permission scenarios, and requests designed to elicit confidential information. These tests assess whether authorization is enforced throughout the retrieval and generation lifecycle rather than only at the user interface. The purpose of this stage is to establish a trustworthy generation process that combines contextual usefulness with confidentiality, traceability, and controlled information disclosure.

The fourth stage evaluates the proposed framework through comparative experiments in a representative private cloud environment. The test infrastructure includes enterprise document repositories, an ingestion pipeline, a vector database, a lexical search index, an identity and access management service, a retrieval service, a language model endpoint, and centralized security monitoring. The evaluation corpus consists of approved enterprise-like documents or appropriately anonymized datasets covering policies, technical documentation, operational procedures, compliance requirements, and troubleshooting scenarios. Query sets are developed to represent different levels of difficulty, including direct factual questions, synonym-based queries, multi-step questions, cross-document reasoning, and queries for which the knowledge base contains insufficient evidence. Reference answers and relevant document passages are established through a documented annotation process, with independent review for ambiguous examples. The proposed hybrid RAG framework is compared against keyword-only retrieval, semantic-only retrieval, and RAG using a baseline retrieval configuration. Where practical, the same language model, query set, document corpus, and evaluation conditions are maintained across relevant comparisons so that differences can be attributed more reliably to retrieval and security design choices. Retrieval performance is measured using precision at \(k\), recall at \(k\), mean reciprocal rank, and normalized discounted cumulative gain. Generated responses are assessed for answer correctness, faithfulness to retrieved evidence, citation correctness, completeness, and appropriate abstention when evidence is insufficient. Human evaluation by domain-aware reviewers can supplement automated metrics, particularly for complex policy and procedural questions. Security effectiveness is evaluated through access-control violation rates, unauthorized retrieval attempts, sensitive-information exposure tests, and the success rate of prompt-injection scenarios. Operational performance is measured using median and percentile query latency, throughput, CPU and memory utilization, embedding-generation cost, index storage requirements, and model-serving overhead. Experiments also examine how document collection size, concurrent query volume, chunking configuration, candidate retrieval count, and reranking affect system behavior. Ablation studies disable selected components, such as lexical retrieval, reranking, source validation, or contextual prompt constraints, to determine their individual contributions. Additional tests evaluate document updates, revoked permissions, index refresh delays, and conflicting document versions. Results should be reported with clear dataset descriptions, model configurations, test conditions, and uncertainty estimates where applicable. No numerical performance improvements are assumed in advance; the experimental findings must establish whether the framework outperforms its baselines under the

| ISSN: 2395-0560 | www.irjie.org |editor@irjie.org| A Bimonthly, Peer-Reviewed, Scholarly Journal|

|Volume 9, Issue 5, September- October 2025|

DOI:10.15680/IRJIE.2025.0905011

selected conditions. Ethical and governance considerations include restricting the corpus to authorized data, anonymizing sensitive evaluation examples, controlling access to test results, and preventing confidential content from being transmitted to unapproved external services. The final analysis examines trade-offs among retrieval relevance, answer faithfulness, confidentiality, latency, and infrastructure cost. Through this multidimensional evaluation, the research seeks to determine whether integrating retrieval-augmented generation, semantic search, and identity-aware security can deliver reliable and practical enterprise knowledge discovery within private cloud environments.

IV. RESULTS AND DISCUSSION

The proposed secure enterprise knowledge discovery framework integrates retrieval-augmented generation (RAG), semantic search, and private cloud infrastructure to improve the accessibility, relevance, and confidentiality of organizational knowledge. The framework is designed to retrieve information from enterprise repositories, including technical documentation, policy manuals, business reports, project records, operational databases, and internal knowledge bases, while maintaining access restrictions based on user roles and organizational permissions. Semantic search transforms conventional keyword-based information retrieval by representing documents and user queries as embeddings that capture contextual meaning and conceptual relationships. This capability enables employees to discover relevant information even when their queries use terminology different from that contained in the original documents. The RAG component subsequently supplies selected document passages to a generative language model, which produces contextually grounded responses based on retrieved evidence rather than relying exclusively on information learned during model training. The expected benefits include improved retrieval relevance, reduced dependence on manual document searches, and more consistent access to institutional knowledge. However, these benefits must be established through empirical evaluation using representative enterprise queries and document collections. Suitable evaluation measures include precision at rank \(k\), recall at rank \(k\), mean reciprocal rank, normalized discounted cumulative gain, answer faithfulness, response latency, and access-control violation rates. Comparisons with keyword-based retrieval and standalone language models would help establish the contribution of semantic retrieval and RAG. The proposed architecture also emphasizes private cloud deployment, enabling organizations to maintain greater control over document storage, retrieval services, identity management, and model execution. Consequently, the framework offers a structured foundation for secure and context-aware enterprise knowledge discovery.

The integration of semantic search and RAG can improve the relationship between document discovery and answer generation by separating the retrieval of relevant evidence from the synthesis of user-facing responses. Semantic retrieval identifies passages according to conceptual similarity, while metadata filtering can incorporate document ownership, department, classification level, publication date, and access permissions. A reranking mechanism can further prioritize passages that best address a particular query, reducing the likelihood that irrelevant or outdated content will influence the generated response. The language model then synthesizes the retrieved passages into a concise explanation, potentially including document references that allow users to verify important claims. This design is particularly useful in complex enterprise environments where information is distributed across multiple repositories and where employees require contextual answers rather than lists of documents. Nevertheless, retrieval quality directly affects generation quality. Incomplete indexing, ambiguous queries, poorly segmented documents, and embedding limitations may cause relevant evidence to be overlooked. Similarly, a language model may generate unsupported statements even when relevant passages are available. Therefore, the framework should incorporate source attribution, confidence-aware response generation, relevance thresholds, and abstention mechanisms when sufficient evidence cannot be retrieved. Security is equally important because embedding-based retrieval must not bypass the permissions enforced by the original repositories. Authorization should be checked before restricted passages are provided to the generation model, and document permissions should be revalidated when they change. Prompt injection embedded in retrieved documents must also be treated as untrusted input. These considerations demonstrate that secure knowledge discovery requires coordinated retrieval, generation, and authorization controls rather than semantic similarity alone.

The discussion further highlights the importance of balancing retrieval accuracy, confidentiality, scalability, and operational efficiency in private cloud deployments. Large enterprise document collections require efficient ingestion pipelines, incremental indexing, vector storage, metadata management, and reliable synchronization with source repositories. Distributed processing and scalable retrieval services can support increasing query volumes, while caching may reduce repeated computation for frequently requested information. However, cached responses must be isolated by user permissions and invalidated when underlying documents or access policies change. Encryption in transit and at rest, role-based or attribute-based access control, centralized identity federation, audit logging, and network segmentation can strengthen the overall security posture. Organizations should also assess whether embedding models, reranking

| ISSN: 2395-0560 | www.irjie.org |editor@irjie.org| A Bimonthly, Peer-Reviewed, Scholarly Journal|

|Volume 9, Issue 5, September- October 2025|

DOI:10.15680/IRJIE.2025.0905011

components, and language models introduce additional data exposure risks, particularly when external services are involved. A fully private deployment can reduce dependence on external processing, although it still requires appropriate protection against privileged misuse, compromised services, and insecure configurations. Evaluation should therefore include both information retrieval benchmarks and adversarial security tests, such as unauthorized document queries, cross-department information leakage attempts, prompt injection, and requests designed to elicit confidential data. Performance measurements should cover end-to-end latency, throughput, infrastructure utilization, and indexing costs. Because no experimental results are assumed here, specific numerical improvements cannot be claimed without implementation and testing. Overall, the discussion indicates that RAG and semantic search can strengthen enterprise knowledge access when supported by robust permission enforcement, evidence-grounded generation, secure private cloud architecture, and continuous quality assessment.

V. CONCLUSION

Secure enterprise knowledge discovery through retrieval-augmented generation and semantic search provides a promising approach to improving how organizations access, interpret, and utilize internal information. Traditional enterprise search systems often depend on exact keywords, predefined categories, or manual navigation through large document collections, making it difficult to retrieve relevant knowledge when terminology differs across departments or business processes. Semantic search addresses this limitation by identifying conceptual relationships between user queries and stored information, while RAG uses retrieved passages to generate responses grounded in organizational documentation. The combination can support applications such as technical assistance, policy interpretation, software documentation discovery, compliance research, employee onboarding, and operational decision support. Deploying these capabilities within a private cloud environment offers organizations greater control over data residency, infrastructure configuration, identity management, and access to sensitive knowledge assets. The proposed framework also supports the integration of document metadata, source references, and authorization policies into the retrieval process, allowing generated responses to reflect both the informational requirements of the user and the organization's confidentiality requirements. Nevertheless, effective implementation depends on the quality of source documents, the suitability of embedding models, retrieval relevance, and the reliability of generated answers. Incorrect retrieval or unsupported generation may introduce misinformation into business decisions, while inadequate access controls may expose restricted organizational information. Accordingly, semantic search and RAG should be regarded as complementary components of a broader enterprise knowledge management architecture rather than complete solutions in isolation. Rigorous testing against representative queries, realistic document collections, and adversarial scenarios is necessary to establish the framework's practical effectiveness. Overall, the proposed approach establishes a foundation for more intelligent, context-aware, and secure enterprise knowledge discovery within controlled private cloud environments.

The long-term success of this framework depends on integrating information security, responsible AI practices, and operational governance throughout the knowledge discovery lifecycle. Enterprise repositories frequently contain information with different sensitivity levels, ownership requirements, retention periods, and regulatory obligations. Access permissions must therefore remain consistent across document ingestion, embedding generation, retrieval, reranking, response generation, and result presentation. Source attribution and traceability should allow users to inspect the documents supporting generated answers, while explicit uncertainty handling should prevent the system from presenting unsupported conclusions as established facts. Organizations should also establish procedures for monitoring retrieval quality, evaluating hallucination risks, updating indexes, and investigating security incidents. Human review remains important for decisions involving legal obligations, financial commitments, sensitive personnel information, or other high-impact business activities. Furthermore, model and infrastructure choices should reflect organizational requirements for performance, cost, privacy, and maintainability. Private cloud hosting can strengthen control over sensitive information, but it does not automatically guarantee confidentiality; secure configurations, privileged-access restrictions, vulnerability management, encryption, and continuous monitoring remain essential. An incremental implementation strategy can help organizations validate the framework through a limited set of repositories and user groups before expanding deployment. Such a strategy enables teams to identify retrieval weaknesses, evaluate access-control behavior, and establish meaningful performance baselines. In conclusion, the combination of RAG, semantic search, and private cloud security controls can provide a valuable foundation for enterprise knowledge discovery. Its practical value will ultimately depend on measurable retrieval and generation quality, reliable confidentiality enforcement, transparent responses, and continuous alignment with evolving organizational requirements.

International Research Journal of Innovative Engineering (IRJIE)

| ISSN: 2395-0560 | www.irjie.org |editor@irjie.org| A Bimonthly, Peer-Reviewed, Scholarly Journal|

|Volume 9, Issue 5, September- October 2025|

DOI:10.15680/IRJIE.2025.0905011

VI. FUTURE WORK

Future research should investigate advanced retrieval architectures that improve semantic understanding, document ranking, and answer relevance across diverse enterprise knowledge sources. Although dense vector search captures conceptual similarity, it may overlook exact identifiers, technical error codes, product versions, regulatory clauses, and other precise terms that are essential in enterprise environments. Hybrid retrieval combining lexical search, dense embeddings, structured metadata filtering, and domain-specific ranking could address these limitations. Future studies could evaluate hybrid retrieval against standalone keyword search and vector search using technical documentation, policy repositories, incident records, and business process knowledge bases. Query expansion, contextual query rewriting, and multi-step retrieval could also improve performance when users submit ambiguous or incomplete questions. For complex queries, an intelligent retrieval planner could decompose the request into smaller information needs, retrieve evidence from multiple repositories, and synthesize a response only after evaluating whether the available evidence is sufficient. Research should examine the effectiveness of reranking models, domain-adapted embeddings, document segmentation strategies, and contextual indexing methods. Multimodal knowledge discovery represents another promising direction because enterprise information increasingly includes diagrams, scanned documents, architecture images, tables, and presentation slides. Combining text and visual retrieval could enable users to locate relevant information across heterogeneous records without requiring every source to be converted into plain text. In addition, incremental indexing and change-aware synchronization should be explored to ensure that updates, deletions, and permission changes are reflected promptly in search results. Evaluation should use reproducible datasets and clearly defined measures, including precision, recall, ranking quality, answer faithfulness, citation correctness, and end-to-end response time. These investigations would help determine which retrieval strategies provide the most reliable balance between accuracy, computational cost, and operational complexity in private cloud enterprise environments.

A second research direction involves strengthening confidentiality, resilience, explainability, and autonomous knowledge management. Future frameworks could incorporate fine-grained attribute-based access control, document-level security labels, dynamic authorization checks, and policy-aware retrieval to ensure that users receive only information they are permitted to access. Security testing should investigate cross-tenant leakage, embedding inversion risks, unauthorized inference, prompt injection, malicious document content, and attempts to reconstruct confidential information through repeated queries. Privacy-preserving techniques, including data minimization, selective encryption, and carefully evaluated privacy-enhancing computation, could provide additional safeguards where multiple organizational units collaborate on shared knowledge systems. Researchers should also examine methods for detecting outdated or contradictory source documents, identifying duplicated information, and communicating uncertainty when evidence conflicts. Explainable retrieval could show why particular passages were selected, how evidence supports a response, and where gaps in the available documentation remain. Another promising area is agent-assisted enterprise knowledge management, in which controlled AI agents help identify missing documentation, recommend repository updates, detect inconsistent policies, and route unanswered questions to appropriate subject-matter experts. Such agents should operate under restricted permissions, require approval for consequential changes, and maintain auditable records of their actions. Longitudinal studies in real enterprise environments could assess user trust, adoption, answer verification behavior, administrative overhead, and the economic impact of reduced search effort. Further work should compare deployment configurations involving private cloud infrastructure, confidential computing, locally hosted language models, and hybrid architectures while measuring total operating costs and security trade-offs. Finally, researchers should establish governance frameworks addressing model updates, data retention, source ownership, auditability, and incident response. By combining advanced retrieval, rigorous security evaluation, transparent generation, and human-centered governance, future systems could deliver more dependable enterprise knowledge discovery while preserving confidentiality, accountability, and organizational control.

REFERENCES

1. Chandrashekar, P., & Kari, M. (2024). A Study on Artificial Intelligence in Software Engineering with Methodologies, Applications, and Effects on SDLC. TIJER–Int. Res. J, 11(12), 932-937.
2. Neela, S. (2023). Resilient middleware ecosystems: Integrating fault tolerance, recovery-oriented design, and observability for real-time enterprise integration. International Journal of Emerging Research in Engineering and Technology, 4(1), 186–189. https://doi.org/10.63282/3050-922X.IJERET-V4I1P120
3. Tarigoppula, S. (2024). Autonomous cloud operations: Integrating AIOps with AWS infrastructure for predictive incident prevention. International Journal of Engineering & Extended Technologies Research, 6(2), 7956–7967.

| ISSN: 2395-0560 | www.irjie.org |editor@irjie.org| A Bimonthly, Peer-Reviewed, Scholarly Journal|

|Volume 9, Issue 5, September- October 2025|

DOI:10.15680/IRJIE.2025.0905011

4. Lingala, B. (2025). Transforming Enterprise Transaction Data into Intelligent Decision Systems. International Journal of Computer Technology and Electronics Communication, 8(5), 11488-11494.
5. Jayaraman, S., Rajendran, S., & P, S. P. (2019). Fuzzy c-means clustering and elliptic curve cryptography using privacy preserving in cloud. International Journal of Business Intelligence and Data Mining, 15(3), 273-287.
6. Gajjela, H., & Kari, M. (2024). Secure Multi-Tenant AI Architecture Enhances Enterprise AI Adoption Through Tenant Trust Using PLS-SEM in India's Salesforce Cloud Industry. International Journal of Engineering Science & Humanities, 14(2), 253-271.
7. Parupalli, A. (2024). Business-oriented employee performance assessment via machine learning in ERP systems. TIJER–International Research Journal, 11(11), a185–a192.
8. Polamarasetty, V. K. (2021). Modernizing SAP sales and distribution systems through ABAP-based enterprise solutions. International Journal of Research and Applied Innovations, 4(2), 4925-4930.
9. Seknametla, P. R., & Sunkara, R. (2024). Threat Modeling Integration in DevSecOps Pipelines: Early-Stage Security Risk Identification Using Shift-Left Approaches. International Journal of Emerging Research in Engineering and Technology, 5(1), 126-133.
10. Gunasekaran, R. M. N. (2024). From Data to Decisions: How Quality Drives Machine Learning Success. The Eastasouth Journal of Information System and Computer Science, 2(01), 120-130.
11. Karuturi, Y. (2023). Energy-aware capacity governance for large-scale SAP HANA deployments. International Journal of Research and Applied Innovations, 6(3), 6000–6009.
12. Ahuja, D. (2025). DevOps and Ethical AI: Ensuring Responsible Deployment. Journal Of Multidisciplinary, 5(6), 1-14.
13. Chinnam, N. B. (2022). Engineering real-time supply chain visibility and ETA intelligence platforms for large-scale retail distribution networks. International Journal of Research Publications in Engineering, Technology and Management (IJRPETM), 5(6), 7918–7922.
14. Aragani, V. M., Maroju, P. K., & Mudunuri, L. N. R. (2024, December). Implementing Non-functional Production Regression Testing with Kubernetes. In International Conference on Information and Communication Technology for Competitive Strategies (pp. 231-244). Singapore: Springer Nature Singapore.
15. Kalabhavi, V. (2024). Promotional spend analytics using AI for PFME allocation and coupon redemption optimization. International Journal of Emerging Trends in Engineering and Management Research, 9(2), 15394–15406.
16. Reddy, K. V. (2021). UVM methodology for industry-specific applications using AI in hardware verification. International Journal of Science, Research and Technology, 4(3), 5731–5744.
17. Chandran, A. S. (2024). Engineering billion-query search systems: Lessons from large-scale metadata platforms. International Journal of Research and Applied Innovations, 7(2), 10510–10519.
18. Uddandarao, D. P., Mahajan, A., & Krishnamsetty, P. (2024). Double machine learning for downstream causal measurement: A hybrid econometric-ML framework for feature valuation. International Journal of Computational and Experimental Science and Engineering, 10(4), 4477–4485.
19. Padmanabham, S. (2025). Security and Compliance in Integration Architectures: A Framework for Modern Enterprises. International Journal of Computing and Engineering, 7(16), 45-55.
20. Mathew, A. R. (2019). Cyber-infrastructure connections and smart gird security. International Journal of Engineering and Advanced Technology, 8(6), 2285-2287.
21. Sugumar, R. (2022). Explainable Deep Learning Framework for Financial Fraud Detection and Intelligent Risk Analysis. International Journal of Emerging Trends in Engineering and Management Research, 7(5), 12528.
22. Mysula, S. (2024). Event-driven integration architecture using publication methods for real-time DealCloud business operations & workflows. International Journal of Research and Applied Innovations, 7(4), 11233–11239.
23. Lenin, S. T., & Venkatasalam, K. (2025). Effective classification of heart disease using convolutional neural networks. Circuits, Systems, and Signal Processing, 44(2), 911-935.
24. Rengarajan, A. (2022). Developing AI Powered Enterprise Integration Platforms Through Cloud Native Observability and Automated Governance. International Journal of Science, Research and Technology, 5(6), 8970-8977.
25. Kundavaram, R. R., & Onteddu, A. R. (2023). Enhancing automotive performance and real-time data analytics through cloud computing. Journal of Computational Analysis and Applications (JoCAAA), 31(4), 1382–1392.
26. Shaik, N. (2022). Beyond vendor lock-in: Building intelligent multi-cloud Kubernetes orchestration. International Research Journal of Innovative Engineering, 6(2), 9985–9996.
27. Bitragunta, S. L. V. (2024). Intelligent power feedback control for motor-generator pairs: A machine learning-based approach. International Journal of Leading Research Publication, 5(12), 1–9.
28. Kopparthi, G. S. (2023). Trustworthy artificial intelligence for high-stakes decision-making: Explainability, human oversight, and system reliability. International Research Journal of Innovative Engineering, 7(4), 12847–12860.

| ISSN: 2395-0560 | www.irjie.org |editor@irjie.org| A Bimonthly, Peer-Reviewed, Scholarly Journal|

|Volume 9, Issue 5, September- October 2025|

DOI:10.15680/IRJIE.2025.0905011

29. Soundappan, S. J. (2021). DataOps: Orchestrating Reliable ML Data Pipelines. International Journal of Research and Applied Innovations, 4(4), 5533-5537.
30. Gopinathan, V. R. (2022). Advanced Predictive Decision Intelligence Using Machine Learning for Strategic Enterprise Management Platforms. International Journal of Research and Applied Innovations, 5(6), 8162-8169.
31. Singh, M. P. (2024). Building Cyber-Resilient Cloud Infrastructure Integrating Zero Trust, Identity Security, and Continuous Threat Monitoring. IACSE-International Journal of Cyber Security (IACSE-IJCS), 5(2), 7-23.
32. Chinnam, N. B. (2022). Engineering real-time supply chain visibility and ETA intelligence platforms for large-scale retail distribution networks. International Journal of Research Publications in Engineering, Technology and Management (IJRPETM), 5(6), 7918–7922.
33. Gadige, C. D. (2023). Scalable CRM automation for high-volume customer onboarding: An event-driven Salesforce architecture. International Journal of Emerging Trends in Engineering and Management Research, 8(6), 14762–14772.
34. Srikanth, D. V. (2025). AI-powered entrepreneurial ecosystems: a computational model for sustainable startup growth in green tech sectors. *Int. J. Environ. Sci, 1068-1077.
35. Jinna, K. R., & Bandari, U. (2023). Progressive delivery for machine-learning inference services: A comparative study of blue-green, canary and shadow strategies with model-aware gates on enterprise container platforms. International Journal of Research Publications in Engineering, Technology and Management, 6(4), 9092–9101.