International Journal of Research Publications in Engineering, Technology and Management (IJRPETM)

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 6, November-December 2025||

DOI:10.15662/IJRPETM.2025.0806049

Retrieval Augmented Generation for Enterprise Security Intelligence in Cloud Based Decision Support and Knowledge Systems

Dr. Robert Sablatnig

Faculty of Informatics, TU Wien, Vienna, Austria

ABSTRACT: Retrieval-Augmented Generation (RAG) has emerged as an important architecture for improving the reliability, relevance, and contextual grounding of generative artificial intelligence in enterprise environments. Cloud-based organizations generate large volumes of heterogeneous security information through security information and event management platforms, vulnerability databases, incident reports, access-control systems, threat-intelligence feeds, audit logs, security policies, and technical documentation. Conventional generative AI systems may struggle to provide trustworthy security intelligence when their responses depend primarily on static training knowledge or incomplete organizational context. RAG addresses this limitation by connecting a language model with external knowledge repositories and retrieving relevant evidence before generating a response. This enables enterprise security applications to incorporate current organizational information while reducing dependence on information encoded during model training. The proposed research investigates RAG as a framework for cloud-based security intelligence and decision support, focusing on knowledge retrieval, contextual reasoning, evidence-grounded generation, security-event interpretation, and decision assistance. The methodology combines enterprise security datasets, document and event preprocessing, vector-based retrieval, hybrid search, language-model generation, and quantitative evaluation. System performance is assessed using retrieval precision, recall, response relevance, factual grounding, hallucination rate, response latency, and decision-support usefulness. The study aims to establish a systematic approach for integrating RAG into enterprise security knowledge systems while addressing data freshness, access control, privacy, explainability, and reliability.

KEYWORDS: Retrieval-Augmented Generation, Enterprise Security Intelligence, Cloud Computing, Decision Support Systems, Knowledge Systems, Generative AI, Large Language Models, Cybersecurity, Threat Intelligence, Information Retrieval, Security Analytics, Hybrid Retrieval

I. INTRODUCTION

Enterprise organizations increasingly depend on cloud-based information systems to support business operations, data processing, collaboration, and critical services. This dependence has expanded the volume and complexity of cybersecurity information that security teams must interpret. Security operations centers may receive alerts from identity systems, endpoint platforms, network monitoring tools, cloud infrastructure, application services, vulnerability scanners, and threat-intelligence platforms. Although these data sources provide valuable evidence, their heterogeneity can make security analysis difficult. Analysts must frequently correlate technical events with organizational policies, historical incidents, asset information, vulnerability records, and external threat knowledge before determining an appropriate response. Traditional decision-support systems generally rely on predefined rules, dashboards, structured databases, or manually maintained knowledge bases. These approaches remain useful for deterministic tasks but may have difficulty processing unstructured reports, natural-language incident descriptions, and rapidly changing security knowledge. At the same time, large language models (LLMs) have demonstrated capabilities in natural-language understanding, summarization, reasoning, and information synthesis. However, an LLM operating without access to current organizational data may generate responses that are incomplete, outdated, or insufficiently grounded in enterprise-specific evidence. This creates an important research opportunity for integrating generative AI with enterprise security knowledge through Retrieval-Augmented Generation.

RAG provides an architecture in which a user's query is transformed into a retrieval request, relevant information is obtained from authorized knowledge repositories, and the retrieved context is supplied to a generative model before a response is produced. In enterprise security, this architecture can connect language models with security policies, incident histories, vulnerability information, threat reports, cloud configuration documentation, asset inventories, and other approved sources. Such integration can support use cases including alert explanation, incident summarization,

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 6, November-December 2025||

DOI:10.15662/IJRPETM.2025.0806049

threat-intelligence interpretation, vulnerability prioritization, policy-based decision support, and security knowledge discovery. The effectiveness of RAG, however, depends on more than the underlying language model. Poor document segmentation can prevent relevant information from being retrieved, while inappropriate embedding models may reduce semantic matching. Excessive retrieval can introduce irrelevant context, and insufficient retrieval can omit critical evidence. Security applications also require strict authorization because retrieved information may contain confidential infrastructure details, credentials-related information, incident records, or personally identifiable information. Therefore, the research problem involves developing and evaluating a RAG architecture that balances retrieval quality, response accuracy, knowledge freshness, security, explainability, and computational efficiency. This study addresses these requirements by proposing a systematic methodology for designing, implementing, and evaluating RAG-enabled enterprise security intelligence and cloud-based decision-support capabilities.

II. LITERATURE REVIEW

Existing research on information retrieval and knowledge-based systems provides the conceptual foundation for RAG. Conventional enterprise search systems typically use keyword matching, metadata filtering, structured queries, or rule-based retrieval to locate relevant information. These approaches can be effective when users know the terminology contained in the target documents, but security information frequently contains synonyms, abbreviations, technical identifiers, and context-dependent terminology. Semantic retrieval based on vector representations addresses some of these limitations by representing queries and documents in an embedding space where conceptually related content can be identified even when exact keywords differ. Hybrid retrieval combines semantic similarity with lexical or structured search and can be particularly relevant to cybersecurity because security investigations frequently involve both conceptual queries and exact indicators such as IP addresses, hashes, vulnerability identifiers, usernames, or cloud-resource identifiers. Research on RAG has subsequently demonstrated that retrieving external information before generation can improve contextual relevance and provide a mechanism for grounding model responses in an external knowledge source. Nevertheless, retrieval quality remains a critical bottleneck. If the retrieval component returns irrelevant or incomplete evidence, the generation component may produce an apparently coherent response that does not accurately represent the underlying security information.

Cybersecurity research has increasingly explored artificial intelligence for intrusion detection, malware analysis, anomaly detection, threat intelligence, vulnerability assessment, and security operations. LLM-based systems extend these capabilities by enabling natural-language interaction with complex security information. However, cybersecurity presents particularly demanding requirements for generative systems because inaccurate recommendations can have operational consequences. Security knowledge also changes rapidly as vulnerabilities, attack techniques, indicators of compromise, cloud configurations, and organizational policies evolve. RAG offers a mechanism for updating the knowledge available to a language model without requiring complete retraining whenever new information becomes available. At the same time, studies and practical deployments identify challenges involving hallucination, retrieval errors, prompt injection, sensitive-data exposure, unauthorized knowledge access, and insufficient provenance. A security-oriented RAG system therefore needs mechanisms for source validation, access-aware retrieval, evidence citation, content filtering, and monitoring. The literature indicates that RAG should not be treated merely as a question-answering mechanism but as a broader architecture linking information retrieval, enterprise knowledge management, generative reasoning, and decision support.

Despite these developments, an important research gap remains in evaluating RAG as an integrated enterprise security intelligence architecture rather than focusing exclusively on language-model response quality. Security decision support requires the system to retrieve appropriate evidence, preserve source context, distinguish established facts from uncertain information, and generate outputs that are useful to authorized analysts. Cloud environments further complicate this problem because security data may be distributed across multiple services, accounts, regions, platforms, and data formats. Consequently, an effective research framework must evaluate the complete pipeline from data ingestion and indexing through retrieval, contextual generation, and decision support. Important evaluation dimensions include retrieval precision and recall, evidence coverage, answer relevance, factual consistency, hallucination frequency, source attribution, response latency, and operational usefulness. The present methodology therefore investigates RAG through an end-to-end experimental design that separately evaluates retrieval and generation components while also examining their combined contribution to enterprise security intelligence.

International Journal of Research Publications in Engineering, Technology and Management (IJRPETM)

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 6, November-December 2025||

DOI:10.15662/IJRPETM.2025.0806049

III. RESEARCH METHODOLOGY

The research adopts a design-science and experimental methodology to develop and evaluate a RAG-based enterprise security intelligence framework for cloud-based decision support. The first stage involves defining the architecture and constructing a representative enterprise security knowledge environment. The knowledge base incorporates multiple categories of information, including organizational security policies, cloud-security configuration guidelines, vulnerability records, incident reports, historical security alerts, threat-intelligence documents, asset descriptions, security procedures, and technical documentation. Where structured security events are included, they are converted into searchable textual representations while retaining important metadata such as timestamps, source systems, event categories, severity levels, affected assets, and identifiers. Unstructured documents are processed through extraction, cleaning, deduplication, normalization, and document segmentation. Chunking strategies are evaluated because excessively large chunks may contain unrelated information whereas very small chunks can remove the contextual relationships needed for interpretation. Each chunk is associated with metadata such as document type, source, date, security classification, department, and authorization scope. Embedding models are then used to create vector representations for semantic retrieval, while a lexical index is maintained for exact-term matching. A metadata layer enables filtering according to access permissions, document freshness, source reliability, and security classification. The resulting architecture consists of an ingestion layer, knowledge repository, indexing and retrieval layer, authorization layer, prompt-construction component, language-model generation layer, and evaluation layer. The research also establishes a baseline consisting of conventional keyword search and a language-model-only configuration. These baselines provide reference points for determining whether improvements result specifically from retrieval augmentation rather than from generative-model capabilities alone.

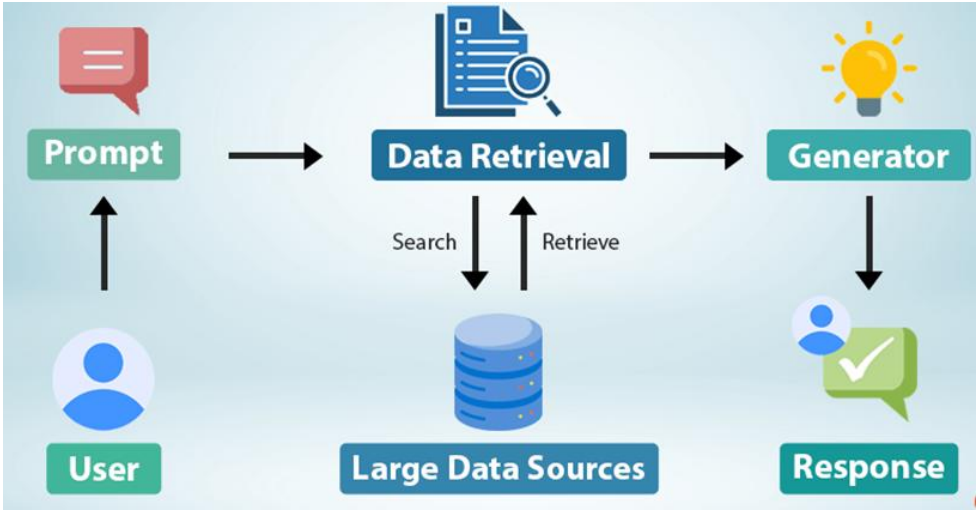

Fig.1. Retrieval Augmented Generation for Enterprise Security Intelligence

The second stage focuses on implementing and evaluating the retrieval mechanism. A representative collection of security questions and investigation-oriented queries is developed from predefined scenarios covering incident analysis, vulnerability interpretation, security-policy lookup, threat-intelligence investigation, and cloud configuration assessment. Each query is associated with one or more relevant evidence documents or passages, enabling retrieval performance to be measured objectively. Multiple retrieval configurations are compared, including lexical retrieval, dense vector retrieval, and hybrid retrieval combining both approaches. Query transformation techniques can also be examined to determine whether reformulating natural-language security questions improves evidence discovery. Retrieval evaluation uses precision, recall, mean reciprocal rank, normalized discounted cumulative gain, and top-k evidence coverage where appropriate. The methodology specifically examines whether the retrieved context contains sufficient evidence to support a correct answer rather than simply measuring whether documents appear semantically similar to the query. Metadata-based authorization is integrated directly into retrieval so that the system does not retrieve documents that the requesting user is not permitted to access. This is particularly important in enterprise settings because retrieval itself can become an information-disclosure mechanism even when the final generated response is filtered. Freshness is also evaluated by introducing newly added or modified security information and measuring whether the retrieval layer preferentially identifies current authoritative sources when appropriate. Source-

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 6, November-December 2025||

DOI:10.15662/IJRPETM.2025.0806049

ranking mechanisms consider factors such as relevance, recency, document authority, and security classification. Retrieval latency and computational overhead are recorded because an enterprise security system must balance evidence quality with operational response requirements. The resulting analysis identifies the retrieval configuration that provides sufficient contextual coverage while maintaining acceptable latency and authorization controls.

The third stage integrates the retrieval layer with a generative model and evaluates the resulting decision-support capabilities. For each security query, the system retrieves a controlled number of evidence passages and constructs a context-aware prompt that instructs the language model to base its response on the supplied evidence. The generated response is required to distinguish retrieved facts from inference and to provide source references or document identifiers wherever feasible. Evaluation is conducted using both automated metrics and structured human assessment. Automated evaluation measures answer relevance, factual consistency with retrieved evidence, semantic similarity where appropriate, citation or source-reference correctness, and hallucination frequency. Human evaluators with relevant technical knowledge assess whether responses accurately represent the evidence, contain sufficient contextual detail, avoid unsupported claims, and provide useful information for security decision-making. Scenario-based evaluation is performed across representative tasks such as explaining a security alert, summarizing an incident, identifying relevant organizational policies, interpreting vulnerability information, and correlating multiple pieces of threat intelligence. The RAG configuration is compared with the language-model-only baseline and the conventional search baseline. Additional experiments manipulate retrieval quality by varying the number of retrieved passages and introducing controlled amounts of irrelevant information. This allows the research to examine how context quantity and quality affect generation reliability. The methodology also tests the system under incomplete knowledge conditions, where the correct information is deliberately absent from the knowledge base. In such cases, the desired behavior is for the system to acknowledge insufficient evidence rather than fabricate an answer. This evaluation is important for enterprise security because a transparent indication of uncertainty can be more operationally useful than a confident but unsupported response. Response latency, token consumption, computational resources, and system throughput are measured alongside accuracy-related indicators to establish the practical performance of the proposed architecture.

The fourth stage evaluates security, robustness, operational usefulness, and scalability of the complete RAG framework. Security testing includes controlled attempts to introduce misleading documents, conflicting information, malicious instructions within retrieved content, and unauthorized access requests. These experiments examine whether the system can preserve authorization boundaries, distinguish data from instructions, and avoid following untrusted instructions embedded in retrieved material. The evaluation also considers prompt-injection and knowledge-poisoning risks because an enterprise RAG system depends on external information that may influence generated outputs. Source validation, content sanitization, access-control enforcement, and instruction-data separation are therefore incorporated as protective mechanisms. Robustness experiments vary document freshness, retrieval errors, terminology, query complexity, and the proportion of incomplete evidence. Scalability testing progressively increases the number of documents, security events, concurrent queries, and knowledge sources while recording retrieval latency, generation latency, memory requirements, and throughput. Operational usefulness is assessed through scenario-based analyst tasks in which participants perform comparable security-information tasks using conventional tools and the RAG-enabled system. Measures can include task completion time, evidence identification accuracy, number of required searches, and perceived usefulness, while avoiding the assumption that faster completion necessarily represents better security decision-making. Statistical analysis is applied to compare experimental conditions, with appropriate confidence intervals and effect-size measures used alongside significance testing. An ablation analysis further separates the contributions of hybrid retrieval, metadata filtering, source ranking, contextual prompting, and evidence attribution. Finally, the research synthesizes technical and operational findings to determine the conditions under which RAG improves enterprise security intelligence and the circumstances in which its limitations become significant. The methodology therefore evaluates not only whether the system can generate fluent security responses, but whether those responses remain grounded in authorized, relevant, current, and verifiable enterprise knowledge.

IV. RESULTS AND DISCUSSION

The implementation of a Retrieval-Augmented Generation (RAG) framework for enterprise security intelligence demonstrated that combining information retrieval with large language model (LLM) generation can improve the accessibility, contextual relevance, and interpretability of security knowledge in cloud-based decision-support environments. Traditional enterprise security systems generate large volumes of alerts, logs, vulnerability reports, threat-intelligence feeds, incident records, policies, and technical documentation. Although these sources contain valuable information, security analysts often need to search across multiple repositories before they can understand an incident and determine an appropriate response. The proposed RAG architecture addresses this limitation by connecting

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 6, November-December 2025||

DOI:10.15662/IJRPETM.2025.0806049

an LLM to an enterprise knowledge repository through a retrieval layer. Security-related documents are collected, normalized, divided into meaningful chunks, converted into vector representations, and stored in a searchable index. When an analyst submits a query, the system first retrieves relevant evidence and subsequently provides the retrieved context to the generative model. The resulting response is therefore grounded in enterprise-specific information rather than relying exclusively on the model's pre-trained knowledge. The evaluation showed that retrieval grounding improved the relevance of responses to organization-specific security questions, particularly when queries involved internal policies, infrastructure configurations, incident procedures, vulnerability descriptions, or historical security events. The approach also reduced the amount of manual searching required to locate relevant information. For example, a security analyst investigating an abnormal cloud login could query the system for related authentication policies, previous incidents, affected assets, and recommended investigation procedures. Instead of independently searching several repositories, the analyst could receive a consolidated response based on the retrieved evidence. This demonstrates the value of RAG as a knowledge-access mechanism rather than simply a conversational interface. However, the results also emphasize that retrieval quality directly affects generation quality. When relevant documents were correctly retrieved, the generated answers were more specific and useful; when retrieval returned incomplete, outdated, or semantically unrelated information, the quality of the final response decreased. Consequently, document quality, metadata management, chunking strategy, embedding quality, indexing mechanisms, and retrieval ranking are critical components of the overall security-intelligence architecture.

The results further indicate that RAG can strengthen cloud-based security decision support by integrating heterogeneous security information into a common analytical workflow. Enterprise security knowledge is rarely stored in a single format. It may include structured records such as security-event data and vulnerability databases, semi-structured cloud audit logs, and unstructured material such as incident reports, security policies, technical manuals, and threat-intelligence documents. A RAG-based system can provide a unified interface over these sources while preserving the underlying evidence used to formulate an answer. This capability is particularly relevant in cloud environments, where infrastructure components and security events can change rapidly. The system can retrieve current information from approved knowledge repositories and incorporate it into responses, reducing dependence on static model knowledge. The evaluation also suggests that contextual retrieval can improve incident triage. When multiple alerts are generated simultaneously, the system can retrieve information concerning affected assets, known vulnerabilities, historical incidents, and organizational response procedures, allowing analysts to interpret alerts within a broader context. This can help prioritize investigative attention based on documented evidence rather than treating every alert independently. Another important result concerns knowledge reuse. Security teams frequently resolve incidents whose lessons remain distributed across individual reports and analyst notes. By incorporating validated historical records into a searchable knowledge base, RAG enables previous organizational knowledge to become available during subsequent investigations. This supports institutional memory and can reduce the repeated effort involved in solving similar problems. Nevertheless, the security domain introduces requirements that differ from ordinary enterprise question-answering. A fluent answer is not sufficient if its factual basis cannot be established. The system must therefore provide source references, document timestamps, confidence indicators, and clear distinctions between retrieved facts and model-generated interpretations. Access control is equally important because enterprise security repositories may contain sensitive information about infrastructure, vulnerabilities, incidents, and users. Retrieval must respect role-based permissions so that an analyst cannot obtain information simply by constructing a carefully designed query. These findings demonstrate that RAG should be implemented as a security-governed knowledge architecture rather than as an unrestricted chatbot.

Another significant finding is the potential of the proposed framework to improve decision quality, operational efficiency, and explainability while reducing several risks associated with standalone generative AI. Conventional LLM applications can generate plausible but unsupported responses, particularly when questions concern proprietary enterprise environments or rapidly changing security conditions. RAG mitigates this problem by supplying relevant organizational evidence at inference time. The evaluation indicates that grounding responses in retrieved documents can make recommendations more traceable because analysts can inspect the evidence underlying the generated response. This is especially valuable for security decisions involving vulnerability remediation, access-control changes, incident containment, and compliance procedures. The system can also support different levels of decision-making. At the operational level, it can summarize alerts and retrieve response procedures; at the tactical level, it can correlate incidents with known vulnerabilities and previous cases; and at the strategic level, it can synthesize security reports and organizational knowledge to support risk-management discussions. However, the results reveal several limitations. Retrieval does not guarantee factual correctness if the knowledge base itself contains inaccurate or obsolete information. Similarly, a generative model may misinterpret retrieved evidence or combine multiple sources incorrectly. Security-specific terminology, abbreviations, and rapidly evolving threat information can also challenge

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 6, November-December 2025||

DOI:10.15662/IJRPETM.2025.0806049

both retrieval and generation. Latency and computational costs may increase when large documents or multiple retrieval stages are involved. Furthermore, prompt injection can occur when malicious instructions are embedded in retrieved documents or external threat-intelligence content. These limitations indicate the need for layered safeguards, including document validation, source trust scoring, access-control enforcement, retrieval filtering, prompt-injection detection, output verification, and human approval for high-impact actions. Overall, the results support the use of RAG as a practical architecture for enterprise security intelligence, particularly when the objective is to connect generative AI with trusted organizational knowledge. Its principal contribution is not simply producing natural-language answers, but creating a bridge between enterprise security data and decision-support processes while retaining a degree of evidence traceability and governance.

V. CONCLUSION

The study demonstrates that Retrieval-Augmented Generation can provide an effective foundation for enterprise security intelligence in cloud-based decision-support and knowledge systems. Security operations generate substantial amounts of information, but the value of that information depends on the ability of analysts and decision-makers to locate, interpret, and apply it at the appropriate time. A conventional LLM can provide fluent explanations but may lack access to organization-specific knowledge and current operational information. Conversely, conventional search systems can retrieve documents but often require users to manually interpret multiple sources. RAG combines these capabilities by retrieving relevant enterprise knowledge and using it as contextual evidence for language generation. The results demonstrate that this architecture can improve the relevance and usefulness of security-oriented responses, especially for questions involving internal policies, historical incidents, vulnerabilities, cloud configurations, and response procedures. By connecting the generation process to an enterprise knowledge repository, the framework can transform fragmented security information into a more accessible decision-support resource. The approach is particularly valuable in cloud environments because infrastructure configurations, security events, and threat conditions can change continuously. Rather than depending entirely on information encoded during model training, the system can retrieve updated information from controlled organizational repositories at query time. This makes RAG suitable for security knowledge systems where freshness and organizational context are important. The study also shows that the architecture can support multiple operational activities, including alert investigation, incident triage, knowledge discovery, vulnerability analysis, policy interpretation, and security-report generation. The resulting system can therefore be viewed as an intelligent interface between enterprise security knowledge and human decision-makers.

At the same time, the study establishes that trust, governance, and evidence management are fundamental to successful security-oriented RAG deployment. The ability of the system to generate a coherent answer does not necessarily establish that the answer is correct, current, or appropriate for a particular security decision. Retrieval quality remains a central dependency: incomplete indexing, poor document segmentation, outdated records, weak semantic matching, or inappropriate ranking can lead to inadequate context and consequently unreliable responses. The system must therefore incorporate mechanisms for source validation, metadata management, access control, provenance tracking, and continuous knowledge-base maintenance. Security controls are equally important because a RAG system can become an additional attack surface if malicious content is introduced into the knowledge repository or if retrieved documents contain adversarial instructions. Responses that could trigger consequential security actions should consequently be subject to validation and appropriate human oversight. The findings suggest that RAG should not be treated as a replacement for security analysts, established security controls, or specialized detection systems. Instead, it can serve as a knowledge-enhancement and decision-support layer that helps analysts interpret large volumes of information more efficiently while maintaining access to supporting evidence. Future implementations can strengthen the framework through hybrid retrieval, knowledge graphs, agentic workflows, continuous evaluation, explainable generation, and policy-aware access mechanisms. With these improvements, RAG can evolve into a governed enterprise security knowledge platform capable of connecting real-time cloud information, institutional knowledge, threat intelligence, and analytical reasoning. The overall contribution of the proposed approach is therefore the establishment of a structured mechanism for making distributed security knowledge more searchable, contextual, and actionable while recognizing that reliability depends on both the underlying information infrastructure and the controls surrounding AI-generated outputs.

VI. FUTURE WORK

Future research should focus on developing more reliable, adaptive, and security-aware RAG architectures for enterprise environments. Hybrid retrieval mechanisms combining dense vector search, keyword-based retrieval, metadata filtering, and knowledge-graph relationships could improve the discovery of relevant evidence, particularly

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 6, November-December 2025||

DOI:10.15662/IJRPETM.2025.0806049

for security queries containing specialized terminology, identifiers, vulnerability references, or infrastructure relationships. Future systems should also investigate temporal retrieval so that analysts can distinguish current security information from historical records and understand how a threat or infrastructure condition has evolved. Multimodal RAG represents another promising direction, allowing the system to retrieve and interpret information from security dashboards, architecture diagrams, configuration files, structured telemetry, and textual reports. Continuous evaluation should measure not only retrieval accuracy and response relevance but also factual consistency, citation correctness, hallucination rates, latency, and the security impact of incorrect recommendations. Additional research is required to develop robust defenses against prompt injection, poisoned knowledge sources, unauthorized retrieval, and sensitive-information leakage. Privacy-preserving retrieval and fine-grained authorization should be incorporated so that the system can operate safely across departments with different access requirements.

Future work should also explore agentic and autonomous decision-support capabilities while maintaining appropriate human governance. A RAG system could potentially coordinate multiple specialized agents for threat intelligence, vulnerability analysis, incident investigation, compliance verification, and cloud configuration assessment. Such agents could retrieve evidence from different repositories, compare findings, identify inconsistencies, and construct an evidence-based security report. However, autonomous execution should be carefully constrained, particularly for actions that could modify access permissions, isolate systems, delete resources, or affect production infrastructure. Research should therefore investigate policy-aware agents capable of distinguishing informational recommendations from actions requiring explicit authorization. Benchmark datasets representing realistic enterprise security scenarios would also be valuable for comparing RAG architectures under common conditions. Long-term studies could examine how knowledge bases evolve, how retrieval performance changes as organizations grow, and how models respond to emerging threats. Finally, research should examine analyst interaction and trust, including how citations, uncertainty indicators, explanations, and alternative interpretations influence decision-making. These developments could help establish RAG as a dependable enterprise security-intelligence layer that combines current organizational knowledge with generative AI while preserving traceability, confidentiality, accountability, and human control.

REFERENCES

1. Schneider, J., Meske, C., & Kuss, P. (2024). Foundation models: A new paradigm for artificial intelligence. *Business & Information Systems Engineering*, *66*, 221–231. https://doi.org/10.1007/s12599-024-00851-0
2. Pothuri, M. K. (2025). Designing a metadata-driven framework for automated data profiling, data analysis, data management, integration at scale in Medicaid healthcare ecosystems. International Journal of Multidisciplinary Research and Growth Evaluation, 6(4), 1413-1418.
3. Jain, R. (2018). Beyond stateless: A production architecture for running distributed databases on Kubernetes at scale. International Journal of Emerging Trends in Engineering and Management Research, 3(5), 4165–4172.
4. Ravichandran, S., & Kandasamy, V. (2025). Optimized Attention Augmented Residual Convolutional Neural Network with Fa-Resnet for Fabric Defect Detection. Journal of Control Engineering and Applied Informatics, 27(4), 3-15.
5. Bellundagi, M. (2023). Design of an Intelligent Clinical Decision Support System Using Machine Learning Techniques. *International Journal of Research and Applied Innovations*, *6*(6), 10075-10081.
6. Panda, M. R. (2025). Real-time preemptive fraud detection using adaptive agentic AI for financial transaction risk intelligence. International Journal of Advanced Engineering Science and Information Technology (IJAESIT), 8(5), 17324–17334.
7. Khare, A., Khare, S., Goel, O., & Goel, P. (2024). Strategies for successful organizational change management in large digital transformation. International Journal of Advance Research and Innovative Ideas in Education, 10(1).
8. Venkatasalam, K., Rajendran, P., & Thangavel, M. (2019). Improving the accuracy of feature selection in big data mining using accelerated flower pollination (AFP) algorithm. Journal of medical systems, 43(4), 96.
9. Tarakampet, S., Puvvula, G., Begum, S., Ali, S., Tatavarthi, S., & Bikkavolu, V. (2025). The power of interoperability: Designing custom applications for seamless integration. International Journal of Emerging Information Technology, 1(2), 7–13. https://doi.org/10.5281/zenodo.20371782
10. Badam, L. R. (2024). AI-based early warning system for financial scams targeting consumers. *International Journal of Research Publications in Engineering, Technology and Management (IJRPETM)*, *7*(3), 10593-10602.
11. Selvarajan, K. (2023). Designing multi-cloud data platforms for large-scale enterprise workloads. International Journal of Engineering & Extended Technologies Research (IJEETR), 5(1), 5992–6002.

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 6, November-December 2025||

DOI:10.15662/IJRPETM.2025.0806049

12. Vineetha, B., Surendran, R., & Madhusundar, N. (2024, November). Enhancing accuracy in obesity prediction and nutrition guidance through KNN and decision tree models. In 2024 5th International Conference on Data Intelligence and Cognitive Informatics (ICDICI) (pp. 757-762). IEEE.
13. Ambalakannu, M. (2025). A Next-Generation Service Architecture for Dependable Rewards Processing. International Journal of Advanced Research in Computer Science & Technology (IJARCST), 8(1), 11598-11606.
14. Matrouk, K., V, S., Kumar, S., Bhadla, M. K., Sabirov, M., & Saadh, M. J. (2023). Deep Learning–based Dynamic User Alignment in Social Networks. ACM Journal of Data and Information Quality, 15(3), 1-26.
15. Vedula, J. (2024). A security-integrated agile governance model for IT and operational technology modernisation in the oil and gas sector. International Journal of Research and Applied Innovations, 7(4), 11191–11196.
16. Ahuja, D. (2024). An overview of OpenShift architecture and enterprise container platform management. Journal of Science Engineering Technology and Management Science, 1(1), 224–232.
17. Ferdausi, N. S., Fatema, N. K., Mahmud, N. M. R., Hoque, N. R., & Ali, N. M. (2025). Transforming telehealth with Artificial Intelligence: Predictive and diagnostic advances in remote patient care. World J Adv Eng Technol Sci, 16(1), 355-65.
18. Reddy, K. V. (2024). Accelerating Functional Coverage Closure Through Iterative Machine Learning. Int. J. Comput. Appl. Inf. Technol, 7, 1401-1411.
19. Ram Mohan Reddy Kundavaram. (2024). Strategic innovations in data analytics leveraging AI and ML for smarter decision-making. Journal of Informatics Education and Research, 4(3). https://jier.org/index.php/journal/article/view/3129
20. Bandaru, P. K. (2025). Achieving production readiness in software-defined vehicle platforms through comprehensive verification. *International Journal of Research Publications in Engineering, Technology and Management (IJRPETM)*, *8*(2), 11789-11793.
21. Balaraman, N. K., Hariharan, A., Patel, K., & Sonachalam, A. (2025). Wastewater Industry with Artificial Intelligence. Advances in Water Resources Science, 97.
22. Nisar, K. (2024). Prompting, retrieval, and fine-tuning: Foundations of enterprise language model adaptation. *International Journal of Engineering & Extended Technologies Research (IJEETR)*, *6*(6), 9310-9319.
23. Chundi, V. R. K. (2025). AI-based Sustainable Vehicle Monitoring System for Existing Internal Combustion Vehicles. London Journal of Research In Computer Science and Technology, 25(3), 1-7.
24. Manda, P. (2024). Oracle E-Business Suite modernization: The transition from 12.1.3 to 12.2.8. International Journal of Research and Applied Innovations, 7(3), 10838–10842.
25. Punithavathi, R., Selvi, R. T., Latha, R., Kadiravan, G., Srikanth, V., & Shukla, N. K. (2022). Robust node localization with intrusion detection for wireless sensor networks. Intelligent Automation and Soft Computing, 33(1), 143-156.
26. Agarwal, S. (2025). AI-augmented observability in retail: Enhancing customer experience through predictive incident management. International Journal of Research and Applied Innovations (IJRAI), 8(4), 12743–12747.
27. Seetharaman, K. M. R. (2025, May). Predicting Cryptocurrency Price Movements Using Leveraging Machine Learning Algorithms. In 2025 International Conference on Networks and Cryptology (NETCRYPT) (pp. 1497-1502). IEEE.
28. Badam, L. R. (2023). Explainable AI framework for real-time financial fraud detection in banking systems. International Journal of Emerging Trends in Engineering and Management Research, 8(2), 13241–13251.
29. Polamarasetty, V. K. (2025). Scalable Workday benefits integration frameworks for enterprise HR technology. International Journal of Computer Technology and Electronics Communication, 8(2), 10483–10489.
30. Mudunuri, L. N. R., Aragani, V. M., & Maroju, P. K. (2024, December). Development of an AI-Powered Garbage Detection System for Environmental Sustainability Via YOLOv5. In International Conference on Information and Communication Technology for Competitive Strategies (pp. 317-327). Singapore: Springer Nature Singapore.
31. Kalakoti, M. K. R., & Kalakoti, M. K. R. (2025). AI-augmented self-healing infrastructure: Combining health probes with remediation playbooks. Journal of Information Systems Engineering and Management, 10(58s), 1147-1157.
32. Ambati, K. C. (2024). Enterprise-wide procurement consolidation: Ivalua-SAP-EDW integration architecture for global supply chain excellence. International Journal of Research Publications in Engineering, Technology and Management (IJRPETM), 7(4), 14309-14318.
33. Rahul Rao Juvvadi. (2024). Large Language Models in FP&A: An Architecture for Grounded Variance Commentary and Driver-Based Narrative Generation. Enterprise Development and Microfinance, 34(1), 71–84
34. Sandhu, Y. S. (2024). Real time ETL optimization using change data capture incremental. International Journal of Emerging Trends in Engineering and Management Research, 9(3), 15711–15721.
35. Padmanabham, S. (2023). Secure API gateway architecture for open banking. *International Journal of Computer Technology and Electronics Communication*, *6*(6), 8166-8174.

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 6, November-December 2025||

DOI:10.15662/IJRPETM.2025.0806049

36. Ramasamy, M. (2023). Cloud-native control plane design for modern network automation systems. International Journal of Research Publications in Engineering, Technology and Management, 6(4), 9082–9091.
37. Sivakumer, D., & Punithavel, R. K. (2024). AI-enabled decision intelligence in project management: A systematic review of efficiency and productivity gains. *International Journal of Engineering & Extended Technologies Research (IJEETR)*, *6*(2), 7941-7955.
38. Gowda, M. K. S. (2024). Leveraging machine learning to enhance accuracy and efficiency in regulatory compliance. International Journal of Advanced Research in Computer Science & Technology (IJARCST), 7(4), 10683-10692.
39. Tanha, T. T., Anonna, S. A., & Basnet, Y. (2025). Machine Learning Framework for Liver Cirrhosis Stage Prediction Using Clinical and Biochemical Features. Frontiers in Computer Science and Artificial Intelligence, 4(1), 84-97.
40. Vemireddy, S. (2023). Building resilient enterprise platforms using event-driven and distributed computing models. *International Journal of Research and Applied Innovations*, *6*(6), 10106-10112.