

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository:<https://orca.cardiff.ac.uk/id/eprint/181392/>

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Briliyant, Obrina , Javed, Amir and Cherdantseva, Yulia 2026. Enhancing cybersecurity log analysis through Retrieval-Augmented Generation. Presented at: 3rd International Conference on Foundation and Large Language Model (FLLM), Vienna, Austria, 25-28 November 2025. Proceedings of the 3rd International Conference on Foundation and Large Language Model (FLLM) 25-28 November 2025, Vienna, Austria. IEEE, pp. 990-995. 10.1109/FLLM67465.2025.11390888

Publishers page: <https://doi.org/10.1109/FLLM67465.2025.11390888>

Please note:

Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See <http://orca.cf.ac.uk/policies.html> for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.



# Enhancing Cybersecurity Log Analysis through Retrieval-Augmented Generation

1<sup>st</sup> Obrina Briliyant  
School of Computer Science  
and Informatics,  
Cardiff University  
Cardiff, United Kingdom  
0000-0002-1054-8112

2<sup>nd</sup> Amir Javed  
School of Computer Science  
and Informatics,  
Cardiff University  
Cardiff, United Kingdom  
0000-0001-9761-0945

3<sup>rd</sup> Yulia Cherdantseva  
School of Computer Science  
and Informatics,  
Cardiff University  
Cardiff, United Kingdom  
0000-0002-3527-1121

**Abstract**—The exponential growth of cyber threats and the corresponding volume of security log data have created unprecedented challenges for security analysts. Traditional log analysis approaches struggle with the scale, complexity, and domain expertise requirements necessary for effective vulnerability detection and incident response. This study addresses these challenges by implementing and evaluating Retrieval-Augmented Generation (RAG) architectures specifically optimized for cybersecurity log analysis. We conducted a comprehensive comparative analysis of three distinct retrieval techniques: base vector similarity search, parent document retrieval, and ensemble retrieval. Our experimental framework utilized Apache server logs and Healthapp logs containing security events, processed through different embedding and chunking strategies. The evaluation employed the RAG Assessment Score (RAGAS) framework to assess precision across multiple local large language models (LLMs). Our methodology revealed critical insights into the selection of local LLM for cybersecurity logs analysis and the performance of three retrieval techniques. The results demonstrate that base vector similarity retrieval achieved optimal overall performance with a score of 0.7482, significantly outperforming parent document retrieval (0.6753) and ensemble techniques (0.6965). Comparative analysis with PDF-based RAG systems revealed that cybersecurity-specialized implementations provide measurable advantages in faithfulness (5.3% improvement) while maintaining competitive performance across other metrics. These findings provide actionable insights for organizations seeking to implement localized AI-augmented cybersecurity log analysis systems in production environments.

**Index Terms**—Security Log Analysis, Retrieval-Augmented Generation, Large Language Model, Retrieval Technique, RAGAS.

## I. INTRODUCTION

Modern cybersecurity operations face an unprecedented challenge in managing and analyzing the massive volumes of log data. According to recent industry reports, organizations generate an average of 2.5 quintillion bytes of data daily, with security-relevant logs comprising a significant portion of this volume [1]. This data explosion, while providing comprehensive visibility into system behavior, has overwhelmed analysis capabilities and created significant operational bottlenecks.

This research was made possible by the support of the Indonesia Endowment Fund for Education Agency (LPDP). Their investment in our work has significantly contributed to the quality and impact of our research findings.

Recent advances in large language model (LLM) and retrieval-augmented generation (RAG) present promising opportunities to address the challenges in cybersecurity log analysis. RAG architectures combine the contextual understanding and reasoning capabilities of LLM with precise information retrieval systems, potentially enabling automated expert-level analysis of security logs at scale [2].

LLM can provide domain-specific interpretation capabilities that rival human experts, and Vector-based retrieval systems enable efficient processing of massive log datasets while identifying relevant information with semantic understanding that surpasses keyword-based approaches [3]. Local deployment of such system also promises a confidential, privacy-preserving environment for organization's sensitive data.

However, the implementation of RAG systems for security log data presents unique challenges that distinguish it from traditional document-based applications. Security logs have distinct structural characteristics, terminology, and analytical requirements that may not align with general-purpose RAG architectures. LLM, even with RAG, can be ineffective, inaccurate and even hallucinate in generating a response [4], [5].

This study presents a novel reproducible evaluation of RAG architectures specifically designed for cybersecurity log analysis. Our research addresses critical knowledge gaps in understanding how different retrieval methodologies, language models, and architectural decisions impact RAG performance in real-world cybersecurity scenarios.

Our primary research contributions include:

- Provides a systematic comparison of three distinct retrieval methodologies—vector similarity, parent document retrieval, and ensemble approaches—evaluated specifically within the cybersecurity domain.
- Presents a detailed analysis of LLM performance for cybersecurity logs analysis, including insights into the trade-offs between domain-specific and general-purpose models.
- Introduce the application of the RAG Assessment Score (RAGAS) evaluation framework to cybersecurity use case, establishing performance benchmarks and evaluation methodologies for similar research.

This paper is structured as follows: Section 2 reviews relevant literature in cybersecurity log analysis automation. Section 3 details our experimental methodology, including dataset preparation, architectural implementations, and evaluation frameworks. Section 4 presents comprehensive results across all evaluation metrics and comparative analysis between different approaches. Section 5 addresses limitations of the current study and outlines future research directions.

## II. RELATED WORKS

The combination of LLMs with RAG techniques improves security log analysis through various retrieval techniques, demonstrating enhanced accuracy in vulnerability detection and incident response when customized for security contexts.

### A. LLM with RAG for Cybersecurity

Hybrid frameworks that combine LLMs with RAG techniques can improve the extraction and analysis of security logs for vulnerability analysis and incident response. Studies report that these methods leverage domain-specific knowledge bases and integrate techniques such as dense retrieval (e.g., Dense Passage Retrieval, Contriever), cosine similarity measures, BM25 combined with reciprocal rank fusion, and knowledge graph integration [6]–[9].

In vulnerability detection, one study noted pairwise accuracy improvements of up to 110% over baselines using BM25 with reciprocal rank fusion, while another achieved accuracy rates of 99% and 97% for exploitation and mitigation tasks, respectively [6], [10].

### B. Retrieval Optimization Approaches

Current research demonstrates that retrieval optimization in RAG-based security log analysis systems significantly enhances vulnerability detection and incident response capabilities. Studies have shown that combining dense vector retrieval techniques (such as Dense Passage Retrieval, Contriever, and ANCE) with traditional sparse retrieval techniques like BM25 and Reciprocal Rank Fusion, results in substantial performance improvements [8], [11], [12]. Hybrid approaches that integrate knowledge graphs with vector-based retrieval have also shown promise, particularly in domain-specific applications, achieving accuracy rates above 97% compared to non-RAG baseline systems operating below 25% accuracy [13], [14]. These results highlight the importance of tailoring retrieval mechanisms to specific security contexts, whether through fine-tuning embeddings or combining multiple retrieval strategies [6], [15].

### C. RAG Performance Analysis

Performance assessment of RAG-enhanced security log analysis systems reveals a sophisticated landscape of metrics that balance traditional ML measures with RAG-specific evaluations. Standard metrics like precision, recall, F1-score, and accuracy remain fundamental, but RAG implementations have introduced additional measures, including Mean Average Precision (MAP) [12], and Retrieval-Augmented Generation Assessment Score (RAGAS) [5] and domain-specific indicators

like attack success rates [16]. These assessments demonstrate consistent improvements over baseline models, particularly when retrieval components are fine-tuned or augmented with knowledge representations [11], [17].

## III. METHODOLOGY

Our experimental approach was designed to address the fundamental question of how RAG architectures can be optimized for cybersecurity log analysis while maintaining rigorous scientific evaluation standards. Our methodology consists of dataset preparation, retrieval architecture, LLM selection, RAGAS evaluation, and comparative analysis. Overall methodology is depicted in Figure 1.

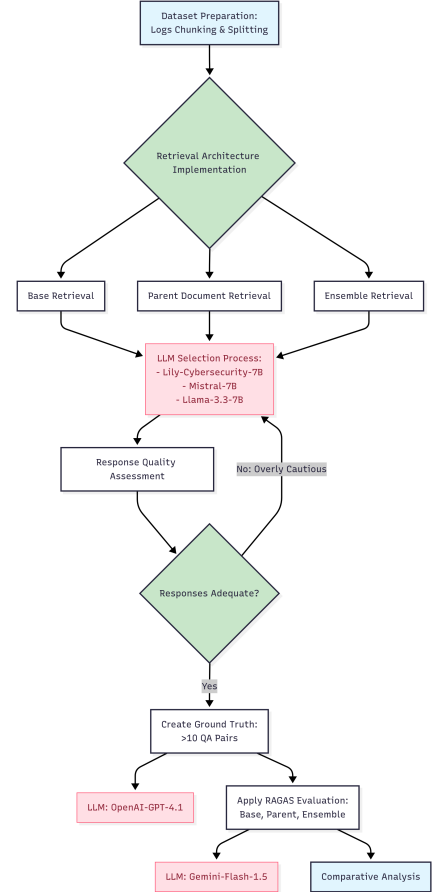


Fig. 1. Methodology to Enhance Cybersecurity Log Analysis.

### A. Dataset Preparation

Our evaluation dataset consists of comprehensive Apache server security logs and a mobile healthcare application logs that capture realistic cybersecurity scenarios encountered in production environments [18], [19]. The logs contain errors, system initialization events, configuration notifications, and various error states that represent common security-relevant events.

The preprocessing pipeline was designed to preserve the semantic and structural integrity of security logs while optimizing them for RAG processing. Our chunking strategy

required particular attention due to the unique characteristics of security logs compared to traditional documents. We implemented `RecursiveCharacterTextSplitter` with parameters optimized through iterative testing to balance context preservation with retrieval precision.

### B. Retrieval Architecture Implementation

Retrieval technique that we evaluate are: Base retrieval, Parent Document Retrieval, and Ensemble Retrieval.

1) *Base Vector Similarity Approach*: The base vector similarity implementation serves as both a performance baseline and a representative of the most straightforward RAG approach commonly used in production environments. Through empirical testing, we determined that 250-token chunks with 50-token overlap provide optimal performance for security log analysis.

2) *Parent Document Retrieval Implementation*: The parent document retrieval (PDR) architecture addresses a fundamental limitation of standard chunking approaches by maintaining hierarchical relationships between document segments.

The parent chunk configuration utilizes 1,500-token segments with 100-token overlap, sized to capture extended sequences of related security events while maintaining manageable processing requirements. Child chunks are configured at 200 tokens with 50-token overlap, optimized for precise matching of specific log entries or event patterns.

3) *Ensemble Retrieval Architecture*: The ensemble retrieval implementation combines the semantic understanding capabilities of vector search with the precision of lexical matching, addressing scenarios where either approach alone might miss relevant security events.

Our ensemble configuration allocates 60% weight to BM25 lexical search and 40% weight to vector similarity search, determined through systematic testing to optimize performance for security log characteristics.

The chunk configuration for ensemble retrieval uses 400-token segments with 75-token overlap, representing a compromise between the requirements of both retrieval components.

### C. Language Model Selection

Our initial approach focused on evaluating domain-specific cybersecurity language models, hypothesizing that specialized training would provide superior performance for security log analysis. The Lily-Cybersecurity-7B [20] model was selected as the primary candidate due to its specific training on cybersecurity datasets.

However, extensive testing revealed significant limitations in the model's practical applicability for operational cybersecurity environments. The model consistently produced overly conservative responses characterized by phrases such as "*it is not possible to definitively determine*" and "*additional context would be needed*". While this cautious approach might be appropriate for certain academic or research contexts, it proved counterproductive for operational cybersecurity analysis where security professionals require definitive, actionable insights for decision-making.

### D. Evaluation Framework Implementation

The evaluation of cybersecurity-focused RAG systems required adaptation of standard evaluation frameworks to address the unique characteristics of security analysis tasks. We employed RAGAS (RAG Assessment Score) framework, which provides comprehensive metrics for evaluating RAG systems across multiple dimensions of performance.

1) *Ground Truth Development*: The development of high-quality ground truth data for cybersecurity log analysis evaluation presents unique challenges due to the specialized knowledge required for accurate security assessment. Our approach involved LLM (OpenAI-GPT-4.1) and **prompt engineering** to create expert-validated question-answer pairs that represent realistic analytical scenarios encountered in operational environments:

```
cyber_prompt = f"""As a cybersecurity analyst, analyze the
following log data and answer the question with
specific, actionable insights.
LOG DATA:
{context}
QUESTION: {question}
ANALYSIS: """
```

The LLM-generated ground truth questions were validated by human expert to provide appropriate depth and specificity for the given security context. This process ensured that our evaluation framework reflects real-world analytical requirements rather than artificial or overly simplified scenarios. The machine-generated human-approved evaluation dataset sample is shown in Table I.

## IV. EXPERIMENT AND RESULT

The systematic evaluation of three retrieval methodologies revealed distinct performance characteristics that have significant implications for cybersecurity operations.

### A. LLM Performance Impact and Selection Insights

The LLM evaluation phase is perhaps the most significant and unexpected finding of this research, fundamentally challenging assumptions about the value of domain-specific models for cybersecurity applications.

The initial local deployment of Lily-Cybersecurity-7B [20], a model specifically trained for cybersecurity domain knowledge, resulted in evaluation scores approaching zero across all RAGAS metrics. Qualitative analysis revealed that the model consistently produced overly cautious, non-committal responses.

The transition to general-purpose LLM (also deployed locally) produced dramatic performance improvements. Llama-3.3-7B-Instruct, selected through systematic evaluation, demonstrated over 700% improvement in overall RAGAS scores compared to the Lily-Cybersecurity-7B. This improvement was driven primarily by the model's willingness to provide definitive technical analysis with specific security recommendations.

TABLE I  
RAG EVALUATION DATASET (SAMPLE)

Question	Ground Truth Answer	Context (Log Excerpt)	Source Document
What specific Apache errors are shown in the logs?	The specific Apache error shown is: [error] mod_jk child workerEnv in error state 6	[Sun Dec 04 04:47:44 2005] [notice] workerEnv.init() ok /etc/httpd/conf/workers2.properties [Sun Dec 04 04:47:44 2005] [error] mod_jk child workerEnv in error state 6 [Sun Dec 04 04:51:08 2005] [notice] jk2_init() Found child 6725 in scoreboard slot 10	apache_2k.log
What timestamps show the most critical issues?	Critical issue at: [Sun Dec 04 04:47:44 2005] Error: [error] mod_jk child workerEnv in error state 6	[Sun Dec 04 04:47:44 2005] [notice] workerEnv.init() ok /etc/httpd/conf/workers2.properties [Sun Dec 04 04:47:44 2005] [error] mod_jk child workerEnv in error state 6	apache_2k.log
What actions are being triggered by the SCREEN_ON intent in the logs?	Actions triggered: Processing broadcast action, Flushing sensor data, Setting today's step data	20171223-22:15:29:633  Step_StandReportReceiver 30002312  onReceive action: android.intent.action.SCREEN_ON 20171223-22:15:29:635 Step_LSC  30002312 processHandleBroadcastAction action:android.intent.action.SCREEN_ON	healthapp_2k.log
At what timestamp does the sensor data flushing occur, and what implications does this have for data integrity?	Timestamp: 20171223-22:15:29:635 Potentially affects data integrity if readings not captured accurately	20171223-22:15:29:635  Step_StandStepCounter 30002312  flush sensor data 20171223-22:15:29:635 Step_SPUtills  30002312 getTodayTotalDetailSteps = 1514038440000##6993##548365...	healthapp_2k.log

### B. Retrieval Augmented Generation Assessment Score (RAGAS) Metrics

RAGAS provides a comprehensive framework for evaluating RAG systems through multiple dimensions of performance. RAGAS evaluates both the retrieval and generation components by measuring how well the system retrieves relevant context and generates accurate, faithful responses.

The RAGAS framework consists of four primary metrics that collectively assess different aspects of RAG performance using the fundamental components: question ( $Q$ ), ground truth ( $GT$ ), generated answer ( $A$ ), and retrieved context ( $C$ ). **Context Precision** measures the proportion of relevant items in the retrieved context relative to the question, calculated as:

$$\text{Context Precision} = \frac{\sum_{k=1}^K \text{Precision}@k \times v_k}{\sum_{k=1}^K v_k} \quad (1)$$

where  $v_k \in \{0,1\}$  indicates whether the  $k$ -th item in the retrieved context  $C$  is relevant to question  $Q$ . **Context Recall** evaluates how well the retrieved context  $C$  captures all relevant information from the ground truth  $GT$  needed to answer the question:

$$\text{Context Recall} = \frac{|GT \cap C|}{|GT|} \quad (2)$$

where ground truth  $GT$  represents the ideal context required for answering question  $Q$ . **Faithfulness** measures whether the generated answer  $A$  is consistent with the retrieved context  $C$ , preventing hallucination:

$$\text{Faithfulness} = \frac{|\text{Claims in } A \text{ supported by } C|}{|\text{Total claims in } A|} \quad (3)$$

Finally, **Answer Relevancy** assesses how well the generated answer  $A$  addresses the original question  $Q$ :

$$\text{Answer Relevancy} = \frac{1}{N} \sum_{i=1}^N \cos(\vec{Q}, Q_i^{\text{artificial}}) \quad (4)$$

where  $\vec{Q}$  is the original question embedding,  $Q_i^{\text{artificial}}$  are embeddings of  $N$  artificially generated questions from answer  $A$ , and cosine similarity measures semantic alignment between the original question and answer relevance.

### C. Retrieval Technique Comparison

The RAGAS evaluation results demonstrate the comparative performance of three retrieval techniques on a cybersecurity log analysis dataset comprising 12 ground truth QA pairs, evaluated using Llama-3.3-7B-Instruct for generation, HuggingFace MiniLM for embeddings, and Gemini-1.5-Pro as the evaluator.

The Base vector similarity retrieval method achieved the highest overall score of 0.7482, excelling particularly in Answer Relevancy (0.9150), Context Recall (0.5972), and Context Precision (0.8636), making it the most balanced approach for cybersecurity applications. While the Ensemble method demonstrated superior Faithfulness (0.7369), indicating better consistency between generated answers and retrieved context, it suffered from lower Context Precision (0.6319), suggesting less focused retrieval.

The PDR (Parent Document Retrieval) method showed moderate performance across all metrics but ranked lowest overall (0.6753), with the weakest Faithfulness score (0.5239). These results indicate that for real-time SOC monitoring and cybersecurity log analysis, the Base retrieval method provides the optimal balance of relevance, precision, and reliability. The RAGAS performance comparison of retrieval techniques is depicted in Table II.

TABLE II  
CYBERSECURITY RAG PERFORMANCE COMPARISON OF RETRIEVAL  
TECHNIQUES

Technique	Faithful.	Ans.Rel.	Ctx.Rec.	Ctx.Prec.	Overall
Base	0.617	<b>0.915</b>	<b>0.597</b>	<b>0.864</b>	<b>0.748</b>
PDR	0.524	0.761	0.583	0.833	0.675
Ensemble	<b>0.737</b>	0.834	0.583	0.632	0.697

#### D. Comparative Analysis with Traditional Plain Document RAG

To contextualize our findings within the broader landscape of RAG applications, we conducted a comprehensive comparison with traditional document-based RAG systems processing plain PDF documents (derived from the work of [21]). This analysis provides insights into the specific value proposition of cybersecurity-specialized RAG implementation.

The comparison result is visualized in Figure 2, where the Base method shows a positive score (+0.035), indicating the Cybersecurity RAG slightly outperforms the Plain Documents RAG. The PDR and Ensemble methods, however, show negative scores (-0.073 and -0.167, respectively), demonstrating that the Plain Documents RAG performs better in these retrievals.

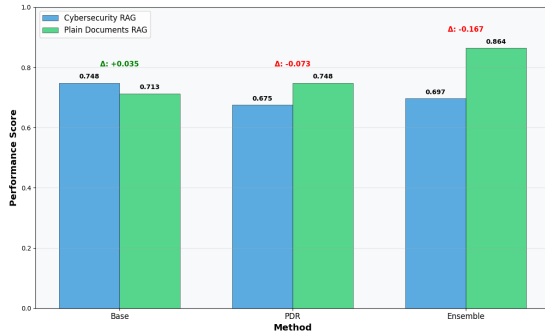


Fig. 2. Overall RAG Performance Comparison by Retrieval Techniques.

The comparison revealed performance differences that highlight the trade-offs between specialized and general-purpose RAG applications. In context precision, traditional document RAG systems achieved slightly superior performance (0.8858 vs 0.8636), suggesting that well-structured documents may be inherently easier to process with precise relevance filtering. However, cybersecurity-specialized systems demonstrated measurable advantages in faithfulness scores (0.7369 vs 0.7000), representing a 5.3% improvement that indicates better adherence to log evidence in security analysis responses. The average performance across all RAGAS metrics is visualized in Figure 3. The overall analysis, visualized in Figure 3 indicates that cybersecurity-specialized RAG systems provide measurable value through improved faithfulness and domain-appropriate analytical approaches, justifying the additional implementation complexity. However, the performance gaps are modest enough that organizations must carefully consider whether cybersecurity-specific optimization efforts align with their operational priorities and resource constraints.

#### E. Discussion

Our findings demonstrate that thoughtful application of RAG technologies can significantly enhance cybersecurity log analysis capabilities, providing security professionals with powerful tools for vulnerability detection, investigation, and response in increasingly complex digital environments. However, success requires careful attention to architectural selection, language model capabilities, and evaluation methodologies that align with operational requirements rather than theoretical sophistication. These are several key insights from our experiment:

- **Cybersecurity Advantage:** Faithfulness (+5.3%) - Better adherence to log evidence.
- **Document Advantage:** Context Recall (+67.4%) - Superior comprehensive information retrieval
- **Specialized Value:** Cybersecurity RAG. provides domain-appropriate analysis despite complexity.
- **Trade-offs:** Cybersecurity specialization improves analytical quality at the cost of broader recall.

While our experimental results provide valuable insights into RAG architecture performance for cybersecurity applications, several limitations must be acknowledged to properly contextualize the findings and guide future research directions.

The evaluation dataset, while representative of common web server security logs, represents only a subset of the diverse log types encountered in comprehensive cybersecurity operations. Modern security environments generate logs from firewalls, intrusion detection systems, endpoint security tools, network devices, and cloud services, each with distinct characteristics and analytical requirements.

#### V. CONCLUSION

This research provides critical insights into RAG architecture optimization for cybersecurity applications through systematic evaluation of retrieval techniques and language model selection strategies.

The most significant discovery involves language model performance, where general-purpose LLM (Llama-3.3-7B) achieved over 700% improvement in RAGAS scores compared to specialized fine-tuned LLM (Lily-Cybersecurity-7B). This counterintuitive finding suggests that technical reasoning capabilities and analytical confidence outweigh domain-specific training for cybersecurity applications.

Among retrieval techniques, Base vector similarity achieved the highest overall performance (0.7482), demonstrating optimal balance for real-time SOC operations.

Comparison with traditional document RAG revealed that cybersecurity systems excel in faithfulness (5.3% improvement) but face challenges in context recall due to the distributed nature of security log information.

#### A. Limitations

Our evaluation focused primarily on web server logs, potentially limiting generalizability across diverse cybersecurity log types including firewall, intrusion detection, and cloud service

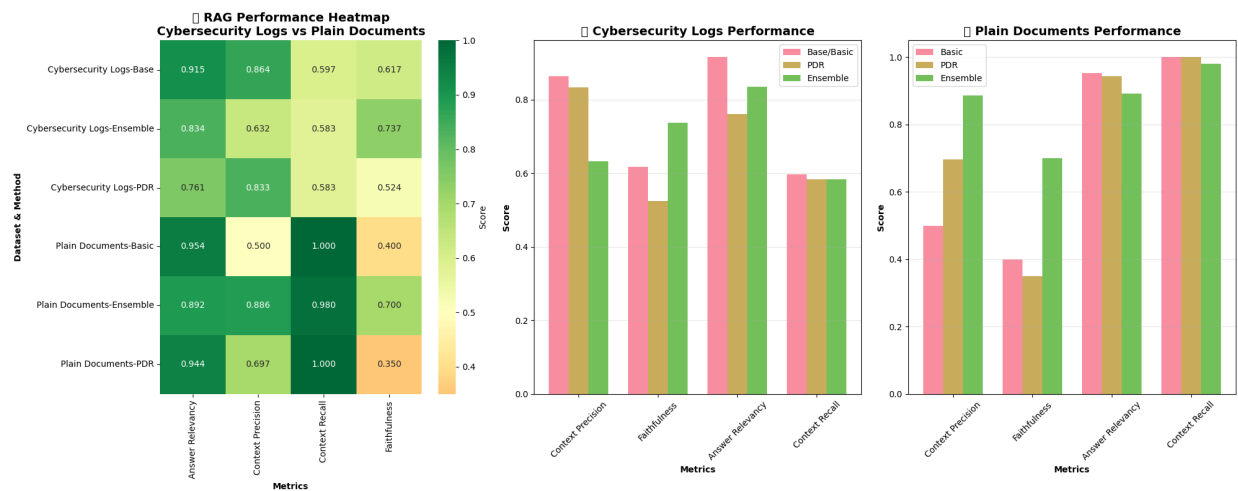


Fig. 3. RAG Performance Comparison for Cybersecurity Logs and Plain Documents.

logs. The temporal scope may not capture long-term performance degradation in evolving threat environments. Additionally, the rapidly advancing AI landscape may introduce new models that alter our observed performance characteristics.

### B. Future Research Directions

Future work should expand evaluation datasets to encompass diverse log types and longer temporal periods to validate architectural robustness. Investigation of fine-tuning approaches that combine general-purpose model reasoning with cybersecurity-specific optimization represents a promising research avenue. Developing adaptive RAG architectures that can dynamically adjust to evolving threat landscapes while maintaining consistent performance metrics warrants further exploration.

### REFERENCES

- [1] IBM, "What is Big Data Analytics?," Apr. 2024.
- [2] ISC2, "Cybersecurity Workforce Study," Sept. 2023.
- [3] P. Zhao, H. Zhang, Q. Yu, Z. Wang, Y. Geng, F. Fu, L. Yang, W. Zhang, J. Jiang, and B. Cui, "Retrieval-Augmented Generation for AI-Generated Content: A Survey," June 2024. arXiv:2402.19473 [cs].
- [4] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," Apr. 2021. arXiv:2005.11401 [cs].
- [5] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "Ragas: Automated Evaluation of Retrieval Augmented Generation," Apr. 2025. arXiv:2309.15217 [cs].
- [6] X. Du, G. Zheng, K. Wang, Y. Zou, Y. Wang, W. Deng, J. Feng, M. Liu, B. Chen, X. Peng, T. Ma, and Y. Lou, "Vul-RAG: Enhancing LLM-based Vulnerability Detection via Knowledge-level RAG," arXiv.org, 2024.
- [7] F. Y. Loumachi, M. C. Ghanem, and M. A. Ferrag, "Advancing Cyber Incident Timeline Analysis Through Retrieval-Augmented Generation and Large Language Models," *Computers*, vol. 14, p. 67, feb 13 2025.
- [8] S. Paul, F. Alemi, and R. Macwan, "Llm-Assisted Proactive Threat Intelligence for Automated Reasoning," arXiv.org, 2025.
- [9] Ms. Reshma Owhal, Viraj Shewale, Aniket Sorate, Mayur Swami, and Dipak Waghmode, "Cybervidya: Rag Infused Cyber Solutions," *International Research Journal on Advanced Engineering Hub (IRJAEH)*, vol. 3, pp. 456–464, mar 15 2025.
- [10] R. Fayyazi, S. H. Trueba, M. Zuzak, and S. J. Yang, "Proverag: Provenance-Driven Vulnerability Analysis with Automated Retrieval-Augmented LLMs," arXiv.org, 2024.
- [11] T. Xue, L. Hao, L. Mingyu, S. Xiaoyan, C. Ping, and D. Jun, "Revprag: Revealing Poisoning Attacks in Retrieval-Augmented Generation through LLM Activation Analysis," 2024.
- [12] G. Byun, S. Lee, N. Choi, and J. D. Choi, "Secure Multifaceted-RAG for Enterprise: Hybrid Knowledge Retrieval with Security Filtering," arXiv.org, 2025.
- [13] D. Roy, X. Zhang, R. Bhawe, C. Bansal, P. Las-Casas, R. Fonseca, and S. Rajmohan, "Exploring LLM-Based Agents for Root Cause Analysis," in *Companion Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering*, pp. 208–219, ACM, jul 10 2024.
- [14] R. C. Barron, V. Grantcharov, S. Wanna, M. E. Eren, M. Bhattarai, N. Solovyev, G. Tompkins, C. Nicholas, K. O. Rasmussen, C. Matuszek, and B. S. Alexandrov, "Domain-Specific Retrieval-Augmented Generation Using Vector Stores, Knowledge Graphs, and Tensor Factorization," in *2024 International Conference on Machine Learning and Applications (ICMLA)*, pp. 1669–1676, IEEE, dec 18 2024.
- [15] L. Ahmed, S. Utsav, S. H., and R. P. Md., "Techniquerag: Retrieval Augmented Generation for Adversarial Technique Annotation in Cyber Threat Intelligence Text," 2025.
- [16] B. Zhang, H. Xin, M. Fang, Z. Liu, B. Yi, T. Li, and Z. Liu, "Traceback of Poisoning Attacks to Retrieval-Augmented Generation," in *Proceedings of the ACM on Web Conference 2025*, pp. 2085–2097, ACM, apr 22 2025.
- [17] Z. Huichi, L. Kin-Hei, Z. Zhonghao, C. Yue, L. Zhenhao, W. Zhaoyang, H. Hamed, and Y. Emine, "Trustrag: Enhancing Robustness and Trustworthiness in Retrieval-Augmented Generation," 2025.
- [18] J. Zhu, S. He, P. He, J. Liu, and M. R. Lyu, "Loghub: A Large Collection of System Log Datasets for AI-driven Log Analytics," Sept. 2023. arXiv:2008.06448 [cs].
- [19] H. Ju, "Reliable Online Log Parsing using Large Language Models with Retrieval-Augmented Generation," in *2024 IEEE 35th International Symposium on Software Reliability Engineering Workshops (ISSREW)*, (Tsukuba, Japan), pp. 99–102, IEEE, Oct. 2024.
- [20] Dataloop, "Lily Cybersecurity 7B V0.2 5.0bpw H6 Ex12 · Models · Dataloop."
- [21] Dataphoenix, "The Art of RAG Evaluation," Jan. 2024.