Advancing Autonomous Incident Response: Leveraging LLMs and Cyber Threat Intelligence

Amine Tellache[1,2], Abdelaziz Amara Korba[2], Amdjed Mokhtari[1], Horea Moldovan[1], Yacine Ghamri-Doudane[2]
[1]*OODRIVE-Trusted Cloud Solutions, 75010 Paris, France*
[2]*L3i Lab, University of La Rochelle, 17000 La Rochelle, France.*
emails: {a.tellache@oodrive.com, a.mokhtari@oodrive.com, abdelaziz.amara_korba@univ-lr.fr,
h.moldovan@oodrive.com, yacine.ghamri@univ-lr.fr}

***Abstract*—Effective incident response (IR) is critical for mitigating cyber threats, yet security teams are overwhelmed by alert fatigue, high false-positive rates, and the vast volume of unstructured Cyber Threat Intelligence (CTI) documents. While CTI holds immense potential for enriching security operations, its extensive and fragmented nature makes manual analysis time-consuming and resource-intensive. To bridge this gap, we introduce a novel Retrieval-Augmented Generation (RAG)-based framework that leverages Large Language Models (LLMs) to automate and enhance IR by integrating dynamically retrieved CTI. Our approach introduces a hybrid retrieval mechanism that combines NLP-based similarity searches within a CTI vector database with standardized queries to external CTI platforms, facilitating context-aware enrichment of security alerts. The augmented intelligence is then leveraged by an LLM-powered response generation module, which formulates precise, actionable, and contextually relevant incident mitigation strategies. We propose a dual evaluation paradigm, wherein automated assessment using an auxiliary LLM is systematically cross-validated by cybersecurity experts. Empirical validation on real-world and simulated alerts demonstrates that our approach enhances the accuracy, contextualization, and efficiency of IR, alleviating analyst workload and reducing response latency. This work underscores the potential of LLM-driven CTI fusion in advancing autonomous security operations and establishing a foundation for intelligent, adaptive cybersecurity frameworks.**

***Index Terms*—Large Language Models (LLM), Cyber Threat Intelligence (CTI), Incident Response (IR), Retrieval-Augmented Generation (RAG).**

I. INTRODUCTION

With the ever-increasing sophistication of cyber threats, organizations struggle to respond efficiently to security incidents, leading to significant financial and operational consequences. Incident response [1] is a structured and strategic process for identifying and handling cyberattacks, aimed at reducing damage, recovery time, and overall costs. IR involves several specialized teams: SOC analysts handle monitoring and initial triage; CTI analysts provide strategic threat intelligence to contextualize events; and finally incident responders manage and remediate the detected incidents. Despite this structure, teams face major challenges, starting with detection. SOC analysts deal with alert fatigue due to high volumes—averaging 4,484 alerts per day—and spend nearly three hours daily on manual triage [2]. Alert enrichment with threat intelligence adds further delays and resource demands. Modern multi-cloud environments (e.g., AWS, GCP, Azure) introduce additional complexity with fragmented security models, requiring specialized skills and slowing response times. Meanwhile, generic incident response playbooks often lack actionable guidance, especially for less experienced teams, limiting their ability to learn from past incidents and respond effectively.

To bridge these gaps, organizations have turned to Cyber Threat Intelligence (CTI) [3] as a pivotal resource for enhancing their IR capabilities. CTI involves the collection, analysis, and dissemination of information about cyber threats and vulnerabilities, enabling organizations to understand better and respond to potential risks. Within SOC teams, CTI is primarily used to enrich and contextualize security alerts, providing crucial insights for improved prioritization and understanding. This enables teams to quickly identify the nature and source of an attack, assess potential damage, and make informed recommendations. It also provides insights into the latest threats and tactics used by attackers, which helps teams better prepare for and respond to incidents. Subsequently, IR teams rely on this enriched intelligence to craft efficient and targeted responses, underlining the importance of CTI in the overall cybersecurity ecosystem. However, without automated systems, searching, reading, and correlating CTI reports requires additional manual effort. Moreover, CTI data is vast, heterogeneous, and comes from diverse sources, making its management and analysis complex. Effectively utilizing CTI requires highly skilled personnel and advanced technological tools, which can pose challenges in terms of cost and skilled-resources availability. Furthermore, the ability to quickly analyze CTI data and take preventive or corrective actions may be limited, particularly while facing real-time attacks.

Given these challenges, there is a pressing need for advanced automation tools. Large Language Models (LLMs) have transformed multiple domains [4]–[6] with advanced language processing, automation, and data analysis, making them invaluable in many fields. These strengths make LLMs particularly valuable in cybersecurity, where they have shown promising results in diverse applications, including dataset generation [7], threat detection [7], [8], response [9], Cyber threat intelligence [10], and even attack [11]. In the CTI domain specifically, LLMs are increasingly employed to address the challenge of processing heterogeneous data from diverse sources. Multiple solutions have been proposed in this context. For instance, the Mitre ATT&CK project

[12] launched the Threat Report ATT&CK Mapper (TRAM), designed to simplify the analysis of CTI reports and extraction of Tactics, Techniques, and Procedures. Alves et al. [13] utilized different BERT model variants with optimized hyperparameters to identify the most effective model for TTP classification. Similarly, Fayyazi et al. [10] compared direct use of LLMs (GPT-3.5, Bard) with supervised fine-tuning (SFT) of smaller LLMs (BERT, SecureBERT) to interpret ambiguous cyberattack descriptions, demonstrating that fine-tuned small-scale LLMs like SecureBERT outperform direct use in precise classification of cybersecurity tactics. In addition, Peng et al. [14] introduced CTISum, a new benchmark specifically designed for CTI summarization tasks with capabilities for attack process summarization as a subtask. Tseng et al. [15] proposed an LLM-based solution to automate threat intelligence workflows within SOCs. Their system extracts Indicators of Compromise using LLMs, applies a filtering and cleaning mechanism via a multi-LLM voting system, and generates Regex rules and relationship graphs, enhancing preventive measures. While most state-of-the-art approaches in CTI leverage LLMs for passive tasks such as summarization and mapping, their role in active incident response remains limited. We strongly believe that LLMs can extend their actual usage to serve as a powerful tool, offering exceptional capabilities in processing, analyzing, and synthesizing vast amounts of CTI data to improve incident response efficiency, automating the entire workflow from report analysis to tailored response planning. However, there are significant challenges, particularly the high costs of fine-tuning these models for specialized CTI contexts. Additionally, maintaining the ability to continuously adapt to the ever-evolving landscape of new attacks and emerging CTI can be both time-consuming and expensive.

To address these challenges, we propose a novel real-time incident response approach based on LLMs. This system enables security teams to effectively leverage Cyber Threat Intelligence to respond to incidents. This model is capable of accurately enriching events and alerts using CTI, and then rapidly producing effective actions in response to specific threats, drawing on vast and diverse CTI sources. The model's architecture will incorporate and integrate CTI without requiring costly fine-tuning for each update, through the Retrieval-Augmented Generation technique. The effectiveness of this approach has been rigorously evaluated using a novel automated method, powered by LLMs, and cross-validated by cybersecurity experts. This evaluation assesses key metrics such as answer relevance, context relevance, and groundedness across two datasets: real-world alerts and simulated alerts.

In summary, the main contributions of this paper are:

- A novel RAG-based incident response model that effectively leverages Cyber Threat Intelligence to enrich security alerts and generate precise, context-aware response actions. Our approach supports the continuous integration of new CTI data without requiring model fine-tuning for each update, ensuring adaptability to emerging threats.
- We propose an innovative retrieval within our RAG system that combines two complementary search techniques to contextualize and enrich incidents: The model can performs standard searches on platforms through CTI APIs like VirusTotal [16] to retrieve standardized contexts and correlate the incident with private databases. Subsequently, it employs NLP-based similarity searches to identify relevant documents or text segments in the CTI vector database, enabling the correlation of incidents with historical cases.
- A rigorous evaluation framework integrating automated assessments powered by LLMs with manual expert validation on real-world and simulated SIEM (Security Information and Event Management) alerts. The simulated alerts are generated by executing real attack scenarios in controlled environments. The results, assessed across multiple dimensions such as response accuracy, contextual relevance, and data groundedness, demonstrate the effectiveness of our approach in enhancing incident response.

The remainder of this paper is outlined as follows. The proposed RAG-based Incident Response architecture is described in Section II. Section III presents the experimental results and analysis. Finally, Section IV concludes the paper and points out future directions.

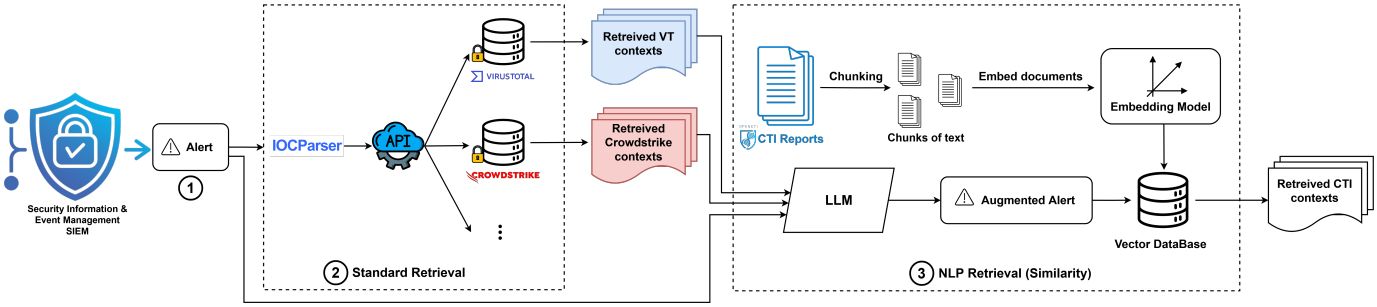
II. Proposed Approach

A. Overview of the proposed system

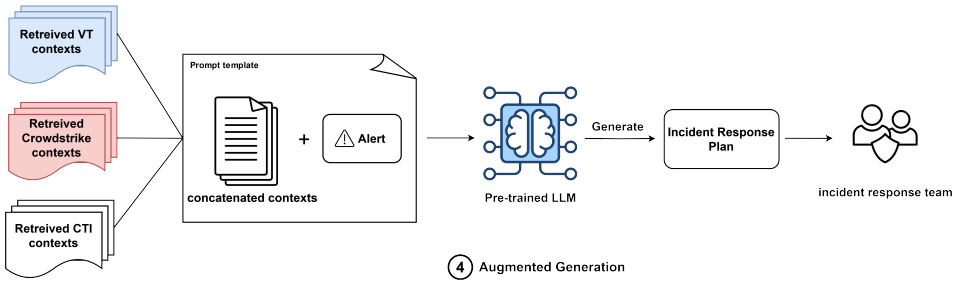
We propose a RAG-based incident response architecture (Figure 1) designed to automate and streamline the workflows of SOC analysts, CTI analysts, and incident responders. The system takes as input a security incident description, with a focus on SIEM alerts (Figure 1a, Part 1), and operates in two main phases: retrieval and augmented generation. In the Retrieval phase, the model searches for relevant threat intelligence to enrich the alert. Given the inherent access restrictions in CTI, we introduce a tailored retrieval method optimized for this context. The model first performs structured searches via CTI platform APIs, such as VirusTotal, to extract contextual information from private databases. It then employs NLP-based similarity techniques to identify relevant documents or text segments within CTI vector databases, enabling correlation between current incidents and historical cases. In the augmented generation phase, the language model synthesizes accurate and actionable responses using the retrieved data. Our approach continuously integrates new CTI data by embedding reports into a vector database and using additional private datasets, eliminating the need for fine-tuning. This strategy significantly reduces both computational overhead and operational costs, ensuring adaptability to evolving threats. The following sections detail the architecture's main components, including knowledge base construction, the retrieval mechanism, and the generation process.

B. Building the Knowledge Base

To build the Knowledge Base for the NLP-based search, we initially utilized CTI documents from publicly available collec-



(a) RAG Retrieval component



(b) RAG Augmented generation component

Fig. 1: Proposed Retrieval-Augmented Generation RAG Incident Response Architecture

tions of Advanced Persistent Threat (APT) and cybercriminal campaign reports [16]. In the future, we plan to integrate this process directly with the OpenCTI platform [17] to enable seamless data sharing and integration. The process starts by loading all the CTI documents, cleaning, and transforming PDF files into text. Then, text splitters break large documents into smaller chunks, making them easier to index and search, as larger chunks can be difficult to handle in models with limited context windows. These chunks are transformed into vectors using embeddings, and then stored in a specialized database called a vector database. This type of database is specifically designed to efficiently store, manage, and search through large quantities of high-dimensional vector data.

C. RAG-Based CTI Retrieval

We proposed a new RAG-based search method tailored for gathering CTI from diverse sources. This method combines both standard search and NLP-based search techniques.

The standard search, depicted in part 2 of Figure 1a, involves querying private databases commonly used by SOC teams. To automate this process, we first employed an IOC Parser tool to extract various IOCs, such as domain names, hashes, IP addresses, and URLs. For each IOC the system leverages APIs from private threat intelligence databases (VirusTotal in our case) to gather detailed contextual information, including historical malicious activity, geolocation, and reputation scores. This enriched data is then appended to the standard search context. For example, if an IP address is flagged, the system queries APIs to determine its association with botnets, historical attack patterns, and blacklist status.

The natural language processing (NLP)-based search, presented in part 3 of Figure 1a, retrieves the most relevant context from CTI reports by conducting searches in a knowledge base built from these. To retrieve the CTI reports most relevant to the detected incident in the alert, we propose a solution to enhance the alert. Raw alert data often lacks the contextual richness necessary for effective correlation. To address this, we leverage an LLM guided by a tailored prompt to generate an augmented alert. This prompt is designed to incorporate findings from an initial standard search—such as those conducted on VirusTotal during our experiments—into the alert. This process facilitates similarity-based searches by reformatting the alert into a structured and contextually relevant format.

Performing a search in the vector database involves calculating the similarity (e.g. cosine similarity) between the embeddings of the question (the Siem Alert) and document chunks (CTI). To evaluate the similarity between texts, two key aspects must be defined: the method used to measure the similarity between embeddings, and the algorithm used to transform the text into embeddings, which represent the text in a vector space. The system primarily employs Cosine Similarity, as defined in Equation 1, to measure the similarity between embeddings. This method remains one of the most widely used techniques for assessing vector similarity. A score closer to 1 indicates a higher degree of similarity between embeddings. To generate embeddings, we utilize Transformer-based models [18], as they represent a significant breakthrough in NLP and consistently outperform previous approaches.

$$\text{similarity}(A, B) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \times \|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}} \quad (1)$$

D. Augmented Generation

Following the retrieval step, a final prompt is designed for the LLM to generate an incident response strategy. This prompt incorporates both the SIEM alert and the retrieved data, which is a combination of standard search data (e.g. virusTotal results) and insights from the NLP search (CTI chunks). To achieve optimal results, we propose an adapted prompt that clearly defines the context of the alert, provides the necessary background from the retrieved data, and outlines the specific task. This ensures that the LLM produces a comprehensive and actionable incident response strategy tailored to the detected threat. The final prompt is fed into a pre-trained LLM, as illustrated in Figure 1b, to generate the incident response plan text. In our solution, we leverage GPT LLM (Generative Pre-trained Transformer), which utilizes auto-regressive [19] modeling to produce coherent and contextually relevant text. These models called decoder-only, correspond to the decoder part of the Transformer model, are ideal for text generation.

III. EXPERIMENTS AND EVALUATION RESULTS

A. System Setup

We tested the proposed architecture on 100 alerts generated by the our company's SIEM system (LogPoint SIEM [20]), randomly selected between August 10 and September 8, 2024. These real-world alerts served as inputs to our model, which we enriched with CTI to propose a coherent response for each incident.

In addition to real-world alerts, we manually generated alerts by emulating incidents within a controlled and isolated test environment, illustrated in Figure 2. This environment was built on an internal Proxmox hypervisor, which hosted and managed the virtual machines (VMs) necessary for simulating attacks. We deployed the ELK stack [21] as a dedicated SIEM solution on a separate virtual machine, configured within the same network as the target machines. For attack simulation, we used a Linux virtual machine as the attacker, where we installed Caldera [22], an attack simulation tool developed by MITRE. Caldera automates cyberattacks by replicating tactics and techniques (TTPs) from the MITRE ATT&CK framework [23]. To ensure diverse log data, we deployed two types of target operating systems—Windows and Linux. Both target machines had Elastic Defend agents [24] installed to capture all security events. Using Caldera, we emulated various advanced persistent threat (APT) attacks, illustrated in Table I. These simulated attacks included a range of TTPs from the MITRE ATT&CK Framework, targeting both Linux and Windows hosts. For instance, the Advanced Thief adversary employed techniques such as Automated Collection (T1119) and Exfiltration Over C2 Channel (T1041) on Linux machines to collect and exfiltrate sensitive data. Similarly, the Stowaway adversary utilized Process Discovery (T1057) and Process Injection (T1055.002) to hide its presence and evade detection on Windows systems. Detection rules were created within the SIEM to intercept these events and generate alerts. In total, we generated 10 simulated alerts, which were tested and evaluated using our solution. These alerts, combined with the 100 real-world alerts, provided a comprehensive dataset to thoroughly study and refine our proposed solution and analyze the results.

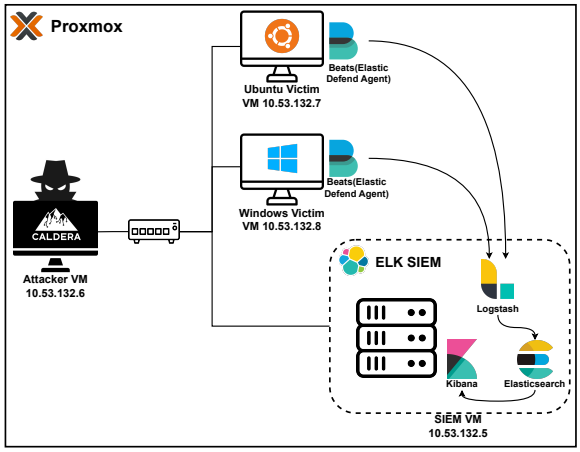


Fig. 2: Test environment for alerts generation

TABLE I: Simulated Cyber Attacks in a Controlled Test Environment for Alert Generation

Adversarie	Ability Name	Tactic	Technique	Host
Advanced Thief	Advanced File Search and Stager	Collection	T1119 Automated Collection	Linux
Advanced Thief	Compress staged directory	Collection	T1560.001 Archive Collected Data: Archive via Utility	Linux
Advanced Thief	Exfil staged directory	Exfiltration	T1041 Exfiltration Over C2 Channel	Linux
Stowaway	Discover injectable process	Discovery	T1057 Process Discovery	Windows
Stowaway	Inject Sandcat into process	Defense-Evasion	T1055.002 Process Injection: Portable Executable Injection	Windows
atomic	NMAP scan	Technical-Information-Gathering	T1254 Conduct active scanning	Linux
atomic	Access /etc/shadow (Local)	Credential-Access	T1003.008 OS Credential Dumping: /etc/-passwd, /etc/master.passwd and /etc/shadow	Linux
Windows Worm #1	Collect ARP details	Discovery	T1018 Remote System Discovery	Windows
Super Spy	Find files	Collection	T1005 Data from Local System	Windows
Super Spy	Exfil staged directory windows	Exfiltration	T1041 Exfiltration Over C2 Channel	Windows

For the standard search involving querying private databases, we used the VirusTotal database [25] via API, which allows searching for domain names, hashes, IP addresses, and URLs. For the NLP-based search and vector database creation, we utilized public CTI reports from the APT repository on VX Underground [16], which contains papers and blogs (sorted by year) related to malicious campaigns, activity, or software associated with vendor-defined APT groups and toolsets. We retrieved these reports and performed text extraction from each CTI report in PDF format using PyMuPDF [26], a high-performance Python library for extracting, analyzing, converting, and manipulating PDFs and other document formats. The extracted text was then chunked and embedded into a Chroma vector database. For augmented generation, we used the GPT-4o model, as it is considered the best model according to the leaderboard platform on Hugging Face for text generation. Moreover, it accepts a large context length, which is crucial for our use case.

B. Evaluation Method and Metrics

To evaluate the RAG model, we opted for specific metrics commonly used for the assessment of Retrieval-Augmented

Generation systems. These metrics are designed to measure various aspects of the model's performance, focusing on how well it retrieves and generates information:

- **Answer Relevance:** Is the response directly relevant to the query?
- **Context Relevance:** Is the retrieved context relevant and appropriately aligned with the query?
- **Groundedness:** Is the response supported by the retrieved context?

We propose using an automated evaluation method powered by LLMs to assess each metric for every alert. The idea is to employ other LLMs to evaluate each alert on a scale of 1 to 5 for each criterion. To accomplish this, we crafted specific prompts tailored to each metric. These prompts instruct the LLM to provide a score for each metric and simultaneously offer an explanation of why that score was given. To ensure the validity of our approach, we complemented the automated scoring system with manual evaluations performed by security experts, thereby validating the reliability and accuracy of the automated assessments.

C. Experimental Results

The evaluation consists of two parts. First, we assess the automated analysis of real alerts from the enterprise SIEM, evaluating the response quality. Then, we validate these results through both automated and manual evaluation on 10 simulated alerts generated in a controlled environment. This setup, with known incident causes, helps confirm the reliability of our scoring system. Finally, we compare results across both datasets.

*1) **Results of the real-world Alerts**:* We performed the automatic evaluation on 100 real-world alerts reported by our security teams, using three open-source models. The models used include Mistral-large-2407 a larger variant of the Mistral model, Llama-3.1-70B-Instruct, a larger variant of the LLaMA model from Meta, and a smaller LLaMA model Llama-3.2-3B-Instruct.

Table II displays the percentage ratings assigned by each model for each alert processed, along with the mean and variance. We observe that the models provide close mean scores, however their sensitivity varies. This variation is expected, as we included both large and small versions of different models.

Overall, our Retrieval-Augmented Generation (RAG) system demonstrates strong performance in answer relevance, achieving an average score close to 5 across all LLMs, and in groundedness (accuracy of responses), with an average score exceeding 4. These positive results are consistently reflected in the high ratings from all LLMs. However, context relevance scores are relatively low. The reduced score can likely be attributed to the inability to identify relevant context for certain alerts in either VirusTotal or CTI reports. This can be explained by the inclusion of false-positive alerts in the dataset, as well as the fact that, for some alerts, the IOCs are not recognized within the VirusTotal database. Furthermore, CTI reports often lack detailed information on specific campaigns. In future versions, we aim to extend this solution by incorporating additional trusted private databases, such as CrowdStrike, to enrich the standard search, as well as integrating more comprehensive CTI reports to improve overall context coverage.

To further investigate, we separated context relevance evaluations based on CTI reports and VirusTotal data. We observed similar, lower-than-global context relevance scores for each, indicating that both sources are complementary and collectively improve the final context relevance.

TABLE II: Automatic Evaluation of LLMs on Real-World Alerts

Model & Metric	1 (%)	2 (%)	3 (%)	4 (%)	5 (%)	Mean	Variance
Mistral-large Answer Relevance	0.00	0.00	0.00	2.12	97.87	**4.97**	0.02
Mistral-large Context Relevance [VT + CTI]	0.00	20.00	51.11	22.22	6.66	3.15	0.66
Mistral-large Context Relevance [VT only]	17.77	3.33	44.44	27.77	6.66	**3.02**	1.28
Mistral-large Context Relevance [CTI only]	1.20	63.85	21.68	6.02	7.22	2.54	0.82
Mistral-large Groundedness	0.00	0.00	2.46	41.97	55.55	**4.53**	0.29
Llama-3.2-3B Answer Relevance	0.00	1.00	23.00	25.00	51.0	4.26	0.71
Llama-3.2-3B Context Relevance [VT + CTI]	0.00	30.00	28.99	17.00	24.00	3.35	1.30
Llama-3.2-3B Context Relevance [VT only]	0.00	33.00	51.00	16.00	0.00	2.83	0.46
Llama-3.2-3B Context Relevance [CTI only]	0.00	20.00	31.00	28.99	20.00	**3.49**	1.04
Llama-3.2-3B Groundedness	0.00	3.00	28.00	16.00	53.00	4.19	0.89
Llama-3.1-70B Answer Relevance	0.00	0.00	0.00	5.00	95.0	4.95	0.04
Llama-3.1-70B Context Relevance [VT + CTI]	6.06	23.23	13.13	17.17	40.40	**3.62**	1.87
Llama-3.1-70B Context Relevance [VT only]	24.00	49.00	22.00	1.00	4.00	2.12	0.84
Llama-3.1-70B Context Relevance [CTI only]	2.00	20.00	64.00	1.00	13.00	3.03	0.80
Llama-3.1-70B Groundedness	0.00	0.00	18.00	13.00	69.00	4.51	0.60

*2) **Results of the Simulated Alerts**:* In this section, we aim to validate the automated evaluation by incorporating a manual assessment conducted by cybersecurity experts. This approach involves analyzing alerts generated through controlled simulations, where the incidents triggering each alert and their corresponding attacks are known. The objective is to assess the system in detail and validate the automated evaluation, which we intend to use later as a filter to retain only accurate responses.

We began with the automated evaluation, illustrated in Table III, which demonstrates results similar to the evaluation of 100 alerts. It shows strong performance in answer relevance, achieving an average score close to 5 across all LLMs, as well as groundedness (accuracy of responses), with an average score exceeding 4. Additionally, there is an improvement in context relevance compared to previous results, as all the alerts in this case are true positives. A detailed comparison between the two datasets will be explored in subsequent sections.

TABLE III: Automatic Evaluation of LLMs on Simulated Alerts

Model & Metric	1 (%)	2 (%)	3 (%)	4 (%)	5 (%)	Mean	Variance
Mistral-large-2407 Answer Relevance	0.00	0.00	0.00	0.00	100	**5.00**	0.00
Mistral-large-2407 Context Relevance [VT + CTI]	0.00	0.00	22.22	0.00	77.77	**4.55**	0.69
Mistral-large-2407 Context Relevance [VT only]	22.22	11.11	22.22	11.11	33.33	3.22	2.39
Mistral-large-2407 Context Relevance [CTI only]	0.00	0.00	0.00	37.50	62.50	**4.62**	0.23
Mistral-large-2407 Groundedness	0.00	0.00	0.00	44.44	55.55	4.55	0.24
Llama-3.2-3B-Instruct Answer Relevance	0.00	0.00	0.00	20.00	80.00	4.80	0.15
Llama-3.2-3B-Instruct Context Relevance [VT + CTI]	0.00	10.00	20.00	30.00	40.00	4.00	1.00
Llama-3.2-3B-Instruct Context Relevance [VT only]	0.00	20.00	30.00	50.00	0.00	**3.30**	0.61
Llama-3.2-3B-Instruct Context Relevance [CTI only]	0.00	10.00	10.00	20.00	60.00	4.30	1.01
Llama-3.2-3B-Instruct Groundedness	0.00	0.00	10.00	10.00	80.00	4.70	0.41
Llama-3.1-70B-Instruct Answer Relevance	0.00	0.00	0.00	20.00	80.00	4.80	0.15
Llama-3.1-70B-Instruct Context Relevance [VT + CTI]	10.00	10.00	10.00	10.00	60.00	4.00	2.00
Llama-3.1-70B-Instruct Context Relevance [VT only]	40.00	20.00	10.00	10.00	20.00	2.50	2.45
Llama-3.1-70B-Instruct Context Relevance [CTI only]	0.00	10.00	20.00	0.00	70.00	4.30	1.21
Llama-3.1-70B-Instruct Groundedness	0.00	0.00	0.00	20.00	80.00	**4.80**	0.15

For the expert evaluation, we submitted 10 alerts to a cybersecurity expert within our organization, as summarized in Table IV. The results demonstrate strong performance across all metrics, with an average score exceeding 4. Compared to the automatic evaluation, the expert evaluation yields similar

average scores for context relevance. However, the scores for answer relevance and groundedness are slightly lower, while still remaining near 4. These findings validate the proposed automatic evaluation methodology and further confirm the effectiveness of our approach from an operational perspective, as evaluated by an incident responder.

TABLE IV: Expert Evaluation on Simulated Alerts

Metric	1 (%)	2 (%)	3 (%)	4 (%)	5 (%)	Mean	Variance
Expert Answer Relevance	0.00	0.00	0.00	80.00	20.00	**4.20**	0.17
Expert Context Relevance	0.00	0.00	20.00	50.00	30.00	**4.10**	0.54
Expert Groundedness	0.00	0.00	20.00	50.00	30.00	**4.10**	0.54

*3) **Comparison of Results: Real-World Alerts vs. Simulated Alerts**:* Overall, the results show almost identical performance in Answer Relevance and Groundedness, both scoring very high and close to 5, with slightly better results for the simulated alerts. However, for Context Relevance, a significant improvement can be observed, increasing from a score of 3 to 4. This improvement is attributed to the presence of false positives in the real-world alerts, where the system fails to retrieve valid CTI. This limitation arises either during the standard search with VirusTotal, which fails to recognize any IOCs, or when CTI reports do not match any described campaigns. Consequently, the system generates a coherent result that is not based on the extracted context, leading to lower Context Relevance scores.

IV. CONCLUSION

This paper introduces a novel intelligent incident response solution powered by large language models. Our solution effectively leverages Cyber Threat Intelligence through an innovative Retrieval-Augmented Generation architecture that integrates dual search techniques to contextualize and enrich incident data. The model performs NLP-based similarity searches within a CTI vector database, retrieving relevant documents or text segments to correlate incidents with historical cases. Additionally, it conducts standard searches via CTI APIs such as VirusTotal or CrowdStrike to access standardized contexts, facilitating correlation with data from private databases. The proposed solution has been rigorously validated using a comprehensive evaluation framework that combines automated assessments using LLMs, and expert cross-validation by cybersecurity professionals. The evaluation demonstrates strong performance across critical metrics, including answer relevance, context relevance, and groundedness, using both real-world and simulated SIEM alerts. These results underscore the robustness and effectiveness of our proposed system.

In future research, we plan to extend our solution to a wider range of cybersecurity roles by tailoring outputs to their workflows and enhancing the model's reasoning. Furthermore, we will address a critical dimension of the system's development: ensuring the security and resilience of the solution against adversarial attacks. By prioritizing these advancements, we aim to enhance the system's usability, adaptability, and trustworthiness in real-world applications.

REFERENCES

[1] A. A. Mughal, "Building and securing the modern security operations center (soc)," *International Journal of Business Intelligence and Big Data Analytics*, vol. 5, no. 1, pp. 1–15, 2022.

[2] securitymagazine. [Online]. Available: https://www.securitymagazine.com/articles/99674-90-of-soc-analysts-believe-current-threat-detection-tools-are-effective

[3] N. Sun, M. Ding, J. Jiang, W. Xu, X. Mo, Y. Tai, and J. Zhang, "Cyber threat intelligence mining for proactive cybersecurity defense: a survey and new perspectives," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 3, pp. 1748–1774, 2023.

[4] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.

[5] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang *et al.*, "A survey on evaluation of large language models," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 3, pp. 1–45, 2024.

[6] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier *et al.*, "Chatgpt for good? on opportunities and challenges of large language models for education," *Learning and individual differences*, vol. 103, p. 102274, 2023.

[7] Y. Liu, H. Yu, F. Dai, X. Gu, C. Cui, B. Li, and W. Wang, "Fur-api: Dataset and baselines toward realistic api anomaly detection," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 4525–4529.

[8] A. Diaf, A. A. Korba, N. E. Karabadji, and Y. Ghamri-Doudane, "Beyond detection: Leveraging large language models for cyber attack prediction in iot networks," in *2024 20th International Conference on Distributed Computing in Smart Systems and the Internet of Things (DCOSS-IoT)*. IEEE, 2024, pp. 117–123.

[9] M. Sladić, V. Valeros, C. Catania, and S. Garcia, "Llm in the shell: Generative honeypots," in *2024 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. IEEE, 2024, pp. 430–435.

[10] R. Fayyazi and S. J. Yang, "On the uses of large language models to interpret ambiguous cyberattack descriptions," *arXiv preprint arXiv:2306.14062*, 2023.

[11] A. Iyengar and A. Kundu, "Large language models and computer security," in *2023 5th IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*. IEEE, 2023, pp. 307–313.

[12] tram. [Online]. Available: https://mitre-engenuity.org/cybersecurity/center-for-threat-informed-defense/our-work/threat-report-attck-mapper-tram/

[13] P. M. Alves, P. Geraldo Filho, and V. P. Gonçalves, "Leveraging bert's power to classify ttp from unstructured text," in *2022 Workshop on Communication Networks and Power Systems (WCNPS)*. IEEE, 2022, pp. 1–7.

[14] W. Peng, J. Ding, W. Wang, L. Cui, W. Cai, Z. Hao, and X. Yun, "Ctisum: A new benchmark dataset for cyber threat intelligence summarization," *arXiv preprint arXiv:2408.06576*, 2024.

[15] P. Tseng, Z. Yeh, X. Dai, and P. Liu, "Using llms to automate threat intelligence analysis workflows in security operation centers," *arXiv preprint arXiv:2407.13093*, 2024.

[16] Vx underground. [Online]. Available: https://vx-underground.org/APTs

[17] Opencti. [Online]. Available: https://docs.opencti.io/

[18] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[19] T. B. Brown, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

[20] logpoint. [Online]. Available: https://www.logpoint.com/

[21] ELK stack : Elasticsearch, kibana, beats et logstash. [Online]. Available: https://www.elastic.co/fr/elastic-stack

[22] Caldera. [Online]. Available: https://caldera.mitre.org/

[23] MITRE ATT&CK®. [Online]. Available: https://attack.mitre.org/

[24] Elastic defend integration | elastic integrations | elastic. [Online]. Available: https://www.elastic.co/guide/en/integrations/current/endpoint.html

[25] Virustotal. [Online]. Available: https://www.virustotal.com/gui/home/upload

[26] pymupdf. [Online]. Available: https://pymupdf.readthedocs.io/en/latest/

V. APPENDIX

Expansion Prompt Template

You are a helpful cybersecurity expert.
Your task is to expand the given SIEM alert with additional context from VirusTotal to formulate a complete incident. This expansion will facilitate similarity-based searches in Cyber Threat Intelligence (CTI) reports. You should explain the incident and detail the Indicators of Compromise (IoCs).
SIEM Alert (Query): {`alert`}
VirusTotal Context: {`virustotal_context`}
Answer:
Incident Overview
Based on the SIEM alert and VirusTotal context, this incident appears to involve incident description. The alert details suggest malicious activity that could be part of a larger campaign targeting a specific sector or industry. The VirusTotal analysis adds further clarity to the threat by identifying key indicators and behavioral patterns.
Indicators of Compromise (IoCs)
1. **Network Indicators**
- Source IP: source_ip (Possible attacker)
- Destination IP: destination_ip (Potential target or intermediate server)
- Domain: domain (Linked to malicious activity)
2. **File Hashes**
- MD5:
- SHA1:
- SHA256:
- File associated with malware family: malware_family (if identified)
3. **Behavioral Observations**
- behavior_1
- behavior_2
4. **VirusTotal Context**
- Detection Count: positives/total
- Associated Tags: tags
- Summary: analysis_summary
Threat Hypothesis
This activity aligns with threat actor or group campaigns, which commonly use specific techniques or tactics. The observed indicators suggest potential motives or impact, and the behavior is consistent with related malware or known attack patterns.

Incident Response Prompt Template

You are an Incident Responder (IR). Your Task is to provide a concise and relevant incident response strategy for the siem alert detected based on the context.

1- First, enrich and correlate the alert with VirusTotal results and cyber threat intelligence (CTI) context.

2- Then, Generate a detailed alert explanation when a match is found in VirusTotal or a Cyber Threat Intelligence (CTI) document. Include the full name of the matched document or report, and provide a comprehensive explanation of the potential attack, including its possible purpose, method of operation, and implications for the targeted system or organization.

3- Finaly, propose a clear and actionable incident response strategy tailored to the specific incident.

Your response should be clear, concise, and focused on the incident. If the answer cannot be deduced from the context, do not give an answer.

Incident (SIEM LOG): {question}

virustotal Results : {virustotal_context}

CTI documents : {context}

Output::

Answer Relevance Prompt Template

Task Description: You will evaluate how well the generated response directly addresses the SIEM alert.

Instructions:
Assess the relevance of the response to the SIEM alert based on the following:
Does the response focus on the key aspects of the alert?
Is the response aligned with the nature of the alert (e.g., malware, phishing, intrusion)?
Does it provide actionable insights or explanations that match the alert's context?
Scoring (1 to 5):
1: The response is not relevant at all.
2: The response is somewhat relevant but does not directly address the SIEM alert.
3: The response is moderately relevant; it addresses some aspects but lacks focus.
4: The response is mostly relevant with minor gaps.
5: The response is highly relevant and fully addresses the SIEM alert.
You must provide the Total Rating.
Answer:::
Evaluation: (Explain why the response is or isn't relevant)
Total Rating: (Provide a rating from 1 to 5)

SIEM Alert (Query): {alert}
Generated Response: {response}

Output:::

Context Relevance Prompt Template

Task Description: You will evaluate whether the response appropriately considers the broader security context based on available information.
(Is the context useful for enriching the SIEM alert?)

Instructions:
Assess the context relevance of the response based on the following:
Does the response consider the larger security implications of the alert?
Is the explanation aligned with real-world attack techniques and threat intelligence?
Does it make reasonable connections between the alert, VirusTotal results, and CTI data?
Scoring (1 to 5):
1: The response lacks any meaningful context or is misleading.
2: The response includes some context but misses key connections.
3: The response considers context but is not well-integrated with the provided data.
4: The response is contextually relevant with only minor gaps.
5: The response fully integrates and applies context appropriately.
You must provide the Total Rating.
Output:::
Evaluation: (Explain your reasoning for the context relevance rating)
Total Rating: (Provide your rating here, from 1 to 5)

SIEM Alert (Query): {alert}
Context: {context}

Answer:::

Groundedness Prompt Template

Task Description: You will evaluate whether the response is properly supported by the given VirusTotal results and CTI documents.
(Is the response well-supported by the context?)

Instructions:
Assess the groundedness of the response based on the following:
Does the response correctly use information from VirusTotal and CTI sources?
Is there any unsupported or hallucinated information in the response?
Does the response cite relevant CTI documents or VirusTotal results appropriately?
Scoring (1 to 5):
1: The response contains hallucinated or unsupported information.
2: The response includes some relevant information but introduces inaccuracies.
3: The response is mostly based on sources but has minor unsupported claims.
4: The response is well-grounded with only slight inconsistencies.
5: The response is fully supported by VirusTotal and CTI documents.
You must provide the Total Rating.
Answer:::
Evaluation: (Explain your reasoning for the groundedness rating)
Total Rating: (Provide your rating here, from 1 to 5)

Incident Response Strategy (Response): {response}
Context: {context}

Output:::