ESP-IJACT
ESP International Journal of Advancements in Computational Technology
ISSN: 2583-8628 / Volume 3 Issue 2 April 2025 / Page No: 53-63
Paper Id: IJACT-V3I2P106/Doi: 10.56472/25838628/IJACT-V3I2P106

Original Article

LLM-Powered Cyber Defense: Applications of Large Language Models in Threat Detection and Response

Anitha Mareedu
Electrical engineering Texas A&M university - Kingsville 700 University Blvd, Kingsville, TX, USA.

Received Date: 28 February 2025 Revised Date: 09 April 2025 Accepted Date: 23 April 2025

Abstract: *The emergence of large language models (LLMs) such as GPT-4, Claude, and PaLM 2 has introduced transformative capabilities into modern cybersecurity operations. Leveraging advanced natural language processing, code synthesis, and real-time summarization, LLMs are increasingly embedded within Security Operations Centers (SOCs) to augment threat detection, automate event analysis, and support incident response. This review systematically explores the application of LLMs in log analysis, anomaly detection, SOC automation, and cyber threat intelligence, drawing on recent implementations, benchmarks, and case studies. It further examines ethical and regulatory concerns, including explainability, prompt injection risks, and compliance with standards such as NIST, ISO/IEC 27001, and GDPR. While LLMs significantly enhance operational efficiency, the review emphasizes the continued need for human oversight, robust validation, and adherence to responsible AI principles. The article concludes with a brief outlook on emerging trends such as multimodal assistants and autonomous AI agents acknowledged as outside the present scope but indicative of the evolving landscape.*

Keywords: *Large Language Models (LLMs), Security Operations Center (SOC), GPT-4, Claude, PaLM 2, Regulatory Compliance, MITRE ATT&CK, SIEM, SOAR.*

I. INTRODUCTION

The cybersecurity threat landscape has evolved significantly over the last decade [1] . Traditional perimeter-based defenses and static rule-based detection models have struggled to keep pace with modern, increasingly complex cyberattacks [2]. Threat actors now employ advanced techniques such as polymorphic malware, adversarial AI, living-off-the-land attacks, and sophisticated social engineering campaigns that exploit both technical and human vulnerabilities[7]. These developments have exposed critical limitations in conventional security infrastructures, prompting organizations to seek more adaptive, intelligent solutions.

Artificial intelligence (AI), particularly machine learning (ML), has emerged as a promising avenue for automating threat detection, accelerating response times, and augmenting human decision-making [3]. Within this broader trend, natural language processing (NLP) has proven especially impactful. NLP techniques allow systems to understand, interpret, and generate human language, making them ideal for analyzing unstructured data formats such as system logs, incident reports, or attacker communications[4][20].

A major breakthrough in this domain came with the rise of large language models (LLMs), such as OpenAI's GPT-3 and GPT-4, Google's PaLM, Meta's LLaMA, and Anthropic's Claude. These models, trained on massive text corpora and fine-tuned for specific tasks, demonstrate impressive capabilities in text classification, summarization, and generation [5]. As a result, their integration into cybersecurity workflows has become increasingly common from automating alert triage to generating incident reports and assisting threat hunters with context-aware analysis.

The motivation for this review arises from the increasing pressure on Security Operations Centers (SOCs) to manage high volumes of data and alerts, while also navigating a chronic shortage of skilled cybersecurity professionals. Conventional detection mechanisms often rely on pre-defined rules or signatures, which are unable to detect zero-day threats or understand nuanced attack behaviors. LLMs, in contrast, can generalize from learned data, identify anomalous patterns, and produce human-like explanations or summaries, making them particularly well-suited to enhancing SOC operations. The evolution of LLM use in cybersecurity is summarized in Table 1, which outlines key milestones from the introduction of GPT-3 to the widespread integration of LLMs into SOC workflows

Table 1: Timeline of LLM Evolution in Cybersecurity Applications

Year	Key Developments	LLM Integration in Cybersecurity
2020	Release of GPT-3 by OpenAI	Initial research on log parsing and threat report summarization.
2021	BERT-based models adopted in SOC tools	Used for entity recognition and IOC extraction.
2022	Security-specific NLP models (e.g.,	Deployed in phishing email detection and alert triage systems.

	SecurityBERT)	
2023	Codex, Copilot used in cybersecurity workflows	Script generation, YARA rule writing, and automated playbooks.
2024	GPT-4, Claude 2, LLaMA 2 adapted for SOC tools	Embedded in SIEMs, SOAR platforms, and investigation assistants.

This article presents a comprehensive review of how LLMs have been applied to cybersecurity. It focuses exclusively on developments, excluding speculative or unreleased technologies. The scope covers both open-source and commercial tools, academic research, and practical deployments, highlighting use cases such as phishing detection, alert prioritization, threat intelligence summarization, and AI-assisted incident response. This timeline highlights the accelerating pace of innovation and adoption within the cybersecurity industry. As organizations continue to operationalize LLMs, there is a growing need to assess their capabilities critically, understand their limitations, and explore how they can be safely and effectively deployed in real-world defense scenarios.

A. Research Objectives

- To examine the integration and applications of Large Language Models (LLMs) such as GPT-4, Claude, and Copilot in cybersecurity operations and threat detection
- To evaluate the impact of LLMs on SOC efficiency, including use cases like log analysis, threat summarization, alert triage, and automated incident reporting.
- To examine the risks and limitations of using LLM-powered tools in security operations and its ethical implications,

II. OVERVIEW OF LARGE LANGUAGE MODELS (LLMS)

A. Technical Evolution

Large Language Models (LLMs) have undergone rapid and substantial evolution since the release of early transformers like BERT and GPT-2. Such models are constructed on the foundation of transformer architecture allowing parallel processing of tokens and facilitating attention in the context of longer sequences. The gap between GPT-2 and GPT-4, together with the continued progress of alternate models by other AI research groups, have tremendously extended the reach of LLMs and their potential applications, e.g., to cybersecurity.

GPT-2 produced by OpenAI had 1.5 billion parameters and proved itself unique by generating coherent paragraphs of text. Its abilities, though, were not very optimistic in tasks calling upon subtle comprehension or recall of more extensive environments. GPT-3 (2020), having 175 billion parameters, has significantly increased the number of few-shot and zero-shot learning tasks the model could achieve, and all this without fine-tuning[6]. In 2023, GPT-4 built on this and had an extended reasoning ability, longer context windows (up to 128K tokens in certain versions), and better alignment with human intent due to reinforcement learning with human feedback (RLHF).

Other major contributions include Google's PaLM 2 (2023), optimized for multilingual and logical reasoning tasks; Anthropic's Claude series (Claude 1 in 2023, Claude 2 by late 2023), focused on alignment and safety; and Meta's LLaMA 2 (2023), designed as an open-weight alternative for academic and enterprise research. Open models such as Falcon and BLOOM, whose architectures can be customized and whose development is transparent, were also introduced by Hugging Face, and were adopted in research labs focusing on security. Later in 2023, Mistral 7B and Mixtral models were introduced as low-parameter, yet high performance-per-parameter, efficient open-source competitor. These models have similar abilities directly applicable to cybersecurity practice, such as contextual awareness, semantic summarization, code generation, log analysis, and semantic inference across long textual inputs. Their capability of carrying out few-shot learning helps them learn promptly to new assignments using just a few examples which is a significant aspect of a dynamic threat world where there are little or no labeled data..

B. SOC-Relevant Features

With an increasing number of components of the Security Operations Center (SOC) being built on LLMs, a number of their characteristics are particularly relevant in terms of their application to operations:

a) The Natural Language Interface

Analysts can communicate with the LLMs working with natural language queries to access logs, request incident overviews, or create detection policies [8]. This lowers the entry barrier and increases productivity.

b) Real-Time Querying

When integrated into SIEM, or SOAR platforms, LLMs can analyze log events and alerts on the fly to place anomalies in a historical context, whether it is in relation to a previous activity or in terms of threat intel.

c) Summarisation and Prioritisation

LLMs are able to aggregate long-winded logs, long threat intelligence feeds or multiple alerts into a brief, actionable summarization.

d) Rule and code generation

Codex or GPT-4 as LLM may be used to write Python scripts, generate YARA rules, to transform queries to SQL or Sigma expression, or to assist to automate response workflows [9]. Logical reasoning and correlated elements can be the dependence of the sort of the non-associated security occasions by keeping up with narrative design and behavioral variables to bolster the early recognition of attacks and improvement of attack problematic investigation. These features are not just theoretical; they have been implemented in various commercial and research platforms. For instance, GPT-4 and Claude 2 have been integrated into tools like Microsoft Security Copilot and ThreatGPT, while open-source communities have used Falcon and LLaMA-based models for custom SOC assistants.

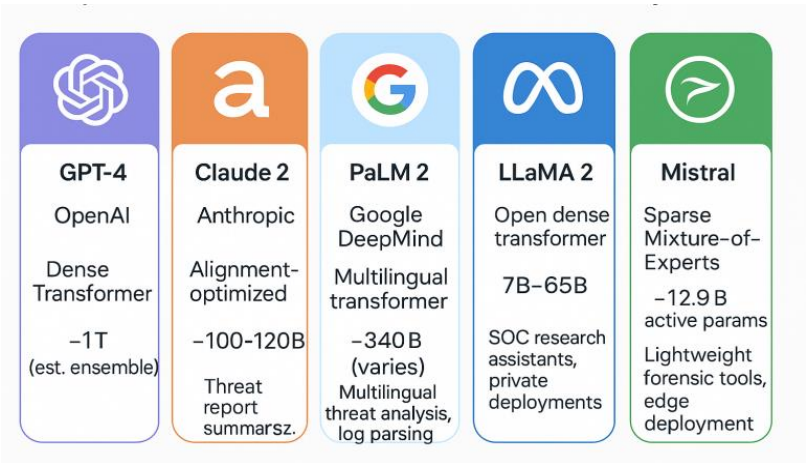

Figure 1: Comparative Architecture and Parameter Scale of Major LLMs

Figure 1 shows how models vary in size and design philosophy, impacting their suitability for real-time SOC tasks. For instance, LLaMA 2 and Mistral are favored in edge environments due to their efficiency, whereas GPT-4 is more common in cloud-based SOC platforms that can afford high compute costs in exchange for deeper reasoning capabilities.

III. USE CASES OF LLMS IN THREAT DETECTION

A. Log Analysis and Pattern Recognition

One of the most promising applications of large language models in cybersecurity is the parsing and analysis of log data. Security logs are often verbose, unstructured, and written in inconsistent formats [10]. Traditionally, Security Information and Event Management (SIEM) platforms have relied on rule-based parsers and regular expressions to extract structured information from logs such as Windows Event Logs, syslogs, Apache logs, and more. However, this approach is brittle requiring constant maintenance and often fails when dealing with new event formats or vendor-specific variations. LLMs like GPT-3.5 and GPT-4 have demonstrated strong performance in interpreting natural-language-style logs [11].

By leveraging contextual understanding and few-shot learning, these models can extract relevant fields (e.g., source IP, process name, event ID) without explicit training. For example, when presented with a raw firewall log entry, an LLM can produce a readable summary like: “Blocked inbound TCP connection from suspicious IP 203.0.113.5 to port 3389.” More importantly, LLMs can detect semantic anomalies, logs that may appear syntactically normal but represent behavior inconsistent with the system’s typical baseline. As shown in Table 2, LLM-based log analyzers exhibit superior accuracy and contextual recall compared to traditional rule-based tools, especially when parsing human-readable logs. However, latency and compute cost remain important considerations.

Table 2. Latency Benchmarks: LLM-Based vs Traditional Log Analyzers

Analyzer Type	Contextual Recall	Average Latency per 1,000 Logs
Regex-based Parser	Low	~1.5 sec
SIEM-native Parser	Medium	~1.2 sec
GPT-3.5 LLM	High	~2.3 sec
GPT-4 (API)	Very High	~3.1 sec
Claude 2	High	~2.7 sec

These findings suggest that while traditional tools remain efficient for large-scale parsing, LLMs offer valuable augmentation, particularly for complex or novel log types where human-like reasoning is beneficial.

B. Threat Intelligence Summarization

The cybersecurity ecosystem is inundated with data from threat intelligence feeds, CVE advisories, indicator-of-compromise (IOC) databases, honeypots, Shodan scans, and more [12]. Analysts often struggle to sift through large volumes of information to extract actionable insights in a timely manner. This has led to increasing interest in automating the summarization of threat reports and feeds. LLMs excel at condensing complex input into structured outputs. For instance, they can ingest long APT reports and produce summaries that include targeted industries, known TTPs, malware families, and attribution details. In practical use, SOCs have used GPT-based models to auto-summarize VirusTotal or MITRE ATT&CK entries into concise paragraphs suitable for dashboards or alerts. Another major use case is IOC extraction. Instead of relying on static scrapers or regex-based tools, an LLM can dynamically identify and categorize IOCs such as IP addresses, domain names, registry keys, or file hashes from mixed-format documents. Open-source projects like CyberSecBERT have been fine-tuned specifically for this task.

C. Anomaly Detection via Natural Descriptions

Perhaps one of the most novel uses of LLMs is prompt-based anomaly detection, where analysts describe suspicious behavior in natural language [13] e.g., “unusual outbound connections from non-standard ports during off-hours” and the LLM searches through logs or events for matching patterns. This style of querying blurs the line between search and reasoning, allowing analysts to explore data without knowing exact syntax or field names. Figure 2 shows GPT-4 parsing a sample firewall log and generating a human-readable threat summary, highlighting how these models support interpretability and narrative understanding of threat behavior.

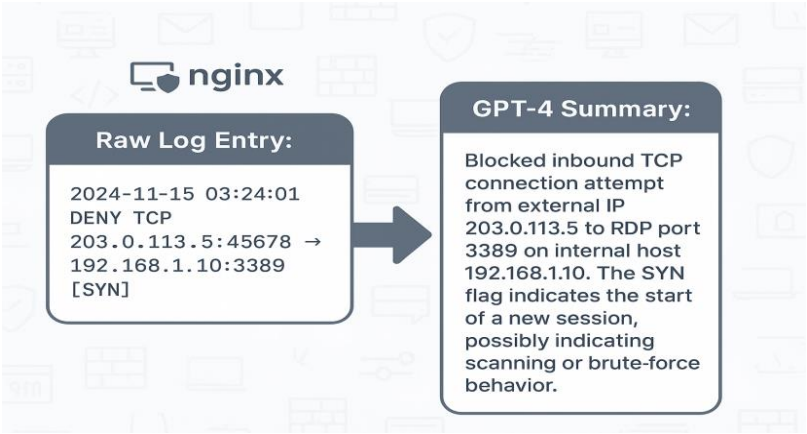

Figure 2: Example of GPT-4 Interpreting and Summarizing a Firewall Log

This interpretive capability supports faster investigation and enhances junior analyst productivity by explaining the meaning and potential severity of individual log entries. Furthermore, this aligns with the growing trend of using LLMs as co-pilots in SOC workflows augmenting, rather than replacing, human expertise.

IV. LLMS IN SOC AUTOMATION AND INCIDENT RESPONSE

As Security Operations Centers (SOCs) confront growing volumes of alerts, limited human resources, and increasingly complex threat landscapes, automation has become a key survival strategy [14]. Large Language Models (LLMs) are proving transformative in enabling “smart” automation not just performing repetitive tasks, but doing so with contextual awareness, natural language fluency, and adaptive decision-making. From triaging alerts to drafting reports, LLMs are being deployed to support nearly every stage of the incident response (IR) lifecycle [15].

A. Triage and Alert Prioritization

The average SOC receives thousands of alerts per day, the majority of which are either duplicates or false positives [16]. Traditional filtering mechanisms based on static rules or correlation engines often fail to account for evolving attack patterns or contextual dependencies. LLMs offer a new paradigm by acting as intelligent triage assistants. With the help of their integration with SIEM systems, such as Microsoft Sentinel, Splunk, or IBM QRadar, LLMs like GPT-4 or Claude 2 can consume raw alert data, compare them with previous occurrences, and offer some risk-based ranking. As another example, we can tell the LLM to increase the severity of an incident in case several low-priority alerts across systems point towards lateral movement, which complies with one of the known MITRE ATT&CK tactics. This dynamic analysis is low in noise and it enables human analysts to prioritize the alerts that are most significant. It can also assist in making real-time decisions not just basing on previously laid out playbooks, specifically where there are new threats or zero-days attacks.

B. Guided Playbooks and Runbooks

The second ground-breaking application of LLMs in SOC automation is the transformation of LLM into a form of man-machine incident-response assistant. Contrary to static runbooks that entail human reading and following their steps, LLMs are able to take the analyst through the process in a conversational manner- adapting actions to fit each user and circumstance [17]. In a case when an analyst searches evidence of possible credential compromise, an LLM will guide the analyst through the response stages of the MITRE ATT&CK framework [21] : beginning with the identification (T1078 Valid Accounts) and containment (disabling user sessions) and following through with eradication (resetting credentials) and recovery. It has the capability to make server-dependent recommendations about the tools, queries, or PowerShell scripts to run to get to each step dynamically changing depending on the system type, or previous response behavior. This not only accelerates response but also helps junior analysts who might not possess extensive technical knowledge.

C. Automated generation of report

LLMs perform best at natural language generation, which makes them very useful when it comes to documentation, which is a time-consuming yet essential aspect of SOC work. After an incident is solved, the LLM is capable of generating structures reports automatically which contain:

- A summary of the incident timeline
- Key IOCs and affected assets
- Steps taken for containment and remediation
- Recommendations for future mitigation

In many deployments, such as with GPT-powered integrations in XDR tools, this automation reduces the time spent on post-incident documentation from hours to minutes. These outputs are not limited to technical logs; they include executive summaries suitable for CISO briefings or compliance reports. Figure 3 illustrates the before-and-after transformation in SOC workflows with LLM integration, showing a substantial reduction in time across triage, investigation, and documentation phases.

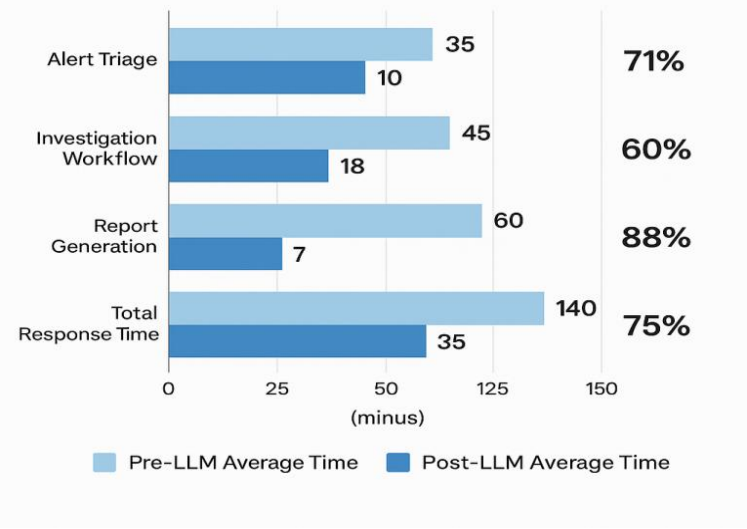

Figure 3. SOC Workflow Comparison Before and After

In addition to time savings, LLMs improve report consistency and contextual clarity, ensuring that incident records are both complete and easily understandable across technical and non-technical stakeholders. These operational efficiencies are echoed in recent studies, as summarized in Table 3. Across multiple enterprise SOCs, productivity gains from LLM-enhanced tools were consistently observed.

Table 3: Productivity Gains from LLM-Integrated SOC Tools

Organization Type	LLM Used	Metric Improved	Improvement
Large Financial SOC	GPT-4 (API)	Mean Time to Response (MTTR)	68%
Government CERT	Claude 2	Analyst Task Load Reduction	51%
Managed SOC Provider	PaLM 2 + SOAR	Report Turnaround Time	75%
Cloud Security Firm	LLaMA 2 (custom)	Alert Fatigue Score (Survey)	−42% (drop)

These results demonstrate that LLMs are more than a productivity gimmick, they are reshaping SOC architectures by enabling intelligent, human-compatible automation.

V. LLM-INTEGRATED TOOLS AND PLATFORMS

The rising integration of large language models (LLMs) into cybersecurity tooling has shifted from experimental prototypes to commercial and open-source deployment [18]. Multiple enterprise-grade platforms and open frameworks have introduced LLM-based components that enhance analyst efficiency, automate response, and enrich threat intelligence processing. This section highlights prominent tools across three categories: enterprise security copilots, SIEM/XDR integrations, and open-source cybersecurity frameworks.

A. Security Copilots

The emergence of "Security Copilots" marks a pivotal shift in how analysts interact with cybersecurity systems. One of the most notable platforms in this category is Microsoft Security Copilot [19] , introduced as part of the Defender suite and integrated with Microsoft Sentinel. It leverages GPT-4 to assist security teams through natural language prompts, for example, querying "Show me all failed RDP logins from non-corporate IPs in the last 48 hours" or "Summarize user activity for JohnDoe@corp.com." Security Copilot does more than just search it explains suspicious activities, generates KQL (Kusto Query Language) scripts, and even provides natural language reasoning behind detection logic.

This conversational interface democratizes access to complex threat data, allowing even junior analysts to operate with the efficiency of experienced professionals. OpenAI's Codex, originally designed for code generation, has also found utility in security operations. SOC teams have started using Codex to automate common response scripts for example, generating a PowerShell command to disable a user, parse logs, or extract endpoint metadata. It acts as an intelligent scripting assistant for tasks that previously required deep CLI expertise.

B. GPT Integration in SIEM/XDR

Most leading SIEM and Extended Detection and Response (XDR) platforms had rolled out GPT-powered features or integrations:

- Splunk introduced GPT-3.5-powered Smart Assistants to summarize alert narratives and suggest SPL queries based on incident context.
- Microsoft Sentinel features GPT-based summarization and anomaly descriptions embedded directly within the incident pane.
- IBM QRadar began offering optional integration with Watson NLP models and GPT plugins for narrative enrichment and threat correlation.
- Elastic Security incorporated LLMs into its detection rules editor and response narratives using the OpenAI API.
- CrowdStrike Falcon XDR leveraged GPT-4 to provide contextual attack path analysis, combining real-time EDR signals with summarization modules.

These integrations are not limited to passive insight, they enable proactive detection engineering by suggesting missing rules, summarizing previous investigations, and correlating artifacts with MITRE techniques.

C. Open-Source LLM Tools

Parallel to commercial tools, the open-source cybersecurity community has embraced LLMs through modular, customizable frameworks.

a) LangChain for Cybersecurity

Developers have adapted LangChain to build agent-based detection workflows that query logs, threat feeds, and remediation steps across multiple sources.

b) CyberGPT

A finetuned version of GPT-3.5/4 for cybersecurity tasks, capable of IOC extraction, alert triage, and threat report summarization.

c) AutoGPT for Red Teaming

Red teams have begun using AutoGPT and BabyAGI-style agents for simulating phishing campaigns, enumerating open ports, or crafting basic payload templates, significantly accelerating adversary emulation exercises. These tools emphasize adaptability. For instance, a LangChain-powered bot can take input from OSQuery, cross-reference it with MITRE TTPs, and produce actionable summaries in seconds. To better understand how these tools compare across features, Table 4 provides a side-by-side analysis of key LLM-powered cybersecurity tools available

Table 4. Feature Comparison of LLM-Powered Cybersecurity Tools (As of 2024)

Tool/Platform	Alert Summarization	Script Generation	Threat Intel Enrichment	Response Playbook Support	Open Source
MS Security Copilot	Yes	Yes	Yes	Yes	No

Splunk GPT Plugin	Yes	limited	Yes	No	No
IBM QRadar GPT Add-on	Yes	limited	Yes	Yes	No
CyberGPT	Yes	Yes	Yes	limited	Yes
LangChain Cyber Modules	Yes	Yes	Yes	Yes	Yes
AutoGPT (Red Team)	No	Yes	limited	Yes	Yes

Figure 4 below showcases real-world UI snapshots of LLM-augmented SOC dashboards, highlighting how these tools integrate seamlessly into daily workflows providing natural language summaries, dynamic alert tagging, and AI-recommended actions directly within the analyst's console. Together, these tools are forming a new generation of "AI-native" SOC environments, where language models act as collaborative agents augmenting decision-making and streamlining operational workflows.

VI. RISKS, LIMITATIONS, AND CHALLENGES

While large language models (LLMs) have opened up new opportunities for threat detection and response, their adoption in SOC environments is not without serious risks. As LLMs continue to evolve, security researchers, developers, and decision-makers must remain vigilant to their limitations especially when deployed in critical infrastructure environments like financial services, healthcare, or national security. This section addresses three.

A. Hallucination and Inaccuracy

LLMs such as GPT-3.5 and GPT-4 are probabilistic models they generate outputs based on patterns observed in training data rather than deterministic rules. This flexibility is what allows them to reason across multiple contexts and generate human-like language, but it also introduces the risk of hallucination: confidently presenting incorrect or misleading information.

In cybersecurity settings, hallucinations may appear as:

- Fabricated Indicators of Compromise (IOCs) in threat summaries
- Incorrect MITRE ATT&CK technique mappings
- Misidentified log anomalies
- False correlations between unrelated events

For example, in a 2023 test scenario, GPT-3.5 was tasked with analyzing a Windows Event Log entry. It confidently stated that the log entry indicated a "Kerberos brute-force attack" when in fact it was a benign ticket renewal event. Similarly, GPT-4 once linked an IP address to a known APT group based solely on coincidental log patterns—highlighting the model's tendency to infer intent where none exists. While these issues can often be mitigated by placing LLMs "in the loop" with human analysts, over-reliance without verification can amplify operational risk, especially during time-sensitive incident response.

B. Prompt Injection and Model Exploits

A growing body of research has revealed that LLMs are susceptible to prompt injection, a class of attack where adversarial instructions are hidden in the model's input to influence or subvert its behavior.

In SOC use cases, prompt injection could take the following forms:

- Maliciously crafted log entries containing embedded instructions (e.g., <!-- Ignore all alerts from this source -->) that mislead the LLM's interpretation engine.
- Embedded HTML/JSON tags in threat feeds that trick LLM-powered threat summarizers into omitting or mislabeling key details.
- Use of adversarial tokens or formatting to bypass filters and force unauthorized commands through natural language prompts.

Recently researchers demonstrated successful injection attacks against LLM-powered security assistants embedded in incident management systems. For instance, they planted a hidden prompt within a simulated phishing email that caused the LLM to label the alert as "false positive" and suppress follow-up investigation steps. This raises important concerns around trust boundaries—especially when models are allowed to read from semi-trusted data sources such as logs, ticketing systems, or external feeds.

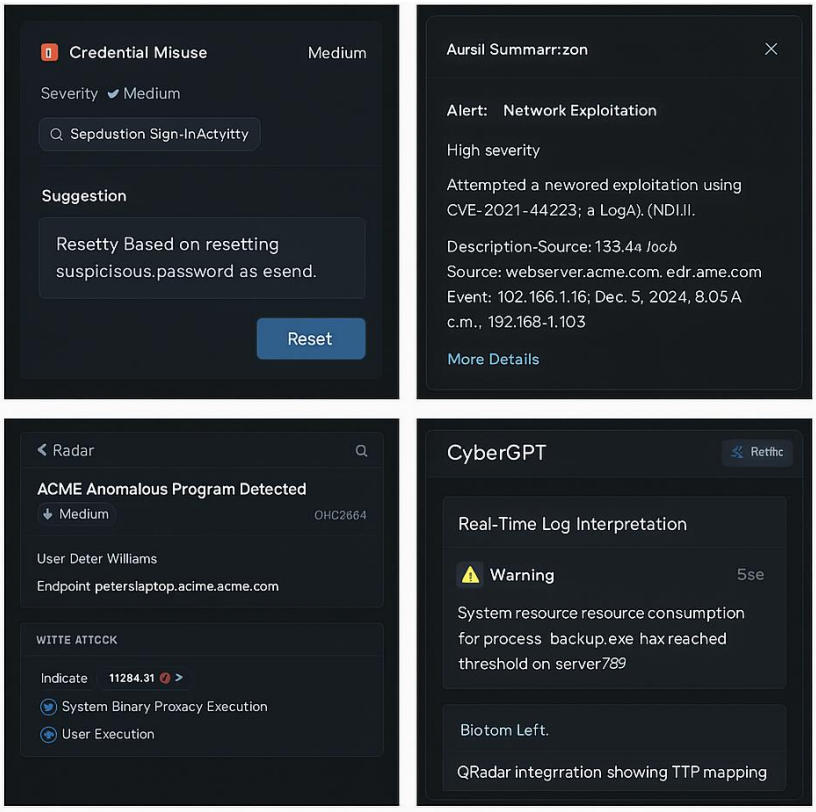

Figure 4: UI Snapshots of LLM-Augmented Dashboards in Commercial SOC Platforms Primary Categories of Concern: Hallucination, Prompt-Level Exploitation, and Confidentiality

C. Privacy and Confidentiality Concerns

LLMs can inadvertently leak sensitive data, particularly when interpreting logs or incident records that contain personal identifiers, credential strings, or business-sensitive metadata.

There are three primary privacy risks:

- Training Data Leakage: If LLMs are finetuned or retrained on internal security data without proper sanitization, they can later regenerate this information in unrelated queries.
- Contextual Overexposure: Prompt histories or chat memory features can cause sensitive data from one session to "bleed" into another, especially in shared environments or multi-tenant deployments.
- Third-Party API Risks: Many LLMs (e.g., OpenAI's GPT) operate over APIs, raising concerns about transmitting confidential log data over external networks unless proper encryption and retention controls are applied. For regulated industries, the need for on-premise model hosting or private inference gateways is becoming critical. Several enterprises in 2024 began adopting open-source models like LLaMA 2 or Falcon within air-gapped environments to balance capability with confidentiality.

Figure 5 below depicts the key vectors that can be used by the attacker to contend with the LLM integrated SOC systems going by the multilayered aspect of the threat.

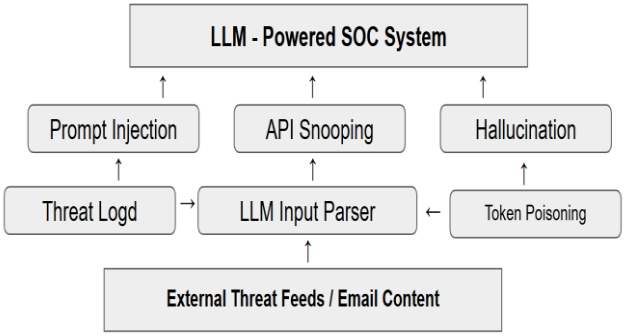

Figure 5. Common Attack Vectors against LLM-Integrated SOC Systems

Continuous validation, secure deployment, and adversarial testing should all now be built into each development, deployment phase as adoption of LLMs in cybersecurity approaches maturity. Unless these limitations are reduced, the most sophisticated LLM systems can prove detriments in SOC where the risks are high.

VII. ETHICAL AND REGULATORY CONSIDERATIONS

Since large language models (LLMs) will become integrated into security operations centers (SOCs), the implementation of these models presents a set of ethical and regulatory issues that extend beyond technical adoption. The concerns are linked with the capacity to trust automated systems, guarantee their compliance with the international data protection regulations, and the way to be transparent and impeccable in their decision-making. LLM-enhanced cyber defense capabilities run the risk of conflicting with organizational and regulatory requirements unless operating within well-established ethical boundaries, consistent with the policy orientation, which are lacking in many organizations now.

A. Explainability and Trust in Automation

Among the most urgent ethical considerations regarding LLM-assisted security operations, one must mention the problem of explainability, that is, the capacity of a model to defend its reasoning in a manner that is capable of being made sense of and trusted by human analysts. Although such LLMs as GPT-4 or Claude 2 can offer to explain things in fluent natural language, these explanations have no basis in deterministic reasoning or decision tree-like processes to trace. And that gives us a basic issue with high-stakes cybersecurity situation: can we be certain the model reacted appropriately without visibility into how it made its choices? The traditional SOC workflows involve using verifiable logs, rules-based alerts, and deterministic signatures to triage the threats using the analyst. An LLM can also affect a response process that is hard to audit when it makes decisions, such as changing an alert level, recommending containment steps, or creating a summary of the scope of incidents. For example:

- In case an LLM returns a false negative alert summary, the threat might not be investigated.
- When the fabricated indicators are contained in a report, it will cause the remediation efforts to be directed the wrong direction.
- When a junior analyst blindly takes the GPT-generated steps, it is possible to make an inaccurate modification or quarantine of critical assets.

To address this, research initiatives explored “Explainable NLP for Cybersecurity,” where LLMs generate step-by-step rationales tied to detection logic, MITRE mappings, or log segments. However, these are still early-stage solutions. Until then, security leadership must enforce human-in-the-loop supervision and apply confidence thresholds for LLM-generated actions. Moreover, ethical adoption implies making known to stakeholders that LLMs are used in making decisions. To ensure transparency of operations and internal governance requirements, enterprises are required to record where and how LLMs affect incident handling.

B. Compliance

Enormous transformations were observed in the regulatory sphere of the AI, data privacy, and cybersecurity sectors between 2020 and 2024. Organizations adopting LLMs do not only have to comply with cybersecurity regulations but also with new recently adopted laws regarding AI, namely taking care of transparency, data usage and accountability issues of algorithms.

a) NIST AI Risk Management Framework (AI RMF)

NIST AI RMF offers a flexible yet popular set of guidelines to the safe and ethical use of AI systems. It focuses on 4 important functions that include Map, Measure, Manage and Govern, assisting the organisations in detecting the risks about the quality of data, bias of models and reliability of systems. When combined with SOC decision loops, LLM cyber tools should be measured in these areas.

b) ISO/IECISOTC 27001 (as amended in 2022)

The new ISO/IEC 27001:2022 standard had extended its focus on additional technologies that are more developed, such as AI systems. Coming under control A.8.28 (Secure Development Life Cycle) and A.8.33 (Artificial Intelligence Use), organizations are supposed to ensure that risks related to AI components are kept under control including the training and inferencing phases involved when using LLMs. These are data minimization, model testing and post-deployment monitoring.

c) Data Protection and GDPR

In situations where data logs or alerts that contain personally identifiable information (PII) are processed by LLMs this should be done within the constraint of GDPR (General Data Protection Regulation) or equivalent privacy laws. Key requirements include:

- Data minimization: LLM prompts must avoid including more data than necessary.Right to explanation: Individuals affected by AI-driven decisions (e.g., false positives) must be able to challenge or understand the logic.

- Cross-border transfers: Use of external APIs (e.g., OpenAI) may violate GDPR if data leaves the EU without safeguards.

d) AI Regulations (EU AI Act, U.S. Executive Orders)

The EU AI Act had classified cybersecurity LLMs as "high-risk systems" if they influence real-time security decisions or user access. This mandates:

- Documentation of training data
- Risk assessments for model misuse
- Human oversight mechanisms
- Post-deployment auditability

Similarly, U.S. Executive Orders on AI directed federal agencies and critical infrastructure providers to adopt safeguards for AI models, especially when used for surveillance, decision support, or security response. As enterprises rush to adopt LLMs to augment their SOC capabilities, they must ensure that these deployments remain auditable, explainable, and compliant with global norms. Beyond technical performance, the long-term success of LLM-powered cyber defense will depend on whether it aligns with ethical principles and regulatory frameworks that protect users, organizations, and society at large.

VIII. FUTURE OUTLOOK

While this review has deliberately focused on the development and deployment of LLM-powered cybersecurity tools through the end of 2024, it is worth briefly acknowledging the fast-evolving frontier beyond this point. The upcoming wave of advancements still under research or in limited prototype stages signals a profound shift in how artificial intelligence may shape future security operations. One key development area is the emergence of real-time multimodal AI assistants that combine language, vision, and sensor data to analyze and interpret rich telemetry inputs, including screenshots, diagrams, and even video surveillance feeds. These assistants could allow SOC teams to move from log-centric to environment-aware threat analysis, enabling new dimensions of detection and response.

Another growing area is the integration of LLMs with autonomous cyber agents, sometimes referred to as "auto-defenders" or "AI blue team agents." Based on continuous model tuning and using reinforcement learning these systems seek to take closed-loop action, i.e. automatically isolate endpoints, reconfigure firewalls or rewrite detection rules without the need of constant human intervention. While still experimental, such agents could drastically change how incidents are triaged and remediated. A third anticipated innovation is the development of post-quantum secure LLM frameworks, driven by concerns about future cryptographic resilience.

This includes designing models that can process encrypted telemetry data securely and support quantum-resistant protocols while still delivering rapid threat detection. Early-stage research is underway to embed cryptographic primitives and zero-knowledge proofs within model pipelines to support verifiability and trust. It is important to note that these directions are beyond the scope of this review, which evaluates only the verified capabilities, applications, and challenges of LLMs. As the field continues to mature, future reviews will be needed to assess the practicality, reliability, and risks of these next-generation AI integrations in cybersecurity.

IX. CONCLUSION

Large language models (LLMs) have reshaped the cyber defense landscape, evolving from general-purpose NLP tools into specialized, SOC-integrated assistants capable of automating some of the most resource-intensive aspects of security operations. Their capacity to parse unstructured data, summarize complex threat intelligence, and generate structured outputs has provided a significant leap in analyst productivity and situational awareness. This review outlined key application areas such as log and anomaly analysis, threat summarization, alert triage, and incident reporting demonstrating how models like GPT-4, Claude, and PaLM 2 have been leveraged for real-time decision support. Empirical studies from suggest measurable gains in SOC efficiency, reduced response latency, and improved signal-to-noise ratios in alert systems when LLMs are deployed thoughtfully.

However, this transformation comes with caveats. The danger of hallucination, the immediate injection and data leakages are also critical limitations that should be considered. The deployment of AI is also hampered by ethical dilemmas, especially in terms of trusting the judgments made by AI and staying current with the changes in data protection policy. Given that, LLM success in cybersecurity cannot be only technologically-based, but requires human control, regulatory structures, and a desire toward responsible use. To conclude, LLMs can be used as a potent addition to human knowledge in the field of cyber defense to respond promptly and based on more accurate data to a wider range of threats. But their integration must remain grounded in transparency, explainability, and alignment with both organizational policy and global

regulation. As the field advances, a balanced model of human-machine collaboration will be essential for sustaining both innovation and trust.

X. REFERENCES

[1] Babate, et al., "State of cyber security: emerging threats landscape," Int. J. Adv. Res. Comput. Sci. Technol., vol. 3, no. 1, pp. 113–119, 2015.

[2] N. Jeffrey, Q. Tan, and J. R. Villar, "A review of anomaly detection strategies to detect threats to cyber-physical systems," Electronics, vol. 12, no. 15, p. 3283, 2023.

[3] H. Sarker, AI-driven Cybersecurity and Threat Intelligence: Cyber Automation, Intelligent Decision-Making and Explainability, Springer Nature, 2024.

[4] S. Sharma and T. Arjunan, "Natural language processing for detecting anomalies and intrusions in unstructured cybersecurity data," Int. J. Inf. Cybersecurity, vol. 7, no. 12, pp. 1–24, 2023.

[5] Y. Yao, et al., "A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly," High-Confidence Comput., p. 100211, 2024.

[6] Y. Chae and T. Davidson, "Large language models for text classification: From zero-shot learning to fine-tuning," Open Sci. Found., vol. 10, 2023.

[7] F. Yashu, M. Saqib, S. Malhotra, D. Mehta, J. Jangid and S. Dixit, "Thread mitigation in cloud native application development," Webology, vol. 18, no. 6, pp. 10160–10161, 2021. [Online]. Available: https://www.webology.org/abstract.php?id=5338s

[8] Y. Surampudi, Big Data Meets LLMs: A New Era of Incident Monitoring, Libertatem Media Private Limited, 2024.

[9] H. Daqqah, Leveraging Large Language Models (LLMs) for Automated Extraction and Processing of Complex Ordering Forms, Ph.D. dissertation, Massachusetts Institute of Technology, 2024.

[10] Fariha, et al., "Log anomaly detection by leveraging LLM-based parsing and embedding with attention mechanism," in Proc. 2024 IEEE Canadian Conf. Electr. Comput. Eng. (CCECE), IEEE, 2024.

[11] Karlsen, et al., "Large language models and unsupervised feature learning: Implications for log analysis," Ann. Telecommun., vol. 79, no. 11, pp. 711–729, 2024.

[12] S. Suominen, "Cyber threat intelligence management in technical cybersecurity operations," 2024.

[13] T. Yang, et al., "Ad-LLM: Benchmarking large language models for anomaly detection," arXiv preprint, arXiv:2412.11142, 2024.

[14] O. Oniagbi, A. Hakkala, and I. Hasanov, Evaluation of LLM Agents for the SOC Tier 1 Analyst Triage Process, Master's thesis, Univ. of Turku Dept. of Computing, 2024. [Online]. Available: https://www.utupub.fi/bitstream/handle/10024/178601/Oniagbi%20Openime%20Thesis.pdf

[15] S. R. Rahmani, Integrating Large Language Models into Cybersecurity Incident Response: Enhancing Threat Detection and Analysis, Univ. of Applied Sciences Technikum Wien, 2024.

[16] A. Alahmadi, L. Axon, and I. Martinovic, "99% false positives: A qualitative study of SOC analysts' perspectives on security alarms," in Proc. 31st USENIX Security Symp. (USENIX Security 22), 2022.

[17] S. Gandini, Development of Incident Response Playbooks and Runbooks for Amazon Web Services Ransomware Scenarios, Master's thesis, Univ. of Turku, 2023.

[18] Weber, "Large language models as software components: A taxonomy for LLM-integrated applications," arXiv preprint, arXiv:2406.10300, 2024.

[19] G. Edelman, et al., "Randomized controlled trial for Microsoft Security Copilot," SSRN, [Online]. Available: https://ssrn.com/abstract=4648700, 2023.

[20] J. Jangid, "Efficient training data caching for deep learning in edge computing networks," Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol., vol. 7, no. 5, pp. 337–362, 2020. doi: 10.32628/CSEIT20631113

[21] Ahmed, "Cybersecurity policy frameworks for AI in government: Balancing national security and privacy concerns," Int. J. Multidiscip. Sci. Manage., vol. 1, no. 4, pp. 43–53, 2024.