# Towards AI-Driven Human-Machine Co-Teaming for Adaptive and Agile Cyber Security Operation Centers

MASSIMILIANO ALBANESE, George Mason University, USA
XINMING OU, University of South Florida, USA
KEVIN LYBARGER, George Mason University, USA
DANIEL LENDE, University of South Florida, USA
DMITRY GOLDGOF, University of South Florida, USA

Security Operations Centers (SOCs) face growing challenges in managing cybersecurity threats due to an overwhelming volume of alerts, a shortage of skilled analysts, and poorly integrated tools. Human-AI collaboration offers a promising path to augment the capabilities of SOC analysts while reducing their cognitive overload. To this end, we introduce an AI-driven human-machine co-teaming paradigm that leverages large language models (LLMs) to enhance threat intelligence, alert triage, and incident response workflows. We present a vision in which LLM-based AI agents learn from human analysts the tacit knowledge embedded in SOC operations, enabling the AI agents to improve their performance on SOC tasks through this co-teaming. We invite SOCs to collaborate with us to further develop this process and uncover replicable patterns where human-AI co-teaming yields measurable improvements in SOC productivity.

CCS Concepts: • **Human-centered computing → Collaborative and social computing**; • **Information systems → Decision support systems**; • **Computing methodologies → Artificial intelligence**; • **Security and privacy → Intrusion detection systems**.

Additional Key Words and Phrases: Cybersecurity, Security Operations Centers, Human-AI Collaboration, Large Language Models

**ACM Reference Format:**
Massimiliano Albanese, Xinming Ou, Kevin Lybarger, Daniel Lende, and Dmitry Goldgof. 2025. Towards AI-Driven Human-Machine Co-Teaming for Adaptive and Agile Cyber Security Operation Centers. 1, 1 (June 2025), 19 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

## 1 Introduction

Security Operations Centers (SOCs) play a critical role in defending organizations against evolving cyber threats. However, SOC analysts face major challenges, including high alert volumes, repetitive tasks, and insufficient automation. Due to the ever-increasing complexity and interconnectivity of modern networks, the challenges of preventing, detecting, and responding to security incidents have surpassed the current capabilities of SOCs[1], placing an overwhelming burden on SOC analysts. Real-time threat intelligence is crucial, but the process of collecting, analyzing, and disseminating information can be time-consuming — as it relies heavily on human analysts to interpret data and make informed decisions — leaving organizations vulnerable to rapidly evolving threats. Additionally, a shortage of skilled cybersecurity professionals limits the ability of organizations to effectively leverage threat intelligence and respond to threats relevant

[1]In the literature, the terms Cyber Security Operation Center (CSOC) and Security Operation Center (SOC) are often used interchangeably, although the latter is more common

Authors' Contact Information: Massimiliano Albanese, George Mason University, Fairfax, VA, USA, malbanes@gmu.edu; Xinming Ou, University of South Florida, Tampa, FL, USA, xou@usf.edu; Kevin Lybarger, George Mason University, Fairfax, VA, USA, klybarg@gmu.edu; Daniel Lende, University of South Florida, Tampa, FL, USA, dlende@usf.edu; Dmitry Goldgof, University of South Florida, Tampa, FL, USA, goldgof@usf.edu.

to them. As these challenges intensify, the demand for more automation in incident response has emerged as a critical focus in cybersecurity research. Leading IT companies are at the forefront of developing tools and orchestration platforms to enhance the efficiency of security incident handling. IBM Resilient, Splunk Phantom, Microsoft Azure Sentinel, Cisco SecureX Orchestration, and Fortinet FortiSOAR are examples of prominent orchestration platforms that support incident response processes. They connect with various security tools and provide a central platform for managing and responding to security incidents. They aim to improve workflows, reduce response times, and increase incident handling efficiency.

Despite significant advancement in tooling support, SOC analyst burnout [52] remains a critical issue that was exacerbated by the COVID-19 pandemic [31]. Several factors contribute to this challenge. First, the sheer volume of alerts generated by security tools — most of which are false positives [3] — creates a significant burden for analysts. This flood of alerts is often plagued by uncertainty and incomplete data, making it difficult to prioritize and respond effectively. SOC analysts typically rely on *runbooks* or standard operating procedures (SOP) to address these alerts. However, the repetitive application of these procedures to predominantly false alarms leads to both fatigue and a lack of fulfillment in the role [52]. Moreover, mainstream security tools often lack the capability to connect important contextual information crucial for comprehending the threat landscape and adapting incident response strategies. Such contextual information may include the specific assets that require protection based on the organization's business needs, the nature of users in the organization's environment, the compliance requirements of the organization, and the political scenario or the economic situation of the organization's country. A human analyst typically considers these contextual factors to decide whether and how to act on an alert. Such reasoning is hard to capture into a typical algorithmic process that can be readily implemented within the traditional framework of building SOC tools. As a result, currently available SOC tools offer limited help to alleviate the pain points in the operations [52]. Sundaramurthy *et al.* have conducted a long-term anthropological study of SOCs spanning more than a decade and multiple corporate and higher education SOCs [52–54]. That work showed that, by becoming an anthropologist and embedding within a SOC, a security researcher can extract tacit knowledge within the SOC environment and create tools that (i) readily fit into the analysts' workflow, (ii) precisely address the pain points that result in burnout, and (iii) expand analysts' capability by proactively hunting and responding to threats [54]. This approach can be explained by the notion of tacit knowledge — a concept introduced by Polanyi [45, 46] — which refers to knowledge held by individuals that has not yet been explicitly articulated.

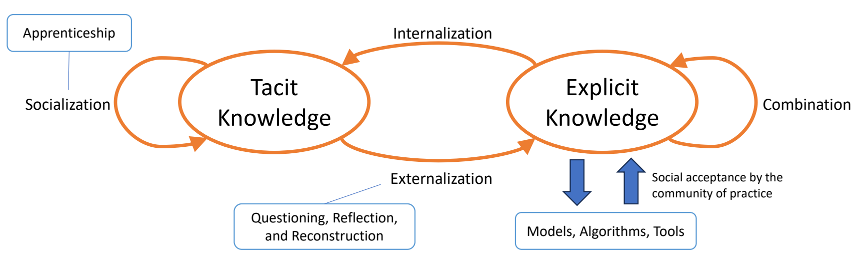


Fig. 1. SECI model for anthropology-guided SOC tool building.

The study in [53] uncovered extensive tacit knowledge among SOC analysts. To access this knowledge, researchers needed to first establish trust with the analysts by embedding themselves in a SOC, performing the same roles as the analysts while observing daily activities through the lens of an anthropologist. This immersive fieldwork, combined with qualitative analysis of fieldnotes, offered deep insights into operational challenges within SOCs that even analysts themselves had not fully recognized. Consequently, the researchers were able to create innovative tools to effectively support the analysts' workflows [52]. This process can be explained by the SECI knowledge model in Fig. 1, first proposed by Nonaka and Takeuchi [44]. Embedded researchers access tacit knowledge by becoming apprentices of SOC analysts and observing their work as it unfolds (*socialization*). Through questioning, reflection, and reconstruction, the researchers externalize the tacit knowledge into explicit forms such as documentation (*externalization*). The accumulated explicit knowledge is combined (*combination*), internalized (*internalization*), and then applied by the apprentices back to their work. This, in turn, forms new tacit knowledge, and the cycle repeats. Working inside the SOC allows researchers to be part of this knowledge conversion cycle, through which tools readily accepted by the analysts can be developed to codify the explicated knowledge and alleviate the burden on the analysts by automating the most repetitive and mundane part of the investigation task.

Traditional SOC tools, such as SIEMs, fail to capture the contextual and tacit knowledge necessary for effective decision-making. To address these limitations, we propose an AI-driven human-machine co-teaming approach that enhances SOC workflows through adaptive AI agents. This research builds upon our prior anthropological studies of SOCs, where researchers embedded within security teams uncovered the importance of tacit knowledge in cybersecurity operations. We extend this work by integrating Large Language Models (LLMs) to serve as AI apprentices, assisting analysts in real-time threat investigations.

The remainder of the paper is organized as follows. Section 2 presents relevant related work, whereas Section 3 introduces our vision and overall approach for human-machine co-teaming in SOCs. Then, Section 4 presents the proposed framework in detail. Section 5 discusses representative case studies and outlines directions for future research. Finally, Section 6 concludes the paper, summarizing our contributions.

## 2 Related Work

While still emerging, the application of LLMs in cybersecurity shows promise in tasks like penetration testing [19, 25, 47], incident response, vulnerability management, policy generation, security training, and detecting threats, malware, and intrusions. Research has primarily focused on classification tasks like identifying network intrusions and detecting malicious URLs [16, 26, 37, 48], with some exploration into synthesizing data into unified reports [49], including integrating global threat intelligence with local organizational knowledge [42]. However, research on leveraging LLMs to conduct SOC analysis and generate actionable outputs, such as creating or refining incident response plans [27], remains limited. While there are several cybersecurity-annotated datasets for training machine learning (ML) models on specific tasks [2, 12, 17, 23, 34], limited work has focused on instruction tuning to distill expert knowledge and enable diverse cybersecurity-specific tasks [36]. Furthermore, the study of continuous learning strategies in this domain remains largely unexplored. Real-world adoption faces challenges, including privacy concerns with proprietary, cloud-based models like ChatGPT, the high computational demands of LLMs, and the need for cybersecurity-specific models that can capture and distill analysts' knowledge into actionable insights aligned with specific environments and the requirements of individual SOCs [16, 26, 37, 48]. Additionally, the probabilistic nature of LLMs introduces variability that contrasts with the deterministic tools SOC analysts typically rely on, raising questions about reliability and trust when identical inputs yield different outputs. This issue is compounded by "hallucinations," where LLMs may provide inaccurate or fabricated information, creating concerns about their dependability for SOC operations [16, 26].

We argue that these perceived deficiencies of LLMs are not insurmountable obstacles to their adoption in SOCs. Rather, they arise from attempting to use generative AI models as traditional algorithmic tools. By adopting a human-centric approach and drawing parallels between human SOC analysts and LLMs, we can view an LLM-based AI agent as a collaborator rather than a deterministic tool [25]. A SOC analyst can use the AI agent to elicit ideas on how to proceed with investigations. The agent can provide supporting evidence for the suggestions, the validity of which human analysts can verify. This human-machine co-teaming is fundamentally different than how an analyst uses traditional SOC tools such as SIEMS. Human analysts do not blindly trust the output of the AI agent but rather must be convinced. The value provided by an AI agent is its ability to churn through vast amounts of data much faster than a human while navigating the nuanced semantics of the data like a human. For investigating cyber incidents, once the relevant data pieces are presented to a human, it is often quite clear what hackers have done and to what effect. However, finding the relevant data pieces across sheer volumes of data may easily overwhelm a human. A machine's upper bound on such bandwidth is much higher than a human's brain, but it must be capable of handling the intricate reasoning to *connect the dots* in the data. How human analysts connect the dots can hardly be specified in a fixed set of rules, and such knowledge is often tacit and hard to explain even by analysts themselves. LLMs present an opportunity to learn such tacit knowledge through data, leading to a human-machine co-teaming where human analysts are gradually liberated from the mundane tasks of processing vast numbers of tickets most of which are false positives [3].

Geoffrey Hinton highlighted parallels between human cognition and AI systems [33]. Like LLMs, human analysts are not immune to inconsistencies, errors, and knowledge gaps. However, SOC analysts mitigate these limitations through rigorous training that blends cybersecurity expertise with organization-specific practices. This training includes explicit knowledge such as regulatory compliance requirements, as well as tacit knowledge gained through experience such as recognizing anomalous patterns in network behavior or interpreting subtle contextual cues from threat intelligence reports. LLMs, like human analysts, can improve their reliability through iterative learning and feedback. By integrating mechanisms to capture both explicit and tacit knowledge, LLMs could evolve into valuable SOC collaborators, supporting analysts in a dynamic threat environment. When working with an LLM-based AI agent, analysts should critically evaluate its suggestions, just as they would evaluate those of a human colleague, requiring evidence-based justification. Viewing an AI agent as a partner rather than a deterministic tool enables SOC teams to utilize its assistance in much the same way they would rely on a human apprentice. Techniques such as prompt engineering, supervised fine-tuning, reinforcement learning from human feedback (RLHF), and retrieval-augmented generation (RAG) can facilitate this adaptation [20, 60]. We envision AI-driven human-machine co-teaming in SOCs, where SOC analysts work alongside AI agents, collaboratively tackling investigation tasks. These agents will enhance analysts' capabilities by internalizing the tacit knowledge required to process nuanced data and help human analysts scale up their work substantially.

In summary, Human-AI collaboration in cybersecurity has been explored in multiple domains, including automated alert triage, threat hunting, and decision support. Existing research on SOC automation highlights the limitations of rule-based systems and the need for adaptive AI-driven approaches. Our work differentiates itself by employing a human-centric co-teaming framework, where LLMs learn from real-world SOC workflows through iterative feedback and adaptation.

## 3 Vision Overview

Our vision for human-machine co-teaming in SOC operations, partly inspired by our anthropological study of SOCs, is illustrated in Fig. 2. We propose an LLM-based AI agent that acts as an apprentice to human SOC analysts, assisting through interactions with subtasks or investigative ideas. The human analyst will evaluate the agent's outputs. If the results are inaccurate or unhelpful, interactions could provide additional guidance or refute the LLM agent's answer with reasons. A human apprentice would have acquired tacit knowledge through such interactions; for an LLM agent, we need a specific learning process to capture such insights and incorporate the knowledge for the future. This learning process will result in an LLM agent specialized to handle the SOC tasks effectively.

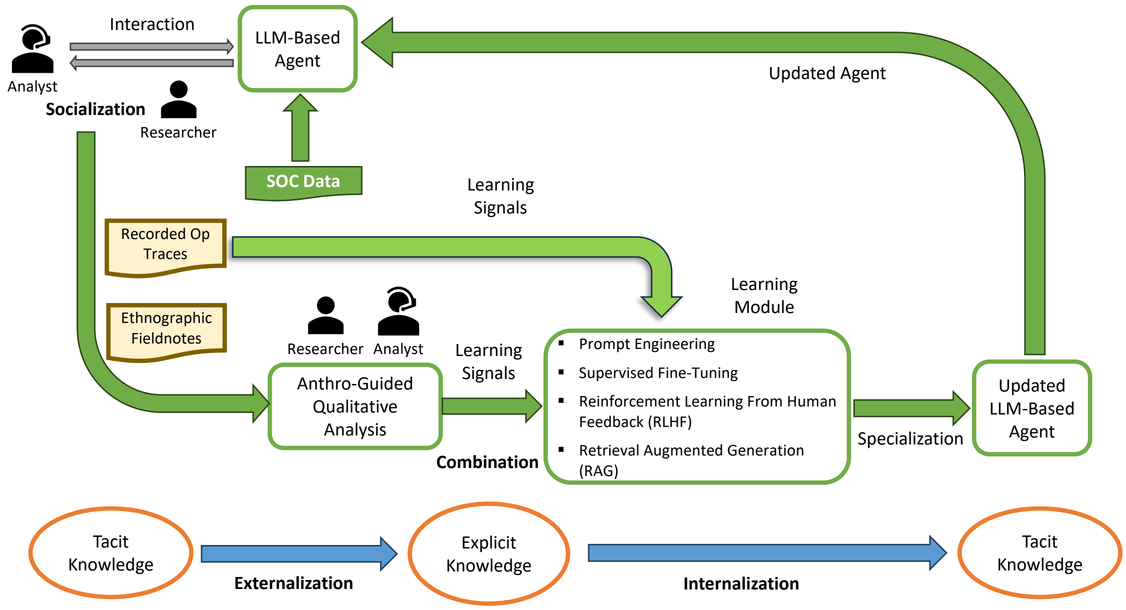


Fig. 2. Our vision for human-machine co-teaming for SOC operations.

The Learning Module utilizes learning signals derived from analyst-LLM interactions. Op traces consist of raw transcripts of conversations between the analyst and the LLM agent. The ethnographic fieldnotes are recorded by a researcher trained in anthropological methods and embedded in the SOC. This researcher observes interactions between the analyst and the LLM, capturing contextual information in fieldnotes that will be analyzed using qualitative research methods to reveal the tacit knowledge hidden in such interactions. Qualitative analysis can extract classifications, assessments, and higher-order concepts that analysts deem important in their work and, through coding, transform this data into annotations [21, 35]. These codes can describe the category or concept, provide relevant examples, and include inclusion and exclusion criteria (all basic parts of qualitative data analysis) [13, 55]. In the initial stages of the learning process, the embedded researcher can also provide a further source of input about gaps between human preferences and AI outputs, which will provide a triangulation between the SOC, the LLM agent, and the research team. Importantly, this triangulation is based on direct observation rather than secondary assessment, thus increasing the validity of what is working and what is missing in ongoing interactions between analysts and the LLM. Op traces, fieldnotes (including scalar descriptions of what is relevant and non-relevant), and analytic codes from anthropological research can all inform learning signals. These learning signals align the LLM agent's behaviors with the analyst's and the SOC's specific goals.

Given the sensitivity of SOC data, LLMs used in this research must be housed on premises, necessitating the use of open-source pre-trained foundation models (e.g., Falcon [8], LLaMA [56], Mistral [29]). These models are trained on broad and publicly available corpora, encompassing general web data, books, and technical documents, which may include some cybersecurity content; however, they lack the domain-specific expertise and operational context required for effective SOC operations. Our prior study shows that proficiency in SOC operations requires specialized tacit knowledge gained through experience. Thus, our proposed learning process is based on data generated through the AI agent's experience in handling real SOC tasks. This is a critical step in specializing the agent to assist analysts effectively. The agents will *learn on the job*, just as how human SOC analysts do, so their behavior will gradually align with the specific needs of the SOC. In this online learning paradigm, the agents will require less human input to produce useful results. The cycle can be seen as an instantiation of the SECI model (see Fig. 1) in the context of human-LLM co-teaming. It equips the LLM agent with the required tacit knowledge to perform the tasks specific to the SOC, developing a robust human-AI partnership to enhance the effectiveness and efficiency of SOC operations, reduce the cognitive load on analysts, and mitigate burnout.

### 3.1 Learning Strategies

Given the constraints that require LLM agents to operate in on-premises environments, it is essential to explore lightweight learning strategies that avoid the intensive resource demands of training new foundation models. Given the rapid development of LLMs and the emergence of new reasoning capabilities, there are many open questions regarding the tasks that these models can perform without any domain- or task-specific training. In this paradigm, where LLMs approach all language processing tasks as generative tasks, a range of learning strategies becomes relevant. Across tasks, our approach begins with prompt engineering, which provides LLM-specific natural language instructions on how to perform a task, serving as the foundational step in all experimentation. By crafting precise prompts, we can evaluate the baseline capabilities of a given LLM, such as Meta's LLaMA, for specific tasks without additional training. This provides critical insight into whether an LLM's inherent capabilities are sufficient or if further refinement is necessary. For tasks where baseline performance proves inadequate, we can employ supervised fine-tuning [41, 60], which involves curating a dataset of input-output pairs tailored to task requirements. This allows the model to specialize in specific tasks by training on the curated dataset. Additionally, we can leverage generative feedback [30], a cutting-edge approach where users provide natural language critiques of system outputs to guide improvements. This feedback enables users to articulate desired refinements and also provides a mechanism for systematically enhancing the LLM by addressing its current limitations. For more nuanced refinements, particularly in tasks involving subjective evaluation criteria, we can integrate reinforcement learning from human feedback (RLHF) [41, 60], which aligns LLM outputs with human preferences by using feedback such as rankings or scores in fine-tuning. In this process, the researcher ranks multiple system outputs (e.g., A is better than B) to guide the model towards outputs better aligned with human preferences and expert judgment. RLHF enables iterative improvement by embedding human-like preferences into the model's decision-making processes. Finally, when tasks require integrating external or internal knowledge sources, we can utilize retrieval-augmented generation (RAG) [20], which integrates external sources to retrieve relevant resources based on user queries and incorporate them into outputs, grounding responses in a predefined knowledge base and ensuring that outputs are informed by up-to-date and domain-specific data. This layered strategy, progressing from prompt engineering to fine-tuning, RLHF, and RAG, provides a flexible, scalable framework to effectively tackle project subtasks and achieve desired outcomes.

### 3.2 Overarching Research Questions

Our work is driven by the following two overarching research questions.

**Research Question 1**. What processes can be designed to enable LLM agents to effectively learn from experience and improve their performance on SOC tasks?
**Research Question 2**. To what extent can these learning processes enhance LLM agents' performance, resulting in productivity gains that substantially outweigh the effort invested in their development?

Our preliminary research on applying LLM in software security pen-testing [47] shows promise for the first research question, so we will leverage this work in our effort. Specifically, after a few iterations of the learning process, we can compare the updated LLM agent's performance against the previous version of the LLM agent on new tasks, utilizing metrics that reflect the agent's helpfulness. One metric could be the number of rounds needed for an analyst to obtain useful information to further the investigation. Our research aims to design specific metrics for an LLM agent based on the specific tasks it is assigned to work on. For the second research question, we can conduct human subject research to obtain feedback from the SOC analysts and use both qualitative and quantitative assessments to measure the return on investment of adopting the LLM agents.

## 4 Framework

Fig. 3 provides an overview of our framework for human-machine co-teaming, based on the vision laid out in Section 3. The framework comprises four primary modules, each comprising several submodules.

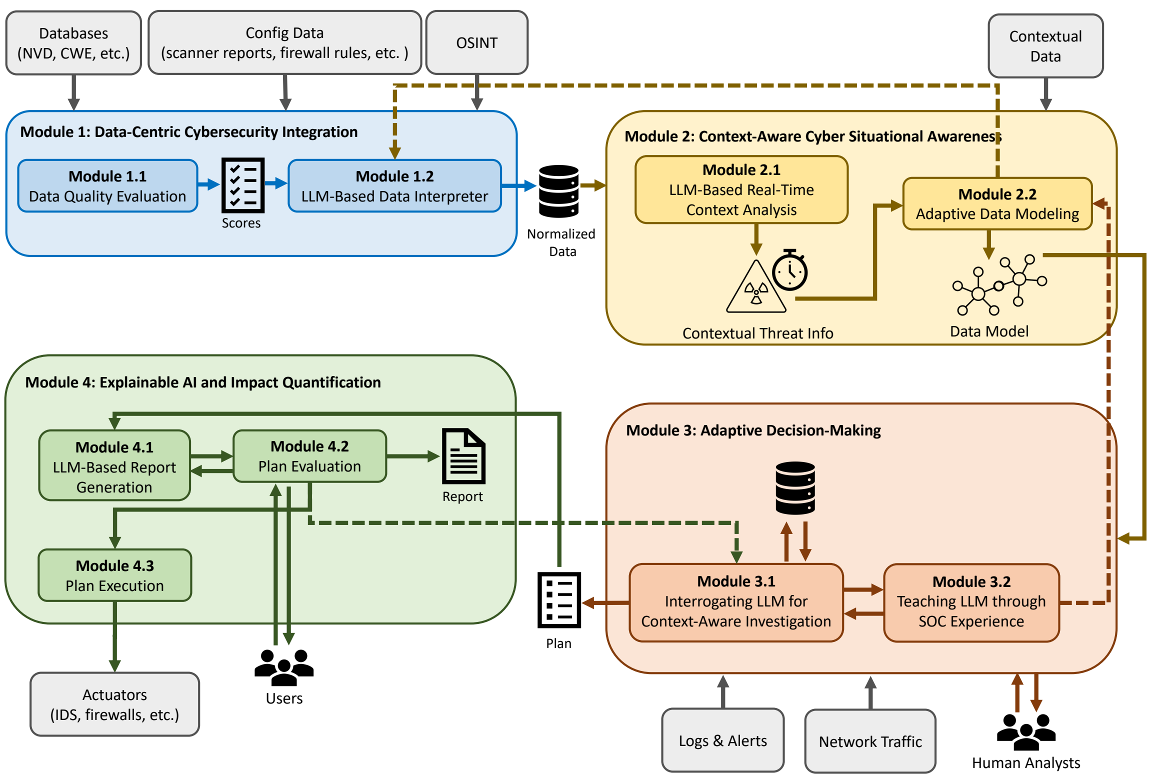


Fig. 3. Overview of the proposed framework.

The framework adopts a data-centric approach that emphasizes cybersecurity integration and interoperability across diverse tools (Module 1). To encourage adoption, the framework complements existing tools and capabilities rather than requiring organizations to overhaul their current systems. By building on established processes, organizations can adopt and integrate the framework with minimal disruption. A key aspect of this approach involves assessing the quality and trustworthiness of data from various sources (Module 1.1). A key innovation is the data normalization mechanism, which leverages LLMs to harmonize data across formats, ensuring tool interoperability and feeding into the next step: context-aware cyber situational awareness (Module 2), which includes an internal graphical data model (Module 2.2). This model is grounded in prior work on attack modeling and configuration security [4, 5] and provides input to an adaptive decision-making process (Module 3), capturing detailed information about security artifacts and their interrelationships. Module 3 utilizes an LLM agent and incorporates data from both Module 1 and Module 2 to analyze and interpret logs, alerts, and network traffic, generating optimal courses of action in response to security events. The LLM agent dynamically adapts to the evolving threat landscape by incorporating real-time context analysis of organizational, economic, and geopolitical data (Module 2.1). This adaptive data modeling improves the accuracy and timeliness of threat assessments by contextualizing security events within their broader environment. Self-awareness mechanisms, based on previous feedback received by analysts in different but analogous situations, enable AI agents to recognize when human expertise is needed — whether to validate findings or contribute to critical decision-making steps. The final plan of action developed through joint human-machine collaboration is passed to the Explainable AI and Impact Quantification module (Module 4). This module utilizes LLMs to generate a human-readable report detailing the recommended course of action (Module 4.1) and its impact on the organization's security posture. The AI-generated explanation is evaluated by various user groups (Module 4.2), and the feedback loops back into the adaptive decision-making process to refine future recommendations. Once validated, the plan is executed by leveraging LLM capabilities from Module 1 to translate it into instructions for the organization's security tools (Module 4.3). The following sections elaborate each module and outline remaining research directions.

### 4.1 Module 1: Data-Centric Cybersecurity Integration

Module 1 adopts a data-centric approach to address the challenges of integrating a vast and growing array of data sources, platforms, and tools to support interoperability and enable large-scale automation.

**Research Questions**. How can LLMs be leveraged to address the challenges of integrating diverse cybersecurity tools within SOCs, particularly by minimizing configuration complexities, overcoming the inflexibility of existing tools, and ensuring interoperability? Additionally, how can these solutions remain robust against incomplete or uncertain data, as well as evolving data formats and APIs?

Modern SOCs handle data in a range of formats used by vulnerability scanners, log analyzers, intrusion detection systems, and external threat intelligence sources. The lack of standardization across these tools presents significant challenges for integration and analysis. This module explores strategies to minimize complexity, thus lowering barriers to adoption, and making the proposed solution robust to incomplete and uncertain data. Given that standardization remains a primary obstacle to integration, we argue that — given the complexity of the problem at hand and the current availability of tools from a large number of vendors — pursuing standardization efforts is not the most effective and efficient way to address the problem, due to several practical challenges like resistance from vendors, dynamic and diverse nature of the cybersecurity domain, regulatory and compliance requirements across industries and regions. To tackle this objective, the framework includes a metrics-based approach for assessing the quality and trustworthiness of data gathered from various sources. An LLM-based mechanism then addresses data interoperability challenges, focusing on the variability in threat and vulnerability report formats and the prevalent use of natural language to describe critical details.

*4.1.1 Module 1.1: Evaluation of Data Quality and Trustworthiness.* The primary objective of this module is to establish a robust mechanism for evaluating the quality, reliability, and trustworthiness of data ingested from diverse cybersecurity tools and platforms. As SOCs increasingly depend on inputs such as network logs, vulnerability scanners, and external threat intelligence, accurately assessing the reliability and completeness of these data sources is essential. Initially, data from various sources is normalized using the mechanisms from Module 1.2. At this stage, the system operates under a neutral trust assumption, treating all sources as equally reliable due to the absence of prior knowledge about their accuracy. Over time, as analysts provide feedback on recommended actions through human-machine cooperation (Module 3) and plan evaluation (Module 4.2), this input is used to infer and adjust the reputation of each data source. It is possible to create a reputation scoring system based on a centrality metric approach similar to PageRank [14], where data sources and validated plans are represented as nodes and feedback is encoded as edges. Unlike traditional PageRank, where all edges transfer importance, this approach allows feedback to be positive or negative, reflecting whether a source contributes accurate or misleading data to a plan, and uses a bipartite graph with two classes of nodes — sources and plans. Scores dynamically adjust based on historical performance and analyst feedback, enabling the system to prioritize reliable sources and refine its trust assessments iteratively. This adaptive scoring system allows continuous improvement as real-world outcomes and feedback are integrated, enhancing both data quality and the effectiveness of decision-making within SOC workflows.

*4.1.2 Module 1.2: LLM-Based Data Interpretation.* An LLM-based agent interprets and normalizes data generated by diverse cybersecurity tools and platforms into a common internal format, translating between formats for seamless interoperability.

*Common data format.* The diverse data formats include structured, semi-structured, and unstructured data, including natural language descriptions. A JSON format is used to normalize cybersecurity threat and vulnerability data, building on the widely used STIX standard [11]. This format leverages key-value pairs, which are both machine-readable and interpretable by LLMs, enabling the extraction and normalization of diverse data types, including categorical variables (e.g., severity levels, attack vectors) and unstructured text (e.g., threat descriptions). The format integrates data quality and trustworthiness scores from Module 1.1, to propagate these metrics to the Module 2 graph modeling.

*Data set curation.* To develop and evaluate an LLM agent, the framework curates a parallel corpus of records that span a range of input-output format combinations. This corpus includes input-output pairs created using existing format conversion tools to provide opportunistic data and manually curated samples spanning format conversions not addressed through available tools to broaden the format conversions. Vendor-provided data-format specifications are incorporated into the LLM instructions to ground the format conversion. These specifications serve as a knowledge source and update path for the LLM agent as formats change or new formats are released.

*LLM development.* The LLM development is treated as a machine translation task, where the LLM translates between data formats. The input includes the current and target format, with the LLM trained to output the data in the desired format. We hypothesize that vendor-provided format specifications will improve performance, generalizability, and handling of unseen formats. The framework accommodates multiple representations (raw text, summaries, structured metadata) for incorporating these specifications to determine which most effectively boosts performance.

*Learning from feedback.* Within this module, in-context learning (prompt-based approaches) and supervised fine-tuning are adopted to develop high-performing LLM agents. Normalized data generated by the agent support the construction of the graph model in Module 2, which feeds into Modules 3 and 4. To refine the LLM agent further, the framework incorporates feedback mechanisms into these subsequent modules, including strategies like RLHF, which uses human feedback to fine-tune models, aligning outputs with desired outcomes. In this context, RLHF iteratively improves the interpreter by leveraging cybersecurity analysts' evaluations of its normalized outputs' accuracy, relevance, and completeness.

### 4.2 Module 2: Context-Aware Cyber Situational Awareness

Module 2 provides advanced solutions for comprehensive and context-aware cyber situational awareness in SOCs and is focused on the following research question.

> **Research Questions**. How can SOCs achieve comprehensive, context-aware cyber situational awareness to enable more informed incident response decisions? How can the context be expanded beyond organizational boundaries to incorporate cyber threat intelligence from external sources and additional dimensions such as economic and geopolitical factors influencing malicious actors?

Traditional approaches focus on internal assets (*knowledge of us*) and partial knowledge of adversaries (*knowledge of them*). The framework extends situational awareness beyond organizational boundaries to include external factors — such as economic conditions, geopolitical tensions, and social dynamics — that influence the motivations and behaviors of malicious actors, ranging from cyber criminals to nation states. Diverse threat intelligence sources are aggregated and structured to create a more holistic view of the cyber threat landscape. By incorporating this multi-dimensional data, SOCs can adapt their defense strategies to evolving threats. Contextual awareness can also improve both proactive and reactive cybersecurity measures, such as predicting attack vectors and prioritizing responses based on real-world implications. This module includes two sub-modules: Module 2.1 focuses on automatically analyzing diverse information streams, including news media, social media, and threat intelligence, to identify SOC-relevant contexts, assess risk, and identify potential targets [57]; Module 2.2 aims to integrate all available information into a comprehensive graphical model.

*4.2.1 Module 2.1: LLM-based Real-Time Context Analysis.* This module provides real-time context analysis capabilities for SOCs by extending Topic Detection and Tracking (TDT) to integrate dynamic risk assessment. Originally a DARPA-sponsored natural language processing (NLP) task, TDT focuses on identifying and tracking emerging events in information streams [7]. The framework builds on TDT to create a near-real-time cybersecurity risk assessment framework that integrates LLMs and enables continuous monitoring of diverse data sources, such as news, social media, and threat intelligence feeds, facilitating the transition from event detection to actionable situational awareness tailored to an evolving, SOC-specific threat landscape. In traditional TDT, identifying and tracking emerging real-world events involves generating structured event representations using predefined ontologies with attributes and relations (e.g., trigger/action, location, or target) or clustering and linking text data to discover and monitor topics, without relying on explicit structuring [1, 10, 15, 59]. We can leverage the language understanding and reasoning capabilities of LLMs through Chain of Thought prompting strategies [58] to focus on ontology attributes relevant to risk assessment. These attributes guide the LLM to consider the contextual information most relevant to SOCs, without requiring strict, predefined templates. For instance, when tracking a geopolitical event, the LLM can be directed to assess statements of hostility (trigger), geographic indicators (location), and the potential threat recipient (target). This attribute-driven focus helps the LLM prioritize relevant details, filter out extraneous information, and identify emergent risks.

The risk posed by an identified event to a specific SOC is framed as a classification problem. Inputs include risk assessment criteria, the event annotation ontology, a representation of the target, and a summary of the event. Outputs are categorical risk labels (e.g., no threat, low, medium, or high) indicating the perceived threat level. These labels inform Module 2.2's adaptive model for dynamic threat assessment. An effective classifier requires a comprehensive SOC representation that accounts for geopolitical, sociopolitical, and technical contexts and generalizable criteria for distinguishing risk levels. It must distinguish between irrelevant content, relevant but non-threatening content, negative commentary lacking genuine risk, and content signaling legitimate threats. The Sony Pictures Entertainment cyber-attack 2 highlights the potential of real-time context analysis for enhancing SOC operations. The framework would process media coverage of the film and North Korea's hostile response, identifying these stories as both relevant and threatening. TDT captures North Korea's public condemnation of the film, particularly in state-run news outlets labeling it anti-propaganda [32]. By leveraging TDT ontology attributes, the LLM could focus on attributes related to the trigger/action (condemnation), sentiment (threatening or hostile), actor (North Korea), and target (Sony Pictures) in assessing relevancy and threat potential. Combining TDT annotation ontology and LLM language understanding yields a framework capable of identifying threats while filtering out irrelevant content, such as negative movie reviews or exaggerated social media posts. By prioritizing actionable insights, the system supports timely and informed responses. This LLM-based approach to TDT produces actionable event summaries and assigns risk labels for the Adaptive Data Modeling in Module 2.2. This enables SOC analysts to monitor emerging risks in real time and adapt defensive strategies in response to geopolitical tensions or other contextual factors, enhancing situational awareness and enabling timely responses. Validation is accomplished through retrospective analysis of prior incidents to uncover patterns and indicators relevant to threat prediction. Using data repositories like the Center for Strategic and International Studies' Significant Cyber Incidents list [22], representations for the applicable SOC are generated, utilizing historical media from tools like the Way Back Machine [9]. Analyzing this data refines the predictive capabilities of real-time systems, helping SOCs adapt to evolving global contexts. The real-time context analysis framework also includes feedback loops from subsequent tasks to refine the LLM-based TDT modeling, further enhancing the accuracy of context-aware incident response.

*4.2.2 Module 2.2: Adaptive Data Modeling.* The goal of Module 2.2 is to develop a dynamic and adaptive data modeling capability that can inform Module 3, supporting real-time risk assessment and effective prioritization of incident response efforts in SOCs. This submodule builds upon the data integration capabilities of Module 1 and the context-aware cyber situational awareness of Module 2.1, using a multi-graph model to capture system components, vulnerabilities, configurations, and external risk factors (e.g., geopolitical conditions). The novelty of this task lies in designing a scalable, continuously updated data model that adapts to the ever-evolving threat landscape and informs automated decision-making processes. The framework adopts a vulnerability-centric approach to data modeling, to enable prioritization of threat responses by focusing on exploitable weaknesses, improving real-time situational awareness and adaptive defenses. This capability builds upon our previous work in developing similar graphical data models [5, 50] and metrics to accurately assess the impact of vulnerability exploits in complex systems [4, 28] and reason about optimal metrics-driven mitigation strategies. Building upon this body of work, we design a multi-graph model, with subgraphs representing different classes of artifacts, including vulnerabilities, system components, configuration parameters, and external risk factors. These risk factors, identified via Module 2.1, include but are not be limited to geopolitical and economic conditions, with nodes representing geopolitical events or economic sanctions, for example, that can influence the likelihood of attacks like nation-state-driven cyber espionage; compliance and regulatory requirements with nodes representing compliance mandates (e.g., GDPR, HIPAA) mapped to configuration or system component nodes to reflect how failure to meet regulatory changes impact system security; and other factors identified through various threat intelligence feeds (e.g., APT activity, emerging attack vectors) that can be used to dynamically adjust the vulnerability subgraph by adding newly discovered vulnerabilities, hypothesizing the presence of zero-day vulnerabilities in seemingly secure but critical components, and adjusting severity scores accordingly. By modeling these external risk factors and mapping them to other constructs in the various subgraphs, the multi-graph model becomes context-aware and capable of evolving in response to changing environmental conditions, such as geopolitical events or new regulatory frameworks.

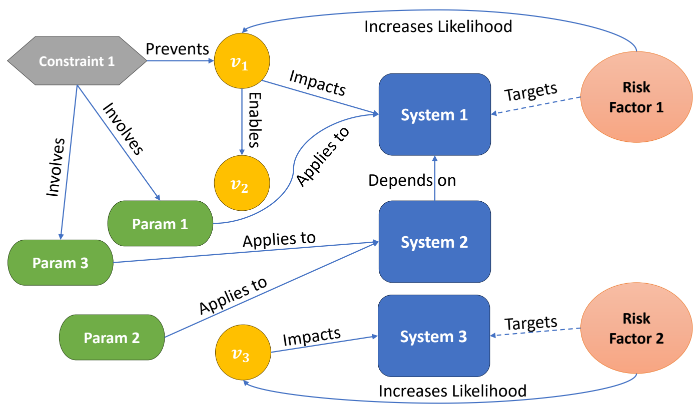


Fig. 4. Example of graphical model.

In our graphical data model, nodes represent key entities that capture different aspects of the system and its risk management environment, as shown in the in the notional example of Fig. 4. Vulnerability nodes (depicted in yellow) capture weaknesses or flaws in the system that can be exploited by threats, offering insights into potential security breaches and multi-step attacks [5, 38, 43] — enabled by the sequential exploitation of multiple dependent vulnerabilities — and their impact on system components. Component nodes (in blue) represent the physical or logical parts of a system at different levels of abstraction (individual components, subsystems, or systems), with the component subgraph capturing functional dependencies among the parts. Risk Factor nodes (in orange) define variables or conditions, identified through Module 2.1, that target specific systems and contribute to the likelihood or impact of security incidents, enabling a contextual assessment of threats and system exposure to external risks. Configuration nodes (in green) document the settings and parameters that govern the operation of the system, reflecting its current state and influencing its vulnerability and resilience to attacks. Constraints among configuration parameters (in gray) capture the role of configuration settings in mitigating vulnerability exposure. Together, these interconnected sets of nodes offer a rich representation of the system's structure, weaknesses, and risk environment, providing a baseline knowledge for LLMs to reason about.

The multi-graph model is designed to scale efficiently with large and complex systems, allowing it to handle a growing number of system components, vulnerabilities, and external factors. Scalability is key to ensuring that the framework remains effective as organizations continue to expand their cybersecurity infrastructure and adopt new tools and technologies. To this aim, the model builds on our previous work on efficiently indexing patterns in graph data [6]. Building on our previous work on vulnerability metrics [4] and quantification of the potential impact of cyber-attacks [28], we define families of metrics to quantify risk and guide the generation of courses of action that can provably reduce risk for the enterprise, paving the way for a novel Metrics-Augmented Generation (MAG). This system of metrics builds upon the two key vulnerability metrics defined in [4]: the exploitation likelihood $\rho(v)$ and the exposure factor $ef(v)$ of a vulnerability $v$, as defined by Equations 1 and 2 below.

$$\rho(v) = \frac{\prod_{x \in X_l^{\uparrow}} \left(1 - e^{-\alpha_x f_x(X(v))}\right)}{\prod_{x \in X_l^{\downarrow}} e^{\alpha_x f_x(X(v))}} \tag{1}$$

$$ef(v) = \frac{\prod_{x \in X_e^{\uparrow}} \left(1 - e^{-\alpha_x f_x(X(v))}\right)}{\prod_{x \in X_e^{\downarrow}} e^{\alpha_x f_x(X(v))}} \tag{2}$$

In these equations, $\mathcal{X}_l^{\uparrow}$, $\mathcal{X}_l^{\downarrow}$, $\mathcal{X}_e^{\uparrow}$, and $\mathcal{X}_e^{\downarrow}$ are sets of variables that influence the exploitation likelihood and exposure factors of vulnerabilities. Specifically, $\mathcal{X}_l^{\uparrow}$ and $\mathcal{X}_l^{\downarrow}$ are sets of variables that, respectively, contribute to increasing and decreasing the likelihood as their values increase. Similarly, $\mathcal{X}_e^{\uparrow}$ and $\mathcal{X}_e^{\downarrow}$ are sets of variables that, respectively, contribute to increasing and decreasing the exposure as their values increase.

Typical examples of variables in $\mathcal{X}_l^{\uparrow}$ include, but are not limited to, a vulnerability's exploitability score, the time since details about the vulnerability were published, and the set of known exploits. In our work, we model the external risk factors discussed earlier as variables in $X_l^{\uparrow}$, allowing us great flexibility in modeling. Examples of variables in $\mathcal{X}_l^{\downarrow}$ include, but are not limited to, the set of known Intrusion Detection System (IDS) rules associated with a vulnerability and the set of available vulnerability scanning plugins. Similarly, examples of variables in $\mathcal{X}_e^{\uparrow}$ include a vulnerability's impact score, and examples of variables in $\mathcal{X}_e^{\downarrow}$ include the set of deployed IDS rules associated with a vulnerability, which can help decrease the impact by detecting the onset of exploitation activity.

In these equations, each variable contributes to the overall likelihood or exposure factor as a multiplicative factor between 0 and 1 that is formulated to account for diminishing returns. Factors corresponding to a variable $X$ in $\mathcal{X}_l^{\uparrow}$ or $\mathcal{X}_e^{\uparrow}$ are of the form $1 - e^{-\alpha_X \cdot f_X(X(v))}$ where $\alpha_X$ is a tunable parameter, $X(v)$ is the value of $X$ for $v$, and $f_X$ is a monotonically increasing function used to convert values of $X$ to scalar values. Similarly, factors corresponding to a variable $X$ in $\mathcal{X}_l^{\downarrow}$ or $\mathcal{X}_e^{\downarrow}$ are of the form $e^{-\alpha_X \cdot f_X(X(v))}$.

A key aspect of our approach is to model and quantify the role of external risk factors. In complex enterprise networks comprising multiple subsystems, each providing distinct services, various risk factors may impact systems differently. To capture these influences, we map each risk factor to the vulnerabilities existing on its target system or subsystem and incorporate these factors into the computation of vulnerability likelihood in Equation (1) as described earlier. In the notional example of Fig. 4, two risk factors respectively affect two different systems, so they can be modeled as variables influencing the likelihood of exposed vulnerabilities $v_1$ and $v_3$ respectively.

### 4.3 Module 3: LLM-Assisted Adaptive Decision-Making

Module 3 uses LLMs to dynamically assess and adapt to the ever-changing risk landscape, and it responds the following research questions.

**Research Questions**. How can SOCs dynamically assess and adapt to the evolving risk landscape, and use LLM-based agents to support real-time risk assessment and prioritize incident response efforts? What are the most effective models for human-machine collaboration in SOCs, to support decision-making and adapt to an ever-changing threat environment?

*4.3.1 Module 3.1: Interrogating LLM for Context-aware Investigation.* This submodule analyzes data from diverse feeds, including network traffic and alerts from various tools, inferences from larger contexts (Module 2.1), and risk indications calculated from the graphical model (Module 2.2), to suggest potential courses of actions both for investigation and remediation. Anthropological fieldwork provides an effective means of capturing tacit analyst knowledge to develop this capability. Embedded fieldworkers carry out participant observation [24, 51] and work as analysts in real-world SOCs. Embedded fieldworkers play the roles of both the researcher and the analyst shown in Fig. 2. In carrying out the investigation, the analysts ask the LLM agent to perform subtasks such as looking for patterns of potential misuse in logs. The LLM is provided with all relevant contextual information from the various data feeds described above. It generates an initial output, including a rationale explaining its decision-making process. Analysts then interact with the LLM by refining its outputs through iterative natural language instructions, guiding the model to better align with their preferences. This refinement process could involve multiple rounds of revisions, where the LLM adjusts its reasoning and suggestions in response to user feedback. If the analysts are unable to achieve a satisfactory result through this interaction alone, they could directly edit the output to produce a revised version. These revised outputs, along with the original inputs, are then used to fine-tune the LLM, gradually aligning its responses with analyst preferences.

This approach builds on the concept of generative feedback [30], which emphasizes learning from natural language critiques to iteratively refine model outputs. Rather than relying on static supervision or RLHF, generative feedback enables LLMs to adapt dynamically through user interactions that provide fine-grained feedback on both the strengths and weaknesses of an initial response. By incorporating iterative refinement and direct human correction, the framework extends this paradigm to better support investigative workflows, allowing the LLM to capture subject matter expertise and institutional preferences over successive interactions. In the early stages, analysts may need to invest more effort in refining outputs, either by providing detailed prompting or by manually editing the model's responses. However, as the LLM incorporates this expertise through feedback-driven learning, the need for direct human intervention is expected to decrease. All interactions are recorded as op traces, as illustrated in Fig. 2. After sufficient improvement, the agent is piloted with production SOC analysts. At this stage, the fieldworkers only play the role of the researcher in Fig. 2, alongside SOC analysts, to observe how human analysts interact with the LLM agent. In addition to the op traces, the embedded researchers also record their observations in fieldnotes. These observations can reveal unspoken suppositions and other dimensions of tacit knowledge not captured in the conversations between human analysts and LLM agents. These tacit dimensions are useful in creating more effective learning signals for LLM agents to be improved through the learning process described below. The continuous evaluation and refinement aim to enhance the precision and effectiveness of human-machine co-teaming for investigating cyber incidents and making the right remediation decisions.

*4.3.2 Module 3.2: Teaching LLM through SOC Experience.* This submodule provides approaches that enable the LLM agents to learn from their experience in the SOC's investigative tasks. The learning signals generated through LLM agents' collaboration with human analysts come from two sources (Fig. 2): 1) recorded op traces, and 2) distilled explicit knowledge through qualitative analysis. To make the learning process sustainable, human analysts bear minimum burden in helping provide the learning signals. The conversations between the LLM agent and human analysts must be organic and for the sole purpose of moving forward with the investigation and finding the most appropriate remediations. In doing so some tacit knowledge is inevitably missed in the recorded op traces. This is where the embedded researcher can help. The researcher, trained in anthropological research methods, records in the fieldnotes any observations that capture the context under which the interactions took place. The analysis of the fieldnotes provide additional learning signals from the explicated tacit knowledge through qualitative analysis. After the learning signals are formed, various learning strategies are applied as explained in Section 3.1. Different strategies are compared to understand their strengths and weaknesses and gain an understanding of when to utilize which strategies to yield the best outcome. After the LLM agent has learned substantial new knowledge, an updated version is deployed into operation, and the process repeats.

### 4.4 Module 4: Explainable AI and Impact Quantification

Module 4 employs LLMs to embed explainable-AI capabilities within SOC workflows, thereby improving transparency and trust in incident-response decisions and providing metrics to quantify the operational impact of AI adoption. It addresses the following research question.

> **Research Questions**. How can we leverage an LLM's ability to articulate reasoning to achieve explainable AI integrated into SOCs to enhance transparency, trust, and understanding of decision-making during incident response? How can the impact of adopting LLMs in SOCs be measured, not only in terms of operational efficiency but also in incident detection and response effectiveness?

*4.4.1 Module 4.1: LLM-based Report Generation.* Once human analysts, assisted by LLM agents, have reached investigative conclusions and formulated remediation strategies, the SOC needs to produce reports for the relevant stakeholders who are impacted by the incident and/or remediations. The framework leverages LLM's capability to transform the analysis results from Module 3.1 into reports designed for various human stakeholders, emphasizing interpretability and transparency. The report provides a narrative explanation of the proposed actions, their rationale, and their alignment with risk metrics. It also includes contextual information, such as the expected impact, underlying assumptions, and data sources, enabling stakeholders to understand and trust the decision-making process. To help the AI develop potential courses of action, a template is developed by the research team with input from SOC team members. Each course of action presents SOC-specific details and clearly links the proposed steps to the particular security issue they address.

*4.4.2 Module 4.2: Plan Evaluation.* In Module 4.2, the focus shifts to evaluating the proposed plan's feasibility and effectiveness through a structured review by human analysts. This evaluation differs from the interactive refinement in Module 3.2 by incorporating a broader, post-generation assessment of the plan, including its alignment with organizational goals and long-term impact. Feedback from this evaluation refines both the analysis process in Module 3.1 and the report generation process in Module 4.1, ensuring iterative improvement and consistency with evolving security contexts.

*4.4.3 Module 4.3: Plan Execution.* Module 4.3 addresses the automation of the plan's implementation, transforming the finalized plan from Module 4.1 into executable artifacts such as scripts, system configurations, or other machine-readable instructions. This task focuses on operationalizing the plan with minimal manual intervention while maintaining oversight mechanisms to ensure alignment with organizational policies. By bridging the gap between planning and action, Module 4.3 enables efficient and reliable execution of risk mitigation strategies, reinforcing the overall AI-assistance framework.

## 5 Discussion and Future Work

While AI-driven co-teaming presents transformative potential, challenges remain, including trust in AI outputs, explainability, and continuous learning from SOC interactions. Future research should explore fine-tuning SOC-specific LLMs, enhancing AI explainability, and integrating human feedback loops for iterative learning.

### 5.1 Representative Case Studies

Cyber-attacks targeting high-profile organizations are often driven by geopolitical tensions, exposing critical gaps in SOCs, particularly in integrating external threat intelligence and leveraging human-machine collaboration for increased efficiency and adaptive response to evolving threats. Two major incidents — the 2014 Sony Pictures Entertainment attack and the 2016 DNC cyber intrusions — illustrate these challenges.

The Sony attack, attributed to North Korea [40] in retaliation for the release of the movie *The Interview* [39], began with a phishing campaign that compromised Sony's network. The attackers deployed Destover malware, wiping data and rendering systems inoperable while simultaneously exfiltrating sensitive data, including unreleased films, emails, and employee records. The breach caused severe operational and reputational damage, highlighting the failure of traditional SOCs to anticipate politically motivated cyber threats. Similarly, the DNC breaches, orchestrated by Russian state-sponsored groups Cozy Bear and Fancy Bear [18], relied on spear-phishing to gain initial access. Cozy Bear persisted within the network for months before Fancy Bear executed more aggressive data exfiltration and public leaks, influencing U.S. politics (see Fig. 5). These attacks underscored the difficulty of identifying and responding to prolonged, stealthy threats in a complex geopolitical landscape.

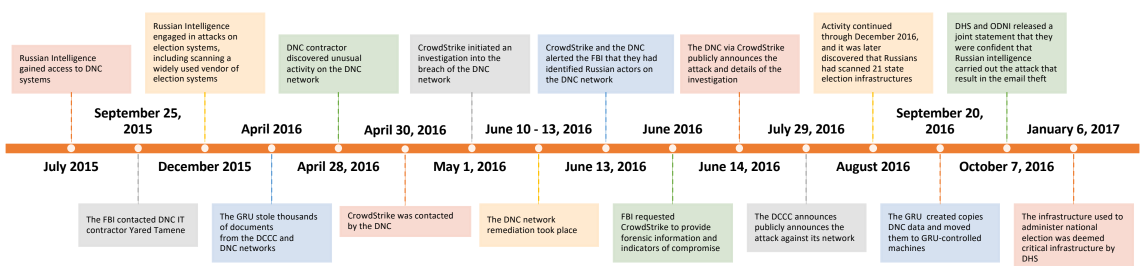


Fig. 5. Timeline of the DNC attack.

Our framework would have significantly enhanced a SOC's ability to detect and mitigate such attacks. By incorporating geopolitical threat signals into SOC workflows, harmonizing disparate data sources, and leveraging LLM-based AI for real-time contextual analysis, our approach would have helped provide analysts with some actionable intelligence well before the attacks escalated. Rather than relying on static detection mechanisms, our adaptive AI-driven co-teaming paradigm would have surfaced geopolitical risks, linked anomalous activity to external threat actors, and generated explainable remediation plans, allowing human analysts to act swiftly and effectively. The Sony and DNC incidents illustrate the urgent need for a human-machine collaboration model that dynamically integrates external intelligence, anticipates emerging threats, and enables proactive defense.

## 6 Conclusions

This paper outlines a vision for AI-driven human-machine co-teaming in Security Operations Centers (SOCs), positioning large language models (LLMs) as collaborative apprentices that learn from and work alongside human analysts. We propose a paradigm in which LLM-based AI agents augment SOC workflows by internalizing tacit knowledge, supporting dynamic decision-making, and alleviating analyst burden without replacing human judgment. By reframing AI not as a deterministic tool but as a learning partner, we open new pathways for adaptive, context-aware SOC operations. We invite the research and practitioner community to explore this human-AI collaboration model, investigate its practical applications, and advance its development toward measurable improvements in SOC productivity and resilience.


## Acknowledgment

This work was partially supported by the National Science Foundation under award no. 2235102, Office of Naval Research under award no. N00014-23-1-2538, and National Security Agency under award No. H98230-22-1-0311. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the above funding agencies.


## References


[1] Z. Akhgari, M. Malekimajd, and H. Rahmani. 2022. Sem-TED: Semantic Twitter Event Detection and Adapting with News Stories. In *Proceedings of the International Conference on Web Research*. 61–69. https://doi.org/10.1109/ICWR54782.2022.9786234

[2] M. N. Al-Mhiqani, R. Ahmad, Z. Z. Abidin, W. Yassin, A. Hassan, K. H. Abdulkareem, N. S. Ali, and Z. Yunos. 2020. A Review of Insider Threat Detection: Classification, Machine Learning Techniques, Datasets, Open Challenges, and Recommendations. *Applied Sciences* 10, 15 (2020), 5208. https://doi.org/10.3390/app10155208

[3] B. A. Alahmadi, L. Axon, and I. Martinovic. 2022. 99% False Positives: A Qualitative Study of SOC Analysts' Perspectives on Security Alarms. In *Proceedings of the 31st USENIX Security Symposium (USENIX Security 2022)*. Boston, MA, USA, 2783–2800. https://www.usenix.org/system/files/sec22-alahmadi.pdf

[4] M. Albanese, I. Iganibo, and O. Adebiyi. 2023. A Framework for Designing Vulnerability Metrics. *Computers & Security* 132 (September 2023), 11 pages. https://doi.org/10.1016/j.cose.2023.103382

[5] M. Albanese and S. Jajodia. 2018. A Graphical Model to Assess the Impact of Multi-Step Attacks. *The Journal of Defense Modeling and Simulation* 15, 1 (2018), 79–93. https://doi.org/10.1177/1548512917706043

[6] M. Albanese, A. Pugliese, V. S. Subrahmanian, and O. Udrea. 2007. MAGIC: A Multi-Activity Graph Index for Activity Detection. In *Proceedings of the 2007 IEEE International Conference on Information Reuse and Integration (IRI 2007)*. Las Vegas, NV, USA, 267–272. https://doi.org/10.1109/IRI.2007.4296632

[7] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. 1998. Topic Detection and Tracking Pilot Study: Final Report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*. 194–218. https://ciir.cs.umass.edu/pubfiles/ir-137.pdf

[8] E. Almazrouei, H. Alobeidli, A. Alshamsi, A. Cappelli, R. Cojocaru, M. Debbah, É. Goffinet, D. Hesslow, J. Launay, and Q. Malartic. 2023. The Falcon Series of Open Language Models. *arXiv* (2023). https://doi.org/10.48550/arXiv.2311.16867

[9] Internet Archive. [n.d.]. Wayback Machine. http://web.archive.org/ [Accessed: Jan. 24, 2025].

[10] M. Asgari-Chenaghlu, M.-R. Feizi-Derakhshi, L. Farzinvash, M.-A. Balafar, and C. Motamed. 2021. Topic Detection and Tracking Techniques on Twitter: A Systematic Review. *Complexity* 2021 (2021), Article ID 8833084. https://doi.org/10.1155/2021/8833084

[11] S. Barnum. 2012. Standardizing Cyber Threat Intelligence Information with the Structured Threat Information Expression (STIX™). The MITRE Corporation. https://www.mitre.org/sites/default/files/publications/stix.pdf

[12] M. Bayer, P. Kuehn, R. Shanehsaz, and C. Reuter. 2024. CySecBERT: A Domain-Adapted Language Model for the Cybersecurity Domain. *ACM Transactions on Privacy and Security* 27, 2 (2024), 1–20. https://doi.org/10.1145/3652594

[13] H. R. Bernard, A. Wutich, and G. W. Ryan. 2016. *Analyzing Qualitative Data: Systematic Approaches*. SAGE Publications. https://uk.sagepub.com/en-gb/eur/analyzing-qualitative-data/book240717

[14] S. Brin and L. Page. 1998. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems* 30 (1998), 107–117. https://www.sciencedirect.com/science/article/abs/pii/S016975529800110X

[15] R. Cantini and F. Marozzo. 2022. Topic Detection and Tracking in Social Media Platforms. In *Proceedings of the 1st International Conference on Pervasive Knowledge and Collective Intelligence on Web and Social Media*. Springer, 41–56. https://doi.org/10.1007/978-3-031-31469-8_3

[16] Y. Chen, M. Cui, D. Wang, Y. Cao, P. Yang, B. Jiang, Z. Lu, and B. Liu. 2024. A Survey of Large Language Models for Cyber Threat Detection. *Computers & Security* (July 2024). https://doi.org/10.1016/j.cose.2024.104016

[17] Y. Chen, Z. Ding, L. Alowain, X. Chen, and D. Wagner. 2023. DiverseVul: A New Vulnerable Source Code Dataset for Deep Learning-Based Vulnerability Detection. In *Proceedings of the 26th International Symposium on Research in Attacks, Intrusions, and Defenses (RAID 2023)*. 654–668. https://doi.org/10.1145/3607199.3607242

[18] CrowdStrike. 2020. CrowdStrike's Work with the Democratic National Committee: Setting the Record Straight. CrowdStrike Blog. https://www.crowdstrike.com/blog/bears-midst-intrusion-democratic-national-committee/

[19] G. Deng, Y. Liu, V. Mayoral-Vilches, P. Liu, Y. Li, Y. Xu, T. Zhang, Y. Liu, M. Pinzger, and S. Rass. 2024. PentestGPT: Evaluating and Harnessing Large Language Models for Automated Penetration Testing. In *Proceedings of the 33rd USENIX Security Symposium (USENIX Security 24)*. Philadelphia, PA, USA, 847–864. https://www.usenix.org/conference/usenixsecurity24/presentation/deng

[20] W. Fan, Y. Ding, L. Ning, S. Wang, H. Li, D. Yin, T.-S. Chua, and Q. Li. 2024. A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. In *Proceedings of the 30th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Barcelona, Spain, 6491–6501. https://doi.org/10.1145/3637528.3671470

[21] M. D. Fetters and E. B. Rubinstein. 2019. The 3 Cs of Content, Context, and Concepts: A Practical Approach to Recording Unstructured Field Observations. *The Annals of Family Medicine* 17, 6 (2019), 554–560. https://pmc.ncbi.nlm.nih.gov/articles/PMC6846267/

[22] Center for Strategic and International Studies (CSIS). [n.d.]. Significant Cyber Incidents. Strategic Technologies Program. https://www.csis.org/programs/strategic-technologies-program/significant-cyber-incidents [Accessed: Jan. 24, 2025].

[23] T.-M. Georgescu. 2020. Natural Language Processing Model for Automatic Analysis of Cybersecurity-Related Documents. *Symmetry* 12, 3 (2020), 354. https://doi.org/10.3390/sym12030354

[24] W. Gunn and L. B. Løgstrup. 2014. Participant Observation, Anthropology Methodology and Design Anthropology Research Inquiry. *Arts and Humanities in Higher Education* 13, 4 (2014), 428–442. https://doi.org/10.1177/1474022214543874

[25] A. Happe and J. Cito. 2023. Getting Pwn'd by AI: Penetration Testing with Large Language Models. In *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE 2023)*. San Francisco, CA, USA, 2082–2086. https://doi.org/10.1145/3611643.3613083

[26] I. Hasanov, S. Virtanen, A. Hakkala, and J. Isoaho. 2024. Application of Large Language Models in Cybersecurity: A Systematic Literature Review. *IEEE Access* 12 (2024), 176751–176778. https://doi.org/10.1109/access.2024.3505983

[27] S. Hays and D. J. White. 2024. Employing LLMs for Incident Response Planning and Review. *arXiv* (March 2024). https://doi.org/10.48550/arxiv.2403.01271

[28] I. Iganibo, M. Albanese, M. Mosko, E. Bier, and A. E. Brito. 2023. An Attack Volume Metric. *Security and Privacy* 6, 4 (July 2023), 23 pages. https://doi.org/10.1002/spy2.298

[29] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, and L. Saulnier. 2023. Mistral 7B. *arXiv* (2023). https://doi.org/10.48550/arXiv.2310.06825

[30] D. Jin, S. Mehri, D. Hazarika, A. Padmakumar, S. Lee, Y. Liu, and M. Namazifar. 2023. Data-Efficient Alignment of Large Language Models with Human Feedback Through Natural Language. Amazon Science. https://www.amazon.science/publications/data-efficient-alignment-of-large-language-models-with-human-feedback-through-natural-language

[31] K. R. Jones, D. A. Brucker-Hahn, B. Fidler, and A. G. Bardas. 2023. Work-From-Home and COVID-19: Trajectories of Endpoint Security Management in a Security Operations Center. In *Proceedings of the 32nd USENIX Security Symposium (USENIX Security 2023)*. Anaheim, CA, USA, 2293–2310. https://www.usenix.org/system/files/usenixsecurity23-jones.pdf

[32] S. Jong. 2015. The Interview as Anti-North Korean Propaganda. *NK News* (March 2015). https://www.nknews.org/2015/03/the-interview-as-anti-north-korean-propaganda/

[33] W. Knight. 2023. Geoffrey Hinton tells us why he's now scared of the tech he helped build. *MIT Technology Review* (May 2023). https://www.technologyreview.com/2023/05/02/1072528/geoffrey-hinton-google-why-scared-ai/

[34] V. Kumar, D. Sinha, A. K. Das, S. C. Pandey, and R. T. Goswami. 2020. An Integrated Rule-Based Intrusion Detection System: Analysis on UNSW-NB15 Dataset and the Real-Time Online Dataset. *Cluster Computing* 23 (2020), 1397–1418. https://doi.org/10.1007/s10586-019-03008-x

[35] D. Lende, A. Monkhouse, J. Ligatti, and X. Ou. 2023. Co-Creation in Secure Software Development: Applied Ethnography and the Interface of Software and Development. *Human Organization* 82, 1 (2023), 13–24. https://doi.org/10.17730/1938-3525-82.1.13

[36] M. Levi, Y. Alluouche, D. Ohayon, and A. Puzanov. 2024. CyberPal.AI: Empowering LLMs with Expert-Driven Cybersecurity Instructions. *arXiv* (August 2024). https://doi.org/10.48550/arXiv.2408.09304

[37] Z. Liu. 2024. A Review of Advancements and Applications of Pre-Trained Language Models in Cybersecurity. In *Proceedings of the International Symposium on Digital Forensics and Security*. San Antonio, TX, USA. https://doi.org/10.1109/isdfs60797.2024.10527236

[38] Z. Liu, C. Wang, and S. Chen. 2008. Correlating Multi-Step Attack and Constructing Attack Scenarios Based on Attack Pattern Modeling. In *Proceedings of the 2008 International Conference on Information Security and Assurance (ISA 2008)*. 214–219. https://doi.org/10.1109/ISA.2008.11

[39] Trend Micro. 2014. The Hack of Sony Pictures: What We Know and What You Need to Know. Trend Micro Security News. https://www.trendmicro.com/vinfo/us/security/news/cyber-attacks/the-hack-of-sony-pictures-what-you-need-to-know

[40] Trend Micro. 2018. A Look into the Lazarus Group's Operations. Trend Micro Security News. https://www.trendmicro.com/vinfo/us/security/news/cybercrime-and-digital-threats/a-look-into-the-lazarus-groups-operations

[41] B. Min, H. Ross, E. Sulem, A. P. Veyseh, T. H. Nguyen, O. Sainz, E. Agirre, I. Heintz, and D. Roth. 2023. Recent Advances in Natural Language Processing via Large Pre-Trained Language Models: A Survey. *Comput. Surveys* 5, 2 (September 2023). https://doi.org/10.1145/3605943

[42] S. Mitra, S. Neupane, T. Chakraborty, S. Mittal, A. Piplai, M. Gaur, and S. Rahimi. 2024. LOCALINTEL: Generating Organizational Threat Intelligence from Global and Local Cyber Knowledge. *arXiv* (January 2024). https://doi.org/10.48550/arxiv.2401.10036

[43] J. Navarro, A. Deruyver, and P. Parrend. 2018. A Systematic Survey on Multi-Step Attack Detection. *Computers & Security* 76 (July 2018), 214–249. https://doi.org/10.1016/j.cose.2018.03.001

[44] I. Nonaka and H. Takeuchi. 1995. *The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation.* Oxford University Press, New York. https://academic.oup.com/book/52097

[45] M. Polanyi. 1966. The Logic of Tacit Inference. *Philosophy* 41, 155 (1966), 1–18. https://doi.org/10.1017/S0031819100066110

[46] M. Polanyi. 1966. *The Tacit Dimension.* Doubleday & Company, Inc., New York.

[47] K. Shashwat, F. Hahn, X. Ou, D. Goldgof, L. Hall, J. Ligatti, S. R. Rajagopalan, and A. Z. Tabari. 2024. A Preliminary Study on Using Large Language Models in Software Pentesting. In *Proceedings of the Workshop on SOC Operations and Construction (WOSOC 2024).* San Diego, CA, USA, 1–7. https://www.ndss-symposium.org/wp-content/uploads/wosoc2024-2-paper.pdf

[48] N. Shenoy and A. V. Mbaziira. 2024. An Extended Review: LLM Prompt Engineering in Cyber Defense. In *Proceedings of the International Conference on Electrical, Computer and Energy Technologies.* Sydney, Australia, 1–6. https://doi.org/10.1109/icecet61485.2024.10698605

[49] G. Siracusano, D. Sanvito, R. Gonzalez, M. Srinivasan, S. Kamatchi, W. Takahashi, M. Kawakita, T. Kakumaru, and R. Bifulco. 2023. Time for aCTIon: Automated Analysis of Cyber Threat Intelligence in the Wild. *arXiv* (July 2023). https://doi.org/10.48550/arXiv.2307.10214

[50] H. Soroush, M. Albanese, M. A. Mehrabadi, I. Iganibo, M. Mosko, J. Gao, D. Fritz, S. Rane, and E. Bier. 2020. SCIBORG: Secure Configurations for the IoT Based on Optimization and Reasoning on Graphs. In *Proceedings of the 8th IEEE Conference on Communications and Network Security (IEEE CNS 2020).* 10 pages. https://doi.org/10.1109/CNS48642.2020.9162256

[51] J. P. Spradley. 2016. *Participant Observation.* Waveland Press. https://www.waveland.com/browse.php?t=689

[52] S. C. Sundaramurthy, A. G. Bardas, J. Case, X. Ou, M. Wesch, J. McHugh, and S. R. Rajagopalan. 2015. A Human Capital Model for Mitigating Security Analyst Burnout. In *Proceedings of the 11th Symposium on Usable Privacy and Security (SOUPS).* Ottawa, Canada, 347–359. https://www.usenix.org/system/files/conference/soups2015/soups15-paper-sundaramurthy.pdf

[53] S. C. Sundaramurthy, J. McHugh, X. Ou, S. R. Rajagopalan, and M. Wesch. 2014. An Anthropological Approach to Studying CSIRTs. *IEEE Security & Privacy* 12, 5 (Sept./Oct. 2014), 41–47. https://doi.org/10.1109/MSP.2014.84

[54] S. C. Sundaramurthy, J. McHugh, X. Ou, M. Wesch, A. G. Bardas, and S. R. Rajagopalan. 2016. Turning Contradictions into Innovations or: How We Learned to Stop Whining and Improve Security Operations. In *Proceedings of the 12th Symposium on Usable Privacy and Security (SOUPS 2016).* Denver, CO, USA, 237–251. https://www.usenix.org/system/files/conference/soups2016/soups2016-paper-sundaramurthy.pdf

[55] D. R. Thomas. 2006. A General Inductive Approach for Analyzing Qualitative Evaluation Data. *American Journal of Evaluation* 27, 2 (2006), 237–246. https://doi.org/10.1177/1098214005283748

[56] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, and F. Azhar. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv* (2023). https://doi.org/10.48550/arXiv.2302.13971

[57] P. Trubowitz and K. Watanabe. 2021. The Geopolitical Threat Index: A Text-Based Computational Approach to Identifying Foreign Threats. *International Studies Quarterly* 65, 3 (May 2021), 852–865. https://doi.org/10.1093/isq/sqab029

[58] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. V. Le, and D. Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, Vol. 35. 24824–24837. https://doi.org/10.5555/3600270.3602070

[59] W. Xiang and B. Wang. 2019. A Survey of Event Extraction from Text. *IEEE Access* 7 (2019), 173111–173137. https://doi.org/10.1109/access.2019.2956831

[60] J. Yang, H. Jin, R. Tang, and J. Tang. 2024. Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond. *Commun. ACM* 67, 1 (January 2024), 70–79. https://doi.org/10.1145/3649506