

DOI: [https://doi.org/10.48009/2\\_iis\\_2026\\_102](https://doi.org/10.48009/2_iis_2026_102)

## The impact of prompt clarity, accuracy, and context on Generative AI performance in cybersecurity operations

Sean Ha, *Robert Morris University*, [sxhst105@mail.rmu.edu](mailto:sxhst105@mail.rmu.edu)

Ping Wang, *Robert Morris University*, [wangp@rmu.edu](mailto:wangp@rmu.edu)

### Abstract

As Generative AI systems become integrated into specialized domains such as cybersecurity, the design, clarity, and accuracy of user interaction prompts for Generative AI virtual assistant chatbots are becoming increasingly critical factors that influence performance, accuracy, usability, and usefulness of the technology. This paper investigates the impact of user input prompt clarity – which is defined as the degree to which a prompt is contextually corrupted or enriched – on the efficacy of Generative AI technologies for cyber defensive operations in a live security setting. Through a series of controlled experiments, we evaluate how variations in prompt clarity and accuracy affect performance and usability for threat detection and incident response. Our findings reveal a correlation between context accuracy and output reliability, particularly in technically dense scenarios that typify cybersecurity operations. The result is a set of practical guidelines for effective prompt composition in cyber operations. This work contributes to a growing body of research on Generative AI interaction, prompt design, construction, and engineering, and the operational deployment of Generative AI in highly technical and specialized domains.

**Keywords:** Generative AI (GenAI), agentic AI, AI assistant, cybersecurity, ChatGPT, operational technology

### Introduction

In the ever-expanding digital landscape, the domain of cybersecurity stands as a critical safeguard. The domain has seen a significant paradigm shift since the mid 2010s, with increasing sophistication from cyber threat actors, such as cybercriminals that leverage ransomware to extort both small and major corporations (Morgan, 2023). The evolution in cyber-attacks is due primarily to an increase in attack surface exposure as a result of the proliferation of connectivity to society’s critical digital assets such as data centers, with ubiquitous accessibility afforded by satellite communications (Tedeschi et al., 2022), 5G/6G technologies (Scalise et al., 2024), and Internet of Things (Alnajim et al., 2023), as remote accessibility has provided a significant boon to business optimization and operational efficiency (Lu et al., 2022).

The 2020s has seen the next step regarding the industry-wide adoption of transformative technologies with the advent of Generative AI (GenAI) due to the advent of Large Language Models (LLMs) and conversational AI “chatbots” (Sedkaoui & Benaichouba, 2024). Regarding cybersecurity, GenAI is a double-edged sword, offering potentially unprecedented boosts in effectiveness for threat detection, incident response, and informed decision-making in high-stakes defensive cyber operations, while equipping cyber adversaries with powerful tools to craft novel, sophisticated, and highly obfuscated attacks (Gupta et al., 2023). The rapid uptake, adoption, and integration of GenAI into both offensive and defensive

cyber operations necessitates a re-evaluation of established cyber defensive strategies and operational workflows to maximize the potential for increased performance, efficiency, and efficacy against cyber threats. (Charfeddine et al., 2024; Hasanov et al., 2024). Industry reports forecast cybersecurity to be one of GenAI's largest and most impactful use cases into the 2030s and beyond (Meeker et al., 2025).

The aim of the study is to explore dominant perceived benefits, risks, and mitigations associated with the use of a GenAI virtual assistant or “chatbot” – a conversational AI that interacts via natural language queries – in assisting human security analysts as part to mitigate a cyber incident. This study contributes to the existing body of work in three ways. First, the study advances the evaluation of GenAI for cybersecurity from benchmarking towards a representative operational cyber incident scenario. Second, it systematically isolates operator-provided context as the core experimental variable regarding GenAI performance. Finally, it evaluates GenAI performance using semantic accuracy and quality measures. Ultimately, the findings will drive industry-wide guidance on how best to adopt and operationalize GenAI in cybersecurity workflows that optimizes the risk-benefit tradeoff.

## Background

Generative AI is defined as the use of AI to produce a variety of forms content such as text, images, music, speech, code, videos, given specific user inputs, defined as “prompts” (Maglaras et al., 2022). The core technology behind Generative AI is LLMs, which make use of predictive natural language processing techniques to automatically generate the requested content (Gupta et al., 2023). The use of GenAI for cybersecurity has rapidly expanded, presenting opportunities for increased performance and efficiency when applied to existing cybersecurity solutions (Charfeddine et al., 2024; Hasanov et al., 2024).

Consider the scenario of live cyber defensive operations, where GenAI can play a significant role in assisting human security analysts, operators, and practitioners in a Security Operations Center (SOC). The role of GenAI in such a scenario has been proposed to encompass a variety of use cases, such as an on-demand virtual assistant providing contextual information for less technically adept or experienced users (Prieto & Blakely, 2023). GenAI can be tasked with tedious, manual tasks, such as sifting through and analyzing massive amounts of network, device, and application log information, thereby carrying out anomaly detection by identifying suspicious patterns such as malicious login attempts, suspicious program executions, or messages that are suggestive of phishing campaigns (Dhanushkodi & Thejas, 2024). This has the benefit of significantly reducing cognitive load for SOC analysts and commercial products have already begun to incorporate GenAI as part of existing security incident and event management (SIEM) or security orchestration and automated response (SOAR) solutions (Bono, 2025; Wang, 2023).

GenAI has further been suggested for use in correlating existing threat intelligence databases and knowledgebases with up-to-date threat reports, as well as combining data from unstructured sources such as social media feeds, technical blogs and industry newsletters (Saddi et al., 2024). Studies have been conducted on GenAI's ability to transform diverse sets of threat intelligence into a singular, coherent narrative, aiding in attribution, and providing actionable insights to threat intelligence practitioners. GenAI has also been proposed for dynamic generation of incident response playbooks, as well as associated reports and audit logs to provide end-to-end traceability and accountability (Prieto & Blakely, 2023).

Major areas of concern with the use of GenAI for cybersecurity include lack of trust, accuracy and explainability, requiring rigorous and comprehensive human review and validation (Choudhury & Asan, 2022). Usability is also a concern, as the lowered barrier for security knowledge may be offset by the domain expertise and knowledge required to craft effective GenAI prompts and rigorously validate GenAI

outputs (Dwivedi, 2023). Finally, privacy is a major concern, as GenAI may unintentionally disclose sensitive operational information in the process of accessing external training data information repositories (Ibrahim & Kashef, 2025). Table 1 provides an overview of GenAI use for cyber defensive operations.

**Table 1. Summary of Adoption and Employment of Generative AI for Cyber Defensive Operations**

| Topic                                                         | Key Findings/Conclusions                                                                                                                                 | References                                                                 |
|---------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------|
| GenAI use cases                                               | Log analysis, contextual CTI mapping (e.g. vulnerability databases), task guidance, policy generation, incident response, matching compliance frameworks | (Charfeddine et al., 2024)                                                 |
|                                                               | Threat detection (phishing), vulnerability assessment, ensuring governance                                                                               | (Hasanov et al., 2024)                                                     |
|                                                               | Malware detection, threat-hunting                                                                                                                        | (Sai et al., 2024)                                                         |
|                                                               | Data aggregation, normalization enrichment, behavioral and predictive analytics, SIEM/SOAR integration                                                   | (Prieto & Blakely, 2023)                                                   |
|                                                               | Real-time anomaly detection via pattern recognition                                                                                                      | (Dhanushkodi & Thejas, 2024)                                               |
|                                                               | Automated CTI gathering across multi-source data                                                                                                         | (Saddi et al., 2024)                                                       |
| Risks of leveraging GenAI for operational defensive workflows | Lack of trust, accuracy, explainability                                                                                                                  | (Choudhury & Asan, 2022)                                                   |
|                                                               | Lack of alignment with regulations                                                                                                                       | (Dwivedi, 2023)                                                            |
|                                                               | Data poisoning                                                                                                                                           | (Hasanov et al., 2024)                                                     |
|                                                               | Quality, reliability, robustness, provability, interpretability, monitorability, and auditability                                                        | (Choi et al., 2024)                                                        |
|                                                               | Lack of trust, accuracy, explainability                                                                                                                  | (Choudhury & Asan, 2022)                                                   |
| Risk mitigation techniques                                    | Policy restrictions, technical access control mechanisms, data protection for privacy, monitoring, human oversight and validation                        | (Charfeddine et al., 2024)<br>(Hasanov et al., 2024)<br>(Sai et al., 2024) |

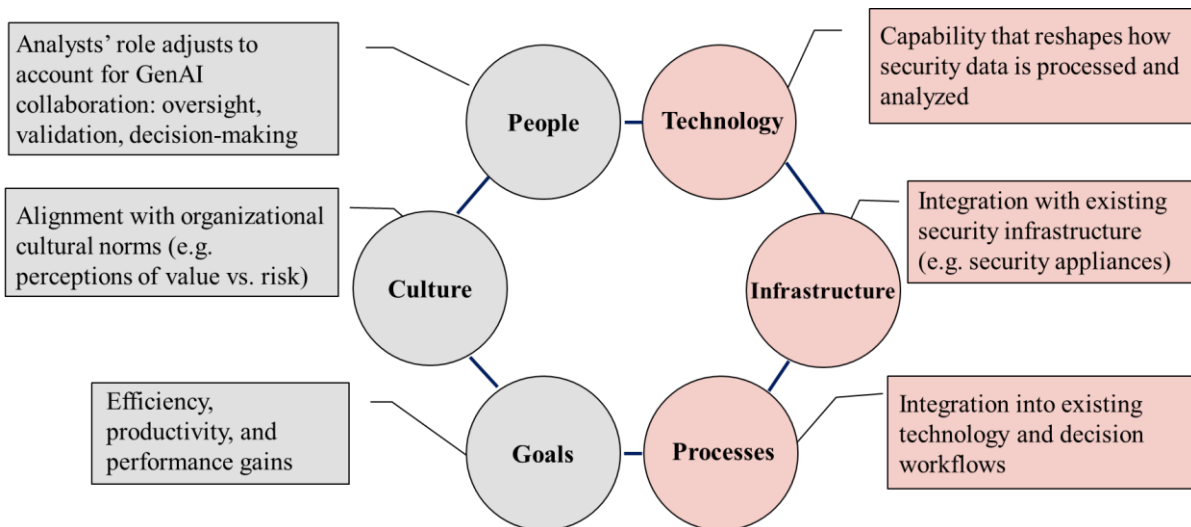
While there have been studies on the risk-benefit tradeoff of adoption of GenAI technology for cybersecurity from a performance benchmarking perspective, there is a lack of evaluating GenAI in live, operational cybersecurity contexts (J. Zhang et al., 2025). For instance, Cheng et al. found that “GenAI has yet to be fully adapted to the complex socio-technical and process-oriented aspects...Furthermore... empirical evidence from industrial contexts [remain] limited” (Cheng et al., 2026, p. 154).

Closing this gap requires formal experimental evaluation in representative deployment scenarios to validate the practical applicability and reliability of GenAI-driven approaches in live operational settings and security workflows. The goal of such an evaluation would be to empirically capture the benefits, opportunities, risks, and limitations of the integration and operationalization of GenAI capabilities for cybersecurity. The measures used in such an evaluation should be informed by the common technology adoption facilitators and blockers – namely, the accuracy, usability, and risk of GenAI.

## Theoretical Model

This study is grounded in socio-technical systems (STS) theory, which posits that an organization's performance and efficacy can be evaluated through the intersection of its technical dimensions and its social dimensions (Trist & Bamforth, 1951). Technical dimensions refer to the actual technology being employed and the work activities directly associated with the technology. Social dimensions refer to more “human” or “actor”-centric aspects of the organization such as policies, social structures, institutional norms, values, and motivations, as well as to the actors themselves. To achieve optimal performance and efficacy, STS theory posits that co-design and joint optimization must be carried out across both dimensions.

In the context of this study, STS theory can be used as a lens for examining how GenAI technology should be operationally realized within existing human-centric workflows and institutional norms to provide the greatest impact while minimizing unintended ill effects. STS theory can be used to critically examine and situate the results and findings of the experiment within the broader context of a generalized human-technology workflow. For instance, the degree to which GenAI can offload cognitive processing, close crucial knowledge and skill gaps, and provide guidance and context that is aligned with cybersecurity best practices and principles can illustrate how institutional values such as perceptions of agency, accountability, responsibility, productivity, and risk, directly shape human-technology interaction and adoption models (Gahrn-Andersen, 2025). Figure 1 illustrates this socio-technical framing by positioning GenAI as a technical capability that interacts with people, culture, goals, processes, and infrastructure.



**Figure 1. STS-theoretic Representation of the use of Generative AI in Cybersecurity Operations.**

## Methodology

This study involves a quantitative analysis of a controlled simulation experiment. The simulation environment comprises a *cyber range* – a repeatable experimentation harness for systematic experimentation – for evaluating GenAI in live cybersecurity operations (Chouliaras et al., 2021; Dehlaghi-Ghadim et al., 2023). The decision to employ a controlled simulation experiment was driven by the goal of constructing an empirically sound understanding of GenAI's practical utility and limitations in live cybersecurity operations (Shadish et al., 2002).

### Cyber Test Range

The simulation environment in Figure 2 is designed to replicate a small-scale cyber-physical system local area network (LAN), comprising information technology (IT) assets in the form of a Windows Active Directory domain and Windows-based workstations, operational technology (OT) assets in the form of a Human-Machine Interface (HMI) workstation and an emulated Programmable Logic Controller (PLC), a monitoring and detection harness, a SIEM appliance for monitoring and alerting, and an adversary emulation harness. All systems were deployed using virtual machines (VMs) within a hypervisor. A comprehensive monitoring workflow was used to log and alert on adversary activity via Zeek (Zeek Project, 2025), Suricata (Open Information Security Foundation, 2025), and Sysmon and Windows event logs (Russovich & Garnier, 2024) which were forwarded to the SIEM (Wazuh Inc, 2025). This testbed design connects enterprise IT assets, OT assets, monitoring tools, SIEM alerting, and adversary emulation in a single controlled environment, allowing GenAI responses to be evaluated against a known attack sequence within grounded operational context.

## Adversary Emulation Scenario

The representative adversary attack follows a structured cyber kill chain driven by real-world threat actor behavior and mapped to tactics and techniques codified in the MITRE ATT&CK framework (The MITRE Corporation, n.d.). The adversary emulation scenario, which is composed of a well-established set of techniques that is representative of contemporary, real-world cyber intrusions, involves an attacker gaining initial access through compromising a user via phishing and web-based connection, conducting network reconnaissance to map the target environment, obtaining Active Directory (AD) domain credentials via an AD ticket compromise technique known as Kerberoasting, remotely logging into the OT workstation using the exploited credentials, and issuing a malicious command to the PLC to disrupt the underlying physical control process and cause physical damage to the equipment (D. Abraham et al., 2025; Knowles et al., 2015; Portase et al., 2024). The attack sequence makes heavy use of living off the land (LoTL) techniques and is directly modeled after the real-life tactics, techniques, and procedures (TTPs) leveraged by cybercriminal groups such as EKANS (Santangelo et al., 2021), DarkSide (Nuce et al., 2021), and Ryuk (Kusuma et al., 2021). The techniques are outlined in Table 2.

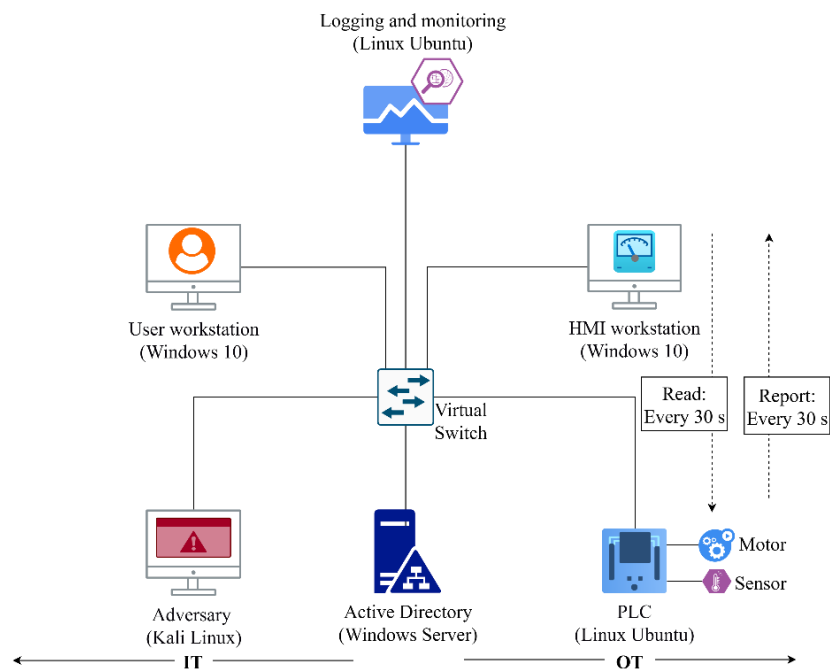


Figure 2. Cyber Test Range

**Table 2. Techniques Comprising Adversary Emulation Sequence**

| Cyber Kill Chain Element | Adversary Technique          | MITRE ATT&CK Mapping | Description                                                             |
|--------------------------|------------------------------|----------------------|-------------------------------------------------------------------------|
| Initial access           | Web Protocols                | T1071                | Payload delivery over HTTP/S protocols                                  |
| Reconnaissance           | Network Service Discovery    | T1046                | Map workstations, Active Directory domain, IT network, and OT network   |
| Exploitation             | Kerberoasting                | T1558                | Active Directory credential compromise                                  |
| Lateral movement         | Remote Services              | T1021                | Remotely log into HMI workstation using stolen credentials              |
| Exploitation             | Process Control Manipulation | T0831                | Issue malicious control commands to impair or damage physical equipment |

## Experimental Workflow

The GenAI chatbot under study is the OpenAI developed *ChatGPT* (OpenAI, 2026). The version of ChatGPT under test is based on the GPT-5 architecture, specifically ChatGPT-5.5. The analyst-GenAI workflow is simulated via a series of prompts designed to enrich the analyst’s knowledge and context of the current security situation. The user interaction prompts employed adhere to the CLEAR prompt framework, which strives for effective text generation by adhering to core principles such as clarity and logic in prompts (Lo, 2023). The core operator prompt used throughout all experimental runs is as follows (paraphrased, omitting formatting instructions):

*“You are a cybersecurity incident response virtual assistant. The following request is for live cybersecurity operations. Given the network architecture and set of attached security logs and alerts, generate a threat and incident response report consisting of*

- the threat activity associated with the security alerts*
- the impact of the threat activity on the operational assets*
- the TTPs, APTs, and broader scale cyber campaigns that could be associated with the threat activity*
- an incident response playbook consisting of suggested immediate mitigation actions, near-term remediation actions, and long-term preventative measures”*

The operator prompt is then systematically manipulated to mirror real-world operational conditions, where security analysts are unlikely to use consistent, precise, or accurate phrasing when interacting with a chatbot virtual assistant using a human language-based interface, such as text or voice (Zhu et al., 2023). This phenomenon is particularly pronounced in the case where an analyst is querying GenAI for guidance in a domain where the analyst has a gap in the specialized knowledgebase of cybersecurity.

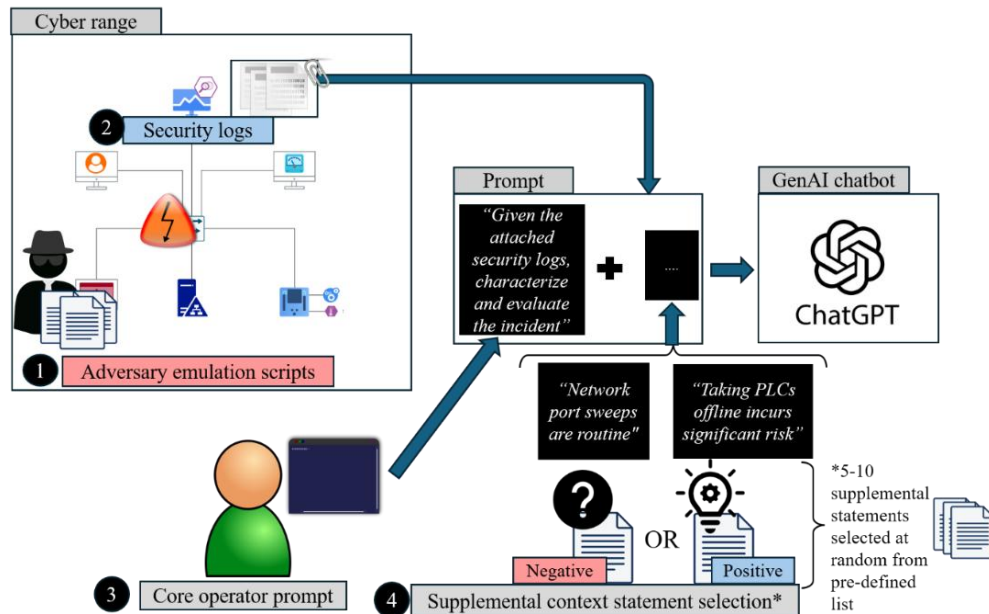
Specifically, the core prompt is supplemented with a set of operator context statements regarding the interpretation of the security logs, the network architecture and services, or guidance on how to manage security incidents. This set of supplemental statements acts as perturbation and will either be “positive” by enriching the core prompt with accurate contextual details, which represents a senior security practitioner that is particularly fluent or versed in cyber operational best practices, or “negative” by adding confusion, distortion, and ambiguity, which represents a junior security practitioner with a gap in domain-specific knowledge. Prior studies on prompt quality and GenAI recommend exploring contextual corruption or enrichment of prompts over superficial prompt “noise” such as typos or arbitrary term swapping (L. Abraham et al., 2025; Alahmari et al., 2025; Jacobsen & Weber, 2025). Examples of supplemental operator context statements used in this experiment are listed in Table 3.

**Table 3. Examples of Corrupted and Enriched Supplemental Operator Context Statements**

| Prompt condition            | Example                                                                                                         |
|-----------------------------|-----------------------------------------------------------------------------------------------------------------|
| Minor contextual corruption | <i>Network port sweeps are routine</i>                                                                          |
| Major contextual corruption | <i>The HMI workstation runs on a general-purpose Linux server and may be rebooted or re-imaged without risk</i> |
| Minor contextual enrichment | <i>Expect Kerberos-related activity - watch for unexpected service ticket requests</i>                          |
| Major contextual enrichment | <i>Taking PLCs offline incurs significant operational risk and coordination with operators</i>                  |

## Experimental Run

An experimental run, summarized in Figure 4, consists of a single execution of the adversary emulation sequence across the cyber range outlined in Table 2; the operational security logs resulting from the execution; the core operator prompt and the supplemental operator context statements provided to the GenAI chatbot; and the generated output from the GenAI chatbot. This study explores five experimental scenarios: a baseline scenario with no supplemental operator context statements and the four scenarios outlined in Table 3 corresponding to varying levels of enriched and corrupted supplemental operator context statements. For a given experimental run, a series of supplemental statements were randomly selected from a pool of 45 total possible supplemental statements that had been pre-defined for “enrichment” or “corruption” categories. “Minor” test scenarios randomly selected 3-5 statements from those that were less verbose, more vague, and less detailed descriptions, whereas “major” test scenarios randomly selected 3-15 statements from those that were highly detailed, more explicit, and more verbose. For each of the five scenarios, 50 independent runs were conducted, for a total of 250 runs.



**Figure 4. Cyber Test Range Experimental Workflow**

## Evaluation Measures and Scoring Procedure

This study employs F1-Score, Embedded Similarity, Hallucination Rate, and Clarification Rate, which serve to capture the quality and usability of the GenAI output and are summarized in Table 4 (Chelli et al., 2024; Dhingra et al., 2019; Ji et al., 2023; T. Zhang et al., 2020). Along with Table 4, Figure 5 illustrates the workflow of the decomposition of the GenAI chatbot output and comparison against the ground-truth reference to calculate quantitative measures related to quality and accuracy of the generated text. The ground-truth reference was created from the adversary emulation sequence and security best practices. The GenAI output was first decomposed into discrete claim statements, which were mapped to corresponding claim statements in the ground-truth reference.

F1-Score was computed on a per-claim statement basis by comparing key facts in each GenAI claim against the corresponding reference claim. True positives were counted when GenAI claim key facts matched reference claim key facts and false negatives were counted when GenAI claim key facts were omitted as compared to reference claim key facts. Unmatched GenAI claim statements were penalized as false positives. Similarly, Embedded Similarity was also computed first on a per-claim statement basis using cosine similarity between matched GenAI and reference claim statements. At the aggregate level, candidate-based precision averaged similarity across GenAI claim statements, while reference-based recall averaged similarity across reference claim statements. Unmatched GenAI claim statements were penalized with a similarity score of zero. Hallucination Rate was computed as the ratio of unmatched GenAI claim statements to total GenAI claim statements. Clarification Rate was computed as the number of additional prompts required to obtain a usable output divided by the total number of prompts.

**Table 4. Quantitative Evaluation Metrics and Measures**

| Metric              | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | Calculation                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
|---------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| F1-Score            | <p>Accuracy based on number of key facts</p> <p><math>R</math>: Set of ground truth claim statements <math>R_i</math><br/> <math>C</math>: Set of GenAI output claim statements where each <math>C_j</math> matches to one <math>R_i</math><br/> <math>U</math>: Set of GenAI output claim statements where each <math>U_k</math> does not match to any <math>R_i</math></p> <p>For each claim statement <math>R_i</math><br/> <math>r_i</math>: set of key facts in <math>R_i</math><br/> <math>c_j</math>: set of key facts in <math>C_j</math>, <math>C_j</math> matched to <math>R_i</math></p> | <p><math>\Sigma TP</math>: Sum across all claim statements <math>R_i</math> of key facts found in <b>both</b> <math>r_i</math> and <math>c_j</math><br/> <math>\Sigma FN</math>: Sum across all claim statements <math>R_i</math> of key facts in <math>r_i</math> but <b>missing from</b> <math>c_j</math><br/> <math>\Sigma FP =  U </math>: Total number of <math>U_k</math> in <math>U</math></p> $Pr = \frac{\Sigma TP}{\Sigma TP + \Sigma FP} \quad Re = \frac{\Sigma TP}{\Sigma TP + \Sigma FN}$ $F_1 = \frac{2(Pr)(Re)}{Pr + Re}$ |
| Embedded Similarity | <p>Accuracy based on semantic similarity</p> <p><math>R</math>: Set of ground truth claim statements <math>R_i</math><br/> <math>C</math>: Set of GenAI output claim statements where each <math>C_j</math> matches to one <math>R_i</math><br/> <math>U</math>: Set of GenAI output claim statements where each <math>U_k</math> does not match to any <math>R_i</math></p>                                                                                                                                                                                                                        | <p><math>sim(C_j, R_i)</math>: cosine similarity</p> $Pr = Avg_{C,U} \left( \begin{matrix} sim(C_j, R_i), C \\ 0, U \end{matrix} \right)$ $Re = Avg_R \left( \begin{matrix} sim(C_j, R_i), R_i \rightarrow C_j \text{ match} \\ 0, R_i \text{ no } C_j \text{ match} \end{matrix} \right)$ $sim_{embedded} = \frac{2(Pr)(Re)}{Pr + Re}$                                                                                                                                                                                                   |
| Hallucination Rate  | <p>Accuracy based on extraneous claims</p> <p><math> C </math>: Total matched claim statements<br/> <math> U </math>: Total unmatched claim statements</p>                                                                                                                                                                                                                                                                                                                                                                                                                                          | $R_H = \frac{ U }{ U  +  C }$                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |

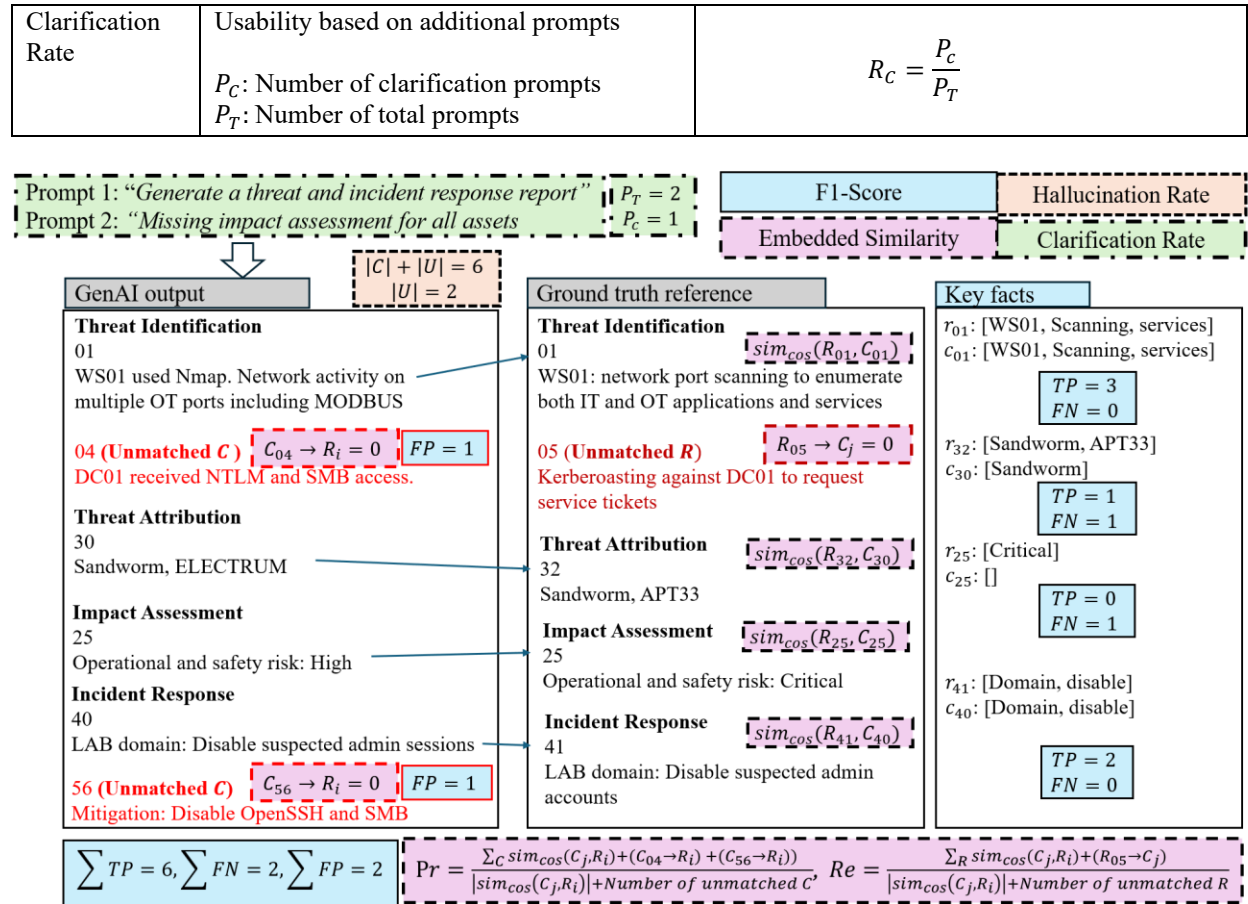


Figure 5. GenAI Output Decomposition and Sample Computational Workflow of Evaluation Measures

## Results and Discussions

Figure 6 shows a small representative sample of Zeek, Suricata, and Wazuh log captures from the test range that resulted from the adversary emulation sequence. The Zeek and Suricata outputs capture adversary activity associated with network traffic. In the Zeek logs, Kerberos (krb\_tgt) exchanges indicate potential malicious Kerberoasting activity, and the clustered network activity for various OT communication protocols (20000 for DNP3, 44818 for Ethernet/IP, 47808 for BACnet, and 502 for MODBUS) are indicative of adversarial network scanning.

The Suricata logs indicate potential transfer of a malicious payload via HTTP and potential PLC compromise via MODBUS control commands. The Wazuh/Sysmon logs indicate potential malicious remote logins to the HMI workstation.

```

zeek
conn.log
172.16.125.77 49956 172.16.125.100 80 tcp http 0.011773 175 0 SH T
172.16.125.100 39102 172.16.125.87 445 tcp gssapi,ntlm,smb 0.006751 1234 1840 RSTR T T 1234 ShCdGar 1 60 9 2212 - 6
172.16.125.100 38374 172.16.125.50 88 tcp krb_tcp 0.000908 0 179 RSTRH T T 0 ^hCdgar 0 0 4 383 - 6
172.16.125.100 38390 172.16.125.50 88 tcp krb_tcp 0.000872 0 1568 RSTRH T T 0 ^hCdgar 0 0 5 1824 - 6
172.16.125.87 49838 172.16.125.200 44818 tcp - 0.000208 0 0 REJ T T 0 Sr 1 52 1 40 - 6
172.16.125.87 49839 172.16.125.200 47808 tcp - 0.000157 0 0 REJ T T 0 Sr 1 52 1 40 - 6
172.16.125.87 49841 172.16.125.200 20000 tcp - 0.000077 0 0 REJ T T 0 Sr 1 52 1 40 - 6
172.16.125.87 49840 172.16.125.200 502 tcp - 0.000285 0 0 OTH T T 0 ScAR 3 132 0 0 - 6
172.16.125.87 49834 172.16.125.50 49669 tcp dce_rpc 13.370012 1074 1160 SF T T 0 ShAddFaf 7 1366 5 1372 - 6
172.16.125.87 49842 172.16.125.200 502 tcp modbus 0.011321 12 0 SH T T 0 ScADF 5 224 0 0 - 6
172.16.125.87 49843 172.16.125.200 502 tcp modbus 0.013376 12 0 SH T T 0 ScADF 5 224 0 0 - 6

Suricata
fast.log
Python SimpleHTTP ServerBanner [**] [Classification: Misc activity] [Priority: 3] {TCP} 172.16.125.100:80 -> 172.16.125.77:49823
ET EXPLOIT Possible GoldenPac Priv Esc in-use [**] [Classification: Attempted Administrator Privilege Gain] [Priority: 1] {TCP}
MODBUS TCP traffic detected [**] [Classification: (null)] [Priority: 3] {TCP} 172.16.125.87:49787 -> 172.16.125.200:502

Wazuh
alerts.log
"description": "Successful Remote Logon Detected - User:\\Administrator - NTLM authentication, possible pass-the-hash attack - Possible RDP connection."
"description": "Lsass process was accessed by C:\\\\Windows\\\\system32\\\\svchost.exe with read permissions, possible credential
"description": "User: LAB\\administrator logged using Remote Desktop Connection (RDP) from ip:172.16.125.100."
"description": "A network share was accessed."

```

Figure 6. Zeek, Suricata, and Wazuh (SIEM) Logs from Adversary Emulation Scenario

Figure 7 shows a sample of the ChatGPT operator interaction prompt as well as a sample of the ChatGPT generated output in response to that prompt.

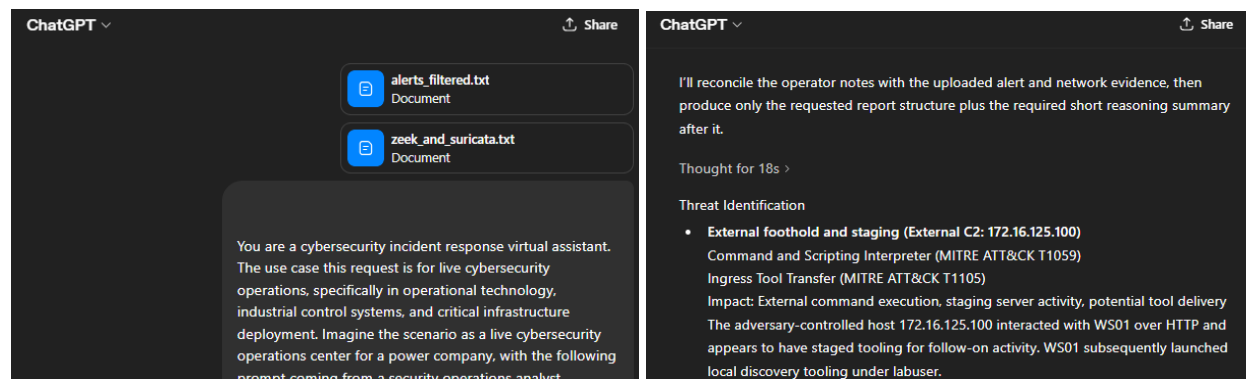


Figure 7. Sample Operator Prompt and Chatbot Response for ChatGPT 5.5

An example of ChatGPT correctly identifying the AD domain exploitation technique as Kerberoasting and correctly evaluating the operational risk to the OT environment as “critical” is as follows:

*Domain Credential Abuse and Kerberoasting (DC01)*  
*Kerberoasting (T1558), Valid Accounts(T1078)*  
*Impact: Service-ticket exposure, privileged account abuse, domain compromise risk*  
*External host contacted DC01 over LDAP and Kerberos. The activity involved Administrator authentication and a service ticket for svc\_web.*  
*Operational and safety risk: Critical*

- *Unauthorized Modbus coil writes could alter PLC-controlled process state.*
- *Remediation: Validate ladder logic/state, restore trusted configuration*
- *Risk profile: High, direct process-control impact, operator coordination required*

In contrast, an example of ChatGPT incorrectly identifying the AD domain exploitation as password guessing and incorrectly evaluating the operational risk to the OT environment as “low” is as follows:

*External Authentication Probing (DC01)*

*Password Guessing (T1110), Valid Accounts (T1078)*

*Impact: Tests account access, identifies usable credentials, enables follow-on domain access*

*An external host initiated remote authentication activity against DC01 using NTLM over SMB.*

*Operational and safety risk: Low*

- *PLC write activity was observed, but operator notes indicate low safety consequence*
- *Remediation for PLC: Validate ladder/configuration state, restore trusted baseline.*
- *Risk profile: Low, PLC can be safely isolated, confirms process integrity*

Table 5 captures the mean and standard deviation of each of the evaluation metrics across all 50 independent runs of each of the five scenarios. The measures were then analyzed via standard one-way and Welch ANOVA tests with Holm correction applied, with the ANOVA results captured in Table 6. According to the ANOVA results, the supplemental context had a statistically significant effect on F1-Score, Embedded Similarity, and Hallucination Rate, but not Clarification Rate, which remained consistently near 0, likely due to the highly structured nature of the prompt to ensure consistency. Clarification Rate may thus be less suited to structured prompts and more suited to variable, interactive GenAI workflows. Tukey HSD and Games-Howell post-hoc tests indicated that the two corruption scenarios reduced F1-Score and Embedded Similarity and increased Hallucination Rate relative to baseline by a statistically significant degree, whereas the enrichment scenarios did not deviate significantly from baseline.

**Table 5. Descriptive Statistics for Evaluation Metrics Across Experimental Prompt Conditions**

|                  | F1-Score |       | Embedded Similarity |       | Hallucination Rate |       | Clarification Rate |       |
|------------------|----------|-------|---------------------|-------|--------------------|-------|--------------------|-------|
|                  | Mean     | SD    | Mean                | SD    | Mean               | SD    | Mean               | SD    |
| Baseline         | 0.450    | 0.046 | 0.511               | 0.035 | 0.202              | 0.076 | 0.01               | 0.071 |
| Minor corruption | 0.363    | 0.076 | 0.454               | 0.055 | 0.242              | 0.083 | 0.01               | 0.071 |
| Major corruption | 0.330    | 0.056 | 0.436               | 0.043 | 0.261              | 0.072 | 0.02               | 0.099 |
| Minor enrichment | 0.466    | 0.046 | 0.537               | 0.030 | 0.175              | 0.072 | 0                  | 0     |
| Major enrichment | 0.489    | 0.044 | 0.526               | 0.027 | 0.159              | 0.062 | 0                  | 0     |

**Table 6. One-way ANOVA Results for Evaluation Metrics Across Experimental Prompt Conditions**

|                     | Selected ANOVA   | F     | Holm-adjusted p | $\eta^2$ | Significant? |
|---------------------|------------------|-------|-----------------|----------|--------------|
| F1-Score            | Welch            | 80.92 | < 0.001         | 0.565    | Yes          |
| Embedded Similarity | Welch            | 62.25 | < 0.001         | 0.514    | Yes          |
| Hallucination Rate  | Standard one-way | 17.51 | < 0.001         | 0.222    | Yes          |
| Clarification Rate  | Standard one-way | 0.88  | 0.474           | 0.014    | No           |

The results demonstrate a statistically significant effect of prompt clarity, accuracy, and contextual richness on chatbot accuracy. Prompts with inaccurate, corrupted operator context served to “nudge” the GenAI chatbot to incorrectly identify and categorize threats, provide an inaccurate assessment of the risk profile and the impact of remediation, and recommend incident response mitigations. The negative impact of inaccurate operator context suggests that ChatGPT places disproportionate emphasis and bias on the legitimacy of operator inputs. Security logs are ambiguous and underspecified by nature; thus, the synthesis of logs into accurate actionable outputs is highly dependent on operator domain expertise or situational context (e.g. asset criticality, expected operations, expected access patterns, intuitive grasp of cyber risk

calculus). Consistent with STS theory, a GenAI driven security operations workflow is therefore ultimately dominated by the expertise of the human operator, with GenAI falling into a supporting, “follower” role whose ability to close any knowledge or skill gaps in the operator’s knowledge is limited by the amount of bias unintentionally introduced by the operator.

The results suggest several practical and safety considerations for GenAI-assisted SOC workflows, such as ensuring trusted sources for contextual GenAI enrichment (e.g. known high-quality context from trusted domain or operations experts), supplemental GenAI usage with accurate, validation operational context via techniques such as via Retrieval Augmented Generation (RAG) or leveraging policy/technical guardrails against using GenAI outputs without formal third-party validation. Prompt quality and accuracy should be managed as part of the workflow through structured prompt templates and analyst training. The results of this study are limited by several key constraints and should be carefully interpreted with these limitations in mind. Such a small-scale scenario cannot fully replicate the unpredictable nature of a live security incident; thus, the findings should not be interpreted as directly predictive of live operations in large-scale SOC environments. The experimental design also abstracts away many operational dimensions of in-situ cybersecurity operations, including human analysts, team coordination, and complex organizational dynamics. This abstraction is consistent with STS-theoretic research that recommends decomposition of complex systems into smaller components that can be tractably studied. Additionally, the evaluation is specific to the GenAI model leveraged in this study: ChatGPT 5.5 and the GPT-5 architecture. Different models with different architectures, training data, and tailored fine-tuning have a high likelihood of yielding different results. This study therefore emphasizes behavioral *patterns* over absolute performance metrics used in traditional benchmarking studies on GenAI. Thus, while model performance will generally improve over time, phenomena such as hallucinations or sensitivity to prompt clarity are likely to persist.

## Conclusions

This study explored how prompt accuracy can impact GenAI performance in a controlled cybersecurity operations scenario. Using a cyber range, adversary emulation techniques, and systematic prompt perturbation, the study found that inaccurate context has the potential to degrade GenAI accuracy and increase hallucinations. These findings reinforce the socio-technical nature of GenAI and suggest that GenAI in cybersecurity settings should be tuned to weight security best practices over operator context and narratives. Future studies should involve live SOC environments and exploring the impact of iterative, real-time prompt refinement and core design interventions (e.g. RAG, analyst-in-the-loop verification).

## Acknowledgements

This research is supported by a grant award from the United States National Science Foundation (NSF) – NSF Grant ID 2234554.

## References

ADDIN ZOTERO\_BIBL {"uncited":[],"omitted":[],"custom":[]} CSL\_BIBLIOGRAPHY Abraham, D., Houmb, S. H., & Erdodi, L. (2025). Cyber-attacks on energy infrastructure—A literature overview and perspectives on the current situation. *Applied Sciences*, 15(17), 9233.

- Abraham, L., Arnal, C., & Marie, A. (2025). Prompt selection matters: Enhancing text annotations for social sciences with large language models. *Journal of Computational Social Science*, 8(3), 1–20. <https://doi.org/10.1007/s42001-025-00388-6>
- Alahmari, S. S., Hall, L., Mouton, P. R., & Goldgof, D. (2025). Large language models robustness against perturbatio. *Scientific Reports*.
- Alnajim, A. M., Habib, S., Islam, M., Thwin, S. M., & Alotaibi, F. (2023). A comprehensive survey of cybersecurity threats, attacks, and effective countermeasures in industrial internet of things. *Technologies*, 11(6), 161.
- Bono, J. (2025). *Security Copilot: Evidence of Productivity Gains in Live Operations*. Microsoft Security.
- Charfeddine, M., Kammoun, H., Hamdaoui, B., & Guizani, M. (2024). ChatGPT's security risks and benefits: Offensive and defensive use-cases, mitigation measures, and future implications. *IEEE Access*, 12, 30263–30310. <https://doi.org/10.1109/ACCESS.2024.3367792>
- Chelli, M., Descamps, J., Lavoué, V., Trojani, C., Azar, M., Deckert, M., Raynier, J.-L., Clowez, G., Boileau, P., & Ruetsch-Chelli, C. (2024). Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: Comparative analysis. *Journal of Medical Internet Research*, 26(1), 1–11.
- Cheng, H., Husen, J. H., Lu, Y., Racharak, T., Yoshioka, N., Ubayashi, N., & Washizaki, H. (2026). Generative AI for requirements engineering: A systematic literature review. *Software: Practice and Experience*, 56(2), 141–170. <https://doi.org/10.1002/spe.70029>
- Choi, S. L., Jain, R., Feng, C., Emami, P., Zhang, H., Hong, J., Kim, T., Park, S., Ding, F., & Baggu, M. (2024). *Generative AI for power grid operations* (Technical Report NREL/TP-5D00-91176; pp. 1–40). National Renewable Energy Laboratory. <https://www.osti.gov/servlets/purl/2477920>
- Choudhury, A., & Asan, O. (2022). Impact of accountability, training, and human factors on the use of artificial intelligence in healthcare: Exploring the perceptions of healthcare practitioners in the US. *Human Factors in Healthcare*, 2. <https://doi.org/10.1016/j.hfh.2022.100021>
- Chouliaras, N., Kittes, G., Kantzavelou, I., Maglaras, L., Pantziou, G., & Ferrag, M. A. (2021). Cyber ranges and testbeds for education, training, and research. *Applied Sciences*, 11(4), 1809.
- Dehlaghi-Ghadim, A., Balador, A., Moghadam, M. H., Hansson, H., & Conti, M. (2023). ICSSIM — A framework for building industrial control systems security testbeds. *Computers in Industry*, 148, 1–14. <https://doi.org/10.1016/j.compind.2023.103906>
- Dhanushkodi, K., & Thejas, S. (2024). AI enabled threat detection: Leveraging artificial intelligence for advanced security and cyber threat mitigation. *IEEE Access*, 12, 173127–173136. <https://doi.org/10.1109/ACCESS.2024.3493957>
- Dhingra, B., Faruqui, M., Parikh, A., Chang, M.-W., Das, D., & Cohen, W. W. (2019). Handling divergent reference texts when evaluating table-to-text generation. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4884–4895.
- Dwivedi, R. (2023). Ten years of cybersecurity governance, risk and compliance: A bibliometric examination of research themes, trends, and influencers. *Issues in Information Systems*, 24(3).

- Gahrn-Andersen, R. (2025). Integrating 4E cognition with science and technology studies: A framework for understanding AI applications. *Frontiers in Artificial Intelligence*, 8, 1–5.
- Gupta, M., Akiri, C., Aryal, K., Parker, E., & Praharaj, L. (2023). From ChatGPT to ThreatGPT: Impact of generative AI in cybersecurity and privacy. *IEEE Access*, 11, 80218–80245. IEEE Access. <https://doi.org/10.1109/ACCESS.2023.3300381>
- Hasanov, I., Virtanen, S., Hakkala, A., & Isoaho, J. (2024). Application of large language models in cybersecurity: A systematic literature review. *IEEE Access*, 12, 176751–176778.
- Ibrahim, N., & Kashef, R. (2025). Exploring the emerging role of large language models in smart grid cybersecurity: A survey of attacks, detection mechanisms, and mitigation strategies. *Frontiers in Energy Research*, 13, 1–23.
- Jacobsen, L. J., & Weber, K. E. (2025). The promises and pitfalls of large language models as feedback providers: A study of prompt engineering and the quality of AI-driven feedback. *AI*, 6(2), 35.
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), 1–38.
- Knowles, W., Prince, D., Hutchison, D., Disso, J. F. P., & Jones, K. (2015). A survey of cyber security management in industrial control systems. *International Journal of Critical Infrastructure Protection*, 9, 52–80. <https://doi.org/10.1016/j.ijcip.2015.02.002>
- Kusuma, R. S., Umar, R., & Riadi, I. (2021). Network forensics against Ryuk ransomware using trigger, acquire, analysis, report, and action (TAARA) method. *KINETIK: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, 6(2), 133–140.
- Lo, L. S. (2023). The CLEAR path: A framework for enhancing information literacy through prompt engineering. *The Journal of Academic Librarianship*, 49(4), 102720. <https://doi.org/10.1016/j.acalib.2023.102720>
- Lu, X., Lu, Y., Luo, L., Lin, W., & Zhang, J. (2022). Remote data acquisition and management technology of power equipment based on internet of things. *Mobile Information Systems*, 2022(1).
- Maglaras, L., Janicke, H., & Ferrag, M. A. (2022). Cybersecurity of critical infrastructures: Challenges and solutions. *Sensors*, 22(14), Article 14. <https://doi.org/10.3390/s22145105>
- Meeker, M., Simons, J., Chae, D., & Krey, A. (2025). *Trends – artificial intelligence* (p. 350) [Trends report]. BOND. [https://www.bondcap.com/report/pdf/Trends\\_Artificial\\_Intelligence.pdf](https://www.bondcap.com/report/pdf/Trends_Artificial_Intelligence.pdf)
- Morgan, S. (2023, October 12). Cybercrime to cost the world \$9.5 trillion USD annually in 2024. *Cybercrime Magazine*. <https://cybersecurityventures.com/cybercrime-to-cost-the-world-9-trillion-annually-in-2024/>
- Nuce, J., Kennelly, J., Goody, K., Moore, A., Rahman, A., Williams, M., McKeague, B., & Wilson, J. (2021, May 11). Shining a light on darkside ransomware operations [Security blog]. *GoogleCloud Blogs*. <https://cloud.google.com/blog/topics/threat-intelligence/shining-a-light-on-darkside-ransomware-operations/>

- Open Information Security Foundation. (2025). *Suricata (Version 8.0.0) [Computer software]*. <https://suricata.io>
- OpenAI. (2026). *ChatGPT (April 2026 version; GPT-5)*. <https://chatgpt.com/>
- Portase, R. M., Colesa, A., & Sebestyen, G. (2024). SpecRep: Adversary emulation based on attack objective specification in heterogeneous infrastructures. *Sensors*, 24(17), 1–21.
- Prieto, I., & Blakely, B. (2023). *Proposed uses of generative AI in a cybersecurity-focused SOAR agent*. 2(1), 386–390.
- Russinovich, M., & Garnier, T. (2024, July 23). *Sysmon v15.15*. Microsoft Learn. <https://learn.microsoft.com/en-us/sysinternals/downloads/sysmon>
- Saddi, V. R., Gopal, S. K., Mohammed, A. S., Dhanasekaran, S., & Naruka, M. S. (2024). Examine the role of generative AI in enhancing threat intelligence and cyber security measures. *2024 2nd International Conference on Disruptive Technologies (ICDT)*, 537–542.
- Sai, S., Yashvardhan, U., Chamola, V., & Sikdar, B. (2024). Generative AI for cyber security: Analyzing the potential of ChatGPT, DALL-E and other models for enhancing the security space. *IEEE Access*, 12.
- Santangelo, G. V., Colacino, V. G., & Marchetti, M. (2021). Analysis, prevention and detection of ransomware attacks on Industrial Control Systems. *IEEE 20th International Symposium on Network Computing and Applications (NCA)*, 1–5.
- Scalise, P., Boeding, M., Hempel, M., Sharif, H., Delloiacovo, J., & Reed, J. (2024). A systematic survey on 5G and 6G security considerations, challenges, trends, and research areas. *Future Internet*, 16(3).
- Sedkaoui, S., & Benaichouba, R. (2024). Generative AI as a transformative force for innovation: A review of opportunities, applications and challenges. *European Journal of Innovation Management*. <https://doi.org/10.1108/EJIM-02-2024-0129>
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton, Mifflin and Company.
- Tedeschi, P., Sciancalepore, S., & Di Pietro, R. (2022). Satellite-based communications security: A survey of threats, solutions, and research challenges. *Computer Networks*, 216, 109246.
- The MITRE Corporation. (n.d.). *MITRE ATT&CK®*. MITRE ATT&CK®. Retrieved September 1, 2025, from <https://attack.mitre.org/>
- Trist, E. L., & Bamforth, K. W. (1951). Some social and psychological consequences of the longwall method of coal-getting: An examination of the psychological situation and defences of a work group in relation to the social structure and technological content of the work system. *Human Relations*, 4(1), 3–38.

- Wang, M. (2023). *A new era of cybersecurity with AI: Predictions for 2024*. Palo Alto Networks. <https://www.paloaltonetworks.com/cybersecurity-perspectives/a-new-era-of-cybersecurity-with-ai>
- Wazuh Inc. (2025). Wazuh (Version 4.12.0) [Computer software]. <https://wazuh.com>
- Zeek Project. (2025). *Zeek (Version 0.17.0) [Computer software]*. <https://zeek.org>
- Zhang, J., Bu, H., Wen, H., Liu, Y., Fei, H., Xi, R., Li, L., Yang, Y., Zhu, H., & Meng, D. (2025). When LLMs meet cybersecurity: A systematic literature review. *Cybersecurity*, 8(1), 55.
- Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). Bertscore: Evaluating text generation with bert. *Proceedings of the International Conference on Learning Representations*, 1–43.
- Zhu, K., Wang, J., Zhou, J., Wang, Z., Chen, H., Wang, Y., Yang, L., Ye, W., Zhang, Y., & Gong, N. (2023). *Promptrobust: Towards evaluating the robustness of large language models on adversarial prompts*. 57–68.