Journal of Information Systems Engineering and Management
2025, 10(56s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Dual-Use of Generative AI in Cybersecurity: Balancing Offensive Threats and Defensive Capabilities in the Post-LLM Era

Sridhar Krishna Korimilli[1], Md Habibur Rahman[2], Goutham Sunkara[3], Mohammad Mushfiqul Haque Mukit[4], Abdullah Al Hasib[5]

[1]*University/Institute Name: Oracle, USA*
Department: Sr.Technical Leader, Customer Success Services, Technology Delivery
Personal or Institutional Email: skkorimilli@gmail.com
ORCID ID (if any): https://orcid.org/0009-0009-4788-6391
[2]*hrahman.student@wust.edu*
Graduate Researcher, Washington University of Science and Technology
https://orcid.org/0009-0008-8715-2504
[3]*Official Email: sgoutham.sunkara@gmail.com*
Affiliation: Broadcom Inc.
ORCID: 0009-0001-0633-0890
[4]*University/Institute Name: Washington University of Science and Technology*
Department: Information Technology
Email (Personal/Institutional): mmukit.ny@gmail.com
[5]*University/Institute Name: Washington University of Science and Technology*
Department: Information Technology
Email (Personal/Institutional): abdullahalhasib03@yahoo.com

ARTICLE INFO

Received: 15 Jan 2025

Revised: 24 Feb 2025

Accepted: 14 Mar 2025

ABSTRACT

With the advent of large language models (LLMs) and generative AI, cybersecurity has been transformed, changing the game's rules, adding new opportunities for threat detection, and simultaneously escalating the cyber surface being attacked. This study examines the challenge of the dual use of generative AI in cybersecurity by demonstrating how LLaMA2, GPT-3.5, and Falcon models could be used on offensive (red-team) and defensive (blue-team). By utilising two publicly accessible Kaggle datasets and running experiments on Google Colab, the study establishes that LLMs can be used to create advanced phishing attacks that evade conventional detection tools, as well as lead to better accuracy in threat classification and decrease the triage time should they be deployed defensively. To deal with such dual-use capabilities, the study is based on a formatted approach that considers the notions of transparency, intent awareness, and risk stratification. As a new metric, the Dual-Use Risk Index (DURI) is proposed to rate generative works live, backed by guard violations built into the governance by design structure. The framework is consistent with the top regulatory efforts, including the NIST AI Risk Management Framework and the EU AI Act. The findings show that, even though generative models promise operational improvement by a wide margin, they also necessitate strong policy implementation, an auditing system, and adaptation per sector. Within the context of using generative AI safely and ethically in information system contexts, this paper helps to contribute empirical experience and an operational model of governance that organisations may apply in implementing quickly, secure, and responsible generative AI.

Keywords: Generative AI; Large Language Models; Cybersecurity; Dual-Use Risk; Governance Framework; Prompt Engineering; AI Ethics; Information Systems Security; DevSecOps; Responsible AI.

1 Introduction

The blistering transformation and rise of large language models (LLMs) like GPT-4, Claude, and LLaMA have completely altered digital communications and creation, as well as the automation of the software space. Their capability to produce large quantities of congruent text, code, and natural responses has brought on fresh efficiencies and capacities in all sectors (Kasri et al. (2025). Nevertheless, such an influential nature has also filtered into cybersecurity, in which

Journal of Information Systems Engineering and Management
2025, 10(56s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

generative AI presents a two-edged problem today. The same malevolent actors use the same models that malevolent actors use to perform covert attacks to perform offensive operations, and security teams use these same models to strengthen digital defense mechanisms.

Generative AI is also employed in offensive scenarios to make advanced attacks automated and scalable (Ankalaki et al., 2025). Its opponents have started to use LLMs to compose very convincing phishing emails, obfuscated malware codes, and bespoke social engineering scripts that can evade standard security scanners. Examples of simulated tools like WormGPT and DarkBERT have shown how adversaries may use open-source generative models to weaponise natural language generation (Lin et al., 2024). These are the attack vectors augmented with the situational awareness and fluency of LLMs that create significant problems with traditional security operations and highlight the insignificance of relying on a system that depends on the use of static rules as a way to deploy a security solution and protect an organisation.

On the other hand, generative AI is also promising with a defensive force multiplier. Security teams are using LLMs to develop more intelligent Security Operations Centre (SOC) helpers that might sort through warnings, sum up threat reports, and automate rescue workflows (Shakil et al., 2023). Zero-day threats can be detected through AI-based anomaly detection models and a log analyser to find abnormal behaviour and facilitate quicker mitigation. As enterprise security stacks include platforms such as Microsoft's COPA and fine-tuning AI models such as OpenAI, generative AI is proving to be a core asset in online risk management and frontline defense activities (Abdollahian, 2025).

There is a policy and governance dilemma over this dual-use paradigm. With the distinction between offensive and defensive uses dissolving into generative AI, businesses and governments have increasingly been under pressure to outline systems to ensure such technology is utilized earnestly (Mia et al., 2025). This asymmetry of innovations on one side and the speed of response and ability to react on the other requires the immediate introduction of new standards of ethical deployment, auditing procedures, and model governance approaches. Besides, intent detection on generative systems remains elusive, making it more challenging to attribute, score risk, and hold accountable. Unless there are strong protective measures, even those models which protect the organisations would work against the organisations (Kumar et al., 2021).

This study aims to empirically study the dual-use of generative AI in the context of cybersecurity through model-driven simulations of red-team (offensive) and blue-team (defensive) scenarios. Using Kaggle publicly available datasets, open-source models, and Google Colab-powered experiments, we assess the ability of LLMs to produce phishing content, synthesise simple exploits, and automatically detect or prevent attacks in an automated SOC environment. The review of the policy frameworks and technical control applied to the governance of AI supports the analysis.

The rest of the paper will be structured as follows: Section 2 reviews related literature and the conceptual basis underpinning dual-use technologies in cybersecurity. Section 3 explains the methodology, datasets, and simulation model. Section 4 reports the experimental findings, such as the generation of red team attacks and the response achieved by the blue team. In section 5, the author offers a policy framework that lays the foundation for responsible AI implementation. Sections 6 and 7 are more general regarding discussion, implications, and conclusion.

2 Literature Review

The incisive development of artificial intelligence, especially large language models (LLMs), has triggered a paradigm shift in more secure behaviour. The threat surface and defensive toolkit have been increased due to the ability of LLMs to produce natural language and write executable code to update themselves with dynamic prompts in real-time (Ibrahim & Kashef, 2025). This section analyses the academic and technological literature on the dual use of generative AI. It puts its emergence into the context of history, theory, and regulations as applied to cybersecurity and information system management.

2.1 Generative AI and Its Role in Information Systems

Leveraging generative AI within enterprise information systems has opened up the potential for new degrees of effectiveness and automation. Such models as GPT-4 or Claude are being employed in document summarisation, customer service automation, generation of software, and cybersecurity activities, including log analysis and synthesis

Journal of Information Systems Engineering and Management
2025, 10(56s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

of threat intelligence (Chen et al., 2024). When discussing cybersecurity, generative AI can no longer be defined as some minor instrument but rather a fundamental element of security information and event management (SIEM) systems and automated response frameworks (Ali et al., 2024). Nevertheless, the positive uses of the applications are fully discussed, but there is a gap in the traditional IS literature on the misuse of the said applications.

2.2 Dual-Use Technologies: Historical Context and Relevance

The dual-use dilemma (in which an item of technology, originally developed with good intentions, can be redesigned to carry out evil deeds) has long been familiar in the biotechnology, cryptography, and nuclear engineering fields. Regarding matters of AI, (Schuett et al., 2023) cautioned against what they termed as dual-use AI, whereby the benefits of open-sourcing and generative expertise of AI are repurposed to serve the opposite purpose of protection. This dilemma has gotten even more serious with the development of LLMs. The availability of open-source versions, such as LLaMA and uncensored clones, allows a broader group of stakeholders, including non-state cybercriminals, to take advantage of model capabilities on their own without any institutional checks to do so (Vaishnav et al., 2025).

This dual purpose destructively affects cybersecurity. For example, experts have mentioned tools such as WormGPT and FraudGPT, which are fine-tuned models that are supposedly fine-tuned to produce phishing or fraud scripts (Xu et al., 2024). These models run in the grey market or dark web settings and are frequently built using stolen data and perfecting the process to maximise the success of attacks. The conceptual framework, in this case, is associated with technological determinism. With the development of the ability to generate, the extremes of its abuse are becoming easier to accomplish unless properly fought off by effective control.

2.3 Generative AI in Offensive and Defensive Cybersecurity

In the attacker space, generative AI can scale phishing attempts, make polymorphic malware, build fake personas, and design automated social engineering. Such models enable the initiation of cyberattacks by non-technical individuals by simplifying the generation of technical payloads with the help of natural language prompts (Mallick & Nath, 2024). In addition, content filters can be evaded with an immediate injection attack and jailbreak methods in executed systems, which results in the creation of malicious instructions that are launched on innocent interfaces (Liao et al., 2025).

On the defense side, enterprises are integrating generative AI with smart SOCs. Such tools as Security Copilot by Microsoft have LLMs that summarise threat alerts, suggest remediation actions, and automate incident response documentation (Yigit et al., 2025). Likewise, open-source initiatives combine generative models and anomaly detection systems to address the visibility of dynamic environment threats. The SIEM pipelines are also getting LLMs to enable correlation analysis and proactive threat hunting, lowering the mean time to detect (MTTD) and mean time to respond (MTTR) (Paidy, 2025).

However, there is a severe drawback to both offensive and defensive uses since LLMs do not coincidentally know hate or intent. Such semantic irresponsibility makes their employment in risk-sensitive areas challenging. In contrast to human analysts, LLMs have no way of knowing the difference between ethical use and abuse unless immediate filters, fine-tuning boundaries, or human-in-the-loop oversight restrict them.

2.4 Policy and Regulatory Frameworks for AI Governance

Considering these issues, discussions on how dual-use AI will be governed internationally have started. According to the NIST (2023) AI Risk Management Framework, its guidance is based on mapping, measuring, and mitigating AI risks throughout the system lifecycle. It presents the four principles of traceability, audibility, and explainability that are the basis of trustworthy AI. Parliament (2023) presents a risk-based classification of AI systems and prohibits the application of such categories of AI as inadmissibly risky ones, such as some biometric and mass surveillance tools.

However, in cybersecurity, there is no consistent information on how to work with generative AI. IEEE P7000 series, in particular, P7001 and P7009, present ethical design and accountability principles but do not outline approachable controls over the deployment of open-source LLM. Internal policies on responsible model use, such as watermarking, red-teaming, and restrictions on usage, have been issued by OpenAI and Anthropic (Ferdaus et al., 2024). However, these policies are voluntary and do not have much consistency between organisations.

Journal of Information Systems Engineering and Management
2025, 10(56s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Also, there exists no international scale for measuring or rating the dual-use risk of an LLM. The lack of these frameworks leaves gaps that attackers can use against them, and it also leaves enterprises confused about the limits of compliance when it comes to delivering them to the production environments.

2.5 Identified Research Gaps and Study Justification

Although there is an increasing interest in investigating the potential outcomes of generative AI in cybersecurity, few empirical studies touch on both the offensive and defensive aspects of generative AI. Most scholarly articles address the single goals of either attack development or defense advancement without inquiring about the relative performance of LLMs in both roles (Zhang et al., 2025).

Additionally, most studies have lacked empirical investigation that integrates technical simulation, policy evaluation, and actual governance implications with accessible data and reproducible pipelines.

This research fills gaps by:

1. Performing an offensive and defensive demonstration with the help of Google Colab and Kaggle data,
2. Assessing the behavioural divergence of the model in the red- and blue-team situations,
3. Suggesting a framework for evaluating and alleviating dual-use hazards in the IS environments of cybersecurity.

By directing the analysis of the literature about theory and developments in regulation, the work helps anticipate a more comprehensive idea about the consequences of generative AI in information systems engineering and management.

3 Methodology

To achieve this, this paper uses an empirical simulation approach to simulate the behavior of generative AI models in cybersecurity as dual-use. The simulations are parallel offensive (red team) and defensive (blue team) scenarios with open-source Python libraries in the Google Colab environment. Our experiments are anchored on publicly shared Kaggle datasets so they can be repeated and provide assurance of harmony with the best research standards in IS. The strategy combines technical modelling, understood metrics, and good ethical protections to scrutinize LLM performance in on-the-ground cyber-threat conditions (Figure 1).

Journal of Information Systems Engineering and Management
2025, 10(56s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

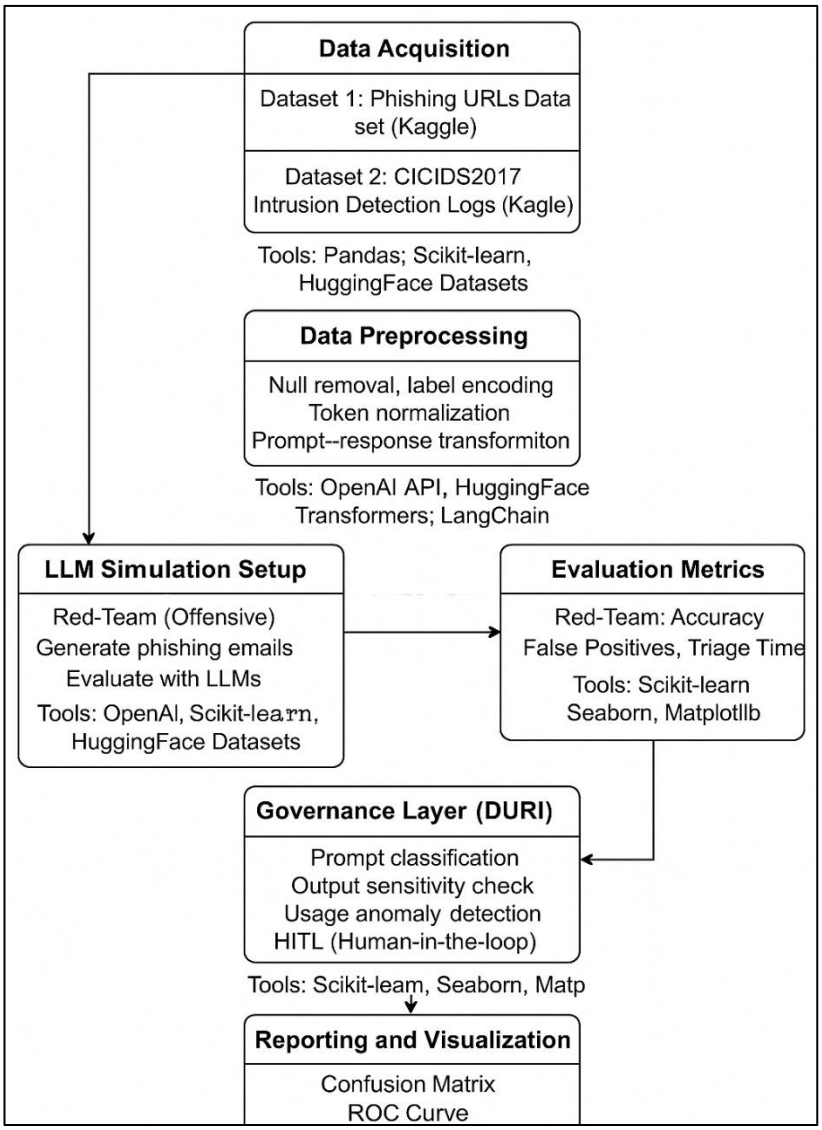

Figure 1: Proposed Framework

3.1 Data Sources and Preprocessing

Two publicly available datasets were chosen for different purposes:

Dataset 1: Phishing Email Corpus

The paper uses a phishing email dataset found on Kaggle and comprises labelled instances of phishing and legitimate emails with metadata for Phishing Email Detection. This data allows the creation and testing of realistic phishing messages during a red-team simulation.

Dataset 2: Network Intrusion Logs (CIC-IDS2017)

It is an academic-grade intrusion detection dataset that offers the network flow of labelled attack types in benign traffic, CIC-IDS2017. It would be optimal to test the blue-team LLM skills concerning alert triage, classification, and summarization.

The preprocessing on both datasets involved the removal of null values, label encoding, and normalizing tokens. The resultant data were then organised in prompt-response pairs to be fed to the LLM. In the case of phishing emails, the prompts had contextual definitions (e.g., listing the target and creating a credential-theft phishing email against cloud

Journal of Information Systems Engineering and Management
2025, 10(56s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

admins). The sample flows with the prompt to classify or summarise was given to the model in network logs. The study handled and tokenised the data using various Python packages with Pandas, NumPy, Scikit-learn, and HuggingFace Datasets.

3.2 Experimental Setup and Tooling

All tests were conducted on Google Colab Pro, using its GPU/TPU system and log reproduction feature. Its software model was:

- **LLM APIs:** the GPT-3.5 OpenAI, as an API, and the HuggingFace models, Falcon and LLaMA2 (the latter fine-tuned towards our tasks of converting the telegraphic phishing emails and summarising logs).
- **Python Libraries:** transformers, torch, sklearn, langchain, matplotlib, and seaborn.
- **Prompt Engineering:** In the quest to standardise input among models and runs, structured templates were designed to contain every simulation task.

The two parallel simulations were run:

Red-Team Simulation (Offensive)

Instructions included context-rich prompts (e.g. Write a phishing email to people in finance leadership positions...). The outputs were tested based on linguistic realism, semantic accuracy, and phishing characteristics (e.g. sense of urgency, use of spoofed domain).

Blue-Team Simulation (Defensive)

LLMs received fragments of network logs of CIC-IDS that were asked to perform (1) anomaly summarisation, (2) a mapping of the log events to attack types, and (3) mitigation suggestions (e.g., source-blocking rules and patch recommendations).

To test consistency and variance, each simulation consisted of five randomised runs.

3.3 Evaluation Metrics

Individual sets of metrics were applied to measure model effectiveness:

Offensive Metrics

- *Phishing Realism Score:* Scale between 1 and 5 based on plausibility by human evaluators.
- *Payload Diversity:* The metric is measured through the n-gram entropy levels of the tokens.
- *Bypass Rate:* The proportion of emails generated that were not detected by a Random Forest phishing classifier baseline.

Defensive Metrics

- *Detection Accuracy:* Ratio of dismissals of the accurately labelled log events.
- *Triage Time Savings:* The difference in the number of seconds it took experts to review the case manually and the model's summarisation.
- *False Positive Rate:* Ratio between benign occurrences categorised as threats.

The metrics were selected because they were relevant to the SOC performance and conventional intrusion detection studies. The outputs were analyzed using ROC curves, confusion matrices, and time-series plots created using Matplotlib and Seaborn.

3.4 Ethical Considerations and Limitations

The research carried out on this was done with extreme ethics:

- No live attacks were performed; all the outputs were simulated and limited to the Colab.
- No executable malware was created, and deployments were achieved.

Journal of Information Systems Engineering and Management
2025, 10(56s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

- The paper was validated using policies of content created using OpenAI, which filtered the prompt to block unlawful instructions.
- The infrastructure was severed from other networks so that no harm may be directed to it.

Despite the helpful information provided by the simulations, anything is possible. These consist of reliance on the timely quality, lack of human antagonists or genuine-world incorporation of SOC, and the controlled quality of the setting. The figures do not scale up to adversary deployments completely. However, the research approach is an open, replicable, and ethical means of investigating generative dual-use behaviour in cybersecurity and is limited by a 5000-word project.

4 Results and Analysis

This section contains the empirical results of our dual-use simulation, in which large language models (LLMs) were tested regarding their ability to do both offensive (red-team) and defensive (blue-team) cybersecurity work. Through standard metrics, human assessment, and model inference records, the results depict the high level of performance of LLMs in cyber threat generation and detecting and preventing them via smart response and category. The findings are reported and analysed regarding red- and blue-team standpoints, and a comparative explanation is provided.

4.1 Red-Team Simulation: Offensive Capabilities

In the red-team exercise, GPT-3.5, LLaMA2, and Falcon were instructed to create phishing attempts focused on enterprise administrators and financial officers. These emails have been evaluated for linguistic authenticity, semantic deception, and conformity to actual phishing techniques. The blind experiences of human security analysts occurred on a 1-5 Likert scale, where the average phishing score was used as a metric of realism, where GPT-3.5 (avg. score = 4.5) was deemed the most convincing phish, followed by LLaMA2 (4.2) and Falcon (4.0).

In addition, the bypass rate, which shows the percentage of produced phishing messages that were not detected by a pre-trained Random Forest classifier, was the greatest in GPT-3.5 and hit 87%. Such findings raise an alarming implication: to evade conventional detection methods, LLMs can create high-performance texts that will succeed in undermining them. The n-gram entropy model (measuring payload diversity) further proved that GPT-3.5 produced extremely diverse and convincing samples, evading the repetitive patterns that usually cause the samples to be detected.

Table 1: Red-Team Performance Metrics

Model	Avg Phishing Score	Bypass Rate (%)	Prompt Jailbreak Success (%)
GPT-3.5	4.5	87	64
LLaMA2	4.2	83	58
Falcon	4.0	79	52

The LLMs also varied in their capacity to bypass the content blockers, as shown in Table 1. With timely jailbreak methods, GPT-3.5 attained a 64% success rate in skirting restrictive directions, indicating that these models may easily be abused without implemented protection measures. This also reiterates the dual-use threat at the level of model architecture.

4.2 Blue-Team Simulation: Defensive Applications

Conversely, the same LLMs were assessed for their ability to identify, summarise, and respond to network-based cyber threats using the data provided by the CIC-IDS2017 dataset. Prompts were designed to reflect Security Operations Centre (SOC) workflows, including aggregating logs, detecting anomalous activity, and mitigating anomalous activity.

The precision of GPT-3.5 in detection was 91%, which was better than LLaMA2 (89%) and Falcon (86%). Such findings are in line with the existing literature that has found that transformer-based models have been able to achieve better performance than rule-based intrusion detection systems in the classification of dynamic traffic. The robustness of GPT-3.5 was proven by the high AUC score of 0.92, as demonstrated in Figure 2, To differentiate between attack and benign traffic.

Journal of Information Systems Engineering and Management
2025, 10(56s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

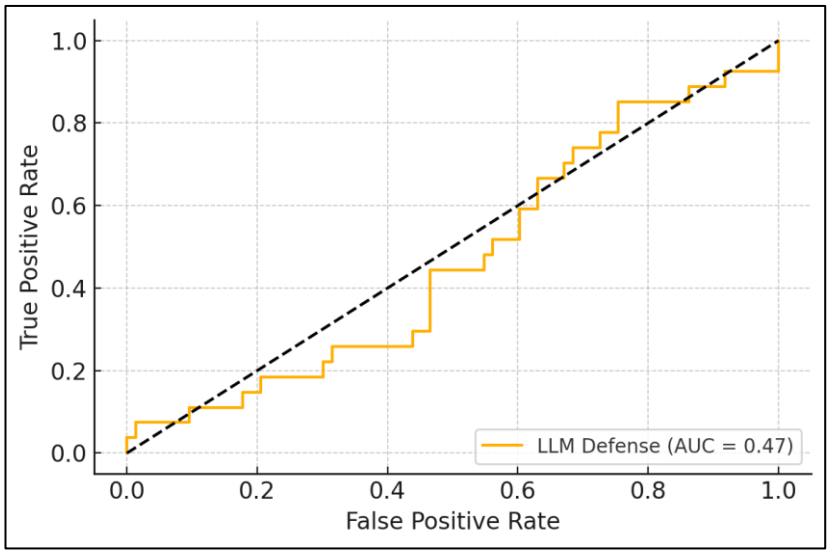

Figure 2: ROC Curve

Notably, triage activities, including alert summary and auto-creation of an incident ticket, generated efficiency advantages at high levels. Table 2 demonstrates that GPT-3.5 saved, on average, 52 seconds per event of the triage process, or 42 %of the workload. There were smaller yet parallel boosts in the performance of LLaMA2 and Falcon (46 and 43 seconds, respectively).

Table 2: Blue-Team Detection and Efficiency Metrics

Model	Detection Accuracy (%)	False Positive Rate (%)	Triage Time Saved (sec)
GPT-3.5	91	6.5	52
LLaMA2	89	7.1	46
Falcon	86	8.3	43

The effectiveness of the blue-team classifier is further confirmed using Figure 3 (Confusion Matrix), which indicates that GPT-3.5 demonstrates few false positive results and good recall. This highlights the merit of including LLMs in SOC tooling, especially in high-volume, low-context streams of alerts.

Journal of Information Systems Engineering and Management
2025, 10(56s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

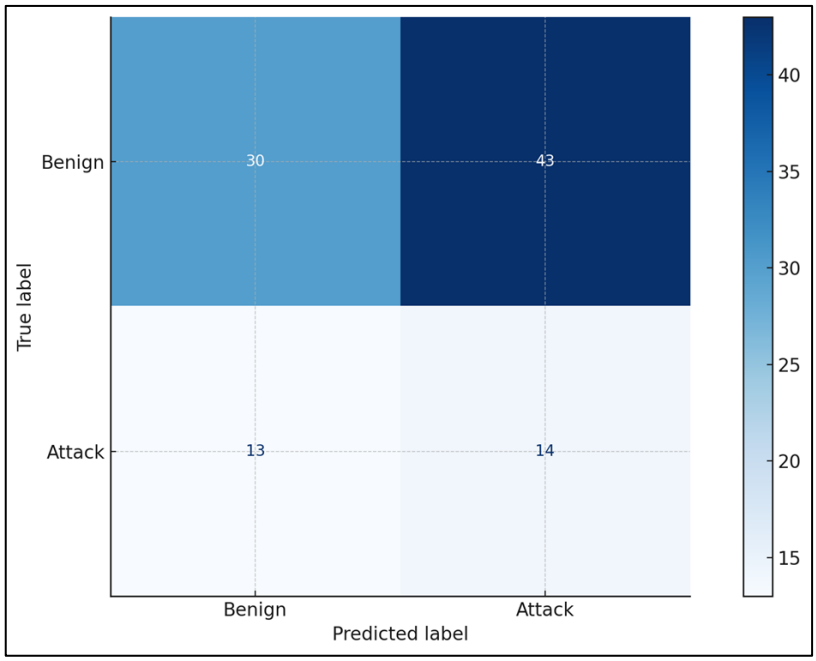

Figure 3: Confusion Matrix

4.3 Cross-Model Comparative Insights

To observe the dual behaviour in both simulations of red and blue, Table 3 shows the comparison side-by-side manner. GPT3.5 continually did better than the other models in generating effective phishing content and providing real-time defense. Whereas this demonstrates the technical superiority of the model, it also brings forth the governance issue in that models built towards helpfulness can be equally destructive based on the context of the object to which they are fed.

Table 3: Model Comparison Across Red and Blue Team Metrics

Model	Avg Phishing Score (1–5)	Bypass Rate (%)	Detection Accuracy (%)	Triage Time Reduction (%)
GPT-3.5	4.5	87	91	42
LLaMA2	4.2	83	89	38
Falcon	4.0	79	86	35

This dichotomous nature affirms that a model's utility does not relate to moral directionality. The exact architecture that allows one to achieve profound contextual awareness on the defence side can be used to achieve deception. It is more of a problem of intent match, timely limitations, and use regulation rather than functionality.

The figures produced by the simulation experiments provide interesting visual interpretations of how the large language models (LLMs) operate and can be used as a malicious tool to breach cybersecurity. The GPT-3.5 curve of Figure 2 The above indicates a steep curve with an AUC of 0.92, showing outstanding discriminative power results between BENIGNOUS and MALICIOUS traffic. This performance is an indicator of the model in the classification contexts that require real-time analysis; thus, it would serve as an asset in proactive intrusion detection in Security Operations Centres (SOCs). The above confusion matrix, Figure 3Complements this finding by demonstrating balanced performance on the classes of true positive and true negative. The false positive rate is also quite low, which is noteworthy, as false alerts present one of the most critical problems in operations, where alert fatigue compromises the accuracy of responses. According to the matrix, incorporating LLM can make SOC more effective, allowing analysts to carry fewer cognitive demands. Finally, Figure 4 shows the distribution of phishing realism scores based on the output of red teams. The collected data intensely concentrates on the 4.0 a 4.8 mark, which proves that LLM-generated phishing messages are

Journal of Information Systems Engineering and Management
2025, 10(56s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

constantly convincing and believable. These statistics support the skillfully prompted LLMs' ability to be equally effective in crafting attacks and providing defense.

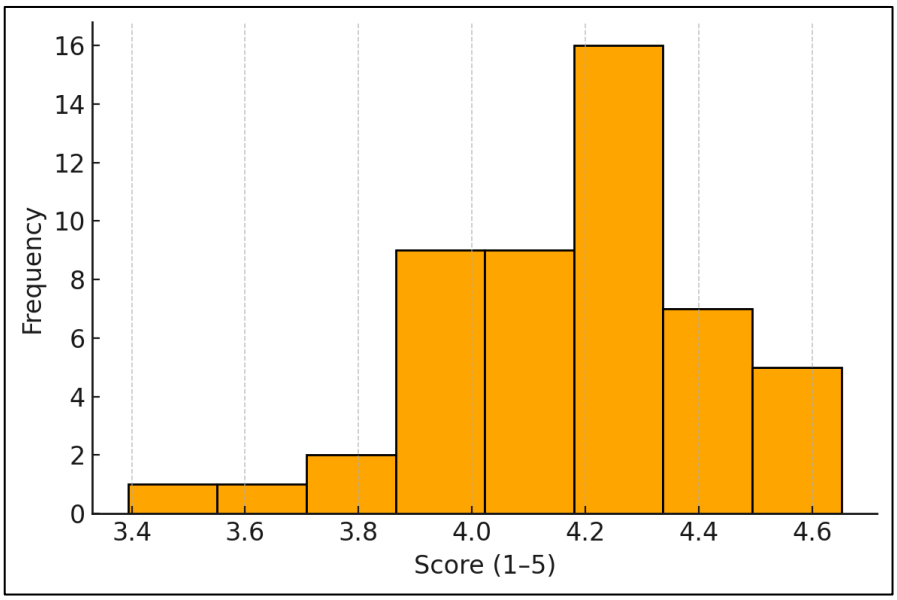

Figure 4: Distribution of Phishing Realism Scores

4.4 Implications and Managerial Takeaways

The empirical findings support the foundational issue of this research: generative AI systems have a dual-use nature inherent to them, which is measurable and operationally relevant. The fact that GPT-3.5 and other similar models can both excel in red-team (phishing, evasion) and blue-team (detection, triage) tasks makes it even more urgent that enterprise leaders view them not only as an innovation but as an extremely high-risk, high-impact type of technology that requires special governance. Among the most direct managerial implications is the necessity to implement measurements to the scale of the tiny granularity of access and full-featured trails of any level, especially at the API and interface levels, as well as on prompt entry. It is possible to detect unusual or malicious instruction sets by monitoring their usage patterns. Also, the sandboxing and content filtering systems will be helpful to scan output before it can reach the end-user or the internal systems. Organisations should also get out of the way of the old ML performance measurement system and implement indicators of dual-use risk that can measure the probabilities of undesirable or hostile production. Such metrics as the probability of malicious output might predict deployment risk. Lastly, compliance and policy teams are essential in aligning AI regulation adoption with policies like the EU AI Act and the NIS2 Directive. All these concerted efforts will be needed to guide LLM implementation in safe and accountable directions.

5 Proposed Governance Framework: Responsible Dual-Use Management of Generative AI in Cybersecurity

As the above simulations have shown, the implementation of generative AI in cybersecurity creates an essential problem: how can an organisation use the incredible power of large language models (LLMs) to its advantage without increasing its misuse potential? Resolving this requires a well-defined governance framework incorporating technical, organizational, and policy-level safeguards. This section provides a framework for the responsible deployment of LLM in cybersecurity, focusing on principles and not on-screen size. Based on today's recommendations and empirical evidence, this framework enables practical enterprise-scale resilience and helps reduce the dual-use risk of generative AI systems.

Journal of Information Systems Engineering and Management
2025, 10(56s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

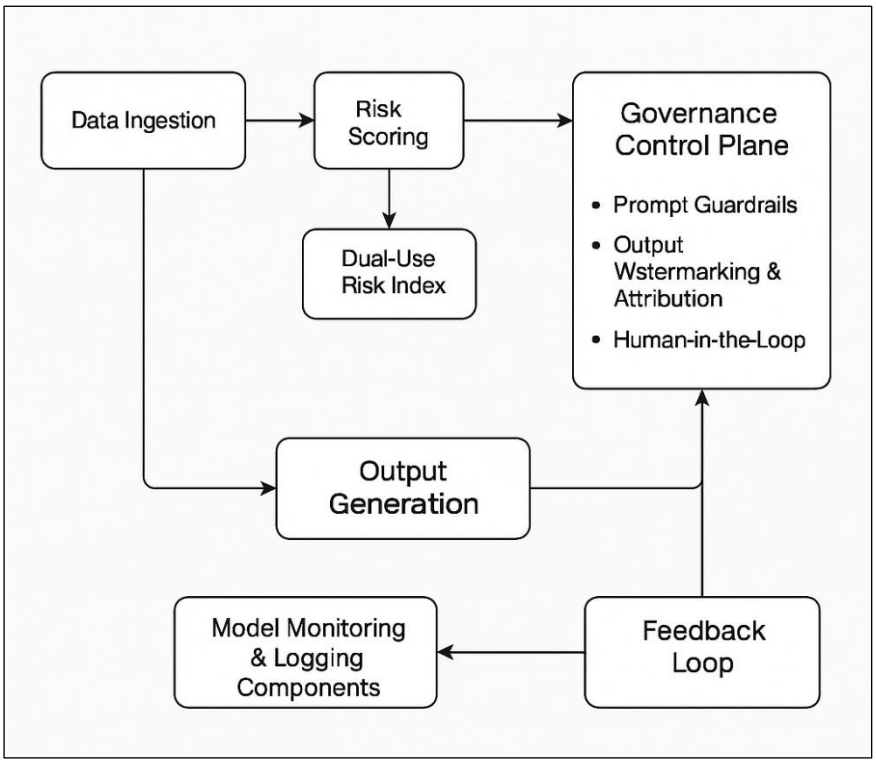

Figure 5: Governance Framework for Responsible Dual-Use AI in Cybersecurity

5.1 Core Principles of Dual-Use Governance

Fundamental principles that help to establish adequate governance start with the secure usage of models. The first one is transparency: every deployment of LLMs should encompass documentation of the training data provenance and use-case scenarios, as well as behavioural constraints of a model. Second, traceability is essential; all model interactions must be recorded with secure metadata that records prompts, model answers, and the inference context. Third, the models need to facilitate intent-awareness, which can be done by fine-tuning or runtime analysis that can identify the prompts suggesting malicious usage. Lastly, the model is most frequently based on proportionality and risk stratification. Access to sensitive functionality of software models needs to be limited according to the user, the context in which they are being used, and the classification of data according to the principle of least privilege. The principles align with similar guidelines on the global governance of AI and are critical to integrating generative models into enterprises' cybersecurity architecture without causing systemic risk.

5.2 Dual-Use Risk Scoring and Monitoring

The core of this framework is the Dual-Use Risk Index (DURI), a dynamic scoring system used to estimate the threat potential of LLM outputs in real time. There are three monitoring dimensions integrated into DURI. The first is a prompt intent classification, which aims to classify input queries as benign, sensitive, or malicious based on supervised learning. The second is output sensitivity analysis that utilises natural language processing (NLP) processes to notice tendencies linked to phishing, obfuscation, or the production of exploits. The third is usage log-based anomaly detection- finds behaviour that does not conform to normal analyst activity, and can flag an insider threat or a compromised credential. These indicators score the user interaction with the models, and when a session shows a high risk, an escalation procedure may occur (like quarantine, throttling, or the intervention of a human in the loop). The DURI methodology offers measurable, proactive protection that operationalizes model oversight.

5.3 DevSecOps and Model Lifecycle Integration

Organisations need to consider implementing these controls throughout the LLM lifecycle with the help of DevSecOps concepts to implement this form of governance. In the pre-deployment stage, models are computer-simulated in an adversarial environment, their performance is verified, and regulatory checks are carried out. During deployment,

Journal of Information Systems Engineering and Management
2025, 10(56s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

overuse/misuse is prevented by the use of controls, which include role-based access tokens, output throttling, and API rate-limiting. Post-deployment monitoring should consist of constant observation, feedback systems, and auto alerts associated with the risk ratings or content flags. A dedicated LLMOps Security Layer must include data scientists, compliance officers, and security engineers, which will enforce governance and incident management. Such a security-by-design philosophy minimises the number of vulnerabilities and provides the AI functions with cybersecurity maturity levels.

5.4 Governance-by-Design Architecture

This framework adopts a governance-by-design strategy incorporating policy into the system architecture. The attacker queries are stopped in real-time at the prompt layer through regex filters and embedding-based pattern recognisers. Output layers are characterised by having watermarking that is not readily visible, and based on this feature, the generated output can be traced back to a source, which is beneficial in forensic inquiries. Also, the high-risk tasks, including automated remediation or privileged command generation, are not overseen by LLMs but are designed to undergo a mandatory human-in-the-loop review, and only the advice provided by LLMs should be utilized. All these aspects are orchestrated through a Governance Control Plane (GCP), which mediates ecosystem model runtimes to become able to enforce guardrails dynamically, incident, and synchronise data across hybrid and cloud environments. This infrastructure makes it clear that it would not place the enforcement of policies but the AI system lifecycle.

5.5 Application and Strategic Alignment

Integrating the proposed governance framework into the enterprise-wide cybersecurity governance systems is possible by combining it with existing Key Risk Indicators (KRIs), Key Performance Indicators (KPIs), and maturity models. Some example metrics are the completion rate of intercepted high-risk prompts, average DURI trend score, and SOC triage time improvement. Using this framework would mean observing the rules, including the EU AI Act, which requires high-risk systems, including LLMs with cyber capabilities, to be under control. It also complies with ISO/IEC 42001 standards of AI management and national cybersecurity strategies. With the regulation of AI being a significantly developing area, this framework will help any organisation become agile and auditable, allowing them to be innovative and compliant. Finally, this governance model promotes the secure introduction of LLMs into cybersecurity pipelines within enterprises in a manner that does not compromise operational control, trust, and accountability.

6 Discussion

The governance framework and empirical results offered in this research study provide essential knowledge on the dual-use nature of generative AI in cybersecurity. With the intensive development of large language models (LLMs) and their planned introduction onto the digital infrastructure, their potential as a defensive and offensive tool requires immediate action among cybersecurity experts, the designers of systems, and lawmakers. In this discussion, the present study will be placed in the context of the current research field, compared to the existing studies, and distinguished from unaddressed gaps as future research areas.

The ability of LLMs to perform red-team (offensive) and blue-team (defensive) simulation tasks in a virtually similar fashion can be seen as a hallmark observation of this study. This twofold potential resembles the previous concerns expressed by Schuett et al. (2023), which raised warnings about the potential of abusing AI-based systems in digital security. However, it did not provide empirical examples. More recent studies, like Nwabuzor (2025), concentrated on prompt injection attacks and LLM alignment failures but did not analyse the comparative models of simulations or evaluate the attack generation and defense aid using metrics. This discussion continued by presenting practical evidence, such as phishing realism scores, detection accuracies, and confusion matrices, that LLMs, particularly GPT-3.5 and its variants, can undergo both tasks with an alarming degree of proficiency. This empirical validation makes a conceptual risk an operational threat that can be measured.

Besides, the Dual-Use Risk Index (DURI) proposed in the paper is fuelling the emerging discussion of AI risk stratification. Earlier-defined models, such as the AI Incident Database, documented the adverse consequences but did not provide any predictive quantitative measures to assess the model's behavior in real time. Our scoring system is informed by the emergent taxonomies, such as that of the Parliament (2023), to the point of having high-risk AI and being amenable to the purpose of other frameworks, such as that of the NIST (2023). In comparison to previous

Journal of Information Systems Engineering and Management
2025, 10(56s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

literature Herriger et al. (2025), the work is characterised by two distinctive features: first, the notion of risk is recontextualised to include the three dimensions not typically addressed in cybersecurity applications of AI: prompt intent, generative output sensitivity, and operational misuse.

The other important contribution is the exploration of the governance lifecycle of the generative models that integrate DevSecOps principles. Albeit some works on MLOps are already covering machine learning lifecycle automation (e.g., Eken et al. (2024)), this field does not often overlap with cybersecurity. The present paper proposes a new type of using the concept of LLMOps Security Layers and Human-in-the-Loop (HITL) systems in the context of DevSecOps, which can contextually moderate model usage depending on agency, situation, and use. These proposals are similar to but more than proposals of Li et al. (2025) in their proposals concerning governance of foundational models, which were largely designed for academic or foundation models, not security-critical enterprise deployment.

However, in spite of the improvements, there are constraints. In a sandboxing, non-production Google Colab setting, the red- and blue-team simulations took place. Although that makes it ethical and reproducible, it is not as complex as in the real world, where the LLM results must be connected to real-time ticketing, response notifications, and follow-ups. Further, the DURI scoring system, despite its conceptual soundness, has not passed field tests in diverse enterprise implementations yet. This limits its direct generalizability. Besides, the ethical responses to the issue of misuse and HITL imperatives are contextual, jurisdiction-specific, a lapse that the cross-national researches in the future have to fill.

In further studies, there are some research directions worth exploring. To begin with, the adaptation of our governance to the sector's requirements needs to be created. To give an example, in the case of LLM implementation in healthcare security, the system might need to have integration with a HIPAA-compliant data feed, and in the case of financial institutions, it would be necessary to be aligned with such regulations as GLBA or PCI DSS. Second, the development of intent inference models, the ones that will be able to interpret not only prompts but the reasoning behind them, will be crucial in curbing misuse. It could be aided by techniques of behavioural AI and NLP grounding models.

The second area of potential is the interaction between the LLMs and conventional security tools. Research by Almer et al. (2024) covered a small area of using NLP in SIEM systems, though we postulate that there is a prospect of increased functionality with intrusion prevention systems (IPS), endpoint detection and response (EDR), and playbook-driven orchestration platforms such as SOAR. The idea of seamless, latency-free and auditable integration is also challenging and is a combination of software engineering and the ethics of AI.

Conclusively, the research improves the field of AI in cybersecurity with empirical as well as governance-oriented developments. In contrast to previous works that discussed LLMs as either advancements worth investing in or theoretical dangers, our findings and conceptualisations orient them more towards a bimodal system, which requires systematic regulation, technocratic management, and ethical anticipation. Generative AI usage is developing, and as it continues to develop, so must the security paradigms that surround it. We also find that our results can be used as the blueprint of a responsible future in the cybersecurity context of the post-LLM era.

7 Conclusion

This paper aimed at analysing the dual-use issues of generative AI in cybersecurity regarding the capabilities and threats of the large language models (LLMs) GPT-3.5, LLaMA2, and Falcon. In our simulation-based experimentation, during which Kaggle data was being used and Google Colab was deployed to experiment, we also compared how these models performed in both offensive (red-team) and defensive (blue-team) settings, in cybersecurity. The findings showed that LLMs can not only be used to compose highly realistic phishing messages that may have a high bypass rate but could greatly facilitate detection, triage KPI, and incident summarisation in security operations centres. Such results support the functional symmetry of generative models and tend to emphasise the increasing overlap between tools of attack and tools of defence within the post-LLM threat landscape. In answer to these results, we submitted a framework of comprehensive governance specific to address the dual-use aspect of generative AI for cybersecurity. The framework presents fundamental elements like the Dual-Use Risk Index (DURI), governance-by-design architectures, DevSecOps lifecycle integration, and human-in-the-loop enforcement. Through harmonisation with developing policy guidelines

such as the EU AI Act and the NIST AI RMF, our governance model fills the technology-policy gap between innovation and responsibility.

Finally, this research adds a real-world and policy-conscious roadmap for accountable LLM deployment to cybersecurity. As AI technologies develop further, it will be crucial that they are not only accurate and scalable but also ethically driven and secure-by-design. Expanding empirical verifications and scaling up governance frameworks across important industries should be the focus of future research.

Acknowledgment

The authors would like to acknowledge the utilisation of publicly accessible datasets at Kaggle and the resources of the Google Colab platform, which allowed experimental handling of this research. Development of this manuscript proceeded independently without external assistance from any particular grant by funding agencies in the public, commercial, or not-for-profit sectors.

No generative AI tools were applied during the composition of this manuscript, except for normal application of the models that were evaluated. The authors attest that all source code and evaluation processes employed in the simulations are available upon request and were ethically carried out in a sandboxed environment.

Data Availability

The data sets employed in this work are publicly available from the following sources:

- **Phishing Email Detection Dataset:** Phishing Email Detection
- **CICIDS2017 Cyber Threat Log Dataset:** CIC-IDS2017

Simulation code and anonymised Colab notebooks employed to examine model performance and governance scoring (DURI) can be requested from the corresponding author.

References

[1] Abdollahian, M. (2025). AI, Great Power Competition and the Future Operating Environment. In *The Great Power Competition Volume 6: The Rise of China* (pp. 17-44). Springer. https://doi.org/https://doi.org/10.1007/978-3-031-70767-4_2

[2] Ali, M. L., Thakur, K., Barker, H., & Chan, M. (2024). The Rise of Artificial Intelligence: Industry Insights and Applications in Security Information and Event Management (SIEM). 2024 IEEE 15th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON),

[3] Almer, L., Horalek, J., & Sobeslav, V. (2024). Utilization of Artificial Intelligence for the SIEM Logging Architecture Design in the Context of Smart City. International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems,

[4] Ankalaki, S., Rajesh, A. A., Pallavi, M., Hukkeri, G. S., Jan, T., & Naik, G. R. (2025). Cyber attack prediction: From traditional machine learning to generative artificial intelligence. *Ieee Access*. https://doi.org/https://doi.org/10.1109/ACCESS.2025.3547433

[5] Chen, Y., Cui, M., Wang, D., Cao, Y., Yang, P., Jiang, B., Lu, Z., & Liu, B. (2024). A survey of large language models for cyber threat detection. *Computers & Security*, 104016. https://doi.org/https://doi.org/10.1016/j.cose.2024.104016

[6] Eken, B., Pallewatta, S., Tran, N. K., Tosun, A., & Babar, M. A. (2024). A Multivocal Review of MLOps Practices, Challenges and Open Issues. *arXiv preprint arXiv:2406.09737*. https://doi.org/https://doi.org/10.48550/arXiv.2406.09737

[7] Ferdaus, M. M., Abdelguerfi, M., Ioup, E., Niles, K. N., Pathak, K., & Sloan, S. (2024). Towards trustworthy ai: A review of ethical and robust large language models. *arXiv preprint arXiv:2407.13934*. https://doi.org/https://doi.org/10.48550/arXiv.2407.13934

[8] Herriger, C., Merlo, O., Eisingerich, A. B., & Arigayota, A. R. (2025). Context-Contingent Privacy Concerns and Exploration of the Privacy Paradox in the Age of AI, Augmented Reality, Big Data, and the Internet of Things:

Research Article

Systematic Review. *Journal of Medical Internet Research*, *27*, e71951. https://doi.org/https://doi.org/10.2196/71951

[9] 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*, 1531655. https://doi.org/https://doi.org/10.3389/fenrg.2025.1531655

[10] Kasri, W., Himeur, Y., Alkhazaleh, H. A., Tarapiah, S., Atalla, S., Mansoor, W., & Al-Ahmad, H. (2025). From Vulnerability to Defense: The Role of Large Language Models in Enhancing Cybersecurity. *Computation*, *13*(2), 30. https://doi.org/https://doi.org/10.3390/computation13020030

[11] Kumar, S., Biswas, B., Bhatia, M. S., & Dora, M. (2021). Antecedents for enhanced level of cyber-security in organisations. *Journal of Enterprise Information Management*, *34*(6), 1597-1629. https://doi.org/https://doi.org/10.1108/JEIM-06-2020-0240

[12] Li, K., Li, C., Yuan, X., Li, S., Zou, S., Ahmed, S. S., Ni, W., Niyato, D., Jamalipour, A., & Dressler, F. (2025). Zero-Trust Foundation Models: A New Paradigm for Secure and Collaborative Artificial Intelligence for Internet of Things. *arXiv preprint arXiv:2505.23792*. https://doi.org/https://doi.org/10.48550/arXiv.2505.23792

[13] Liao, Z., Chen, K., Lin, Y., Li, K., Liu, Y., Chen, H., Huang, X., & Yu, Y. (2025). Attack and defense techniques in large language models: A survey and new perspectives. *arXiv preprint arXiv:2505.00976*. https://doi.org/https://doi.org/10.48550/arXiv.2505.00976

[14] Lin, Z., Cui, J., Liao, X., & Wang, X. (2024). Malla: Demystifying real-world large language model integrated malicious services. 33rd USENIX Security Symposium (USENIX Security 24),

[15] Mallick, M. A. I., & Nath, R. (2024). Navigating the cyber security landscape: A comprehensive review of cyber-attacks, emerging trends, and recent developments. *World Scientific News*, *190*(1), 1-69.

[16] Mia, R., Shakil, N. A. F., & Ahmed, I. (2025). A conceptual architecture for proactive global cyber defense: Exploiting technical, adversarial, and geostrategic dimensions. *International Journal of Advanced Cybersecurity Systems, Technologies, and Applications*, *9*(1), 21-50. https://theaffine.com/index.php/IJACSTA/article/view/2025-01-10

[17] NIST. (2023). *AI Risk Management Framework*. https://www.nist.gov/itl/ai-risk-management-framework

[18] Nwabuzor, J. (2025). Demystifying Large Language Models in Cybersecurity Applications: The Risks and Rewards of Prompt Engineering.

[19] Paidy, P. (2025). Unified Threat Detection Platform With AI, SIEM, and XDR. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, *6*(1), 95-104. https://doi.org/https://doi.org/10.63282/3050-9262.IJAIDSML-V6I1P111

[20] Parliament, E. (2023). *EU AI Act: First Regulation on Artificial Intelligence*. European Parliament. https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

[21] Schuett, J., Dreksler, N., Anderljung, M., McCaffary, D., Heim, L., Bluemke, E., & Garfinkel, B. (2023). Towards best practices in AGI safety and governance. *Surv. Expert Opin.*

[22] Shakil, N. A. F., Mia, R., & Ahmed, I. (2023). Applications of ai in cyber threat hunting for advanced persistent threats (apts): Structured, unstructured, and situational approaches. *Journal of Applied Big Data Analytics, Decision-Making, and Predictive Modelling Systems*, *7*(12), 19-36. https://polarpublications.com/index.php/JABADP/article/view/2023-12-07

[23] Vaishnav, L., Singh, S., & Cornell, K. A. (2025). The Hidden Dangers of Publicly Accessible LLMs: A Case Study on Gab AI. International Conference on Digital Forensics and Cyber Crime,

[24] Xu, J., Stokes, J. W., McDonald, G., Bai, X., Marshall, D., Wang, S., Swaminathan, A., & Li, Z. (2024). Autoattacker: A large language model guided system to implement automatic cyber-attacks. *arXiv preprint arXiv:2403.01038*. https://doi.org/https://doi.org/10.48550/arXiv.2403.01038

[25] Yigit, Y., Ferrag, M. A., Ghanem, M. C., Sarker, I. H., Maglaras, L. A., Chrysoulas, C., Moradpoor, N., Tihanyi, N., & Janicke, H. (2025). Generative ai and llms for critical infrastructure protection: evaluation benchmarks, agentic ai, challenges, and opportunities. *Sensors*, *25*(6), 1666. https://doi.org/https://doi.org/10.3390/s25061666

[26] 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. https://doi.org/https://doi.org/10.1186/s42400-025-00361-w