Leveraging Generative AI Models for Multidomain Network Security

Rasa Brūzgienė[1][a], Šarūnas Grigaliūnas[2][b], Ilona Veitaitė[1][c], Renata Danielienė[1][d], Kęstutis Driaunys[1][e], Paulius Astromskis[1][f], Živilė Nemickienė[1][g], Dovilė Vengalienė[1][h], Rokas Stankūnas[1][i], Ieva Andrijauskaitė[3][j] and Neringa Gaubienė[3][k]

[1]*Kaunas Faculty, Vilnius University, Muitinės Str. 8, 44280 Kaunas, Lithuania*

[2]*Department of Computer Sciences, Kaunas University of Technology, Studentu Str. 50, 51368 Kaunas, Lithuania*

[3]*Faculty of Law, Vilnius University, Saulėtekio al. 9, 10222 Vilnius, Lithuania*

Keywords: GenAI, Network Security, TechSec, OpSec.

Abstract: The rise of Generative Artificial Intelligence (GenAI) presents new opportunities for enhancing network security across technical, operational, human, and physical domains. This paper proposes a GenAI-driven network security framework that integrates OSI-layer–aware threat analysis with mandatory human oversight to support CISO decision-making. The framework is empirically evaluated using NetFlow datasets representing port scanning, ICMP flooding, and SPAM attacks. Six GenAI models are assessed using explainability index, hallucination rate, and token efficiency metrics. The results demonstrate that while certain models achieve high analytical accuracy and explainability, performance varies significantly in efficiency and hallucination behavior. The study further discusses legal and regulatory implications of deploying GenAI in security-critical environments, highlighting the necessity of human-in-the-loop control for accountable and reliable network security operations.

1 INTRODUCTION

The contemporary cyber-threat landscape is characterized by rapidly increasing scale, complexity, and impact, with attacks such as ransomware and distributed denial-of-service posing growing risks to organizational and critical infrastructure security (Cloudflare, Inc., 2025). The European Commission (EC) has acted on these trends by requiring critical infrastructure organizations to implement the NIS2 (Network and Information Systems) Directive to reduce incidents. In today's context of cyber threats, the Chief Information Security Officer (CISO) (European Parliament and Council of the European Union, 2022) has a unique responsibility, especially in view of the multidimensional nature of network security (encompassing technical (TechSec), operational (OpSec), human (HumSec) and physical (PhySec) aspects). In this context, the emergence of Generative Artificial Intelligence (GenAI) technology, with its capabilities to process natural language, generate various types of content and analyze complex data, opens new opportunities for the CISO and security teams to strengthen network security, subject to organizational context and risk-based assessment (Wen, 2024b).

Advanced machine learning algorithms can identify patterns and predict potential threats in real-time, allowing organizations to respond both proactively and reactively, and significantly reducing the time needed to respond to vulnerabilities (Wen, 2024a). AI systems use advanced machine learning technologies to detect anomalies and potential threats, improving the accuracy of threat detection by learning from historical data and evolving with new attack strategies. Integrating human intelligence into AI models through human-participation systems can improve

[a] https://orcid.org/0000-0002-0816-8700
[b] https://orcid.org/0000-0001-9268-9244
[c] https://orcid.org/0000-0001-9046-0788
[d] https://orcid.org/0000-0003-3308-0919
[e] https://orcid.org/0000-0002-8456-123X
[f] https://orcid.org/0000-0003-2205-3165
[g] https://orcid.org/0000-0002-4857-0112
[h] https://orcid.org/0000-0002-9113-9301
[i] https://orcid.org/0009-0002-0893-5452
[j] https://orcid.org/0009-0007-6315-0427
[k] https://orcid.org/0009-0002-6756-2246

Brūzgienė, R., Grigaliūnas, Š., Veitaitė, I., Danielienė, R., Driaunys, K., Astromskis, P., Nemickienė, Ž., Vengalienė, D., Stankūnas, R., Andrijauskaitė, I. and Gaubienė, N.
Leveraging Generative AI Models for Multidomain Network Security.
DOI: 10.5220/0013948100004052

In *Proceedings of the 18th International Conference on Agents and Artificial Intelligence (ICAART 2026) - Volume 5*, pages 3909-3916
ISBN: 978-989-758-796-2; ISSN: 2184-433X

decision-making processes in areas of threat detection and incident management. The application of GenAI in critical infrastructure organizations must also consider current challenges, as it can be difficult to assess the reliability of decisions made and the potential for hallucinations, as the model operates in a "black box" approach. In addition, the deployment of GenAI models in safety-critical environments raises legal and ethical issues.

This paper proposes a framework that integrates GenAI-driven network security with human oversight, grounded in a layered network security analysis method. The approach leverages domain-specific reasoning, contextual threat interpretation, and compliance with legal and regulatory standards to ensure both technical and ethical robustness. The proposed framework is evaluated using a dataset consisting of security domain-specific questions and NetFlow traffic data, enabling realistic and targeted testing of GenAI performance in various cyber threat detection. The experimental evaluation focuses on key metrics including accuracy, explainability, resource usage, and hallucination rate, to comprehensively assess both the effectiveness and efficiency of GenAI models. By analyzing the behavior of these models, the authors contribute to a deeper understanding of GenAI's role in cybersecurity, particularly focusing on network security. It also addresses the legal and ethical implications of the deployment of GenAI. Ultimately, the proposed framework emphasizes the importance of combining automated intelligence with human oversight to ensure accountability, reliability, and the responsible development of future cyber defense systems.

The novelty of this work lies in the empirical evaluation of multiple state-of-the-art GenAI models within a unified multidomain network security framework. Unlike prior studies that focus on single security domains or conceptual architectures, this paper combines Technical, Operational, Human, and Physical security perspectives with OSI-layer–aware threat analysis. Furthermore, the study introduces a hallucination-aware evaluation methodology and assesses both analytical effectiveness and efficiency of GenAI models using real NetFlow-based cyberattack scenarios, complemented by legal and regulatory considerations relevant to CISO-level decision-making.

The paper is structured as follows: after the Introduction (section 1), a review of related works is presented (Section 2). The Section 3 details the proposed GenAI-driven network security framework. The Section 4 presents the experimental use cases and analyses the evaluation results. The Section 5 is devoted to the summary of the GenAI models' behavior and discussion on GenAI as the "must-have" tool in the CISO toolbox. Finally, the Section 6 concludes the paper with a summary of the main obtained results of this work and a discussion of future research directions.

2 RELATED WORKS

Recent studies explore the use of Generative AI and large language models in network security for threat detection, incident analysis, and decision support, highlighting their potential for explainable and context-aware security operations. A key trend in this literature is the emphasis on using LLMs not only for detecting threats but also for enhancing human oversight and operational effectiveness. Several works introduce the concept of chain-of-thought prompting, rationale logging, and "ask-the-LLM" interfaces to make model outputs more transparent and actionable for security analysts (Houssel et al., 2024; Balasubramanian et al., 2023). This is complemented by metrics and frameworks proposed for evaluating the coherence and domain alignment of explanations provided by LLMs, specifically focusing on their ability to mirror domain-specific rules and support compliance requirements.

The integration of up-to-date threat intelligence is commonly addressed through Retrieval-Augmented Generation (RAG) techniques and, more recently, knowledge graphs and GraphRAG frameworks. These approaches aim to overcome the limitations of static LLMs by grounding model reasoning in real-time data such as Common Vulnerabilities and Exposures (CVE) feeds, Adversarial Tactics, Techniques, and Common Knowledge (MITRE ATT&CK) mappings, and dynamically generated ontologies (Kasri et al., 2025). Some papers also discuss multi-modal data integration, addressing the challenges of combining different types of operational, technological, and physical context within a unified security analysis pipeline (Bui et al., 2024).

Despite the proliferation of conceptual frameworks and pilot deployments, there is a notable shortage of comprehensive empirical validation. Few works perform large-scale, real-world evaluations or fully assess regulatory compliance, with most studies relying on prototypes, case studies, or benchmarks using limited datasets (Houssel et al., 2024). Comprehensive surveys, such as those by Motlagh *et al.* and Ghanem *et al.*, provide taxonomies and highlight research gaps, particularly around standardizing evaluation metrics, adversarial robustness, and real-world regulatory mapping (Ghanem and Ali, 2025; Motlagh

et al., 2024).

In contrast to existing studies, this work provides a multidomain, empirical comparison of multiple GenAI models for network security, incorporating hallucination analysis, efficiency metrics, and regulatory context within a single evaluation framework.

3 GenAI-DRIVEN NETWORK SECURITY FRAMEWORK

The proposed framework that is visually conceptualized in Fig. 1 shows how GenAI models work with security teams in a comprehensive and structured interaction between network protocol layers, AI capabilities, human security roles and functional security domains. This framework provides a systemic approach to enhancing cybersecurity through structured AI–human collaboration. Its architecture is centered on the interaction between GenAI models and the Chief Information Security Officer (CISO), who is responsible for human oversight, policy definition, and regulatory compliance. This central coordination layer interfaces with three core components: network infrastructure layers, security function domains, and GenAI evaluation metrics. The framework incorporates the OSI model layers (L1–L7), representing the hierarchical network stack from physical hardware to application-level services. These layers are semantically mapped to GenAI reasoning processes, enabling automated yet context-aware analysis across different protocol abstractions. The experimental analysis focused on OSI layers L3, L4, and L7; support for additional OSI layers is conceptual and was not empirically validated in this study. While GenAI provides detailed analytical insights, all final security decisions remain under human control.

The proposed framework delineates four primary security domains (Technical Security (TechSec), Operational Security (OpSec), Human Security (HumSec) and Physical Security (PhySec)) each of which focuses on different security areas, that is, network security through threat detection, NetFlow analysis, vulnerability assessment, compliance monitoring and *etc*. GenAI models contribute to both threat detection (e.g., DDoS, port scanning, spam, malware) and anomaly detection (e.g., behavioral deviations, statistical outliers, machine learning-based predictions). The effectiveness and efficiency of the GenAI functions are evaluated against three key performance metrics:

- explainability index (EI) that measures the clarity C_s, relevance R_s, and explicitness of GenAI-generated insights E_s where each component is scored on a normalized scale based on predefined qualitative criteria assessing comprehensibility, task alignment, and reasoning transparency:

$$EI = \frac{C_s + R_s + E_s}{3} \tag{1}$$

- hallucination rate (HR) that quantifies the frequency of incorrect or fabricated outputs N_h in relation to the total outputs N_t, where a hallucination is defined as a factual claim or inference not supported by the provided data or established domain knowledge:

$$HR = \frac{N_h}{N_t} \tag{2}$$

- token efficiency (TE) that assesses the ratio of useful insight to token consumption, linking computational cost to analytical value:

$$TE = \frac{I_s}{T_c}, \tag{3}$$

where I_s is human-assigned insight score, T_c is token count.

Human-in-the-Loop mechanism is embedded within the process, ensuring that AI-generated outputs are subject to human validation, thereby mitigating risks of hallucinations and reinforcing accountability. The main principle is that AI enhances human decision-making in cybersecurity rather than replacing human judgment entirely.

The proposed framework accounts for various threat types - including port scanning, ICMP flooding, email spam, ransomware, data breaches, and DDoS attacks - demonstrating the applicability of GenAI across diverse threat vectors and emphasizing the multidomain nature of modern cybersecurity operations.This structured interaction between network layers, security domains, GenAI analysis, and human oversight enables consistent, explainable, and legally accountable decision-making in complex network security environments.

4 EXPERIMENTAL EVALUATION AND ANALYSIS OF THE RESULTS

To assess the practical utility and performance of state-of-the-art GenAI models in cybersecurity, we conducted a comprehensive experimental evaluation using six leading GenAI systems: GPT-4o, Claude Sonnet 4, Gemini 2.5 Flash, Grok 3, Microsoft Copilot, and DeepSeek R1. The selected GenAI models were chosen based on their state-of-the-art

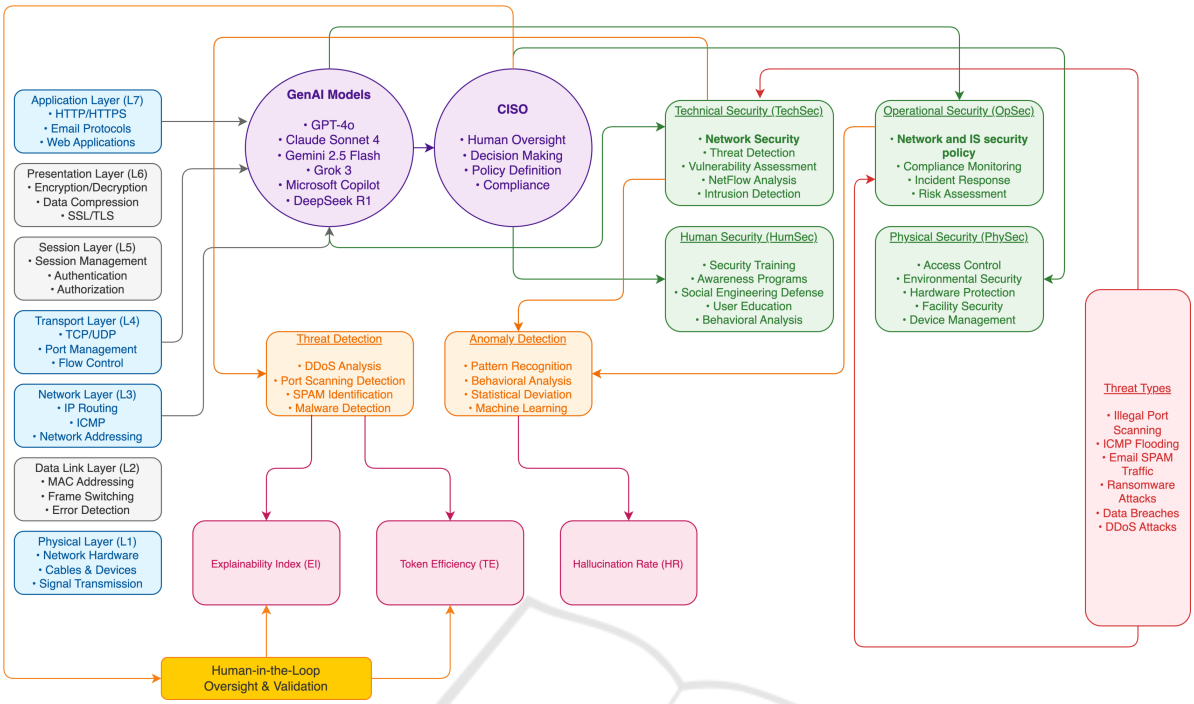

Figure 1: Concept of the proposed GenAI-driven network security framework.

performance, diverse architectural foundations, and broad integration capabilities, ensuring a representative evaluation of current GenAI capabilities across operational, technical, human, and physical security domains. The evaluation was designed to capture both the functional effectiveness and operational efficiency of these models in network security applications.

The experiments were structured into two complementary parts:

- multidomain assessment of GenAI functional effectiveness, that was focused on evaluating the models' functional capabilities in network security across four security domains;
- GenAI efficiency analysis via cyberattack detection during which each model was evaluated on its ability to detect and interpret cyber threats using structured NetFlow datasets representing various attack types.

This dual-layered experimental design enables a nuanced comparison of generative models both in terms of security domain-specific effectiveness and technical performance in active threat environments.

4.1 Multidomain Assessment of GenAI Functional Effectiveness

Experimental efforts were conducted to evaluate how effectively GenAI can support the responsibilities of an IT security professional, such as a CISO, with a primary focus on the Technical (TechSec) and Operational (OpSec) security domains.The Human Security (HumSec) and Physical Security (PhySec) domains are included in the framework design but were not empirically evaluated in the current study and are considered directions for future work. The research examined selected GenAI models by measuring their capabilities in responding to network security-related tasks such as data analysis, anomaly detection, and compliance with regulations. The functional capabilities and performance of the GenAI model were evaluated for each network security related category based on the following criteria:

- ability to analyze files in various formats - i.e., Can it process a .pcap or .log file to identify network security breaches or incidents?;
- accuracy of analysis - i.e., Can it identify key network events from a network traffic file? Can it assess the organization's network security compliance with legal regulations and the standards on which the organization's security policy is based on?;
- extraction of structured information - i.e., Can it convert network traffic data into a structured table? Can it generate a statistical report on the changes and updates to the network security policy?;

- search and filtering capabilities - i.e., Can it filter network records by IP addresses, ports, or time? Can it find the provisions for the secure use of email?;
- speed and performance - i.e., What is the model's analysis speed (events per second) for large network traffic files? Can it process multiple network traffic files simultaneously?;
- anomaly detection - i.e., Can it detect potential intrusions from network traffic data? Can it determine the time period during which unusual network activity occurred? Can it detect security policies that have not been updated for more than 18 months?;
- data visualization - i.e., Can it generate visualizations of network activity? Can it generate a visualization of the network security policy's compliance with legal regulations and standards?

A systematic evaluation framework was established, grounded in Lithuanian and EU cybersecurity laws, to define assessment benchmarks and domain-specific needs. The models were subjected to repeated testing using provided security policy files and standardized prompts, with expert reviewers assessing their outputs for precision, consistency, and contextual appropriateness. Token consumption was analyzed to gauge token efficiency, highlighting that a higher number of tokens used does not necessarily indicate superior performance. To account for response quality, human insight scores (on a 1–10 scale) were assigned using predefined criteria evaluating relevance to the task, technical correctness, and practical usefulness of each model's output. Token efficiency was calculated as the ratio of the final agreed insight score to token count (Table 1). Within the broader context of network security, specific aspects of the operational and technical security domains were examined in greater detail. In particular, the network and information systems security policy falls under the OpSec domain, while network security pertains to the TechSec domain.

In the OpSec category, Claude demonstrated the highest token efficiency (TE = 0.00431), indicating that it delivered a relatively high level of insight with minimal token expenditure. This performance was followed by Microsoft Security Copilot (TE = 0.00230) and Grok (TE = 0.00125), both of which balanced insightful content with moderate token use. Despite producing highly insightful responses, GPT-4o (TE = 0.00057) and Gemini (TE = 0.00027) were penalized by high token usage, reflecting lower efficiency. DeepSeek, with a TE of 0.00665, also showed strong performance relative to its lightweight nature.

Table 1: Tokens used by GenAI models and human-assigned insight scores.

GenAI model	OpSec: Net & IS security policy		TechSec: Net security	
	Tokens	Human score	Tokens	Human score
GPT-4o	10586	6	3993	7
Claude	1391	6	1399	6
Gemini	10970	3	15856	3
Microsoft Copilot	3908	9	3012	8
Grok	8019	10	5433	10
DeepSeek	1053	7	950	2

In the TechSec domain, Claude once again emerged as the most efficient (TE = 0.00429), closely followed by Microsoft Security Copilot (TE = 0.00266) and Grok (TE = 0.00184). GPT-4o improved its relative position compared to OpSec, scoring TE = 0.00175. Notably, Gemini yielded the lowest TE (0.00019), suggesting a poor tradeoff between insight and token consumption despite its high output length. DeepSeek, although extremely low on token use, scored poorly (TE = 0.00211) due to a low insight rating of 2 in TechSec tasks.

Although Claude frequently acknowledged its limitations, it consistently demonstrated awareness of its functional scope. In contrast, Microsoft Security Copilot provided direct and specific responses, often addressing tasks with clear functional assertions. GPT-4o typically employed structured reasoning, presenting methodical problem-solving strategies and elaborating on its analytical process. Grok distinguished itself by delivering in-depth analyses of the provided inputs and offering actionable, well-defined remediation strategies. Gemini, while generally cautious, tended to request additional context and responded with procedural, step-by-step guidance tailored to each query. Conversely, DeepSeek produced brief responses that, although concise, often lacked sufficient informational depth or specificity.

4.2 GenAI Efficiency Analysis by Detection of Cyber Threat

GenAI models were evaluated for their efficiency in detecting and interpreting cyber threats within structured NetFlow traffic generated in a controlled laboratory environment. Labeled NetFlow datasets representing three distinct types of cyber threats: illegal port scanning, ICMP flooding, and email SPAM traffic. Each model analyzed these datasets to identify malicious patterns over OSI layers and infer the nature of the threats based on flow attributes such as

IP addresses, ports, protocols, timestamps, and traffic volumes. The goal was to benchmark the models' capabilities in a realistic network security context using several thousand NetFlow records per attack scenario, where explainability and analytical precision are critical.

Experimental testing began with the development of a structured prompt designed to support the systematic analysis of cyberattack scenarios. The prompt formulation process involved defining the topical scope of the prompt and a series of testing protocols. The generated outputs were evaluated at the question-group level for technical accuracy, linguistic coherence, and suitability for academic discourse, with hallucinations recorded per evaluation group. Where configurable, GenAI model temperature settings were kept at default values to minimize variability in response generation across models. The designed prompt included questions divided into four groups:

- Group 1: threat detection (What anomalous patterns or deviations from typical traffic behaviour can be observed in the NetFlow dataset? Does the data reveal indicators of known attack types? Which source or destination IP addresses display irregular or suspicious activity across multiple flows? Is there evidence of concentrated high-volume traffic from a single source?);
- Group 2: attribute-based threat indicators (Which flow attributes correlate strongly with potentially malicious activity? Can suspicious flows be characterised by traffic directionality, payload size, or TCP flag behaviour? Based on flow patterns, which IP addresses are most likely acting as attackers and which as victims? What services or ports are repeatedly targeted, and what does this suggest about the attacker's intent or tactics?);
- Group 3: target profiling and exposure (Which internal hosts or network segments are receiving the majority of suspicious or unsolicited traffic? Does the behaviour suggest targeting of specific protocols, services or system types? Are any devices being systematically scanned across a range of ports or IPs?);
- Group 4: timeline and temporal analysis (When does the anomalous or malicious activity begin and end, based on flow timestamps? Can the flow data be used to reconstruct a sequence of attack phases? Are there recurring time-based patterns that indicate automation or scheduled tasks? How does the behaviour of suspicious sources change over time - do they escalate, shift targets, or alter tactics?)

Previous studies on NetFlow-based intrusion detection using traditional machine-learning and signature-based approaches report high detection accuracy but limited explainability and adaptability to evolving attack semantics. The evaluation of illegal scanning detection tasks (Fig. 2) revealed that Grok 3 achieved the highest average effectiveness index score of 0.984, and similarly Claude (0.978), while also maintaining a zero hallucination rate, indicating both accuracy and reliability in threat interpretation.

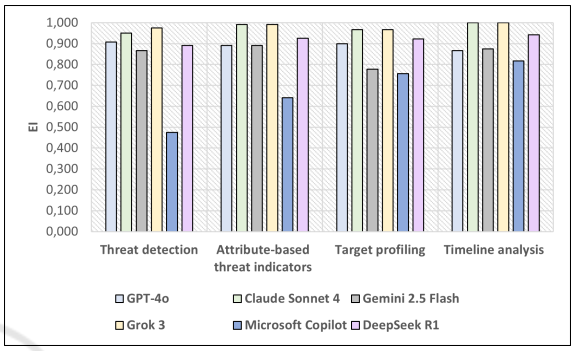

Figure 2: Explainability index in detecting illegal scanning of a network.

In contrast, Gemini 2.5 Flash showed a lower EI of 0.858 and introduced hallucinations in Group 3 (0.33) and Group 4 (0.25), reflecting occasional misinterpretations (Fig. 3).

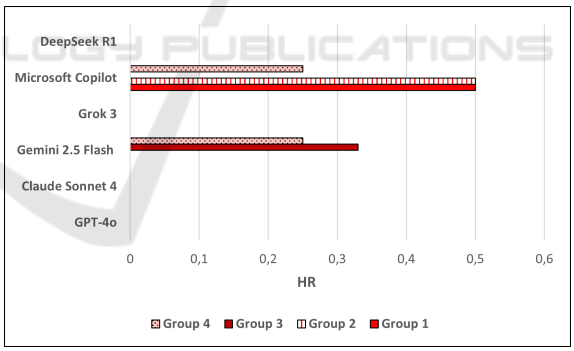

Figure 3: Hallucination rate in detecting illegal scanning.

Microsoft Copilot performed poorly with an EI of 0.667, but exhibited hallucinations in Groups 1 and 2 (0.5) and Group 4 (0.25), suggesting a trade-off between analytical depth and factual consistency.

In the ICMP flooding detection task (Fig. 4), Claude Sonnet 4 achieved the highest EI of 1, (similarly lower results of EI for GPT-4o, and Grok 3) along with a zero hallucination rate, demonstrating both accuracy and reliability.

In contrast, DeepSeek R1 showed a significantly lower EI of 0.529 and a high hallucination rate of up to 0.75 in multiple groups (Fig. 5). Microsoft Copilot

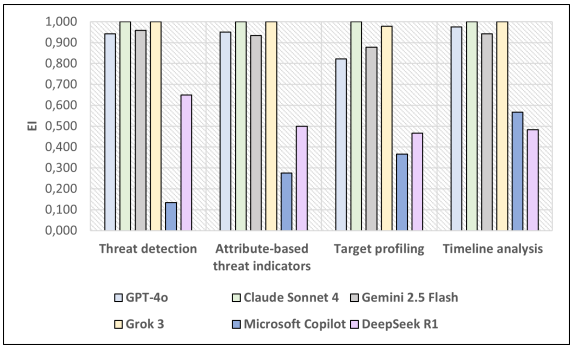

Figure 4: Explainability index in detecting ICMP flooding.

also exhibited lower performance with an EI of 0.333 and hallucination rates ranging from 0.5 to 0.75, suggesting occasional inconsistencies in its analytical responses.

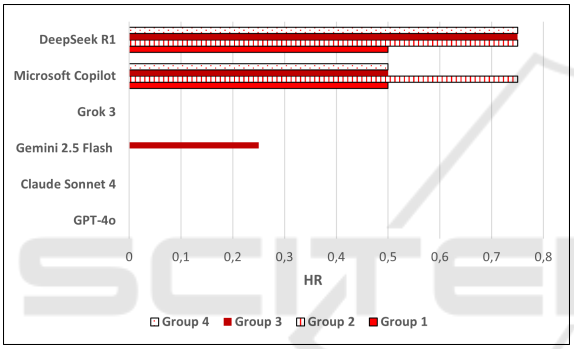

Figure 5: Hallucination rate in detecting ICMP flooding.

In the SPAM detection task (see Fig. 6,7), Claude Sonnet 4 achieved a perfect EI of 1.0 with a zero hallucination rate, while Grok 3 followed closely with an EI of 0.987, also maintaining complete factual consistency.

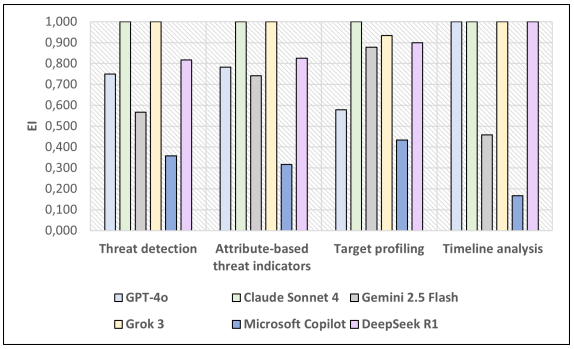

Figure 6: Explainability index in detecting SPAM traffic.

DeepSeek R1 performed well with an EI of 0.884 and no hallucinations, whereas GPT-4o scored a lower EI of 0.791 and exhibited a hallucination in Group 4 (0.25). Gemini 2.5 Flash and Microsoft

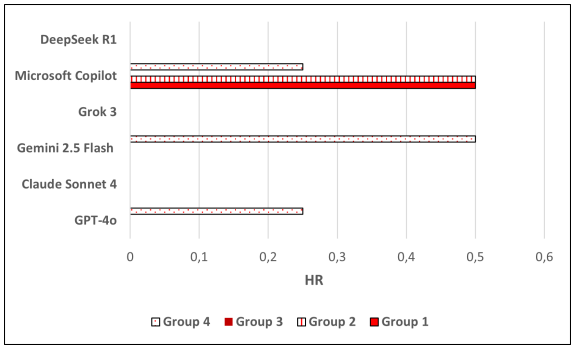

Figure 7: Hallucination rate in detecting SPAM traffic.

Copilot demonstrated the weakest performance, with EIs of 0.647 and 0.311 respectively, alongside moderate hallucination rates across several groups.

The experimental results demonstrate that Claude Sonnet 4 and Grok 3 can effectively detect and explain cyber threats across multiple security domains with high accuracy and low hallucination rates, validating the framework's potential for enhancing multidomain network security through AI-human collaboration.

5 DISCUSSION ON GenAI MODELS' BEHAVIOR AND COMPLIANCE TO LEGISLATION

The evaluation of six GenAI models revealed nuanced behavior patterns with respect to accuracy and alignment with CISO-level operational demands. While the models demonstrated strong capabilities in identifying cyber threat patterns and generating context-aware responses, the evaluation was limited to offline analysis scenarios and did not include real-time processing, adversarial robustness testing, or scalability assessment. Model performance also varied across security contexts; in particular, hallucination rates remained a limiting factor in high-stakes environments such as threat intelligence and regulatory interpretation, where factual precision is paramount. These limitations were most evident in real-world operational scenarios.

From a legal perspective, the deployment of GenAI in security operations aligns with emerging regulatory standards. Article 21(1) of the NIS2 Directive mandates that entities adopt "state-of-the-art" technical, operational, and organisational measures proportionate to the risks. Similarly, GDPR Articles 25 and 32 require controllers and processors to integrate such measures into their data protection prac-

tices. While neither legal framework precisely defines "state of the art", authoritative interpretations such as the ENISA describe it as the best available and proven technology on the market. In this context, GenAI may be considered part of the evolving 'state of the art' spectrum, positioned between novel research and emerging security practice.

Jurisdictions like Australia, while using different terminology such as "reasonably practicable" under the Security of Critical Infrastructure Act 2018, essentially impose similar expectations: the deployment of the best available safeguards relative to risk. GenAI, given its current effectiveness and relatively low implementation burden, likely satisfies both the European and Australian thresholds. In certain risk contexts, failure to evaluate the potential applicability of GenAI-based solutions could, depending on circumstances, raise questions regarding compliance with 'state of the art' or 'reasonable care' standards. Therefore, GenAI represents an increasingly relevant instrument in the CISO toolbox, whose adoption should be guided by proportionality, risk assessment, and regulatory context. Its integration into cybersecurity strategy is not only technically justified but increasingly necessary to meet evolving legal compliance and operational effectiveness standards.

6 CONCLUSION AND FUTURE WORK

This work confirms that GenAI models can significantly support cybersecurity operations across technical, operational, human, and physical security domains. Evaluated models demonstrated the ability to analyze NetFlow traffic, detect various types of cyber threats, and provide structured, explainable insights. The experimental results revealed that while some models offer high analytical depth, they also vary in efficiency and hallucination rates. GenAI has proven valuable in augmenting CISO-level decision-making, aligning with legal standards. However, human oversight remains essential due to current limitations in accuracy and reliability.

Future research directions include the integration of real-time packet streaming and adversarial robustness testing, scalability analysis with alignment to EU AI Act compliance requirements.

ACKNOWLEDGEMENTS

This paper was funded by the project "Research on Cyber Resilience Through Application of Generative Artificial Intelligence in Chief Information Security Officer Operations", which has received funding from the Research Council of Lithuania (LMTLT) (agreement No S-ITP-24-13).

REFERENCES

Balasubramanian, P., Seby, J., and Kostakos, P. (2023). Transformer-based llms in cybersecurity: An in-depth study on log anomaly detection and conversational defense mechanisms. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 3590–3599.

Bui, M., Boffa, M., Valentim, R. V., Navarro, J. M., Chen, F., Bao, X., Houidi, Z. B., and Rossi, D. (2024). A systematic comparison of large language models performance for intrusion detection. *Proceedings of the ACM on Networking*, 2(CoNEXT4):1–23.

Cloudflare, Inc. (2025). Cloudflare signals report. Cybersecurity trends report, Cloudflare. Accessed: 2025-05-08.

European Parliament and Council of the European Union (2022). Directive (eu) 2022/2555 of the european parliament and of the council of 14 december 2022 on measures for a high common level of cybersecurity across the union (nis2 directive). Directive, European Union. Accessed: 2025-05-10.

Ghanem, M. C. and Ali, A. (2025). Beyond detection: Large language models and next-generation cybersecurity. In *Shifra*, pages 81–97. Shifra.

Houssel, P. R. B., Singh, P., Layeghy, S., and Portmann, M. (2024). Towards explainable network intrusion detection using large language models. In *2024 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT)*, pages 67–72.

Kasri, W., Himeur, Y., Alkhazaleh, H. A., Tarapiah, S., Atalla, S., Mansoor, W., and Al-Ahmad, H. (2025). From vulnerability to defense: The role of large language models in enhancing cybersecurity. *Computation*, 13(2):30.

Motlagh, F. N., Hajizadeh, M., Majd, M., Najafi, P., Cheng, F., and Meinel, C. (2024). Large language models in cybersecurity: State-of-the-art.

Wen, F. (2024a). The new trend of the integration of artificial intelligence and blockchain in network security. *Academic Journal of Computing & Information Science*, 7(3):38–42.

Wen, S. (2024b). The power of generative AI in cybersecurity: Opportunities and challenges. In *Proceedings of the 4th International Conference on Signal Processing and Machine Learning*, volume 48 of *Applied and Computational Engineering*, pages 31–39, Oxford, UK. EWA Publishing.