The Dual Role of Large Language Models in Network Security: Survey and Research Trends

Haiyun Liu
University of South Florida
Tampa, Florida, USA
haiyunliu@usf.edu

Jiahao Xue
University of South Florida
Tampa, Florida, USA
jiahao@usf.edu

Shangqing Zhao
University of Oklahoma
Norman, Oklahoma, USA
shangqing@ou.edu

Yao Liu
University of South Florida
Tampa, Florida, USA
yliu21@usf.edu

Zhuo Lu
University of South Florida
Tampa, Florida, USA
zhuolu@usf.edu

Abstract

Large language models (LLMs) have profoundly shaped various domains, including several types of network systems. With their powerful capabilities, LLMs have recently been proposed to enhance network security. However, the development of LLMs can introduce new risks due to their potential vulnerabilities and misuse. In this paper, we are motivated to review the dual role of LLMs in network security. Our goal is to explore how LLMs impact network security and ultimately shed light on how to evaluate LLMs from a network security perspective. We further discuss several future research directions regarding how to scientifically enable LLMs to assist with network security.

CCS Concepts

• **Networks → Network security**; • **Computing methodologies** → *Natural language processing*.

Keywords

Large language model (LLM), network security, adversarial attacks

1 Introduction

The rapid advancements in large language models (LLMs), such as GPT-4, BERT, and LLaMA, have revolutionized natural language processing (NLP) [11]. As network environments evolve with the integration of cloud computing, Internet of Things (IoT) ecosystems, and ubiquitous wireless connectivity, LLMs have extended their role beyond traditional NLP applications to enhance network security [44]. For instance, they can assist systems in detecting unauthorized access, analyzing traffic patterns, and mitigating potential threats across various network infrastructures [26].

The disadvantages shall not be neglected alongside these benefits. LLMs may expand attack surfaces due to their inherent vulnerabilities [18]. For example, attackers can exploit LLM-based intrusion detection systems (IDS) by injecting adversarial samples that mislead models into misclassifying malicious traffic [6, 32]. Additionally, adversaries can misuse LLMs as an attack tool to target existing networks [38], such as leveraging LLMs to analyze Wi-Fi traffic patterns and identify weak encryption schemes.

In this paper, we present a systematic analysis of the dual impact of LLMs on network security. In particular, we discuss the following three major topics:

RQ_1: How can LLMs be leveraged to enhance network security?
RQ_2: What security vulnerabilities arise from integrating LLMs into modern networks?
RQ_3: How can adversaries exploit LLMs to conduct adversarial attacks on network systems?

By addressing these questions, this work provides insights into the benefits and risks of using LLMs for network security.

The remainder of this paper is structured as follows: Section 2 provides an overview of LLMs. Section 3 explores the key applications of LLMs to enhance network security and answers the question RQ_1. Section 4 discusses the potential risks associated with users leveraging LLMs in networks and answers the question RQ_2. Section 5 analyzes how attackers exploit the capabilities of LLMs to conduct attacks on networks and answers the question RQ_3. We discuss future research directions for securing LLMs in network security in Section 6 and conclude this paper in Section 7.

2 Overview of LLMs

In this section, we provide a structured overview of LLMs, examine their connection to network security as well as their development for network security and the potential impacts.

2.1 LLMs and Network Security

LLMs are advanced AI systems built on deep learning architectures and trained on vast amounts of text data. They excel at understanding and generating human-like text, recognizing patterns, reasoning in context, and adapting to different tasks. These capabilities allow LLMs to process large-scale information efficiently, extract meaningful insights, and assist in the decision-making process, making them widely applicable across various fields [39, 43, 54].

Given the fact that network security fundamentally relies on data analysis, pattern recognition, and real-time decision-making, LLMs naturally intersect with this field due to their ability to process large-scale information, recognize complex patterns, and generate context-aware insights [43, 54]. This alignment makes them valuable for various security applications. However, this connection extends beyond just enhancing security. While LLMs can strengthen network defenses, they also introduce new challenges. Their integration into security systems expands the attack surface, and their capabilities could be misused by adversaries to create new risks

Table 1: Evolution of LLMs and their impacts on network security.

LLM Development Stage	Positive Impact	Negative Impact
Early Language Models (Pre-2018)	Improved keyword filtering and log analysis for security systems.	Easily evaded through synonym substitution and keyword obfuscation; ineffective against phishing.
Transformer Breakthrough (2018-2020)	Better phishing detection, anomaly identification, and NLP-based authentication.	Enabled more convincing phishing attacks and misinformation.
Rise of Large-Scale LLMs (2021-Present)	Advanced threat intelligence, malware detection, and penetration testing.	Automated phishing, deepfake scams, and polymorphic malware; privacy risks.
Future LLMs (2025+)	Real-time threat prediction and enhanced AI-driven security strategies.	Autonomous AI-driven cyberattacks, deepfake fraud, and adaptive malware.

[43, 54]. This dual role highlights the deep and complex relationship between LLMs and network security, where they both serve as powerful tools for defense and as potential threats to sophisticated cyberspace.

2.2 Evolution of LLMs

LLMs have evolved significantly, shaping network security in both positive and negative ways. The key milestones in LLM development and their dual impact on network security are as follows:

Early Language Models (Pre-2018): Early language models, such as N-grams and Word2Vec, improved keyword-based filtering and log analysis in security systems [43]. However, their reliance on simple statistical patterns made them vulnerable to evasion through synonym substitution or keyword obfuscation. Due to the lack of contextual understanding, they struggled against social engineering attacks, such as phishing emails that rely on deceptive language rather than specific keywords.

Transformer Breakthrough (2018-2020): The introduction of Transformer-based models like BERT and GPT-2 enhanced security tools with deeper contextual [54]. They improved phishing detection, anomaly identification, and NLP-based authentication. They are able to analyze messages that include more sophisticated keywords to reduce false negatives. However, the same advancements enabled attackers to generate more convincing phishing content and misinformation, making social engineering tactics harder to detect and counter.

Rise of Large-Scale LLMs (2021-Present): As models like GPT-3 and GPT-4 advanced, AI-driven cybersecurity solutions improved threat intelligence, malware detection, and penetration testing [54]. However, these advancements have also introduced new risks. Attackers leveraged LLMs to automate phishing, generate deepfake scams, and create polymorphic malware that evades traditional security measures. Additionally, privacy concerns arose as models trained on vast datasets may expose sensitive information.

Future LLMs (2025+): Looking ahead, LLMs are likely to be more predictive and explainable, which improves real-time threat detection and AI-driven security strategies [54]. However, their growing sophistication may pose serious security challenges because they enable autonomous cyberattacks, deepfake fraud, and adaptive malware. As the role of AI in cybersecurity expands, balancing innovation with regulation will be crucial.

This evolution brings both positive and negative impacts, which are summarized in Table 1.

3 Applications of LLMs to Enhance Network Security

With their powerful capabilities, LLMs have been proposed to be deployed in network systems to enhance security. They strengthen defense mechanisms and automate security processes across various applications, including threat detection, incident response, automated security management, malware analysis, and vulnerability assessment, as summarized in Fig. 1. In the following subsections, we focus on discussing the most common applications to address the previously mentioned RQ_1.

3.1 Threat Detection

Threat detection (TD) is a key application of LLMs in network security that identifies malicious activities, abnormal behaviors, and cyberattacks. There are four major types of TD.

LLM-Based TD: LLMs like GPT-4, BERT, and LLAMA3 have been applied to intrusion detection by analyzing network data. For example, the work in [55] applied in-context learning to enhance IDS in wireless networks, achieving 95% accuracy without additional model fine-tuning. 6G-XSec, an LLM-powered framework for wireless threat detection in 6G/OpenRAN, is introduced in [50]. It leverages LLMs and anomaly detection techniques to achieve 100% accuracy in identifying cellular attacks. The study of [49] proposed LLMIF, an LLM-powered framework for wireless threat detection in IoT networks. It improves Zigbee vulnerability detection, achieving 55.2% higher coverage and uncovering 11 security flaws. In the work of [53], it proposed LuaTaint, an LLM-enhanced static analysis system for wireless IoT vulnerability detection. It combines taint analysis with LLM-assisted false alarm pruning, and identifies 111 vulnerabilities in 2447 firmware samples with 89.29% precision.

Hybrid ML-LLM TD: Combining LLMs with ML/DL improves detection performance. For example, [19] developed IoV-BERT-IDS, a hybrid IDS leveraging BERT. It achieves state-of-the-art accuracy on BoT-IoT, CICIDS, Car-Hacking, and IVN-IDS datasets. The

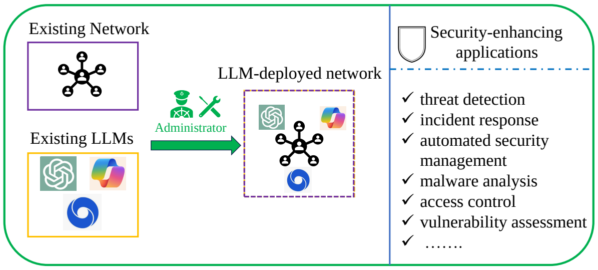

Figure 1: Applications of LLM-deployed networks for security enhancement.

work in [16] introduced BARTPredict, a framework integrating fine-tuned BART and BERT for intrusion detection. It achieves 98% accuracy on the CICIoT2023 dataset.

Federated and Distributed TD: LLMs support privacy-preserving detection in federated learning-based IDS. A transformer-based federated IDS was proposed in [3]. It combines a transformer encoder with a Gaussian Mixture Model (GMM) to enhance intrusion detection and achieve 95.6% accuracy. The study in [2] developed a BERT-based federated IDS for 5G networks. It achieves up to 97.12% accuracy in independent and identically distributed (IID) scenarios and remains robust under non-IID conditions.

Explainable TD: LLMs enhance the interpretability of IDS. For example, IDS-Agent is an explainable LLM-based IDS with an F1-score of 0.97 [36]. An LLM-driven explanation framework for IDS is studied in [59]. Semantic rule trees and chain-of-thought reasoning are used to improve anomaly detection transparency [48].

3.2 Incident Response

Incident response involves generating decisions autonomously or executing actions based on security incidents.

Incident Detection and Response Enhancement: LLMs are increasingly used to improve incident detection by automating anomaly recognition and response mechanisms. The study of [41] provided a systematic mapping study of intrusion response that addresses key challenges and research gaps. The work in [40] explored how LLMs can detect zero-day exploits in cloud networks to improve security over traditional rule-based monitoring.

Root Cause Analysis for Cloud Incidents: The complexity of cloud environments necessitates automated approaches for root cause analysis. An in-context learning-based approach with GPT-4 demonstrates superior performance over fine-tuned models [58]. The work in [13] introduced RCACopilot. It is an LLM-driven system that matches incidents to handlers, collects diagnostic data, and predicts root causes. Additionally, [33] proposed LLexus, an AI agent that transforms troubleshooting guides into executable plans, enhancing efficiency in incident resolution.

Automated Incident Management and Response: AI-powered automation is transforming incident response workflows. Xpert, an LLM-based system, was proposed in [30] to generate optimized queries for incident analysis, streamlining investigation processes. Another study in [10] evaluated the ability of forensic pipelines to detect AI-generated threats, proposing improvements for handling text-based cyber incidents.

3.3 Automated Security Management

Automated security management uses LLMs to optimize security policies, automate threat mitigation, and manage access control before an attack occurs. The related work is summarized as follows.

Intent-Driven Security Management and Automated Network Configuration: LLMs enhance intent-based security management by improving security configurations and reducing human errors. An LLM-driven intent-based networking framework can translate high-level security intents into executable network policies [21]. The study in [27] introduced an LLM-driven framework for wireless security, using federated fine-tuning to enhance privacy, policy automation, and misconfiguration detection in dynamic wireless networks.

Vulnerability Detection and Risk Mitigation: LLMs exhibit strong vulnerability detection capabilities. An instruction fine-tuned open-source LLM for wireless network analysis in 5G, called Mobile-LLaMA, was developed in [31]. It enhanced IP routing, packet inspection, and performance evaluation, aiding vulnerability detection and risk mitigation in wireless communications. Side-channel risk [57] in wireless LLM services may enable user data inference from encrypted traffic with 50 to 92% accuracy, highlighting security concerns.

Firewall Automation and Secure Communication: LLMs contribute to firewall automation and secure communication. An adaptive firewall framework in [1] dynamically adjusted security rules to prevent data leakage and adversarial attacks. LLM-Twin, a semantic encryption framework, could secure communication in beyond-5G IoT networks [24].

Automated Access Control Policy Enforcement: LLMs automate access control policy enforcement by translating natural language policies into structured rules. For example, [34] developed an LLM-driven method to extract key policy elements, improving readability and enforcement. LLM can be applied in Chain-of-Thought reasoning to automate security control validation and reduce reliance on manual audits [5].

4 Risks of Deploying LLMs in Network Systems

While LLMs enhance network security, their deployment also expands the attack surface, making them susceptible to adversarial threats. Attackers can exploit the vulnerabilities of LLM-integrated network systems through various ways, including prompt injection, membership inference, data poisoning, model extraction, as summarized in Fig. 2. In the following subsections, we analyze representative strategies that are being used to attack LLM-deployed networks and address RQ_2.

4.1 Prompt Injection

Prompt injection attacks manipulate inputs to bypass controls, exfiltrate data, or generate misleading responses, potentially causing unauthorized access, data breaches, and misinformation. These threats pose significant risks to network security and trust in automated systems.

Direct and Contextual Injection: Attackers issue direct instructions or embed adversarial prompts within seemingly benign content to override LLM constraints. The work in [9] tested 144 injection attempts across 36 LLMs, reporting a 56% success rate in

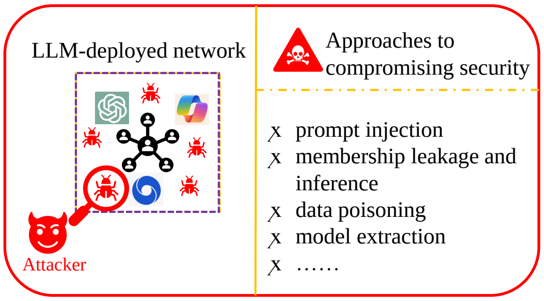

Figure 2: Attackers exploit the inherent vulnerabilities of LLMs to launch attacks on LLM-deployed network system.

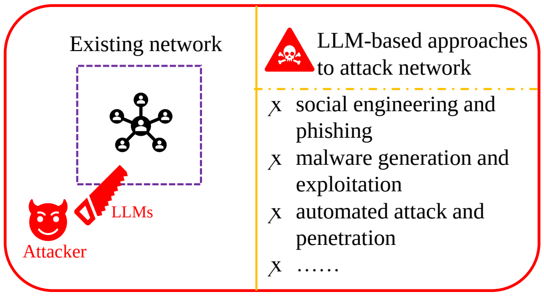
Existing network
LLM-based approaches to attack network
x social engineering and phishing
x malware generation and exploitation
x automated attack and penetration
x ……
LLMs
Attacker

Figure 3: Exploiting LLM for attacks on existing Networks.

bypassing content filters. Attackers can prompt injections in email metadata to manipulate LLM-generated summaries, which leads to misinformation [42].

Multi-Step and Persistent Injection: Adversaries use multi-step approaches or exploit LLM memory to achieve their objectives. The research in [37] proposed a chaining mechanism that gradually coaxes LLMs into revealing restricted content, achieving an 82% success rate against GPT-4o. Besides, [15] highlighted how adversarial prompts injected in earlier interactions could persist across sessions, influencing future responses.

Propagation-Based Injection: In interconnected LLM environments, malicious prompts propagate across agents, leading to systemic failures. For example, [35] introduced *Prompt Infection*, in which an injected adversarial instruction self-replicates across multiple LLMs to disrupt automated workflows.

4.2 Membership Leakage & Inference

Membership leakage and inference attacks expose sensitive training data, such as network logs and security incidents, compromising confidentiality in intrusion detection and malware classification models.

Score-Based Attacks: These attacks exploit statistical differences in model confidence scores and loss values. The study in [20] proposed a self-prompt calibration method to achieve high accuracy without external reference models. The noisy neighbor method developed in [23] amplifies confidence differences by perturbing input embeddings.

Feature-Based Attacks: Different attacks utilize aggregated model outputs (e.g., gradients, logits) for better inference. For example, [52] exploited wireless LLM vulnerabilities by using model output patterns to infer undocumented 5G security features, which exposes risks in wireless communication protocols. The side-channel leaks in wireless LLM services are exploited by [46]. Their method achieves 50 to 92% accuracy in inferring user attributes from encrypted traffic patterns.

Training Method-Specific Attacks: Some attackers target vulnerabilities in specific training paradigms. The attack method in [51] achieved about a 95% success rate against LLaMA models via in-context learning attacks. PREMIA in [17] showed that Direct Preference Optimization models are more vulnerable than Proximal Policy Optimization models in preference data attacks.

Ensemble and Hybrid Attacks: These methods combine multiple attack strategies. LiRA, LOSS-based, and Min-k attacks can be integrated with an XGBoost classifier to achieve state-of-the-art performance across LLM architectures [47].

4.3 Data Poisoning

Data poisoning can mislead LLMs, weaken network defense systems, and manipulate AI-driven security responses. This can result in reduced detection accuracy, the evasion of malicious activities, and the potential disruption of critical security operations, ultimately compromising network integrity and resilience.

In-Context Learning Poisoning: This type of attack manipulates the examples used for in-context learning, causing the model to make biased or incorrect predictions. For example, [25] introduced the ICLPoison framework to demonstrate how minor perturbations in context examples can significantly alter LLM predictions. Similarly, [14] presented a case study on biomedical models to reveal how poisoned training data can manipulate clinical decision-making outputs.

Fine-Tuning Stage Poisoning: These attacks occur during the fine-tuning phase of LLMs to embed hidden vulnerabilities that can be exploited later. The work in [28] explored how poisoning LLMs during this phase leads to unstable or biased text generation, while [29] discussed the susceptibility of parameter-efficient fine-tuning methods to poisoning. Additionally, [12] introduced the concept of Jail-Tuning to show how adversaries can bypass safety mechanisms by injecting harmful behavior during fine-tuning.

Retrieval-Augmented Poisoning: Retrieval-augmented generation systems depend on external data sources, making them susceptible to poisoning. Attackers can manipulate retrieved documents to inject imperceptible changes that lead to incorrect or misleading responses from LLMs [56].

5 Exploiting the Capabilities of LLMs for Network Security Attacks

The misuse of LLMs can amplify the scale, sophistication, and automation of cyberattacks. Such threats are more evasive and difficult to detect. Fig. 3 summarizes common ways for attackers to leverage the capabilities of LLMs against existing networks, including social engineering, phishing, malware generation, exploitation, automated attack, and penetration. In the following sections, we discuss the recent strategies to misuse LLMs and address the question Q_3.

Social Engineering and Phishing: LLMs enable highly convincing phishing and social engineering attacks. Afane et al. [4] showed how LLMs generate targeted, adaptive phishing messages. Alotaibi et al. [7] demonstrated how prompt engineering bypasses content filters to craft deceptive narratives. Singer et al. [45] further revealed how LLMs support scalable manipulation and automated fraud.

Malware Generation and Exploitation: LLMs have been misused to automate malware creation, lowering the barrier for developing sophisticated threats. They enable polymorphic malware, automated exploit generation, and advanced obfuscation to evade detection. RatGPT [8] demonstrated the creation of ransomware, keyloggers, and remote access trojans with undetectable in-memory payloads. Alotaibi et al. [7] showed how prompt engineering can manipulate LLMs to produce malicious scripts, such as shellcode injection and privilege escalation. DeepLocker and similar AI-driven malware [22] highlight how machine learning enhances evasion and adaptive attack strategies.

Automated Attack and Penetration: LLMs streamline cyberattacks by automating reconnaissance, exploitation, lateral movement, and data exfiltration. Singer et al. [45] proposed an LLM-based framework for optimizing multi-stage attacks. LLMs also support command-and-control automation [8], enabling adaptive execution and evasion. Gabrian et al. [22] showed how AI enhances persistence, obfuscation, and decision-making in attack automation.

6 Future Research Directions

LLMs offer powerful capabilities for enhancing cybersecurity, but they also introduce new risks, such as adversarial attacks, privacy concerns, and interpretability challenges. To fully realize their potential, future research should focus on the following key areas:

Defending Against LLM-Based Attacks: The increasing sophistication of LLM-powered attacks, such as prompt injection, model inversion, and adversarial exploitation, necessitates the development of robust defense mechanisms. Future research should focus on designing adaptive security frameworks that detect and mitigate these attacks in advance, leveraging techniques such as adversarial training, anomaly detection, and reinforcement learning-based security policies.

Enhancing Interpretability and Transparency: The black-box nature of LLMs poses significant challenges in security-critical applications. Enhancing the interpretability of LLM-driven security systems can improve trust, compliance, and effectiveness. Researchers shall explore explainable AI techniques, interpretable neural network architectures, and visualization tools that provide insights into decision-making processes within LLMs. It enables more transparent cybersecurity solutions.

Privacy-Preserving and Secure Deployment: Given that LLMs process sensitive network data, privacy-preserving mechanisms must be incorporated to safeguard user information. Researchers should investigate federated learning, homomorphic encryption, and differential privacy techniques to enable secure training and inference without compromising confidentiality. Additionally, decentralized and trust-aware frameworks can further enhance the secure deployment of LLMs in distributed network environments.

Combating Model Hallucination and Data Integrity: LLMs can generate plausible but incorrect information, a phenomenon known as model hallucination. This presents risks in security-critical applications where misinformation could impact decision-making. Future studies shall develop methods to validate model outputs, incorporate factual verification mechanisms, and improve model robustness against erroneous generations.

Adversarial Testing and Red-Teaming: To ensure the resilience of LLMs in network security, systematic adversarial testing and red-teaming approaches should be developed. This includes the creation of benchmarking datasets, standardized attack simulations, and automated security testing pipelines to evaluate the robustness of LLM-driven security systems under real-world adversarial conditions.

Automating Cybersecurity Responses: Integrating LLMs into real-time cybersecurity response systems can enhance incident mitigation and containment strategies. However, ensuring the reliability and accuracy of automated security responses is crucial. Future work shall focus on hybrid AI-human-in-the-loop frameworks, self-learning security orchestration, and autonomous threat-hunting methodologies to balance automation with human oversight in cybersecurity operations.

Optimizing Resource Efficiency and Sustainability: Training and deploying LLMs demand high computational resources, raising sustainability and cost concerns. Future work should explore lightweight models, energy-efficient architectures, and resource-aware deployment for accessible, eco-friendly security solutions.

Ethical Guidelines and Regulation: The rapid adoption of LLMs in cybersecurity calls for ethical guidelines and regulations to prevent misuse. New research should address legal, ethical, and societal concerns, including bias, accountability, and AI-driven threats. Collaboration among policymakers, academia, and industry is key to ensuring responsible AI use in network security.

7 Conclusion

In this paper, we surveyed the literature on how the rapidly growing LLMs have dual impacts on the increasingly complex field of network security. We categorized the relevant work into three key topics: (1) LLMs-enhanced network security, (2) security risks associated with LLM deployment, and (3) adversarial use of LLMs for cyber-attacks. We discussed future research directions on securing LLMs for network security applications since some concerns remain in this area.

References

[1] S. Abdelnabi, A. Gomaa, E. Bagdasarian, P. O. Kristensson, and R. Shokri. Firewalls to secure dynamic LLM agentic networks. *arXiv preprint*, 2025.

[2] F. Adjewa, M. Esseghir, and L. Merghem-Boulahia. Efficient federated intrusion detection in 5G ecosystem using optimized BERT-based model. In *IEEE WiMob*, 2024.

[3] F. Adjewa, M. Esseghir, and L. Merghem-Boulahia. LLM-based continuous intrusion detection framework for next-gen networks. *arXiv preprint*, 2024.

[4] K. Afane, W. Wei, Y. Mao, J. Farooq, and J. Chen. Next-generation phishing: How LLM agents empower cyber attackers. In *IEEE BigData*, 2024.

[5] M. Ahmed, J. Wei, and E. Al-Shaer. Prompting LLM to enforce and validate CIS critical security control. In *ACM SACMAT*, 2024.

[6] A. Alotaibi and M. A. Rassam. Adversarial machine learning attacks against intrusion detection systems: A survey on strategies and defense. *Future Internet*, 2023.

[7] L. Alotaibi, S. Seher, and N. Mohammad. Cyberattacks using ChatGPT: Exploring malicious content generation through prompt engineering. In *IEEE ICETSIS*, 2024.

[8] M. Beckerich, L. Plein, and S. Coronado. RatGPT: Turning online LLMs into proxies for malware attacks. *arXiv preprint*, 2023.

[9] V. Benjamin, E. Braca, I. Carter, H. Kanchwala, N. Khojasteh, C. Landow, Y. Luo, C. Ma, A. Magarelli, R. Mirin, et al. Systematically analyzing prompt injection vulnerabilities in diverse LLM architectures. *arXiv preprint*, 2024.

[10] A. Bhandarkar, R. Wilson, A. Swarup, M. Zhu, and D. Woodard. Is the digital forensics and incident response pipeline ready for text-based threats in LLM era? *arXiv preprint*, 2024.

[11] G. O. Boateng, H. Sami, A. Alagha, H. Elmekki, A. Hammoud, R. Mizouni, A. Mourad, H. Otrok, J. Bentahar, S. Muhaidat, et al. A survey on large language models for communication, network, and service management: Application insights, challenges, and future directions. *arXiv preprint*, 2024.

[12] D. Bowen, B. Murphy, W. Cai, D. Khachaturov, A. Gleave, and K. Pelrine. Data poisoning in LLMs: Jailbreak-tuning and scaling laws. *arXiv preprint*, 2024.

[13] Y. Chen, H. Xie, M. Ma, Y. Kang, X. Gao, L. Shi, Y. Cao, X. Gao, H. Fan, M. Wen, et al. Automatic root cause analysis via large language models for cloud incidents. In *ACM EuroSys*, 2024.

[14] A. Das, A. Tariq, F. Batalini, B. Dhara, and I. Banerjee. Exposing vulnerabilities in clinical LLMs through data poisoning attacks: Case study in breast cancer. *medRxiv*, 2024.

[15] E. Derner, K. Batistič, J. Zahálka, and R. Babuška. A security risk taxonomy for prompt-based interaction with large language models. *IEEE Access*, 2024.

[16] A. Diaf, A. A. Korba, N. E. Karabadji, and Y. Ghamri-Doudane. BARTPredict: Empowering IoT security with LLM-driven cyber threat prediction. *arXiv preprint*, 2025.

[17] Q. Feng, S. R. Kasa, H. Yun, C. H. Teo, and S. B. Bodapati. Exposing privacy gaps: Membership inference attack on preference data for LLM alignment. *arXiv preprint*, 2024.

[18] O. Friha, M. A. Ferrag, B. Kantarci, B. Cakmak, A. Ozgun, and N. Ghoualmi-Zine. LLM-based edge intelligence: A comprehensive survey on architectures, applications, security and trustworthiness. *IEEE OJ-COMS*, 2024.

[19] M. Fu, P. Wang, M. Liu, Z. Zhang, and X. Zhou. IoV-BERT-IDS: Hybrid network intrusion detection system in IoV using large language models. *IEEE Trans. Veh. Technol*, 2024.

[20] W. Fu, H. Wang, C. Gao, G. Liu, Y. Li, and T. Jiang. Membership inference attacks against fine-tuned large language models via self-prompt calibration. In *NeurIPS*, 2024.

[21] A. Fuad, A. H. Ahmed, M. A. Riegler, and T. Čičić. An intent-based networks framework based on large language models. In *IEEE NetSoft*, 2024.

[22] C.-A. Gabrian. Unveiling the dark side: How hackers exploit artificial intelligence for cyber warfare.

[23] F. Galli, L. Melis, and T. Cucinotta. Noisy neighbors: Efficient membership inference attacks against LLMs. *arXiv preprint*, 2024.

[24] S. Guo, Y. Wang, N. Zhang, Z. Su, T. H. Luan, Z. Tian, and X. Shen. A survey on semantic communication networks: Architecture, security, and privacy. *IEEE Commun. Surv. Tutorials*, 2024.

[25] P. He, H. Xu, Y. Xing, H. Liu, M. Yamada, and J. Tang. Data poisoning for in-context learning. *arXiv preprint*, 2024.

[26] J. Jang-Jaccard and S. Nepal. A survey of emerging threats in cybersecurity. *Elsevier JCSS*, 2014.

[27] F. Jiang, L. Dong, S. Tu, Y. Peng, K. Wang, K. Yang, C. Pan, and D. Niyato. Personalized wireless federated learning for large language models. *arXiv preprint*, 2024.

[28] S. Jiang, S. R. Kadhe, Y. Zhou, L. Cai, and N. Baracaldo. Forcing generative models to degenerate ones: The power of data poisoning attacks. *arXiv preprint*, 2023.

[29] S. Jiang, S. R. Kadhe, Y. Zhou, F. Ahmed, L. Cai, and N. Baracaldo. Turning generative models degenerate: The power of data poisoning attacks. *arXiv preprint*, 2024.

[30] Y. Jiang, C. Zhang, S. He, Z. Yang, M. Ma, S. Qin, Y. Kang, Y. Dang, S. Rajmohan, Q. Lin, et al. Xpert: Empowering incident management with query recommendations via large language models. In *IEEE/ACM ICSE*, 2024.

[31] K. B. Kan, H. Mun, G. Cao, and Y. Lee. Mobile-LLaMA: Instruction fine-tuning open-source LLM for network analysis in 5G networks. *IEEE Network*, 2024.

[32] H. Kheddar. Transformers and large language models for efficient intrusion detection systems: A comprehensive survey. *arXiv preprint*, 2024.

[33] P. Las-Casas, A. G. Kumbhare, R. Fonseca, and S. Agarwal. LLexus: an AI agent system for incident management. *ACM SIGOPS*, 2024.

[34] S. Lawal, X. Zhao, A. Rios, R. Krishnan, and D. Ferraiolo. Translating natural language specifications into access control policies by leveraging large language models. In *IEEE TPS-ISA*, 2024.

[35] D. Lee and M. Tiwari. Prompt infection: LLM-to-LLM prompt injection within multi-agent systems. *arXiv preprint*, 2024.

[36] Y. Li, Z. Xiang, N. D. Bastian, D. Song, and B. Li. IDS-agent: An LLM agent for explainable intrusion detection in IoT networks. In *NeurIPS OWA*, 2024.

[37] F. Mohammad Ali Pour and M. Rashidi. Web LLM attacks: Unveiling the future of cyber threats. *Available at SSRN 5049058*, 2024.

[38] M. Mozes, X. He, B. Kleinberg, and L. D. Griffin. Use of LLMs for illicit purposes: Threats, prevention measures, and vulnerabilities. *arXiv preprint*, 2023.

[39] C. Parsing. Speech and language processing. *Power Point Slides*, 2009.

[40] K. Patil and B. Desai. Leveraging LLM for zero-day exploit detection in cloud networks. *Asian American Research Letters Journal*, 2024.

[41] A. Rezapour, M. GhasemiGol, and D. Takabi. A systematic mapping study on intrusion response systems. *IEEE Access*, 2024.

[42] S. Rossi, A. M. Michel, R. R. Mukkamala, and J. B. Thatcher. An early categorization of prompt injection attacks on large language models. *arXiv preprint*, 2024.

[43] B. Sabir, F. Ullah, M. A. Babar, and R. Gaire. Machine learning for detecting data exfiltration: A review. *ACM Computing Surveys*, 2021.

[44] O. Santos, S. Salam, and H. Dahir. The AI revolution in networking, cybersecurity, and emerging technologies. *Addison-Wesley Professional*, 2024.

[45] B. Singer, K. Lucas, L. Adiga, M. Jain, L. Bauer, and V. Sekar. On the feasibility of using LLMs to execute multistage network attacks. *arXiv preprint*, 2025.

[46] M. Soleimani, G. Jia, I. Gim, S.-s. Lee, and A. Khandelwal. Wiretapping LLMs: Network side-channel attacks on interactive LLM services. *Cryptology ePrint Archive*, 2025.

[47] Z. Song, S. Huang, and Z. Kang. EM-MIAs: Enhancing membership inference attacks in large language models through ensemble modeling. *arXiv preprint*, 2024.

[48] F. Tang, X. Wang, X. Yuan, L. Luo, M. Zhao, and N. Kato. Large language model (LLM) assisted end-to-end network health management based on multi-scale semanticization. *arXiv preprint*, 2024.

[49] J. Wang, L. Yu, and X. Luo. Llmif: Augmented large language model for fuzzing IoT devices. In *IEEE S&P*, 2024.

[50] H. Wen, P. Sharma, V. Yegneswaran, P. Porras, A. Gehani, and Z. Lin. 6G-XSec: Explainable edge security for emerging OpenRAN architectures. In *ACM HotNets*, 2024.

[51] R. Wen, Z. Li, M. Backes, and Y. Zhang. Membership inference attacks against in-context learning. In *ACM CCS*, 2024.

[52] T. Wray and Y. Wang. 5G specifications formal verification with over-the-air validation: Prompting is all you need. In *IEEE MILCOM*, 2024.

[53] J. Xiang, L. Fu, T. Ye, P. Liu, H. Le, L. Zhu, and W. Wang. LuaTaint: A static analysis system for web configuration interface vulnerability of internet of things device. *IEEE Internet Things J*, 2024.

[54] Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang. A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, 2024.

[55] H. Zhang, A. B. Sediq, A. Afana, and M. Erol-Kantarci. Large language models in wireless application design: In-context learning-enhanced automatic network intrusion detection. *arXiv preprint*, 2024.

[56] Q. Zhang, B. Zeng, C. Zhou, G. Go, H. Shi, and Y. Jiang. Human-imperceptible retrieval poisoning attacks in LLM-powered applications. In *ACM FSE*, 2024.

[57] R. Zhang, S. S. Hussain, P. Neekhara, and F. Koushanfar. REMARK-LLM: A robust and efficient watermarking framework for generative large language models. In *USENIX*, 2024.

[58] X. Zhang, S. Ghosh, C. Bansal, R. Wang, M. Ma, Y. Kang, and S. Rajmohan. Automated root causing of cloud incidents using in-context learning with GPT-4. In *ACM FSE*, 2024.

[59] N. Ziems, G. Liu, J. Flanagan, and M. Jiang. Explaining tree model decisions in natural language for network intrusion detection. *arXiv preprint*, 2023.