

Large Language Models in Cybersecurity: Applications, Vulnerabilities, and Defense Techniques

Niveen O. Jaffal
 Birzeit University
 Palestine
 njaffal@birzeit.edu

Mohammed Alkhanafseh
 Birzeit University
 Palestine
 mkhanafeh@@birzeit.edu

David Mohaisen*
 University of Central Florida
 USA
 mohaisen@ucf.edu

ABSTRACT

Large Language Models (LLMs) are transforming cybersecurity by enabling intelligent, adaptive, and automated approaches to threat detection, vulnerability assessment, and incident response. With their advanced language understanding and contextual reasoning, LLMs surpass traditional methods in tackling challenges across domains such as IoT, blockchain, and hardware security. This survey provides a comprehensive overview of LLM applications in cybersecurity, focusing on two core areas: (1) the integration of LLMs into key cybersecurity domains, and (2) the vulnerabilities of LLMs themselves, along with mitigation strategies. By synthesizing recent advancements and identifying key limitations, this work offers practical insights and strategic recommendations for leveraging LLMs to build secure, scalable, and future-ready cyber defense systems.

KEYWORDS

LLMs, Cybersecurity, Applications, Attacks, Defenses

1 INTRODUCTION

Advances in machine learning and deep learning, particularly with the advent of transformer architectures [1], have driven the development of LLMs. These advanced Natural Language Processing (NLP) systems, characterized by their extensive parameterization, are trained on foundational tasks such as masked language modeling and autoregressive prediction. These training paradigms enable LLMs to process human language effectively, analyzing contextual semantics and probabilistic relationships across massive text datasets. LLMs exhibit four essential characteristics: (i) a deep understanding of the natural language context; (ii) the ability to generate human-like text; (iii) advanced contextual awareness, particularly in knowledge-intensive applications; and (iv) strong instruction following capabilities that support problem solving and decision making. Prominent LLMs, including BERT [2], GPT-3.5 [3] and GPT-4 [4], PaLM [5], Claude [6], and Chinchilla [7], have demonstrated exceptional performance in various NLP tasks, such as language understanding, text generation, and reasoning. The adaptability of these models is particularly notable as they enable breakthroughs in downstream applications with minimal fine-tuning. These include open domain question answering [8], dialogue systems [9], and program synthesis [10], among others. Using rich linguistic representations and robust reasoning capabilities, LLMs are reshaping how NLP challenges are addressed, paving the way for transformative advancements across various domains.

The increasing complexity and sophistication of cyber threats require innovative and adaptive approaches to strengthen cybersecurity mitigation [11–27], and LLMs have found many applications in the security domain [28]. For instance, LLMs are known to generalize from human languages to other domains with minor modifications, making them ideal for automating the generation of security rules, associating cyber threats with one another, and even discovering new phenomena and threats unseen before. Moreover, the fact that such models are deployed in a rather hostile environment with both benign and malicious users makes them ideal for manipulation, giving rise to the idea of understanding the robustness of such models through a finer understanding of their attack surface and mitigations to minimize such surface. This survey offers a comprehensive analysis of the transformative impact that LLMs could have on the field of cybersecurity by identifying the various applications of such tools, enumerating their vulnerabilities, and highlighting key defense techniques. In compiling this survey, we hope to shed light on the existing literature and identify gaps in such literature that could open the door for future work.

Although there have been several surveys in this space, ours stands out in multiple ways. First, it is the first survey to integrate the application landscape in cybersecurity with the attack surface and defense mechanisms, providing a comprehensive enumeration of a broader spectrum of applications. Second, given the evolving nature of this domain, our survey encompasses a range of studies across previously unexplored dimensions.

1.1 Research Questions

To effectively position our work in the context of previous works, we formulate several broad research questions that we aim to address through this effort. **RQ 1. What are the key cybersecurity domains where LLMs address specific tasks and challenges within these domains and how?** The first research question focuses on the scope and nature of security tasks in which LLMs have been applied, with the aim of categorizing and understanding the breadth of security challenges addressed in different security domains. By analyzing previous studies, this question seeks to provide a detailed inventory of the various security tasks that utilize LLM, offering insight into their adaptability, effectiveness, and impact within each domain. **RQ 2. What vulnerabilities are associated with LLM in cyber security applications and what strategies can be implemented to mitigate these risks and protect models?** The second research question investigates the vulnerabilities of LLMs and explores defense techniques to improve their security. By analyzing potential attack vectors and prevention techniques, it provides a comprehensive understanding of the challenges and

*Corresponding author

safeguards necessary to improve the resilience of these models in cybersecurity applications.

1.2 Contributions

The primary objective of this survey is to explore the future of cybersecurity through the lens of generative Artificial Intelligence (AI) and LLMs, encompassing all the main critical aspects of the cyber domain. To this end, the key contributions of this survey are summarized as follows:

- **Applications.** We enumerate and evaluate the diverse applications of LLMs in cybersecurity, including hardware design security, intrusion detection, malware detection, and phishing prevention, while analyzing their capabilities in these different contexts to address how LLMs are used.
- **Vulnerabilities and Mitigation.** We systematically enumerate and examine the vulnerabilities in LLMs with respect to their implications for security applications. Our exploration of such an attack surface encompasses aspects like prompt injection, jailbreaking attack, data poisoning, and backdoor attacks. Moreover, we enumerate and assess the defense techniques that are in place or could be further deployed to reduce these risks and improve LLMs security for those critical applications.
- **Potential Challenges.** We identify the possible potential challenges that arise in the use of LLMs for specified cybersecurity tasks, calling for more attention and research actions from the community.

1.3 Organization

The remaining sections are structured as follows. Section 3 summarizes recent surveys on LLMs and their applications in various cybersecurity domains, highlighting previous work and positioning this survey within the literature on LLM applications and security. Section 4 explores the security domains enhanced by LLM innovations, analyzing their impact on cybersecurity applications and their effectiveness in addressing complex challenges. Section 5 investigates LLM vulnerabilities, categorizing threats and countermeasures while highlighting the need for proactive mitigation strategies to ensure secure deployment. Section 6 discusses the challenges and limitations of integrating LLMs into cybersecurity, considering both practical and theoretical aspects. Finally, Section 7 highlights the potential of LLMs in cybersecurity, summarizes key findings and defense strategies, and outlines research directions.

2 METHODOLOGY

To conduct this survey, we prioritize key factors in selecting relevant studies, including timeliness, coverage, diversity, and relevance. For timeliness, we focus on research published within the last four years, emphasizing the most recent three. This selection includes 15 studies from 2021, 25 from 2022, 64 from 2023, and 68 from 2024. For coverage, we ensure a broad range of applications, attack vectors, and defense mechanisms, encompassing 114 studies on applications, 12 on attacks, and 31 on defense techniques. To enhance diversity, we include studies from both AI (71 studies) and security research venues (101 studies). Figure 1 visualizes this distribution, illustrating the evolution of research over time, the dominance of applications,

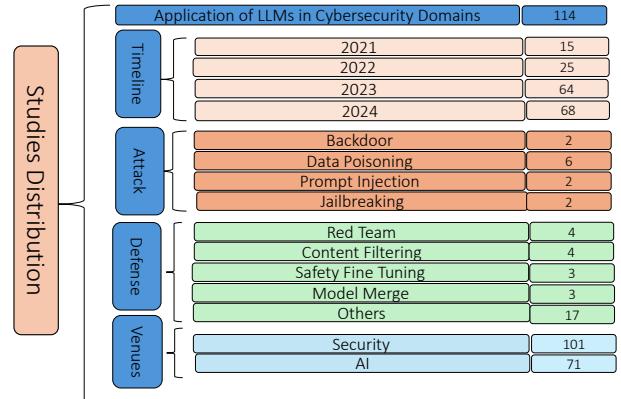


Figure 1: Comprehensive Survey Studies Analysis

attacks, and defenses in selected studies, and the representation of AI and security venues. This structured approach ensures a comprehensive analysis.

3 RELATED REVIEWS

We present in this section several previously selected surveys that make significant contributions; these surveys encompass a detailed and thorough examination of key aspects related to LLM, including the datasets utilized for training and fine-tuning these models, the vulnerabilities inherent to these models, and the strategies proposed for their mitigations. In addition, they highlight the innovative methodologies employed by LLMs to address complex security challenges, offering a comparative analysis of their efficacy in addressing issues across domains such as cloud computing, the Internet of Things (IoT), hardware, Blockchain, software, and system security.

As shown in Table 1, this work provides a comprehensive analysis of LLM applications in cybersecurity, distinguishing itself from previous work by addressing critical gaps across five key security domains. Although existing surveys, such as [31] and [32], explore general LLM applications and vulnerabilities, they often lack detailed insights into specific domains such as IoT, cloud, hardware, and blockchain security. Furthermore, our bridges gaps in underexplored domains such as IoT and cloud environments, providing a more holistic view of LLM deployment in cybersecurity. This work advances the state of the art by integrating broader perspectives with practical insights, positioning it as a critical addition to the evolving body of research. Finally, the work covers a range of recent advances and papers that are more timely and were not covered in previous studies and surveys.

4 SECURITY DOMAINS AND TASKS EMPOWERED BY LLM-BASED INNOVATIONS

The increasing complexity of cybersecurity threats has driven the adoption of LLMs across various security domains. This survey categorizes LLM applications into eight key domains: network, software and system, information and content, hardware, blockchain, cloud,

Table 1: A summary of the related work of LLMs applications. Highlights study contribution, Datasets used by LLM for security use case, Vulnerabilities associated with LLMs and defense techniques, Optimized techniques for LLMs, and domains: ① Internet of Things, ② Cloud, ③ Hardware, ④ Blockchain, ⑤ Software and System Security.

Ref	Year	Study Area	Contribution	D	V	O	Domains				
				①	②	③	④	⑤			
[29]	2023	LLMs in SoC	Surveys LLM potential, challenges, and outlook for SoC security.	X	✓	✓	X	X	✓	X	✓
[30]	2024	LLMs for blockchain	Covers LLM use in auditing, anomaly detection, and vulnerability repair.	X	✓	✓	X	X	✓	✓	X
[31]	2024	LLM applications	Explores use cases, dataset limitations, and mitigation strategies.	✓	X	✓	X	X	X	✓	✓
[32]	2024	GenAI in cybersecurity	Reviews GenAI attack applications in cybersecurity.	✓	✓	X	X	X	X	✓	✓
[33]	2024	LLM in cybersecurity	Highlights LLM capabilities for solving key cybersecurity issues.	✓	✓	X	✓	X	X	✓	X
[34]	2024	Survey	Surveys 42 models, their roles in cybersecurity, and known flaws.	✓	✓	✓	X	✓	✓	X	✓
Ours	2025	Apps, vulns., and defenses	Presents LLM use across domains, identifies vulnerabilities, and proposes defenses.	✓	✓	✓	✓	✓	✓	✓	✓

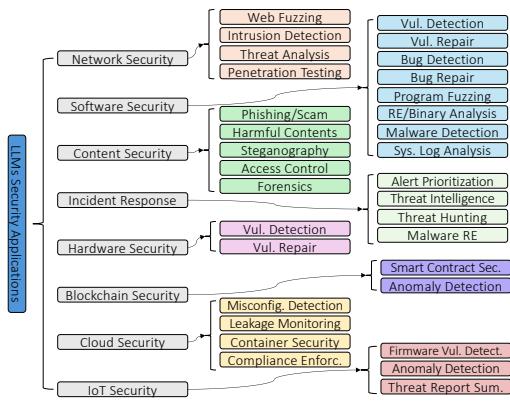


Figure 2: Cybersecurity Domains and LLM Based Applications

incident response and threat intelligence, and IoT security. This classification provides a structural perspective on the diverse applications of LLMs in cybersecurity domains, offering deeper insights into their contributions to strength security robustness across real-time applications and domains. Figure 2 visualizes the distribution of LLM applications across cybersecurity domains. Moreover, Tables 2 and 2 present LLM-based solutions for various cybersecurity tasks and use cases. LLMs have proven to be instrumental in improving efficiency, accuracy, and adaptability, marking a significant leap in modern cybersecurity defense strategies.

4.1 LLMs in Network Security

This section delves into the applications of LLMs in the field of network security. LLMs have proven to be versatile and powerful tools for addressing a wide range of tasks in this domain, including web fuzzing, traffic and intrusion detection, threat analysis, and penetration testing, which we review in the following subsections.

TAKEAWAY 1. *LLMs enhance network security by improving intrusion and anomaly detection through in-context learning and graph-based techniques. In CTI, they automate intelligence extraction for real-time threat monitoring. Tools like GPTFuzzer refine web fuzzing by generating targeted test cases, while PentestGPT streamlines penetration testing through automated reconnaissance and exploit generation. These advancements boost efficiency, accuracy, and adaptability in network security operations.*

4.1.1 Web Fuzzing. Web fuzzing is a mutation-based testing technique that incrementally generates test cases by leveraging coverage feedback obtained from instrumented web applications. Given the critical importance of security in web applications, fuzzing is vital in identifying potential vulnerabilities. For example, Liang *et al.* [35] introduced GPTFuzzer, a tool built on an encoder-decoder architecture. GPTFuzzer effectively generates payloads targeting Web Application Firewalls (WAFs) to identify and test Structured Query Language injection (SQLi), Cross-Site Scripting (XSS), and Remote Code Execution (RCE) attacks. This is achieved through reinforcement learning, fine-tuning, and applying a KL divergence penalty, which helps overcome local optima during the payload generation process.

Similarly, Liu *et al.* [36] utilized an encoder-decoder architecture to design SQL injection detection test cases for web applications, translating user input into new test scenarios. Meng *et al.* [37] expands the scope by employing LLMs to generate structured and sequential test inputs for network protocols that lack machine-readable formats. These advancements demonstrate the growing potential of machine learning and fuzzing in enhancing the security testing process for web applications.

4.1.2 Traffic and Intrusion detection. Detecting network traffic and intrusions is crucial for network security and management. LLMs have become powerful tools for intrusion detection, demonstrating versatility across traditional web applications, IoT ecosystems, and in-vehicle networks. These models effectively learn complex patterns in malicious traffic, identify deviations in user behavior that signal anomalies, and interpret the intent behind intrusions and abnormal activities.

A marked improvement is the effort of Liu *et al.* [57], who used LLM as a tool to extract features from malicious URLs while expanding the detection to user-level contexts. Zhang *et al.* [58] demonstrated that employing GPT-4 for in-context learning can achieve an intrusion detection accuracy of more than 95% with a limited amount of labeled data, thus eliminating the need for fine-tuning. Houssel *et al.* [7] explored the explainability of LLMs in the context of NetFlow-based Network Intrusion Detection Systems (NIDS), showing that they can augment traditional methods and improve interpretability by using tools such as Retrieval Augmented Generation (RAG). LLMs promise real-time analysis and response, enabling zero-day vulnerability discovery and identification of emerging attack patterns through continuous learning. Additionally, they scale efficiently to handle large network traffic, making them ideal for

Table 2: LLMs-based Models for Cybersecurity (Part I)

Author	Ref.	Year	Dataset	Task	Key Contributions	Challenges	Optimized Technique
Meng <i>et al.</i>	[37]	2024	PROFUZZ	Web Fuzzing	Enhanced state transitions and vuln. discovery.	Weak for closed protocols.	GPT-3.5 for grammar and seed tuning.
Liu <i>et al.</i>	[38]	2023	GramBeddings, Mendeley	Traffic Detection	CharBERT improves URL detection.	High compute, few adversarial tests.	CharBERT + attention + pooling.
Moskal <i>et al.</i>	[39]	2023	Sim. forensic exp.	Threat Analysis	Refined ACT loop via sandboxed LLM agents.	Weakness with complex net/env.	FSM + chained prompts.
Temara <i>et al.</i>	[40]	2023	None	Pentesting	Multi-tool consolidation via LLM.	Pre-2021 limit, prompt sensitivity.	Case-driven extraction.
Tihanyi <i>et al.</i>	[41]	2023	Form AI	Vuln. Detection	Benchmarked form AI dataset.	Limited scope, costly verification.	Zero-shot + ESBMC.
Sun <i>et al.</i>	[42]	2024	Solidity, Java CWE	Firmware Vuln.	LLM4Vuln detects misuse + zero-days.	No cross-language support.	Prompt + retrieval + reasoning.
Meng <i>et al.</i>	[43]	2023	RISC, MIPS	HW Vuln.	NSPG w/ HS-BERT for spec extraction.	Doc scarcity, manual labels.	Fine-tuned HS-BERT + MLM.
Du <i>et al.</i>	[44]	2023	Synthetic SFG	Bug Localization	Graph-based CodeBERT + contrastive loss.	Low scale, limited eval.	HMCBL w/ neg. sampling.
Joyce <i>et al.</i>	[45]	2023	VirusTotal (2006-23)	Malware Feature Learn.	AVScan2Vec for AV task embedding.	Mix of benign/malware, retrain needs.	Token + label masking w/ Siamese fine-tuning.
Labonne <i>et al.</i>	[46]	2023	Email spam datasets	Phishing Detect.	Spam-T5 outperforms baseline.	Costly training, weak domain generality.	Prefix-tuned Flan-T5.
Malul <i>et al.</i>	[47]	2024	KCFs	Misconfig. Detection	GenKubeSec w/ UMI for high accuracy.	Unseen configs, labeled data, scaling.	Fine-tuned LLM + few-shot.
Kasula <i>et al.</i>	[48]	2023	NSL-KDD, cloud logs	Data Leakage	Real-time detection in dynamic clouds.	Low generalization.	RF + LSTM.
Bandara <i>et al.</i>	[49]	2024	PBOMs	Container Sec.	DevSec-GPT tracks vulns via blockchain.	LLM overhead.	Llama2 + JSON schema.
Nguyen <i>et al.</i>	[50]	2024	Compliance data	Compliance	Olla Bench tests LLM reasoning.	Scalability, adaptability.	KG + SEM.
Ji <i>et al.</i>	[51]	2024	Incident reports	Alert Prioritization	SEVENLLM for multi-task response.	Limited multilinguality.	Prompt-tuned LLM.
Gai <i>et al.</i>	[52]	2023	Ethereum txns	Anomaly Detect.	BLOCKGPT ranks real-time anomalies.	False positives.	EVM tree + transformer.
Ahmad <i>et al.</i>	[53]	2023	MITRE, OpenTitan	HW Bug Repair	LLM repair in Verilog, beats CirFix.	Multi-line bug limits.	Prompt tuning + testbench.
Tseng <i>et al.</i>	[54]	2024	CTI Reports	Threat Intel	GPT-4 extracts IoCs + Regex rules.	Extraction accuracy.	Segmentation + GPT-4.
Scanlon <i>et al.</i>	[55]	2023	Forensics	Forensic Tasks	GPT-4 used in education + analysis.	Output inconsistency.	Prompt + expert check.
Martin <i>et al.</i>	[56]	2023	OpenAI corpus	Bias Detect.	ChatGPT shows political bias.	Ethical balance.	GSS-based bias reverse.

high throughput settings. Their integration with methods such as graph-based anomaly detection and reinforcement learning increases their ability to detect complex threats. These capabilities highlight the crucial role of LLMs in modern intrusion detection, providing robust, scalable, and advanced digital infrastructure protection.

4.1.3 Cyber Threat Intelligence (CTI). CTI is now a vital element in risk management, as highlighted by recent studies [59]. The rise in CTI reports requires automated tools for efficient creation and assessment. In network threat analysis, LLMs are critical, particularly in CTI generation and analysis, enhancing decision making processes. The CTI generation involves extracting intelligence from diverse sources, such as books, blogs, and news, and transforming it into structured reports. An example is CVEDrill by Aghaei *et al.* [60], which helps to formulate prioritized cybersecurity risk reports and predict their influence on systems. Similarly, Moskal *et al.* [39] explored the role of ChatGPT in automating responses to network threats, demonstrating its utility in handling basic attack scenarios.

LLMs improve CTI reports by providing real-time updates to monitor evolving cyber threats. Their ability to handle large datasets, identify threat patterns, and merge intelligence from multiple sources improves the efficiency and effectiveness of cybersecurity efforts.

With capabilities such as integrating live threat data, modeling attack scenarios, and offering predictive insights, LLMs play a vital role in proactive threat management, advancing CTI reporting, and supporting informed, data-driven decisions in cybersecurity.

4.1.4 Penetration Testing. Penetration testing is identified by simulated attacks on computer systems to assess their defenses and remains essential for organizations in combating cyber threats. Traditionally, penetration testing involves three key stages: information gathering, payload creation, and vulnerability exploitation. Recent advances emphasize the crucial role of LLMs in automating and improving these steps. Temara *et al.* [40] utilized LLMs to optimize data collection by retrieving key information about the target, such as IP addresses, domain info, vendor technologies, and Secure Socket Layer (SSL) / Transport Layer Security (TLS) certificates. Likewise, Sai Charan *et al.* [61] explored LLMs in the creation of malicious payloads, noting that models such as ChatGPT can produce more precise payloads, underscoring the challenge of dual use of the technology. Additionally, Happe *et al.* [62] advanced the Linux privilege escalation automation with LLM, with practical guidance for privilege escalation in penetration tests.

PentestGPT [63] is a cutting-edge automated penetration testing tool that uses LLMs, demonstrating remarkable abilities in a benchmark with 13 scenarios and 182 subtasks. Its effectiveness comes from three self-interacting modules: inference, generation,

and parsing, which enhance task management and response to complex tests. Investigations are underway into the use of LLMs for adversarial simulations and custom exploit scripts, valuable for practical testing. Future developments could include reinforcement learning and adaptive features in real-time to enhance the simulation of dynamic threats and improve the accuracy and prediction of testing. However, the risk of misuse underscores the need for responsible use and strict ethical oversight.

TAKEAWAY 2. Challenges and Open Directions. *Despite their advantages, LLMs for network security face key challenges, including vulnerability to adversarial attacks, raising concerns about robustness and security. Real-time adaptability requires efficient continuous learning while maintaining scalability in high-throughput environments. Ethical risks, particularly dual use concerns, require safeguards against malicious exploitation. In addition, enhancing explainability and transparency in security-critical applications is vital to trust and informed decision making. Addressing these issues is crucial to fully harnessing LLMs for network security.*

4.2 LLMs in Software and System Security

The complexity in software systems coupled with the increasing number of reported vulnerabilities require advanced security solutions. LLMs have emerged as powerful tools for automating vulnerability, malware detection, program fuzzing, bug repair, reverse engineering, binary analysis, and system log analysis. These tasks are important for ensuring software reliability, improving automated security assessments, and mitigating security threats. Furthermore, their ability for processing and analyzing a large amount of code and large-scale system logs outperform traditional approaches, specially in enabling real-time anomaly detection and proactive threat mitigation. In this section, we explore and examine how LLMs improve software and security tasks by introducing advanced techniques that improve accuracy, efficiency, and scalability.

TAKEAWAY 3. *LLMs enhance software and system security by enabling real-time vulnerability detection, automated bug repair, precise binary analysis, adaptive fuzzing test, reverse engineering, and system log analysis. Tools such as LATTE, Zeroleak, and Repiolt demonstrate their robustness in mitigating security threats, while fine-tuning and contrastive learning techniques enhance bug detection precision. Moreover, LLM-driven fuzz testing and binary analysis optimize vulnerability assessment methodologies, in addition, their integration into system log analysis facilitates real-time anomaly detection and proactive cyber threat response.*

4.2.1 Vulnerability Detection. The growing number of reports in Common Vulnerabilities and Exposures (CVEs) highlights the rise in software vulnerabilities, increasing the risk of breaches, and posing major economic and social threats. As such, detecting vulnerabilities has become a critical requirement for protecting software systems and ensuring societal and economic stability.

Recent advances have showcased the potential of LLMs in vulnerability detection tasks, particularly for static code analysis, where they demonstrate superior performance compared to traditional methods such as graph neural networks or rule-based approaches [64–66]. The GPT series of models stand out in their ability to identify vulnerabilities effectively [66–69]. However, challenges persist,

as LLMs can generate false positives due to subtle variations in function names, variable usage, or library modifications. Notable advances include LATTE, proposed by Liu *et al.* [?], which combines LLMs with automated binary taint analysis. LATTE addresses the limitations of traditional taint analysis methods that rely heavily on manual customization of taint propagation and vulnerability inspection rules. Impressively, LATTE identified 37 previously undiscovered vulnerabilities in real firmware, showcasing the practical impact of LLMs in vulnerability detection. Tihany *et al.* [41] further demonstrated the utility of LLMs by using them to generate FormAI, a large-scale vulnerability labeled dataset. However, their study also revealed a critical concern: more 50% of the code generated by LLMs contained vulnerabilities, raising significant security risks for automated code generation processes. These findings underscore the dual-use nature of LLMs, highlighting both their potential to advance vulnerability detection and the need for rigorous oversight to mitigate associated risks.

4.2.2 Vulnerability Repair. The rapid increase in detected vulnerabilities, coupled with the increasing complexity of modern software systems, has made manual vulnerability remediation an extremely time-consuming and resource-intensive task for security professionals [70]. Advancements in LLMs and related architectures have demonstrated promising capabilities in automating vulnerability repair tasks. The T5 model, built on an encoder-decoder framework, has shown superior performance in generating effective fixes for vulnerabilities [71, 72]. However, challenges persist in maintaining the functional correctness of repaired code [73], with LLM performance varying between programming languages, particularly with limited capabilities observed in repairing Java vulnerabilities [74]. Several innovative approaches have emerged to address these challenges. Alrashedy *et al.* [75] developed an automated vulnerability repair tool that integrates feedback from static analysis tools, enabling iterative improvement of fixes. Tol *et al.* [76] proposed ZeroLeak, a technique that uses LLM to identify and mitigate side-channel vulnerabilities in applications, demonstrating the capabilities of the models to address complex security challenges. Charalambous *et al.* [77] combined LLMs with Bounded Model Checking (BMC) to ensure the accuracy of the corrected code, mitigating functionality issues often seen after automated vulnerability fixes.

4.2.3 Bug Detection. Software and hardware failures, commonly termed bugs, can cause program malfunctions or unexpected results. Beyond affecting performance, some bugs can be manipulated by attackers to create security vulnerabilities, highlighting the critical need for bug detection to maintain software system safety and reliability. LLMs have emerged as effective tools for automating bug detection. They can generate code lines relative to the original to detect potential bugs. LLMs also utilize feedback from static analysis tools, enhancing the accuracy and precision of bug identification [78, 79]. Fine-tuning is crucial for adapting LLMs to bug detection, allowing error identification without test cases by using annotated datasets [80, 81]. Du *et al.* [44] and Li *et al.* [82] use contrastive learning to train LLMs to differentiate between correct and faulty code, increasing error detection in complex codebases. Fang *et al.* [83] introduced Represent Them All (RTA), a platform-independent representation method that combines contrastive learning and custom

fine-tuning. This technique excels in bug detection and predicts bug priority and severity, highlighting the potential of comprehensive representation methods for software quality.

4.2.4 Bug Repair. LLMs efficiently automate bug fixes by generating precise code, improving development speed but risking security concerns like vulnerabilities [84]. Addressing unresolved bugs is crucial and requires automated repairs in modern software engineering. LLMs are adept at creating repair patches for software defects. Architectures, such as Repilot [85], use encoder-decoder frameworks for accurate repairs, leveraging LLMs' grasp of code semantics to produce high-quality patches on par with traditional techniques [86]. Fine-tuning enhances LLMs' real-world repair capabilities with domain-specific datasets for better language and task handling, delivering reliable fixes. Interactive feedback systems, such as ChatGPT [87], further refine repair precision, supporting effective patch development through iterative validation and a deep understanding of software semantics.

4.2.5 Program Fuzzing. Program fuzzing, fuzz testing, or simply fuzzing, is an automated software testing technique that generates input to uncover unexpected behaviors, such as crashes. Various effective fuzzing tools have successfully detected bugs and security flaws in real systems [88]. The incorporation of LLMs into fuzzing has significantly improved the generation of test cases. Traditional methods often rely on predefined patterns, which limits their effectiveness. In contrast, LLMs can produce diverse and contextually suitable test cases for different programming languages and system features [89]. LLMs employ advanced strategies, such as repetitive and iterative querying [90], to improve the creation of test cases. These methods allow LLMs to create test cases that:

- Identify vulnerabilities: LLMs can analyze previous bug reports to create inputs that uncover similar issues in new or updated systems [91].
- Create variations: They generate various test cases related to sample inputs, ensuring coverage of potential edge cases [92].
- Optimize compilers: By examining compiler code, LLMs craft programs that trigger specific optimizations, revealing compilation process flaws [93].
- Divide testing tasks: A dual-model interaction lets LLMs separate tasks like test case generation and requirements analysis for efficient parallel processing.

The adaptability of LLMs to create intelligent and tailored test inputs has transformed fuzz testing, making it more effective at finding complex bugs. As these models advance, integrating feedback loops, real-time testing, and domain-specific knowledge will further improve system security and robustness.

4.2.6 Reverse Engineering and Binary Analysis. Reverse engineering involves analyzing artifacts such as software or hardware to discern their functionality, which can be used defensively or maliciously. It is crucial for security in vulnerability analysis, malware investigation, and intellectual property protection. Although LLMs excel at automating reverse engineering by identifying software functions and extracting essential data, Xu *et al.* [94] showcased their ability to restore variable names from binaries through iterative query propagation. Moreover, Armengol *et al.* [95] paired

type inference engines with LLMs to disassemble executables and generate source code, simplifying the binary-to-source translation process. LLMs are crucial in binary program analysis, improving comprehension of low-level code structures and behaviors. Significant progress includes:

- DexBert: Proposed by Sun *et al.* [96], this tool characterizes the binary bytecode of the Android system, improving the specific binary analysis of the Android ecosystem.
- SYMC Framework: Developed by Pei *et al.* [97], this framework uses group theory to preserve the semantic symmetry of the code during analysis. The approach has demonstrated exceptional generalization and robustness in diverse binary analysis tasks.
- Authorship Analysis: Song *et al.* [98] applied LLMs to address software authorship analysis challenges, enabling effective organization level verification of Advanced Persistent Threat (APT) malicious software.

LLMs also improve the readability and usability of decompiler outputs, helping reverse engineers interpret and understand binary files more effectively. By improving decompiler-generated code, these models reduce manual effort and increase the efficiency of reverse engineering processes. As LLM technologies advance, their integration with reverse engineering tools is expected to further enhance functionality. Future directions may include real-time disassembly, automated malware deobfuscation, and dynamic binary analysis using hybrid techniques that combine LLMs with symbolic execution or formal verification methods.

4.2.7 Malware Detection. The increasing complexity and volume of malware require advanced detection approaches. Signature-based and heuristic methods often fall short against novel camouflaged malware because of the sophisticated evasion techniques used by attackers, such as encryption, polymorphism, and metamorphism. LLMs are effective tools for detecting semantic and structural malware features, thus boosting detection. AVScan2Vec technique proposed by Joyce *et al.* [45], converts antivirus scan reports into vectors, enabling efficient handling of large malware datasets and excelling in tasks like classification and clustering. Using semantic patterns in antivirus data, this method improves scalability and accuracy, presenting a new approach to malware analysis. LLMs have been explored for their role in both detecting and analyzing malware development and prevention. As noted by Botacin [99], LLMs can combine functionalities to create modular malware components, helping to evolve malware variants. Although LLMs cannot independently generate complete malware from prompts, their ability to create elements supports countermeasure development, highlighting the importance of responsible LLM usage to avoid misuse.

4.2.8 System Log Analysis. Examining extensive log data from software systems manually is impractical due to its complexity and volume. Deep learning techniques have been proposed for anomaly detection within logs. They face challenges such as managing high dimensional, noisy data, tackling class imbalances, and achieving generality [100]. Recently, researchers have used the advanced language comprehension capabilities of LLMs to enhance anomaly detection in system logs. LLMs surpass conventional deep learning

models in both accuracy and interpretability [101]. Their adaptability can be further optimized by fine tuning for specific types of logs [102] or by adopting strategies based on reinforcement learning [103], which enables precise identification of anomalies specific to particular domains. LLMs prove beneficial in analyzing cloud server logs [57]. By integrating reasoning with log data, they effectively deduce the root causes of issues within cloud services. This demonstrates the impact of LLMs in the realm of system log analysis, providing a scalable and intelligent approach to detect and address anomalies in intricate environments.

TAKEAWAY 4. Challenges and Open Directions. *Despite their advantages, LLMs face key challenges, including a high false positive rate in detection and functional correctness in automated bug repair. Research indicates that 50% of LLM-generated code contains exploitable threats, raising concerns. Additionally, scalability challenges in high-throughput environments and ethical risks associated with the dual use of AI necessitate strict oversight and policies. Future research should focus on developing real-time security techniques, improving adversarial defenses, and integrating hybrid AI models with formal verification and reinforcement learning to enhance LLM-driven security solutions' security, reliability, and interpretability.*

4.3 LLMs in Information and Content Security

With phishing, misinformation, manipulation, and cybercrime, LLMs are powerful for enhancing information and content security through context-aware learning, fine-tuned detection models, and advanced response mechanisms that lead to enabling real-time threat identification, fraud prevention, and content moderation with greater accuracy and scalability. This section explores how LLMs redefine information and content security across distinct tasks including phishing and scam detection, harmful content identification, steganography, access control, and digital forensics. It also addresses challenges such as bias, adversarial risks, and ethical concerns.

TAKEAWAY 5. *LLMs have transformed information and content security through prompt-based learning and fine-tuned models. They accurately detect phishing attempts and scams while proactively disrupting fraud through automated scam engagement. In content moderation and online safety, LLMs enhance the identification of harmful misinformation. In steganography, they facilitate covert data embedding and advance steganalysis using few-shot learning and natural language ciphertext encoding for secure communication. LLMs also support file identification, incident response, and evidence extraction in digital forensics. Tools like PassGPT strengthen authentication by generating high-entropy passwords and evaluating their strength.*

4.3.1 Phishing and Scam Detection. Network deception involves intentionally adding false or misleading information, which threatens user privacy and property security. Typical attack methods include emails, SMS, and web ads that are used to direct users to phishing sites or harmful links [65]. LLMs can produce deceptive content on a large scale with certain prompts. However, LLMs-generated phishing emails usually have lower click-through rates than manually created ones, highlighting the limitations of automated methods [104]. LLMs are highly effective at identifying phishing emails. They use prompt-based strategies with website data or fine-tuned models suited to email traits to reach high effectiveness in phishing

detection. They also excel at spotting spam, which often includes phishing. Labonne *et al.* [46], show that LLMs significantly outperform traditional machine learning in spam detection, confirming their prowess in this area. Beyond detection, LLMs offer innovative uses against scams. As shown in [105], LLMs can mimic human interactions with scammers automatically, wasting their time and resources. This reduces the efficiency of scammers and lessens the impact of scam emails. These abilities show the potential of LLMs in enhancing phishing and scam detection, providing scalable, intelligent, and proactive cybersecurity defenses.

4.3.2 Harmful Contents Detection. Social media platforms face criticism for worsening polarization and weakening public discourse. Harmful content, often mirroring users' political opinions, can lead to toxic discussions and harmful behavior. LLMs help detect harmful content in three main areas: identifying extreme political positions [106], monitoring crime-related discourse [92], and identifying fake social media accounts or bots [107]. Although LLMs can identify this content, their interpretations often reflect their internal biases, highlighting the complexities of dealing with intricate social and political topics. Martin *et al.* [56] significantly contributed by creating a large-scale dataset of harmful and benign discourse for 13 minority groups using LLM. Validation showed that human annotators struggled to distinguish between LLM-generated and human-authored discourse, indicating LLMs' potential to improve harmful content detection, aiding efforts against toxic behavior.

4.3.3 Steganography. Anderson [108] described the embedding of secret data within regular information carriers to ensure that hidden content remains secure. This is crucial for secure communication. Recent advancements use LLMs to improve steganography and steganalysis techniques. Wang *et al.* [109] presented a novel steganalysis method using LLMs and few-shot learning. By integrating small labeled datasets with auxiliary unlabeled data, this technique addresses the scarcity of labeled samples, greatly improving detection in low-data scenarios. This marks a significant advancement in language-based steganalysis. Bauer *et al.* [110] concurrently showed how the GPT-2 model could encode ciphertext into natural language cover texts. This feature allows users to determine the outward appearance of the ciphertext, making it possible to discreetly transfer sensitive information across public forums. Such developments show the contributions of LLMs in contemporary steganography, offering secure techniques for embedding data, as well as improved methods for detecting potential misuse.

4.3.4 Access control. Access control is crucial for cybersecurity, designed to limit the actions of authorized users within a system. Despite the new authentication technologies, passwords are still the primary method of enforcing access control [111]. PassGPT, an advanced password generation system that uses LLMs, presents an innovative strategy for crafting passwords that adhere to user-specified constraints. This technique excels beyond conventional methods, including those based on Generative Adversarial Networks (GANs), by generating a more diverse collection of unique passwords. Furthermore, PassGPT improves the effectiveness of password strength evaluators, underscoring the promise of LLMs to pioneer and strengthen access control measures [112].

4.3.5 Forensics. Digital forensics play a crucial role in cybercriminal prosecution by ensuring that evidence extracted from digital devices is allowed in court [113]. This domain protects the integrity and enhances the effectiveness of cybercrime investigations. Scanlon *et al.* [55] conducted an evaluation of LLMs within the domain of digital forensics, focusing on tasks such as file identification and responding to incidents. The research concluded that, although LLMs should not be considered independent tools, they offer valuable assistance in particular forensic situations.

TAKEAWAY 6. Challenges and Open Directions. *Despite their advantages, LLMs face challenges in information and content security. In phishing and scam detection they struggle to replicate human crafted phishing strategies, requiring adversarial training and adaptive threat response. In harmful content detection, LLMs require bias mitigation and a context-aware model to improve accuracy. Their role in steganography enables covert data embedding, increasing the need for advanced steganalysis. In access control, they demand stronger authentication frameworks. While in digital forensics, stronger datasets and legal frameworks are needed. Ensuring ethical, unbiased, and secure deployment remains critical in this domain.*

4.4 LLMs in Hardware Security

SoC architectures, essential for modern computing, integrate multiple IP cores but introduce security challenges, as a weakness in any single core can compromise the entire system. Although software and firmware updates can address many issues, some vulnerabilities cannot be patched this way, necessitating rigorous security measures during the initial design phase. This section provides an overview of LLM applications in hardware security, with a particular focus on their role in detecting and mitigating vulnerabilities. Through these capabilities, LLMs demonstrate their potential to enhance the security of system-on-chip (SoC) architectures.

TAKEAWAY 7. *LLMs enhance hardware security by automating vulnerability detection, security verification, and repair mechanisms in SoC architectures. Using NLP driven analysis, they identify threats within hardware design documents and link weaknesses to Common Weakness Enumerations (CWEs) while enforcing security assertions. Additionally, LLMs enhanced hardware vulnerability repair by generating secure hardware code and leveraging large vulnerability corpora to improve patching automation. These advancements make LLMs powerful, scalable, and proactive solutions to strengthen SoC security.*

4.4.1 Hardware Vulnerability Detection. LLMs are being increasingly used to detect hardware vulnerabilities by scrutinizing security aspects embedded in hardware development documents. In an illustration of their capabilities, Meng *et al.* [43] leveraged HS-BERT, a model trained on a variety of hardware architecture documents, including Reduced Instruction Set Computing (RISC-V), Open Source Reduced Instruction Set Computing (OpenRISC), and Microprocessor without Interlocked Pipeline Stages (MIPS), facilitated the discovery of eight security vulnerabilities in the OpenTitan SoC design. This demonstrates LLMs' proficiency in dissecting complex hardware configurations to pinpoint significant security issues. Building on this advancement, Paria *et al.* [114] extended the utility of LLMs by identifying vulnerabilities within user-specified SoC designs. Their innovative strategy links identified vulnerabilities

to CWEs, makes corresponding security assertions, and enforces security measures to counteract potential threats. These cutting-edge advances demonstrate the importance of LLMs in improving hardware security by streamlining the process of detecting and averting vulnerabilities in advanced hardware infrastructures.

4.4.2 Hardware Vulnerability Repair. LLMs play an essential role in the security verification of SoC. They handle a variety of tasks, including the insertion of vulnerabilities, their assessment, and verification, as well as the development of strategies for their mitigation [29]. By utilizing comprehensive data on hardware vulnerabilities, LLMs provide suggestions for repairs, which significantly enhance the effectiveness and precision of security assessments and mitigation efforts. Nair *et al.* [115] demonstrated that LLMs can identify hardware vulnerabilities while generating code and can produce hardware code that prioritizes security. In their study, they utilized LLMs to design hardware that effectively addresses ten identified CWEs. In a complementary study, Tan *et al.* [116] constructed an extensive corpus detailing hardware security vulnerabilities and evaluated the proficiency of LLM in automating the repair of these vulnerabilities.

TAKEAWAY 8. Challenges and Open Directions. *LLMs enhance hardware vulnerability detection and repair, but face challenges in accurately interpreting complex SoC architectures, which require deep contextual understanding. The generalization of LLMs across diverse hardware designs is a limitation, as models trained on specific architectures may not effectively identify vulnerabilities in unfamiliar systems. Additionally, linking detected vulnerabilities to CWEs and generating security assertions require further refinement to ensure precision and reliability in automated mitigation efforts. Open research directions include developing specialized LLMs tailored for hardware security, improving adaptive security assertion generation, and integrating LLMs with formal verification methods to improve the accuracy of security enforcement.*

4.5 LLMs in Blockchain Security

Blockchain technology has revolutionized decentralized finance, digital identity, and secure transactions by providing a transparent ledger system. However, its growing adoption has also introduced critical security challenges, particularly in smart contract vulnerabilities and transaction anomalies. As blockchain ecosystems become increasingly complex, the need for intelligent, scalable, and proactive security mechanisms has become essential. LLMs present a transformative solution. In this section, we explore how LLMs impact blockchain security, particularly in smart contract security and the identification of transaction anomalies. Examining their substantial capabilities highlights the potential of LLMs to transform vulnerability management, thereby strengthening blockchain systems by mitigating risks and enhancing overall security.

TAKEAWAY 9. *LLMs enhance blockchain security by enabling smart contract vulnerability detection and real-time transaction anomaly identification. Frameworks like GPTLENS employ a dual-phase approach to generate and prioritize threat scenarios, reducing false positive rates and improving verification accuracy. Unlike rule-based models, LLMs dynamically identify abnormal transactions without predefined constraints, allowing for a broader spectrum of anomalies,*

improving intrusion detection efficiency, and minimizing the need for manual analysis.

4.5.1 Smart Contract Security. Blockchain applications are heavily based on smart contracts, but the construction of these contracts can result in vulnerabilities that pose significant risks, including the potential for substantial financial loss. LMs offer the potential to automate the detection of these vulnerabilities, though their performance is frequently hampered by common errors and a limited understanding of context [59, 117]. To address these challenges, frameworks such as GPTLENS [92] employ a dual-phase approach. This process begins with the creation of a wide range of potential vulnerability scenarios and then proceeds with the evaluation and prioritization of these scenarios to minimize false positive detections. Sun *et al.* [42] contributed to the field of smart contract security by incorporating LLMs into the analysis of the program to detect logical vulnerabilities effectively. They structured their approach by categorizing vulnerabilities into specific scenarios and attributes, using LLMs for detection, and confirming findings with static analysis. This development shows the ability of LLMs to improve smart contract security, but there remains a need to reduce false positives and increase accuracy.

4.5.2 Transaction Anomaly Detection. Real-time detection of intrusions within blockchain transactions is challenging due to the extensive search space and the substantial amount of manual analysis required. Conventional methods, such as reward-based and pattern-based models, rely on designated rules or predetermined patterns to pinpoint profitable or suspicious transactions. However, these methods often fail to detect a wide range of anomalies [118]. LLMs offer a versatile and generalizable solution for real-time detection of anomalies. Gai *et al.* [52] demonstrated that LLMs are capable of dynamically identifying anomalies in blockchain transactions as they occur. Unlike traditional methods, LLMs are not constrained by predefined rules or limited search spaces, which enables them to detect a broader array of abnormal transactions. This flexibility highlights the potential of LLMs to advance anomaly detection in blockchain, delivering more efficient and comprehensive solutions.

TAKEAWAY 10. Challenges and Open Directions. *Despite advancements in LLMs for blockchain security, they still face challenges such as contextual limitations, false positives, and limited understanding of contract logic. Scalability, computational efficiency, and generalization among blockchain protocols also remain significant hurdles. Future research should focus on improving the contextual reasoning of LLMs, integrating formal verification, and developing hybrid AI-driven security frameworks that combine symbolic execution, reinforcement learning, and deep learning-based anomaly detection to improve the accuracy and robustness of LLMs in blockchain security.*

4.6 LLMs in Cloud Security

The dynamic nature of cloud environments requires real-time threat intelligence, automated security enforcement, and proactive anomaly detection to ensure system integrity and data protection. The integration of LLMs into cloud security has significantly improved threat detection, security monitoring, and data leakage prevention.

By employing advanced NLP techniques, these models have improved automation and overall efficiency in addressing complex security challenges, such as misconfigurations, data leaks, compliance issues, and container security.

TAKEAWAY 11. *LLMs have transformed cloud security by enhancing threat detection, misconfiguration analysis, data leakage monitoring, container security, and compliance enforcement. In misconfiguration detection, tools like GenKubeSec provide automated reasoning and high-precision detection in Kubernetes environments. Data leakage monitoring benefits from AI-driven models such as Secure Cloud AI, which improves real-time detection. Frameworks like DevSec-GPT enhance vulnerability tracking and compliance validation. Additionally, OllaBench and PRADA demonstrate the effectiveness of LLMs by automating regulatory compliance and integrating Zero Trust security models across multi-cloud environments.*

4.6.1 Misconfiguration Detection. Misconfiguration detection plays an essential role in ensuring the security and stability of systems in cloud native settings. Recent innovations have incorporated machine learning and LLMs to effectively identify, pinpoint, and address these misconfigurations. A noteworthy advancement was made by Mitchel *et al.* [119], who developed a tool designed to utilize system call data obtained from Linux kernels operating within Kubernetes clusters. This tool applies anomaly detection techniques, including Principal Component Analysis (PCA), to detect hidden attacks that capitalize on misconfigurations. Pranata *et al.* [120] proposed a framework integrating metamorphic testing with PCA to identify misconfigurations in cloud-native applications, enhancing scalability and reducing developer effort. Malul *et al.* [47] developed GenKubeSec, an LLM-based system effective in detecting Kubernetes configuration file misconfigurations, offering automated reasoning and solutions. GenKubeSec achieved a precision of 0.990 and a recall of 0.999, surpassing traditional rule-based tools and using a UMI for standardized evaluations.

4.6.2 Data Leakage Monitoring. Data leakage poses a significant threat to security in cloud computing by compromising the confidentiality and integrity of sensitive information. Issues typically stem from misconfigured hypervisors, inadequate dashboard authentication, and insecure VM replication, especially during operations like data migration. Ariffin *et al.* [121] recommended using Wireshark for packet analysis to monitor data flows during VM migration and dashboard authentication, revealing significant risks in the lack of TLS encryption, which exposed credentials. Vaidya *et al.* [122] suggested perturbation and fake object injection techniques to better detect unauthorized access by embedding decoy elements. Advanced AI-driven frameworks are revolutionizing data leakage monitoring. Kasula *et al.* [48] introduced “Secure Cloud AI,” a hybrid model combining Random Forest and LSTM networks for real-time anomaly detection. Their method achieved 94.78% accuracy in identifying malware and efficiently classifying network traffic anomalies, demonstrating the necessary scalability and adaptability for dynamic cloud settings. However, challenges persist in scaling for large cloud installations, detecting complex multi-layered attacks, and achieving cross-platform compatibility. Future work should enhance AI integration, real-time encryption, and anomaly detection to strengthen defenses against data leakage.

4.6.3 Container Security. Integrating LLMs into container security has shown significant progress in securing native cloud environments by automating vulnerability detection, optimizing container management, and improving pipeline integrity. For example, the DevSec-GPT [49] framework uses Meta's Llama2 LLM to create Pipeline Bills of Materials (PBOMs) for container security, analyzing vulnerabilities and development data to ensure comprehensive tracking and prevent supply chain attacks via blockchain traceability. In runtime and security management, LLMs improve real-time anomaly detection by analyzing logs and configurations. Lanka *et al.* [38] employed LLMs to analyze data from the decoy system, detecting malicious patterns and attacker tactics. The RAG model quickly identified threats in containers by matching commands with adversary data. Additionally, LLMs facilitated compliance checks through JSON schema generation for vulnerability scans on platforms such as GitHub Actions and Kubernetes. Automated documentation updates helped meet regulatory standards, while blockchain features, such as NFT tokenization, improved data provenance and auditability.

4.6.4 Compliance Enforcement. LLMs transform cloud security compliance by automating regulations and boosting efficiency. Nguyen *et al.* [50] introduced OllaBench, an evaluation tool with 24 cognitive behavioral theories to test the reasoning of LLM in compliance. OllaBench revealed that GPT-4o and Claude are effective in regulatory automation, providing key information to compliance teams. Henze *et al.* [123] proposed PRADA, a cloud storage system using LLMs for transparent data management. It achieves compliance by tagging and routing data by attributes, addressing issues like localization and encryption in distributed systems. PRADA notably improved the management of Data Handling Requirements (DHRs), providing a scalable compliance solution for multi-cloud environments. Dye *et al.* [124] highlighted the critical role of LLMs in integrating Zero Trust Architectures (ZTA) with Attribute-Based Access Control (ABAC) systems for compliance. Cloud providers such as AWS utilized LLMs to automate policy generation, monitor privilege escalation, and adhere to Federal Risk and Authorization Management Program (FedRAMP) and National Institute of Standards and Technology (NIST) standards, enhancing least privilege access control and increasing security in hybrid and public clouds.

TAKEAWAY 12. Challenges and Open Directions. Despite progress in applying LLMs to cloud security, major challenges persist: scalability in large environments, detecting multilayered attacks, and maintaining cross-platform compatibility. Data leakage prevention demands tighter AI integration, real-time encryption, and adaptive anomaly detection. Container security calls for efficient, real-time anomaly detection and compliance checks. Compliance enforcement remains hindered by real-time auditing, localization, and evolving regulations. Future research should target self-learning models, Zero Trust architectures, and automated security frameworks to improve scalability, adaptability, and resilience.

4.7 LLMs in Incident Response and Threat Intelligence

As cyber threats become more advanced, traditional methods of incident response and threat intelligence often struggle with scalability, speed, and accuracy. LLMs are emerging as transformative tools in incident response and threat intelligence by automating cybersecurity data analysis, enhancing decision making, and improving the speed and accuracy of threat detection. Their advanced pattern recognition allows security teams to identify anomalies and threats within vast amounts of unstructured data, reducing manual effort and response time.

TAKEAWAY 13. *LLMs are reshaping incident response and threat intelligence through automation and improved precision. SEVENLLM cuts down false positives in SIEMs via multitask learning for smarter alerting. HunGPT boosts threat analysis with interpretable anomaly detection and structured knowledge extraction. DISASLLM advances malicious code detection, while MALSIGHT generates readable summaries of malware behavior. Together, these tools streamline reverse engineering and speed up intelligence workflows, showing LLMs' growing role in cybersecurity.*

4.7.1 Alert Prioritization. LLMs are revolutionizing alert prioritization in incident response and threat intelligence, supporting efficient context-driven security operations. Molletti *et al.* [125] showcased the ability of LLM agents to manage large security data sets, alleviate alert fatigue, and highlight key threats through advanced NLP. These agents seamlessly integrate with SIEM systems, boosting threat detection and response. Ji *et al.* [51] introduced SEVENLLM, an optimized LLM framework using selected bilingual data. Specializing in analyzing and prioritizing security alerts, SEVENLLM employs multitask learning and SEVENLLM Bench evaluations. This approach improves Indicator of Compromise (IoC) detection and refines alert prioritization by evaluating severity and impact. LLMs significantly improve alert management by reducing false positives and enabling quick responses; however, challenges persist, such as model interpretability and adaptation to evolving threats, underscoring their impact on modern cybersecurity.

4.7.2 Automated Threat Intelligence Analysis. Incorporating LLMs into cybersecurity has greatly improved automated threat analysis by reducing the manual work involved with unstructured CTI reports. Tseng *et al.* [54] presented an AI agent using LLMs such as GPT-4 to automatically extract IoC from CTI reports and create regex patterns for Security Information and Event Management (SIEM) systems. The agent also constructs relationship graphs to depict connections between IoCs, streamlining incident response, and reducing dependency on human intervention. Using LLMs like Llama 2 and Mistral 7B, Fieblinger *et al.* [126] generated Knowledge Graphs (KGs) from CTI reports. Their approach involves fine-tuning and prompt engineering to derive triples, which are employed in link prediction. KGs offer better-structured data representation, enhancing decision making and threat prediction. HuntGPT, presented by Ali and Kostakos [127], showcases the capabilities of LLM in cybersecurity. This dashboard combines GPT-3.5 with XAI frameworks such as SHAP and LIME, delivering understandable threat intelligence. It emphasizes detected anomalies and clarifies their context, boosting trust and enabling dynamic cybersecurity.

processes. These innovations highlight the role of LLM in automating threat tasks, speeding up response, and enhancing detection accuracy in security operations.

4.7.3 Threat Hunting. LLMs are revolutionizing threat hunting by automating the analysis of intricate cybersecurity data. Schwartz *et al.* [128] developed LLMCloudHunter, a framework using LLMs like GPT-4o to create OSCTI detection rules. It achieved 92% precision and 98% recall, improving rule generation for SIEM systems and boosting threat detection in cloud environments. Mitra *et al.* [129] presented LOCALINTEL, a system that merges global threat intelligence (e.g., CVE, CWE) with local knowledge using RAG, reducing Security Operations Center (SOC) analysts, LLMs achieved a RAGAS score of 0.9535. Their ability to automate threat detection and response promises scalable, accurate, proactive cybersecurity. Future work should improve real-time adaptability and integration with evolving threats.

4.7.4 Malware Reverse Engineering. The integration of LLMs into malware reverse engineering has transformed the ability to analyze complex binary malware. DISASLLM, introduced by Rong *et al.* [130], leverages an LLM-based classifier fine-tuned in assembly code to efficiently identify valid instruction boundaries within hidden executables. This approach significantly outperforms traditional disassembly tools by integrating the semantic understanding of LLMs, thereby improving the detection of malicious code segments. Complementing this, MALSIGHT, as proposed by Lu *et al.* [131], employs LLMs like MalT5 to iteratively generate human-readable summaries of malware functionality from malicious source code and benign pseudocode. This method improves the usability, accuracy, and completeness of malware behavior descriptions, effectively bridging the semantic gap introduced by obfuscation techniques. Furthermore, Patsakis *et al.* [132] demonstrated the capabilities of LLMs such as GPT-4 to clarify real-world malware campaigns such as Emotet. The study shows that these models can obtain actionable information, like C2 server configurations, from heavily obfuscated scripts. Although local LLMs have accuracy issues, cloud-based tools like GPT-4 excel in understanding malware payloads.

TAKEAWAY 14. Challenges and Open Directions. *Despite their influence in this domain, LLMs face key challenges. In alert prioritization, contextual accuracy is required to minimize false positives and misclassifications in SIEM systems. Threat intelligence automation struggles with processing hidden data and adapting to new patterns in real time. Effective threat hunting requires deeper integration with security monitoring data and retrieval-augmented intelligence. Future research must focus on improving contextual LLM reasoning, integrating Explainable Artificial Intelligence (XAI) for transparent threat analysis, and improving real-time adaptability.*

4.8 LLMs in IoT Security

The rapid expansion of IoT ecosystems has introduced significant security challenges due to resource constraints, diverse architectures, and evolving cyber threats. Traditional security solutions struggle with scalability, real-time anomaly detection, and firmware vulnerability management, making IoT devices prime targets for cyberattacks. LLMs have emerged as transformative solutions that enable automated threat detection and efficient processing. The

following subsections explore the role of LLMs in enhancing IoT security, focusing on firmware vulnerability detection, behavioral anomaly detection, and automated threat report summarization.

TAKEAWAY 15. *LLMs have significantly advanced IoT security by enhancing firmware vulnerability detection, behavioral anomaly detection, and automated threat report summarization. Frameworks such as LLM4Vuln improve firmware vulnerability analysis by integrating retrieval-augmented generation (RAG) and prompt engineering, enhancing reasoning across various programming languages. The UVSCAN framework employs NLP-driven binary analysis, excelling in detecting API misuse. In behavioral anomaly detection, an Intrusion Detection System Agent (IDS-Agent) combines reasoning pipelines, memory retrieval, and external knowledge to detect zero-day attacks. These advancements integrate LLM reasoning, NLP frameworks, and efficient learning models, providing robust, scalable, and resource-efficient solutions for IoT firmware vulnerabilities.*

4.8.1 Firmware Vulnerability Detection. LLMs have notably improved IoT firmware vulnerability detection by translating abstract security needs into actionable analyzes. Sun *et al.* [42] introduces LLM4Vuln, which separates the reasoning abilities of LLMs for accurate vulnerability detection in languages like Solidity and Java. The framework improves performance using advanced prompt engineering and RAG to integrate current vulnerability knowledge. Zhao *et al.* [133] noted that the UVSCAN framework supplements this by translating high-level API specifications into binary-level analysis in IoT firmware using NLP-driven methods. It excels at identifying API misuse, causality errors, and return value issues, offering scalability across various architectures, such as RISC and MIPS. Furthermore, Li *et al.* [134] introduce Binary Neural Networks (BNNs) to enhance resource efficiency in IoT settings, allowing lightweight on-device learning for vulnerability detection with maintained accuracy.

4.8.2 Behavioral Anomaly Detection. With the rapid increase in IoT devices, security challenges have intensified, making the detection of behavioral anomalies vital. LLMs have transformed this field by leveraging reasoning and contextual understanding. Li *et al.* [135] introduced IDS-Agent, an intrusion detection system powered by LLMs, which combines reasoning pipelines, external knowledge, and memory retrieval to detect malicious traffic accurately. In benchmarks such as the Army Cyber Institute for Cybersecurity of Things 2023 (ACI-IoT'23) and the Canadian Institute for Cybersecurity Internet of Things 2023 (CIC-IoT'23), IDS-Agent achieved a 61% recall in identifying zero-day attacks, outperforming traditional machine learning methods and providing better interpretability. Su *et al.* [136] applied GPT-4o and domain-adapted BERT to IoT time series data, detecting behavioral shifts for threat identification. These models scaled efficiently in resource-limited IoT networks, minimized false alarms, and provided actionable insights to address incidents.

4.8.3 Automated Threat Report Summarization. The rise of IoT devices has increased the volume and complexity of threat data, emphasizing the need for automated processing. LLMs effectively transform unstructured IoT threat reports into actionable insights. Feng *et al.* [137] developed IoTShield, an LLM-based framework for assessing IoT vulnerability reports, extracting critical details such

as exploit parameters and severity ratings, and generating custom signatures for IDS to enhance defense accuracy. Building on this, Baral *et al.* [138] integrated XAI with LLMs to produce personalized threat reports tailored to the expertise of analysts. Their system balances technical complexity with user-friendliness, improving decision making and response efficiency. These advancements underscore the role of LLMs in optimizing IoT threat intelligence processes and addressing challenges related to scale, complexity, and interpretability in modern IoT security environments.

Security Domains	Security Tasks	Total
Network Security	Web fuzzing (3) Traffic and intrusion detection (3) Cyber threat analysis (3) Penetration test (4)	13
Software and System Security	Vulnerability detection (9) Vulnerability repair (8) Bug detection (7) Bug repair (4) Program fuzzing (6) Reverse engineering and binary analysis (5) Malware detection (2) System log analysis (5)	46
Information and Content Security	Phishing and scam detection (4) Harmful contents detection (4) Steganography (3) Access control (2) Forensics (2)	15
Hardware Security	Hardware vulnerability detection (2) Hardware vulnerability repair (3)	5
Blockchain Security	Smart contract security (4) Transaction anomaly detection (3)	7
Cloud Security	Misconfiguration detection (3) Data leakage monitoring (3) Container security (2) Compliance enforcement (3)	11
Incident Response and Threat Intel.	Alert prioritization (2) Automated threat intelligence analysis (3) Threat hunting (2) Malware reverse engineering (3)	10
IoT Security	Firmware Vulnerability Detection (3) Behavioral Anomaly Detection (2) Automated Threat Report Summarization (2)	7

Table 3: Security Domains and Related Tasks

Table 3 categorizes key cybersecurity domains along with their associated tasks and corresponding counts, highlighting the potential of LLMs to enhance security operations. We examine the applications of LLMs across 32 security tasks spanning eight distinct security domains. Furthermore, this classification serves as a foundational reference for the role of LLMs in cybersecurity research, facilitating more adaptive and intelligent security.

TAKEAWAY 16. Challenges and Open Directions. *Firmware vulnerability detection still requires improved cross-architecture generalization to enhance accuracy in various IoT environments. Behavioral anomaly detection struggles with reducing false positives in complex IoT ecosystems while maintaining efficient resource utilization in constrained environments. Automated summarization of threats and reports demands better contextual understanding to generate actionable insights. Future research should focus on developing energy-efficient*

LLM models for IoT, creating lightweight architectures, and improving real-time intrusion detection through adaptive learning.

5 VULNERABILITIES AND DEFENSES

Existing researches works have classified the vulnerabilities and challenges associated with LLMs into distinct domains. Security and privacy risks include misinformation [139], trustworthiness concerns [140], hallucinations [141], and significant resource consumption [142]. These risks focus on the need for robust measures to mitigate potential security attacks. Security is mainly intended to protect systems by preventing unauthorized access, modification, malfunction, or denial of service to legitimate users during regular operations [143]. On the other hand, privacy aims to protect personal information and ensure that individuals retain control over who can access their sensitive data [144]. In this work, our objective is to systematically examine LLM vulnerabilities by focusing on security attacks through a goal-oriented approach. The subcategories include Backdoor Attacks, Data Poisoning, Prompt Injection, and Jailbreaking, each paired with the corresponding defense techniques. This structure highlights the challenges and innovative countermeasures in securing LLMs, providing a clear framework for understanding and addressing these risks. The following subsections first explore key primary defense techniques against security attacks on LLMs, followed by an analysis of major security attack types with their defense methods, focusing on a secure and reliable LLM deployment.

5.1 Defenses Against Attacks on LLMs

This subsection outlines key defense techniques proposed to enhance the robustness and safety of LLMs, belonging to three main categories. The first focuses on preventing LLMs from generating harmful output using a set of rules and constraints at the input or output levels. Red teaming and content filtering play a critical role in intercepting and blocking potentially harmful interactions before they occur, ensuring that LLMs adhere to ethical and safety standards. The second category is concerned with modifying LLMs internal mechanisms or representations to improve LLM robustness and safety. Safety fine-tuning and model merging are key techniques within this category, working to make the model more resilient to adversarial attacks and misalignments, while strengthening its safety protocols through model optimization. The third category integrates the strengths of these defense strategies to offer a more comprehensive protection framework.

(1) **Red Team Defenses.** This is an effective technique for simulating real-world attack scenarios to identify LLM vulnerabilities. The process begins with attack scenario simulation, where researchers test LLM responses to issues such as abusive language. This is followed by test case generation using classifiers to create scenarios that help eliminate harmful outputs. Finally, the attack detection process assesses the susceptibility of LLMs to adversarial attacks. Continuous updates to security policies, refinement of procedures, and strengthening of technical defenses ensure that LLMs remain robust and secure against evolving threats. Ganguli *et al.* [145] proposed an efficient AI-assisted interface to facilitate large-scale Red Team data collection for

further analysis. Additionally, their research explored the scalability of different LLM types and sizes under Red Team attacks and their ability to reject various threats.

Challenges. This technique poses several challenges, such as its resource-intensive nature and the need for skilled experts to effectively simulate complex attack strategies [146]. Red Teaming is still in its early stages with limited statistical data. However, recent advancements, such as leveraging automated approaches for test case generation and classification, have improved scalability and diversity in Red Teaming efforts [147, 148].

- (2) **Content Filtering.** This technique encompasses input and output filtering to protect the integrity and appropriateness of LLM interactions by identifying harmful inputs and outputs. Recent advancements have introduced two key approaches: rule-based and learning-based systems. Rule-based systems rely on pre-defined rules or patterns, such as detecting adversarial prompts with high perplexity values [149]. To neutralize semantically sensitive attacks, Jain *et al.* [150] utilize paraphrasing and re-tokenization techniques. On the other hand, learning-based systems employ innovative methods, such as an alignment-checking function designed by Cao *et al.* [151] to detect and block alignment-breaking attacks, enhancing security.

Challenges. Despite advancements, this technique still suffers from significant limitations that hinder its effectiveness and robustness. One key issue is the evolving nature of adversarial prompts designed to bypass detection mechanisms. Rule-based filters, while simple, struggle to remain effective against increasingly sophisticated attacks. Learning-based systems, which rely on large and diverse datasets, may fail to fully capture harmful content variations and struggle with balancing sensitivity and specificity. This imbalance can lead to false positives (flagging benign content) or false negatives (missing harmful content). Additionally, the lack of explainability in machine learning-based filtering decisions complicates transparency and acceptance, underscoring the need for more interpretable and trustworthy models.

- (3) **Safety Fine-Tuning.** This widely used technique customizes pre-trained LLMs for specific downstream tasks, offering flexibility and adaptability. Recent research, such as that presented by Xiangyu *et al.* [152], found that fine-tuning with a few adversarially designed training examples can compromise the safety alignment of LLMs. Additionally, even benign, commonly used datasets can inadvertently degrade this alignment, highlighting the risks of unregulated fine-tuning. Researchers have proposed data augmentation and constrained optimization objectives to address these challenges. Data augmentation enriches the training dataset with a diverse range of samples, including adversarial and edge cases, helping the model generalize safety principles more effectively. Constrained optimization applies additional loss functions or restrictions during training to guide the model toward prioritizing safety without sacrificing task performance. Bianchi *et al.* [153] and Zhao *et al.* [154] validate the effectiveness of this technique, suggesting that incorporating a small number of safety-related examples during fine-tuning improves the safety of LLMs without reducing their practical effectiveness. Together, these defensive techniques enhance safety throughout the generative process, reducing

the risk of adversarial attacks and unintentional misalignments while preserving the adaptability and effectiveness of fine-tuned LLMs for real-world tasks.

Challenges. While widely used, integrating safety-related examples into fine-tuning can limit the generalization of LLMs to benign, non-malicious inputs or degrade performance on specific tasks. Additionally, the reliance on manually curated safety examples or prompt templates demands considerable human expertise and is subject to subjective biases, leading to inconsistencies and affecting model reproducibility across different use cases. Moreover, fine-tuning itself can both enhance safety through the introduction of safety-focused examples and expose the model to new vulnerabilities when adversarial examples are incorporated.

- (4) **Model Merging.** The model merging technique combines multiple models to enhance robustness and improve performance [155]. This method complements other defense strategies by significantly strengthening the resilience of LLMs against adversarial manipulations. By leveraging the diversity of multiple models, it creates a more robust system capable of handling adversarial inputs while maintaining generalization across various tasks. When merging models, different fine-tuned models initialized from the same pre-trained backbone share optimization trajectories while diverging in specific parameters tailored to different tasks. These diverging parameters can be merged through arithmetic averaging, allowing the model to generalize better over domain inputs and perform multi-task learning. This idea has been proven effective in fields like Federated Learning (FL) and Continual Learning (CL), where model parameters from different tasks are combined to mitigate conflicts. Zou *et al.* [156] and Kadhe *et al.* [157] applied model merging techniques to balance unlearning unsafe responses while minimizing over-defensiveness.

Challenges. Despite its potential, model merging faces significant challenges due to a lack of deep theoretical investigation in two key areas. First, the relationship between adversarially updated model parameters derived from unlearning objectives and the embeddings associated with safe responses remains unclear. There is a risk that adversarial training with a limited set of harmful response texts could lead to overfitting, making the model more susceptible to new, unseen jailbreaking prompts. This limited approach may fail to generalize effectively to novel adversarial threats, undermining the robustness of LLMs. Second, controlling the over-defensiveness of merged model parameters presents a significant challenge. While merging aims to improve resilience, it does not provide a clear methodology for preventing over-defensive behavior, where the model may excessively restrict certain types of outputs, limiting its usability and flexibility. This lack of control could hinder the model's ability to balance safety with task performance, as overly defensive behavior may result in missed or overly cautious responses, impacting user experience and task accuracy. These challenges highlight the need for further research into the theoretical foundations of model merging to ensure its effectiveness and versatility as a defense mechanism.

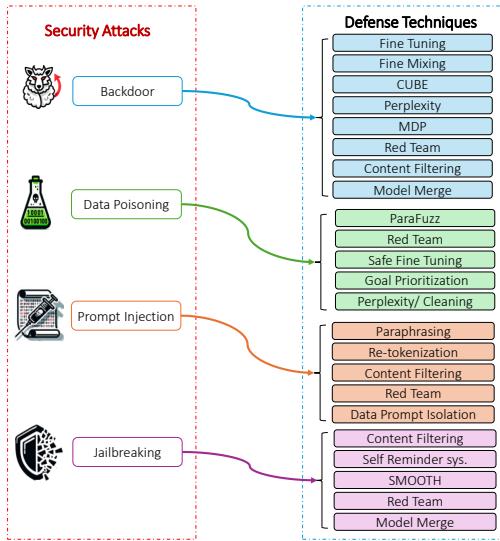


Figure 3: Security Attacks and Defense Techniques

5.2 Adversarial Attacks

Adversarial attacks are a key vulnerability in LLMs, involving input manipulation to trigger errors or unintended output [70]. These attacks exploit the sensitivity of a model to minor input changes to deceive it. In DNNs, such attacks disrupt operations by altering input data, and this also applies to LLMs, with potential effects such as spreading misinformation or creating biased content [158].

In the following, we review those attacks and defenses. At a high level, Table 4 highlights various defense techniques proposed to mitigate LLMs against four key security vulnerabilities (backdoor attacks, jailbreaking, data poisoning, and prompt injection). The authors below explore these techniques, which are used to safeguard LLMs from security attacks by addressing their limitations and challenges. In more detail, Figure 3 presents a structured overview of security attacks targeting LLMs alongside corresponding defense techniques. By integrating these advanced security frameworks, LLMs can achieve enhanced robustness and reliability for ensuring safe and trustworthy deployment in real-world applications.

5.2.1 Data Poisoning. Adversaries pose a serious threat to LLMs by deliberately altering training datasets. Through the insertion of misleading samples, they create subtle distortions that bias the model, leading to errors in prediction and decision-making [159]. Adversaries can manipulate Deep Neural Networks (DNNs) to serve malicious purposes by corrupting training data. Research indicates that poisoned data can be discreetly added to datasets during training or fine-tuning of Language Models (LMs), exploiting vulnerabilities in data pipelines [160]. These risks are particularly pronounced when using external or unverified dataset sources, highlighting the need for strict data curation and validation to mitigate such threats.

Defense Techniques. Several defense techniques have been employed to protect LLMs from poisoning attacks, including data validation, filtering, cleaning, and anomaly detection [161]. Yan *et al.* [162] introduced ParaFuzz, a framework designed to detect poisoned samples in NLP models. This technique employs fuzzing,

a software testing methodology, to identify poisoned samples as outliers by analyzing the interpretability of model predictions. ParaFuzz provides a robust method for filtering malicious data during training, noting that poisoned data often diverges from benign distributions and takes longer for a model to learn. Additionally, dataset curation techniques, as highlighted by Contiella *et al.* [163], emphasize the importance of removing near-duplicate poisoned samples, identifying known triggers, and isolating anomalies in training datasets. This approach has proven effective against attacks such as AutoPoison [164] and TrojanPuzzle [165]. Fine-pruning has also emerged as an effective defense against data poisoning attacks [166], while perplexity filtering and query rephrasing are specifically employed to mitigate white-box attacks such as Agent-Poison [167]. Despite these advancements, continuous research is necessary to refine and optimize defenses against evolving threats in modern LLMs.

5.2.2 Backdoor Attacks. In a backdoor attack, poisoned samples are introduced into a model to embed hidden malicious functionality. These attacks allow the model to perform normally on benign inputs but behave maliciously on specific, poisoned inputs. Backdoor attacks in LLMs can be categorized as input-triggered, prompt-triggered, instruction-triggered, and demonstration-triggered [168]. Adversaries use techniques such as injecting triggers into training data, modifying prompts to elicit malicious output, exploiting fine-tuning processes with poisoned instructions, or subtly altering demonstrations to manipulate model behavior while embedding hidden vulnerabilities without detection [169].

Defense Techniques. To counter backdoor attacks in LMs, various advanced mitigation techniques have been proposed. One such approach is *Fine Mixing*, introduced by Zhang *et al.* [170], which employs a two-step fine-tuning process that merges backdoor-optimized weights with pre-trained weights, followed by refinement on a clean dataset. This method also integrates *Embedding Purification* (E-PUR) to neutralize backdoors in word embeddings, improving model robustness. Another technique, *CUBE* (Clustering-Based Unsupervised Backdoor Elimination) by Cui *et al.* [171], leverages the HDBSCAN density clustering algorithm to identify and separate poisoned samples from clean ones based on distinct clustering patterns. Additionally, *Masking Differential Prompting* (MDP) by Xi *et al.* [172] offers an efficient and adaptable defense for prompt-based LMs by exploiting the increased sensitivity of poisoned samples to random masking, which causes significant variations in their probability distributions. While these techniques enhance security, further research is needed to assess their effectiveness against advanced backdoor attacks in LLMs such as GPT-4 and Llama-3.

5.3 Prompt Hacking

Prompt hacking involves manipulating input prompts to influence the output of LLMs. By crafting precise and intentional prompts, attackers aim to direct model responses toward specific objectives, which may include generating unintended or harmful outcomes. Since LLMs operate through interaction-based systems where user queries drive their outputs, carefully designed prompts can exploit the model's underlying mechanisms, overriding safeguards and producing misleading, malicious, or unexpected results.

Technique	Backdoor	Jailbreaking	Data Poisoning	Prompt Injection
ParaFuzz	✗	✗	✓	✗
CUBE	✓	✗	✗	✗
Masking Differential Prompting	✓	✗	✗	✗
Self Reminder System	✗	✓	✗	✗
Content Filtering	✓	✓	✓	✓
Red Team	✓	✓	✓	✓
Safety Fine-Tuning	✓	✗	✓	✗
A Goal Prioritization	✗	✓	✗	✗
Model Merge	✓	✗	✓	✗
Prompt Engineering	✓	✓	✓	✓
Smooth	✗	✓	✗	✗

Table 4: Comparison of Techniques for Mitigating LLM Vulnerabilities

5.3.1 Jailbreaking Attacks. Jailbreaking attacks involve bypassing software restrictions imposed by manufacturers or service providers, granting users elevated access to system functionality. While commonly associated with Apple’s iOS [173], similar practices exist for Android and other systems. Jailbreaking grants users privileged access to core functions and the file system, allowing for unauthorized application installations, bypassing regional locks, and performing advanced system manipulations [174]. However, it introduces significant risks, such as compromised security, loss of functionality, and potentially irreversible damage to the device.

Defense Techniques. Several defense techniques have been developed to mitigate jailbreaking attacks on LLMs. Kumar *et al.* [175] introduced *substring safety filtering*, which analyzes and filters input prompts to block harmful or unintended responses, providing robust defense despite its increased complexity for longer inputs. Wu *et al.* [176] developed a *self-reminder system* that directs LLMs toward safe behaviors, improving context-specific responses and reducing jailbreak success rates, particularly in role-playing scenarios. Jin *et al.* [177] proposed a *goal prioritization* method that prioritizes safety over utility in response generation, thereby reducing harmful content risks. Additionally, Robey *et al.* [178] introduced the *Smooth LLM* framework, which applies randomized smoothing by perturbing input prompts and aggregating outputs to lower the success rate of instruction-based attacks on models such as Llama-2 and Vicuna, enhancing model defenses against emerging threats.

5.3.2 Prompt Injection. Prompt injection manipulates LLMs to generate attacker-desired outputs by bypassing safety mechanisms [179]. Carefully crafted prompts enable adversaries to override original commands or execute malicious actions. This vulnerability facilitates harmful content generation, including data leakage, unauthorized access, hate speech, disinformation, and other security breaches [180]. In prompt injection attacks, adversaries may directly instruct the LLM to bypass filtering mechanisms or process compromised inputs. Additionally, attackers can pre-inject harmful prompts into web content, which the LLM may inadvertently process, making these attacks difficult to detect and mitigate.

Defense Techniques. Prompt injection defenses are categorized into prevention-based and detection-based methods, with ongoing research efforts enhancing their effectiveness. Prevention-based defenses, as described in [181], seek to block injected tasks through techniques such as *paraphrasing*, *re-tokenization* [150], and *data*

prompt isolation [182]. Paraphrasing disrupts the sequence of injected data, while re-tokenization breaks down infrequent tokens in compromised prompts, mitigating malicious instructions. Detection-based defenses, such as those proposed by Wang *et al.* [109], assess prompt integrity through response-based or prompt-based evaluations. *Perplexity-based detection* [183], for instance, identifies compromised triggers by analyzing quality degradation and increased perplexity. Despite these advancements, Liu *et al.* [184] found that traditional prevention and detection techniques remain inadequate against optimization-based attacks, such as *Judge Deceiver* [185], highlighting the need for continuous innovation in this field.

Table 5 presents an analysis of security techniques employed across various approaches to safeguard LLMs from security attacks. This analysis highlights the diverse methodologies adopted in existing research and provides insights into their effectiveness and robustness in protecting LLMs.

6 LIMITATIONS AND FUTURE DIRECTIONS

While LLMs have shown significant potential in addressing cybersecurity challenges, several inherent limitations hinder their broader adoption and effectiveness in security tasks. One major limitation is the lack of interpretability, as the black-box nature of LLMs prevents users from understanding how models make security-critical decisions. This undermines trust and transparency, which are essential for deployment in sensitive domains. The lack of transparency is particularly concerning due to the increased risks posed by AI-generated content (AIGC) [187], including privacy breaches, the spread of misinformation, and the production of vulnerable code [188]. Moreover, LLMs’s use in cybersecurity domains such as network, hardware, blockchain, and content security is constrained by the lack of high-quality, domain-specific datasets needed for effective fine-tuning [189]. Future research should focus on developing explainability tools to clarify model decisions, collaborating with domain experts to curate relevant datasets, and refining models to integrate cybersecurity-specific knowledge. Expanding LLM capabilities to handle multimodal inputs such as voice, images, and videos would enhance their contextual understanding in security settings [94]. Addressing these challenges will enable LLMs to better support cybersecurity efforts, offering deeper insights and facilitating the development of secure, automated solutions.

Table 5: Comprehensive Analysis of Security Defense Techniques and Related Approaches. Abbreviations: RT (Red Team), CF (Content Filtering), SFT (Safty Fine Tuning), MM (Model Merge), CE (CUBE), GP (Goal Prioritization), FM (Fine Mixing), P (Perplexity), PF (ParaFuzz), SF (Substring Filtering), S (Smooth), DPI (Data Prompt Isolation), PH (Paraphrasing), SR (Self Reminder), C (Cleaning), CU (Curation), MDP (Masking Differential Prompting), R (Re-tokenization), PI (Prompt Injection).

Author	Ref.	RT	CF	SFT	MM	CE	GP	FM	P	PF	SF	S	DPI	PH	SR	C	CU	MDP	R	PI
Alone <i>et al.</i>	[149]	✗	✓	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Bianchi <i>et al.</i>	[153]	✗	✗	✓	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
Cao <i>et al.</i>	[151]	✗	✓	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Contiella <i>et al.</i>	[163]	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗
Cui <i>et al.</i>	[171]	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Ribeiroet <i>et al.</i>	[148]	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Ganguali <i>et al.</i>	[145]	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Gonen <i>et al.</i>	[183]	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Jain <i>et al.</i>	[150]	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✓	✗
Jin <i>et al.</i>	[177]	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Kupmar <i>et al.</i>	[175]	✗	✓	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
Liu <i>et al.</i>	[186]	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗
Perez <i>et al.</i>	[147]	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Xiangyu <i>et al.</i>	[152]	✗	✗	✓	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Robey <i>et al.</i>	[178]	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
Schulhoff <i>et al.</i>	[182]	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓	✓	✗	✗	✗	✗	✗	✗
Shan <i>et al.</i>	[161]	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗
Wang <i>et al.</i>	[109]	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗
Wu <i>et al.</i>	[176]	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗
Xi <i>et al.</i>	[172]	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗
Yan <i>et al.</i>	[162]	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
Zhang <i>et al.</i>	[170]	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗
Zhao <i>et al.</i>	[154]	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Zou <i>et al.</i>	[156]	✗	✗	✗	✓	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗

Despite substantial advancements in understanding security vulnerabilities in machine learning models, research on mitigating backdoor attacks in LLM-based tasks, such as text summarization and generation, remains limited [190]. Mitigating backdoor attacks is crucial for ensuring robust defenses and secure LLM deployment. While poisoning attacks on ML models have been well studied [191], advanced techniques such as ProAttack [133] and BadPrompt [192] are not yet fully addressed.

To enhance LLM security and build robust systems, further research is required across various tasks and architectures. Current defense measures, such as dataset cleansing [193] and anomaly detection, often disrupt the development process by inadvertently removing critical data. Similarly, methods like early stopping after specific training epochs offer moderate protection but frequently lead to reduced model performance [194]. Adaptive defense mechanisms are necessary to strike a balance between securing models and maintaining their utility.

Despite their potential, LLMs remain inherently vulnerable to security risks, posing critical challenges for their application in cybersecurity tasks. One key avenue for future research is enabling LLMs to autonomously detect and resolve vulnerabilities within their architectures. Such self-repair capabilities would enhance resilience while reducing dependence on external restrictions. Achieving this requires a dual focus: automating cybersecurity tasks using LLMs while concurrently implementing robust self-protection

mechanisms to mitigate model-specific risks. Current research highlights significant shortcomings in LLM defenses. For example, while safety features in existing models, such as those in ChatGPT, can prevent simple attacks, multi-step exploits continue to compromise these systems [56]. Moreover, newer AI systems, such as Bing AI, have demonstrated even greater susceptibility to advanced attack strategies [195]. Techniques inspired by human reasoning, such as self-reminder mechanisms [196], have been proposed to mitigate these vulnerabilities; however, their effectiveness in handling complex queries remains underexplored. Addressing this gap requires deeper insights into how these mechanisms influence model reasoning and further refinements to optimize defense strategies without compromising performance.

Defending LLMs against vulnerabilities requires addressing both intrinsic weaknesses and external adversarial threats. Given the impracticality of manually auditing training data sets, alternative strategies have been used, such as personal information filtering and restrictive terms of use, to mitigate the inclusion of sensitive content. Advanced techniques like MDP [172] have shown promise in countering earlier backdoor attacks. However, these methods struggle with complex NLP tasks, such as paraphrasing and semantic similarity, and do not fully address emerging threats like BadPrompt [192] and BToP [197]. The lack of comprehensive evaluation metrics, such as perplexity, further complicates the assessment of attack and defense strategies. Additionally, the high computational

demands of LLM training in federated learning (FL) environments create additional obstacles to scalable and adaptive defenses.

Organizations such as OpenAI implement defense measures, yet these do not fully eliminate risks like jailbreaking. Malicious datasets and prompts still allow attackers to bypass security mechanisms and generate harmful content [198]. These persistent vulnerabilities underscore the urgent need for flexible and scalable defense strategies that can adapt to evolving attack techniques.

In summary, securing LLMs requires a holistic strategy to improve robustness, scalability, and effectiveness across various applications. Future studies should focus on developing interpretability frameworks, autonomous protection mechanisms, multimodal capabilities, and improved evaluation metrics to create more secure and reliable AI-driven cybersecurity solutions.

7 CONCLUSION

LLMs are at the forefront of transforming cybersecurity, offering innovative solutions to address increasingly complex challenges. This survey analyzed LLM applications on 32 security tasks covering eight domains, including blockchain, hardware, IoT, and cloud security, with a focus on task-specific applications such as vulnerability detection, malware analysis, and threat intelligence automation. Their versatility and significant impact were demonstrated in various cybersecurity applications.

This study initially explored the landscape of LLM applications, categorizing security tasks within each domain while highlighting their potential in modern cybersecurity. Furthermore, we examined LLM vulnerabilities, such as adversarial attacks and prompt injection, and identified mitigation strategies, including re-tokenization, perplexity-based detection, prompt isolation, and other defense mechanisms aimed at enhancing security.

This survey lays the foundation for future research, providing key insights for integrating LLMs into secure cybersecurity frameworks. By addressing existing challenges, LLMs can evolve to develop robust solutions that counter emerging cyber threats and safeguard critical digital infrastructure.

REFERENCES

- [1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023. [Online]. Available: <https://arxiv.org/abs/1706.03762>
- [2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019. [Online]. Available: <https://arxiv.org/abs/1810.04805>
- [3] OpenAI, "Gpt-3.5," <https://platform.openai.com/docs/models/gpt-3-5>, 2022, accessed: January 11, 2025.
- [4] ——, "Gpt-4," <https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo>, 2023, accessed: January 11, 2025.
- [5] C. Chen, J. Su, J. Chen, Y. Wang, T. Bi, J. Yu, Y. Wang, X. Lin, T. Chen, and Z. Zheng, "When chatgpt meets smart contract vulnerability detection: How far are we?" 2024. [Online]. Available: <https://arxiv.org/abs/2309.05520>
- [6] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, C. Chen, C. Olsson, C. Olah, D. Hernandez, D. Drain, D. Ganguli, D. Li, E. Tran-Johnson, E. Perez, J. Kerr, J. Mueller, J. Ladish, J. Landau, K. Ndousse, K. Lukosuite, L. Lovitt, M. Sellitto, N. Ellilage, N. Schieber, N. Mercado, N. DasSarma, R. Lasenby, R. Larson, S. Ringer, S. Johnston, S. Kravec, S. E. Showk, S. Fort, T. Lanham, T. Telleen-Lawton, T. Conerly, T. Henighan, T. Hume, S. R. Bowman, Z. Hatfield-Dodds, B. Mann, D. Amodei, N. Joseph, S. McCandlish, T. Brown, and J. Kaplan, "Constitutional ai: Harmlessness from ai feedback," 2022. [Online]. Available: <https://arxiv.org/abs/2212.08073>
- [7] P. R. B. Houssel, P. Singh, S. Layeghy, and M. Portmann, "Towards explainable network intrusion detection using large language models," 2024. [Online]. Available: <https://arxiv.org/abs/2408.04342>
- [8] A. Abdallah and A. Jatwt, "Generator-retriever-generator approach for open-domain question answering," 2024. [Online]. Available: <https://arxiv.org/abs/2307.11278>
- [9] J. Ou, J. Lu, C. Liu, Y. Tang, F. Zhang, D. Zhang, and K. Gai, "Dialogbench: Evaluating llms as human-like dialogue systems," 2024. [Online]. Available: <https://arxiv.org/abs/2311.01677>
- [10] R. Aguinaga-Kang, M. Gumin, D. H. Han, S. Morris, S. J. Yoo, A. Ganeshan, R. K. Jones, Q. A. Wei, K. Fu, and D. Ritchie, "Open-universe indoor scene generation using llm program synthesis and uncurated object databases," 2024. [Online]. Available: <https://arxiv.org/abs/2403.09675>
- [11] A. Mohaisen and O. Alrawi, "Unveiling zeus: automated classification of malware samples," in *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013, Companion Volume*, L. Carr, A. H. F. Laender, B. F. Lóscio, I. King, M. Fontoura, D. Vrandecic, L. Arroyo, J. P. M. de Oliveira, F. Lima, and E. Wilde, Eds., International World Wide Web Conferences Steering Committee / ACM, 2013, pp. 829–832.
- [12] A. Mohaisen, O. Alrawi, and M. Mohaisen, "AMAL: high-fidelity, behavior-based automated malware analysis and classification," *Comput. Secur.*, vol. 52, pp. 251–266, 2015.
- [13] J. Jang, H. Kang, J. Woo, A. Mohaisen, and H. K. Kim, "Andro-dumpsys: Anti-malware system based on the similarity of malware creator and malware centric information," *Comput. Secur.*, vol. 58, pp. 125–138, 2016.
- [14] F. Shen, J. D. Vecchio, A. Mohaisen, S. Y. Ko, and L. Ziarek, "Android malware detection using complex-flows," in *37th IEEE International Conference on Distributed Computing Systems, ICDCS 2017, Atlanta, GA, USA, June 5-8, 2017*, K. Lee and L. Liu, Eds., IEEE Computer Society, 2017, pp. 2430–2437.
- [15] J. Choi, A. Anwar, H. Alasmary, J. Spaulding, D. Nyang, and A. Mohaisen, "Iot malware ecosystem in the wild: a glimpse into analysis and exposures," in *Proceedings of the 4th ACM/IEEE Symposium on Edge Computing, SEC 2019, Arlington, Virginia, USA, November 7-9, 2019*, S. Chen, R. Onishi, G. Ananthanarayanan, and Q. Li, Eds., ACM, 2019, pp. 413–418.
- [16] A. Abusnaina, A. Khormali, H. Alasmary, J. Park, A. Anwar, and A. Mohaisen, "Adversarial learning attacks on graph-based iot malware detection systems," in *39th IEEE International Conference on Distributed Computing Systems, ICDCS 2019, Dallas, TX, USA, July 7-10, 2019*, IEEE, 2019, pp. 1296–1305.
- [17] H. Alasmary, A. Khormali, A. Anwar, J. Park, J. Choi, A. Abusnaina, A. Awad, D. Nyang, and A. Mohaisen, "Analyzing and detecting emerging internet of things malware: A graph-based approach," *IEEE Internet Things J.*, vol. 6, no. 5, pp. 8977–8988, 2019.
- [18] F. Shen, J. D. Vecchio, A. Mohaisen, S. Y. Ko, and L. Ziarek, "Android malware detection using complex-flows," *IEEE Trans. Mob. Comput.*, vol. 18, no. 6, pp. 1231–1245, 2019.
- [19] H. Alasmary, A. Abusnaina, R. Jang, M. Abuhamad, A. Anwar, D. Nyang, and D. Mohaisen, "Soteria: Detecting adversarial examples in control flow graph-based malware classifiers," in *40th IEEE International Conference on Distributed Computing Systems, ICDCS 2020, Singapore, November 29 - December 1, 2020*, IEEE, 2020, pp. 888–898.
- [20] A. Anwar, H. Alasmary, J. Park, A. Wang, S. Chen, and D. Mohaisen, "Statically dissecting internet of things malware: Analysis, characterization, and detection," in *Information and Communications Security - 22nd International Conference, ICICS 2020, Copenhagen, Denmark, August 24-26, 2020, Proceedings*, ser. Lecture Notes in Computer Science, W. Meng, D. Gollmann, C. D. Jensen, and J. Zhou, Eds., vol. 12282, Springer, 2020, pp. 443–461.
- [21] A. Abusnaina, A. Anwar, S. Alshamrani, A. Alabduljabbar, R. Jang, D. Nyang, and D. Mohaisen, "Systemically evaluating the robustness of ml-based iot malware detectors," in *51st Annual IEEE/IFIP International Conference on Dependable Systems and Networks, DSN 2021, Taipei, Taiwan, June 21-24, 2021 - Supplemental Volume*, IEEE, 2021, pp. 3–4.
- [22] ——, "Systematically evaluating the robustness of ml-based iot malware detection systems," in *25th International Symposium on Research in Attacks, Intrusions and Defenses, RAID 2022, Limassol, Cyprus, October 26-28, 2022*, ACM, 2022, pp. 308–320.
- [23] J. Choi, A. Anwar, A. Alabduljabbar, H. Alasmary, J. Spaulding, A. Wang, S. Chen, D. Nyang, A. Awad, and D. Mohaisen, "Understanding internet of things malware by analyzing endpoints in their static artifacts," *Comput. Networks*, vol. 206, p. 108768, 2022.
- [24] A. Abusnaina, M. Abuhamad, H. Alasmary, A. Anwar, R. Jang, S. Salem, D. Nyang, and D. Mohaisen, "DL-FHMC: deep learning-based fine-grained hierarchical learning approach for robust malware classification," *IEEE Trans. Dependable Secur. Comput.*, vol. 19, no. 5, pp. 3432–3447, 2022.
- [25] A. Abusnaina, A. Anwar, M. Saad, A. Alabduljabbar, R. Jang, S. Salem, and D. Mohaisen, "Exposing the limitations of machine learning for malware detection under concept drift," in *Web Information Systems Engineering - WISE 2024 - 25th International Conference, Doha, Qatar, December 2-5, 2024, Proceedings, Part II*, ser. Lecture Notes in Computer Science, M. Barhamgi, H. Wang, and X. Wang, Eds., vol. 15437, Springer, 2024, pp. 273–289.
- [26] D. Zapzalka, S. Salem, and D. Mohaisen, "Semantics-preserving node injection attacks against gnn-based ACFG malware classifiers," *IEEE Trans. Dependable*

- Secur. Comput.*, vol. 22, no. 1, pp. 549–560, 2025.
- [27] A. Alghamdi and D. Mohaisen, “Through the looking glass: Llm-based analysis of AR/VR android applications privacy policies,” in *International Conference on Machine Learning and Applications, ICMLA 2024, Miami, FL, USA, December 18-20, 2024*, M. A. Wani, P. Angelov, F. Luo, M. Ogihara, X. Wu, R. Precup, R. Ramezani, and X. Gu, Eds., IEEE, 2024, pp. 534–539. [Online]. Available: <https://doi.org/10.1109/ICMLA61862.2024.00078>
- [28] M. Kharma, S. Choi, M. AlKhanafseh, and D. Mohaisen, “Security and quality in llm-generated code: A multi-language, multi-model analysis,” *CoRR*, vol. abs/2502.01853, 2025. [Online]. Available: <https://doi.org/10.48550/arXiv.2502.01853>
- [29] D. Saha, S. Tarek, K. Yahyaei, S. K. Saha, J. Zhou, M. Tehranipoor, and F. Farahmandi, “Llm for soc security: A paradigm shift,” 2023. [Online]. Available: <https://arxiv.org/abs/2310.06046>
- [30] Z. He, Z. Li, S. Yang, A. Qiao, X. Zhang, X. Luo, and T. Chen, “Large language models for blockchain security: A systematic literature review,” 2024. [Online]. Available: <https://arxiv.org/abs/2403.14280>
- [31] H. Xu, S. Wang, N. Li, K. Wang, Y. Zhao, K. Chen, T. Yu, Y. Liu, and H. Wang, “Large language models for cyber security: A systematic literature review,” 2024. [Online]. Available: <https://arxiv.org/abs/2405.04760>
- [32] Y. Yigit, W. J. Buchanan, M. G. Tehrani, and L. Maglaras, “Review of generative ai methods in cybersecurity,” 2024. [Online]. Available: <https://arxiv.org/abs/2403.08701>
- [33] D. M. Divakaran and S. T. Peddinti, “Llms for cyber security: New opportunities,” 2024. [Online]. Available: <https://arxiv.org/abs/2404.11338>
- [34] M. A. Ferrag, F. Alwahedi, A. Battah, B. Cherif, A. Mechri, and N. Tihanyi, “Generative ai and large language models for cyber security: All insights you need,” 2024. [Online]. Available: <https://arxiv.org/abs/2405.12750>
- [35] H. Liang, X. Li, D. Xiao, J. Liu, Y. Zhou, A. Wang, and J. Li, “Generative pre-trained transformer-based reinforcement learning for testing web application firewalls,” *IEEE Transactions on Dependable and Secure Computing*, vol. 21, no. 1, pp. 309–324, 2024.
- [36] M. Liu, K. Li, and T.-A. Chen, “Deepsql: deep semantic learning for testing sql injection,” *Proceedings of the 29th ACM SIGSOFT International Symposium on Software Testing and Analysis*, 2020. [Online]. Available: <https://api.semanticscholar.org/CorpusID:218870218>
- [37] R. Meng, M. Mirchev, M. Böhme, and A. Roychoudhury, “Large language model guided protocol fuzzing,” in *31st Annual Network and Distributed System Security Symposium, NDSS 2024, San Diego, California, USA, February 26 – March 1, 2024*. The Internet Society, 2024. [Online]. Available: <https://www.ndss-symposium.org/ndss-paper/large-language-model-guided-protocol-fuzzing/>
- [38] P. Lanka, K. Gupta, and C. Varol, “Intelligent threat detection–ai-driven analysis of honeypot data to counter cyber threats,” *Electronics*, vol. 13, no. 13, p. 2465, 2024.
- [39] S. Moskal, S. Laney, E. Hemberg, and U. O'Reilly, “Llms killed the script kiddie: How agents supported by large language models change the landscape of network threat testing,” *CoRR*, vol. abs/2310.06936, 2023. [Online]. Available: <https://doi.org/10.48550/arXiv.2310.06936>
- [40] S. Temara, “Maximizing penetration testing success with effective reconnaissance techniques using chatgpt,” *CoRR*, vol. abs/2307.06391, 2023. [Online]. Available: <https://doi.org/10.48550/arXiv.2307.06391>
- [41] N. Tihanyi, T. Bisztray, R. Jain, M. A. Ferrag, L. C. Cordeiro, and V. Mavroeidis, “The formai dataset: Generative AI in software security through the lens of formal verification,” in *Proceedings of the 19th International Conference on Predictive Models and Data Analytics in Software Engineering, PROMISE 2023, San Francisco, CA, USA, 8 December 2023*, S. McIntosh, E. Choi, and S. Herbold, Eds., ACM, 2023, pp. 33–43. [Online]. Available: <https://doi.org/10.1145/3617555.3617874>
- [42] Y. Sun, D. Wu, Y. Xue, H. Liu, H. Wang, Z. Xu, X. Xie, and Y. Liu, “Gptscan: Detecting logic vulnerabilities in smart contracts by combining GPT with program analysis,” in *Proceedings of the 46th IEEE/ACM International Conference on Software Engineering, ICSE 2024, Lisbon, Portugal, April 14-20, 2024*. ACM, 2024, pp. 166:1–166:13. [Online]. Available: <https://doi.org/10.1145/3597503.3639117>
- [43] X. Meng, A. Srivastava, A. Arunachalam, A. Ray, P. H. Silva, R. Psiakis, Y. Makris, and K. Basu, “Unlocking hardware security assurance: The potential of llms,” *CoRR*, vol. abs/2308.11042, 2023. [Online]. Available: <https://doi.org/10.48550/arXiv.2308.11042>
- [44] Y. Du and Z. Yu, “Pre-training code representation with semantic flow graph for effective bug localization,” in *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/FSE 2023, San Francisco, CA, USA, December 3-9, 2023*, S. Chandra, K. Blincoe, and P. Tonella, Eds., ACM, 2023, pp. 579–591. [Online]. Available: <https://doi.org/10.1145/3611643.3616338>
- [45] R. J. Joyce, T. Patel, C. Nicholas, and E. Raff, “Avscan2vec: Feature learning on antivirus scan data for production-scale malware corpora,” in *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security, AISec 2023, Copenhagen, Denmark, 30 November 2023*, M. Pintor, X. Chen, and F. Tramèr, Eds., ACM, 2023, pp. 185–196. [Online]. Available: <https://doi.org/10.1145/3605764.3623907>
- [46] M. Labonne and S. Moran, “Spam-t5: Benchmarking large language models for few-shot email spam detection,” *CoRR*, vol. abs/2304.01238, 2023. [Online]. Available: <https://doi.org/10.48550/arXiv.2304.01238>
- [47] E. Malul, Y. Meidan, D. Mimran, Y. Elovici, and A. Shabtai, “Genkubesec: Llm-based kubernetes misconfiguration detection, localization, reasoning, and remediation,” *CoRR*, vol. abs/2405.19954, 2024. [Online]. Available: <https://doi.org/10.48550/arXiv.2405.19954>
- [48] V. K. Kasula, A. R. Yadulla, B. Konda, and M. Yenugula, “Fortifying cloud environments against data breaches: A novel ai-driven security framework,” *World Journal of Advanced Research and Reviews*, vol. 24, no. 01, pp. 1613–1626, 2024.
- [49] E. Bandara, S. Shetty, R. Babu Mukkamala, A. Rahman, P. B. Foytik, X. Liang, K. D. Zoysa, and N. W. Keong, “Devsec-gpt: generative-ai (with custom-trained meta's llama2 llm), blockchain, nft and pbom enabled cloud native container vulnerability management and pipeline verification platform,” 2024 *IEEE Cloud Summit*, pp. 28–35, 2024. [Online]. Available: <https://api.semanticscholar.org/CorpusID:271867592>
- [50] H.-c. I. Cybersecurity, “Ollabench: Evaluating llms' reasoning for human-centric interdependent cybersecurity.”
- [51] H. Ji, J. Yang, L. Chai, C. Wei, L. Yang, Y. Duan, Y. Wang, T. Sun, H. Guo, T. Li et al., “Sevenllm: Benchmarking, eliciting, and enhancing abilities of large language models in cyber threat intelligence,” *arXiv preprint arXiv:2405.03446*, 2024.
- [52] H. Luo, J. Luo, and A. V. Vasilakos, “BC4LLM: A perspective of trusted artificial intelligence when blockchain meets large language models,” *Neurocomputing*, vol. 599, p. 128089, 2024. [Online]. Available: <https://doi.org/10.1016/j.neucom.2024.128089>
- [53] B. Ahmad, S. Thakur, B. Tan, R. Karri, and H. Pearce, “Fixing hardware security bugs with large language models,” *CoRR*, vol. abs/2302.01215, 2023. [Online]. Available: <https://doi.org/10.48550/arXiv.2302.01215>
- [54] P. Tseng, Z. Yeh, X. Dai, and P. Liu, “Using llms to automate threat intelligence analysis workflows in security operation centers,” *arXiv preprint arXiv:2407.13093*, 2024.
- [55] M. Scanlon, F. Breitinger, C. Hargreaves, J. Hilgert, and J. Sheppard, “Chatgpt for digital forensic investigation: The good, the bad, and the unknown,” *Forensic Sci. Int. Digit. Investig.*, vol. 46, no. Supplement, p. 301609, 2023. [Online]. Available: <https://doi.org/10.1016/j.fsidi.2023.301609>
- [56] J. L. Martin, “The ethico-political universe of chatgpt,” *J. Soc. Comput.*, vol. 4, no. 1, pp. 1–11, 2023. [Online]. Available: <https://doi.org/10.23919/jsc.2023.0003>
- [57] R. Liu, Y. Wang, H. Xu, Z. Qin, Y. Liu, and Z. Cao, “Malicious URL detection via pretrained language model guided multi-level feature attention network,” *CoRR*, vol. abs/2311.12372, 2023. [Online]. Available: <https://doi.org/10.48550/arXiv.2311.12372>
- [58] 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,” 2024. [Online]. Available: <https://arxiv.org/abs/2405.11002>
- [59] P. Chen, L. Desmet, and C. Huygens, “A study on advanced persistent threats,” 09 2014, pp. 63–72.
- [60] E. Aghaei, E. Al-Shaer, W. Shadid, and X. Niu, “Automated cve analysis for threat prioritization and impact prediction,” 2023. [Online]. Available: <https://arxiv.org/abs/2309.03040>
- [61] P. V. S. Charan, H. Chunduri, P. M. Anand, and S. K. Shukla, “From text to mitre techniques: Exploring the malicious use of large language models for generating cyber attack payloads,” 2023. [Online]. Available: <https://arxiv.org/abs/2305.15336>
- [62] A. Happé, A. Kaplan, and J. Cito, “Llms as hackers: Autonomous linux privilege escalation attacks,” 2024. [Online]. Available: <https://arxiv.org/abs/2310.11409>
- [63] G. Deng, Y. Liu, V. Mayoral-Vilches, P. Liu, Y. Li, Y. Xu, T. Zhang, Y. Liu, M. Pinzger, and S. Rass, “Pentestgpt: An llm-empowered automatic penetration testing tool,” 2024. [Online]. Available: <https://arxiv.org/abs/2308.06782>
- [64] V. L. A. Quan, C. T. Phat, K. V. Nguyen, P. T. Duy, and V.-H. Pham, “Xgv-bert: Leveraging contextualized language model and graph neural network for efficient software vulnerability detection,” 2023. [Online]. Available: <https://arxiv.org/abs/2309.14677>
- [65] C. Thapa, S. I. Jang, M. E. Ahmed, S. Camtepe, J. Pieprzyk, and S. Nepal, “Transformer-based language models for software vulnerability detection,” 2022. [Online]. Available: <https://arxiv.org/abs/2204.03214>
- [66] S. Ullah, M. Han, S. Pujar, H. Pearce, A. Coskun, and G. Stringhini, “Llms cannot reliably identify and reason about security vulnerabilities (yet?): A comprehensive evaluation, framework, and benchmarks,” 2024. [Online]. Available: <https://arxiv.org/abs/2312.12575>
- [67] A. Khare, S. Dutta, Z. Li, A. Solko-Breslin, R. Alur, and M. Naik, “Understanding the effectiveness of large language models in detecting security vulnerabilities,” 2024. [Online]. Available: <https://arxiv.org/abs/2311.16169>
- [68] X. Liu, Y. Tan, Z. Xiao, J. Zhuge, and R. Zhou, “Not the end of story: An evaluation of ChatGPT-driven vulnerability description mappings,” in *Findings of the Association for Computational Linguistics: ACL 2023*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds., Toronto, Canada: Association for

- Computational Linguistics, Jul. 2023, pp. 3724–3731. [Online]. Available: <https://aclanthology.org/2023.findings-acl.229/>
- [69] C. Zhang, H. Liu, J. Zeng, K. Yang, Y. Li, and H. Li, “Prompt-enhanced software vulnerability detection using chatgpt,” 2024. [Online]. Available: <https://arxiv.org/abs/2308.12697>
- [70] Z. Zhang, J. Yang, P. Ke, F. Mi, H. Wang, and M. Huang, “Defending large language models against jailbreaking attacks through goal prioritization,” 2024. [Online]. Available: <https://arxiv.org/abs/2311.09096>
- [71] M. Fu, C. Tantithamthavorn, T. Le, V. Nguyen, and D. Phung, “Vulprepair: a t5-based automated software vulnerability repair,” in *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ser. ESEC/FSE 2022. New York, NY, USA: Association for Computing Machinery, 2022, p. 935–947. [Online]. Available: <https://doi.org/10.1145/3540250.3549098>
- [72] Q. Zhang, C. Fang, B. Yu, W. Sun, T. Zhang, and Z. Chen, “Pre-trained model-based automated software vulnerability repair: How far are we?” 2023. [Online]. Available: <https://arxiv.org/abs/2308.12533>
- [73] H. Pearce, B. Tan, B. Ahmad, R. Karri, and B. Dolan-Gavitt, “Examining zero-shot vulnerability repair with large language models,” 2022. [Online]. Available: <https://arxiv.org/abs/2112.02125>
- [74] Y. Wu, N. Jiang, H. V. Pham, T. Lutellier, J. Davis, L. Tan, P. Babkin, and S. Shah, “How effective are neural networks for fixing security vulnerabilities,” in *Proceedings of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis*, ser. ISSTA’23. ACM, Jul. 2023, p. 1282–1294. [Online]. Available: <http://dx.doi.org/10.1145/3597926.3598135>
- [75] K. Alrashedy, A. Aljasser, P. Tambwekar, and M. Gombolay, “Can llms patch security issues?” 2024. [Online]. Available: <https://arxiv.org/abs/2312.00024>
- [76] M. C. Tol and B. Sunar, “Zero leak: Using llms for scalable and cost effective side-channel patching,” 2023. [Online]. Available: <https://arxiv.org/abs/2308.13062>
- [77] Y. Charalambous, N. Tihanyi, R. Jain, Y. Sun, M. A. Ferrag, and L. C. Cordeiro, “Dataset for : A new era in software security: Towards self-healing software via large language models and formal verification (version 1),” <https://doi.org/10.5281/zenodo.8026525>, Jun. 2023, accessed on YYYY-MM-DD. [Online]. Available: <https://doi.org/10.5281/zenodo.8026525>
- [78] M. Jin, S. Shahriar, M. Tufano, X. Shi, S. Lu, N. Sundaresan, and A. Svyatkovskiy, “Inferfix: End-to-end program repair with llms,” 2023. [Online]. Available: <https://arxiv.org/abs/2303.07263>
- [79] H. Li, Y. Hao, Y. Zhai, and Z. Qian, “The hitchhiker’s guide to program analysis: A journey with large language models,” 2023. [Online]. Available: <https://arxiv.org/abs/2308.00245>
- [80] J. Lee, K. Han, and H. Yu, “A light bug triage framework for applying large pre-trained language model,” in *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering*, ser. ASE ’22. New York, NY, USA: Association for Computing Machinery, 2023. [Online]. Available: <https://doi.org/10.1145/3551349.3556898>
- [81] A. Z. H. Yang, R. Martins, C. L. Goues, and V. J. Hellendoorn, “Large language models for test-free fault localization,” 2023. [Online]. Available: <https://arxiv.org/abs/2310.01726>
- [82] T.-O. Li, W. Zong, Y. Wang, H. Tian, Y. Wang, S.-C. Cheung, and J. Kramer, “Nuances are the key: Unlocking chatgpt to find failure-inducing tests with differential prompting,” 2023. [Online]. Available: <https://arxiv.org/abs/2304.11686>
- [83] S. Fang, T. Zhang, Y. Tan, H. Jiang, X. Xia, and X. Sun, “Representthemall: A universal learning representation of bug reports,” 2023 *IEEE/ACM 45th International Conference on Software Engineering (ICSE)*, pp. 602–614, 2023. [Online]. Available: <https://api.semanticscholar.org/CorpusID:259859780>
- [84] N. Perry, M. Srivastava, D. Kumar, and D. Boneh, “Do users write more insecure code with ai assistants?” in *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS’23. ACM, Nov. 2023, p. 2785–2799. [Online]. Available: <http://dx.doi.org/10.1145/3576915.3623157>
- [85] Y. Wei, C. S. Xia, and L. Zhang, “Copiloting the copilots: Fusing large language models with completion engines for automated program repair,” in *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ser. ESEC/FSE’23. ACM, Nov. 2023. [Online]. Available: <http://dx.doi.org/10.1145/3611643.3616271>
- [86] C. S. Xia and L. Zhang, “Automated program repair via conversation: Fixing 162 out of 337 bugs for \$0.42 each using chatgpt,” in *Proceedings of the 33rd ACM SIGSOFT International Symposium on Software Testing and Analysis*, ser. ISSTA’24. ACM, Sep. 2024, p. 819–831. [Online]. Available: <http://dx.doi.org/10.1145/3650212.3680323>
- [87] ——, “Conversational automated program repair,” 2023. [Online]. Available: <https://arxiv.org/abs/2301.13246>
- [88] M. Boehme, C. Cedar, and A. Roychoudhury, “Fuzzing: Challenges and reflections,” *IEEE Software*, vol. 38, pp. 79–86, 2020. [Online]. Available: <https://api.semanticscholar.org/CorpusID:221737349>
- [89] Y. Deng, C. S. Xia, H. Peng, C. Yang, and L. Zhang, “Large language models are zero-shot fuzzers: Fuzzing deep-learning libraries via large language models,” 2023. [Online]. Available: <https://arxiv.org/abs/2212.14834>
- [90] C. Zhang, Y. Zheng, M. Bai, Y. Li, W. Ma, X. Xie, Y. Li, L. Sun, and Y. Liu, “How effective are they? exploring large language model based fuzz driver generation,” in *Proceedings of the 33rd ACM SIGSOFT International Symposium on Software Testing and Analysis*, ser. ISSTA’24. ACM, Sep. 2024, p. 1223–1235. [Online]. Available: <http://dx.doi.org/10.1145/3650212.3680355>
- [91] Y. Deng, C. S. Xia, C. Yang, S. D. Zhang, S. Yang, and L. Zhang, “Large language models are edge-case fuzzers: Testing deep learning libraries via fuzzgpt,” 2023. [Online]. Available: <https://arxiv.org/abs/2304.02014>
- [92] S. Hu, T. Huang, F. Ilhan, S. F. Tekin, and L. Liu, “Large language model-powered smart contract vulnerability detection: New perspectives,” in *5th IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications, TPS-ISA 2023, Atlanta, GA, USA, November 1-4, 2023*. IEEE, 2023, pp. 297–306. [Online]. Available: <https://doi.org/10.1109/TPS-ISA58951.2023.00044>
- [93] C. Yang, Y. Deng, R. Lu, J. Yao, J. Liu, R. Jabbarvand, and L. Zhang, “Whitefox: White-box compiler fuzzing empowered by large language models,” *Proceedings of the ACM on Programming Languages*, vol. 8, no. OOPSLA2, p. 709–735, Oct. 2024. [Online]. Available: <http://dx.doi.org/10.1145/3689736>
- [94] X. Xu, Z. Zhang, Z. Su, Z. Huang, S. Feng, Y. Ye, N. Jiang, D. Xie, S. Cheng, L. Tan, and X. Zhang, “Symbol preference aware generative models for recovering variable names from stripped binary,” 2024. [Online]. Available: <https://arxiv.org/abs/2306.02546>
- [95] J. Armengol-Estabé, J. Woodruff, C. Cummins, and M. F. P. O’Boyle, “Slade: A portable small language model decompiler for optimized assembly,” 2024. [Online]. Available: <https://arxiv.org/abs/2305.12520>
- [96] T. Sun, K. Allix, K. Kim, X. Zhou, D. Kim, D. Lo, T. F. Bissyandé, and J. Klein, “Dexbert: Effective, task-agnostic and fine-grained representation learning of android bytecode,” 2023. [Online]. Available: <https://arxiv.org/abs/2212.05976>
- [97] K. Pei, W. Li, Q. Jin, S. Liu, S. Geng, L. Cavallaro, J. Yang, and S. Jana, “Exploiting code symmetries for learning program semantics,” 2024. [Online]. Available: <https://arxiv.org/abs/2308.03312>
- [98] Q. Song, Y. Zhang, L. Ouyang, and Y. Chen, “Bimllm: Binary authorship verification with flow-aware mixture-of-shared language model,” 2022. [Online]. Available: <https://arxiv.org/abs/2203.04472>
- [99] M. Botacini, “Gpthreats-3: Is automatic malware generation a threat?” 05 2023, pp. 238–254.
- [100] B. Jakub and J. Branišová, “A dynamic rule creation based anomaly detection method for identifying security breaches in log records,” *Wireless Personal Communications*, vol. 94, 06 2017.
- [101] S. Shan, Y. Huo, Y. Su, Y. Li, D. Li, and Z. Zheng, “Face it yourselves: An llm-based two-stage strategy to localize configuration errors via logs,” in *Proceedings of the 33rd ACM SIGSOFT International Symposium on Software Testing and Analysis*, ser. ISSTA’24. ACM, Sep. 2024, p. 13–25. [Online]. Available: <http://dx.doi.org/10.1145/3650212.3652106>
- [102] E. Karlsen, X. Luo, N. Zinchir-Heywood, and M. Heywood, “Benchmarking large language models for log analysis, security, and interpretation,” 2023. [Online]. Available: <https://arxiv.org/abs/2311.14519>
- [103] X. Han, S. Yuan, and M. Trabelsi, “Loggpt: Log anomaly detection via gpt,” 2023. [Online]. Available: <https://arxiv.org/abs/2309.14482>
- [104] F. Heiding, B. Schneier, A. Vishwanath, J. Bernstein, and P. S. Park, “Devising and detecting phishing: Large language models vs. smaller human models,” 2023. [Online]. Available: <https://arxiv.org/abs/2308.12287>
- [105] E. Cambiasso and L. Caviglione, “Scamming the scammers: Using chatgpt to reply mails for time and resources,” 2023. [Online]. Available: <https://arxiv.org/abs/2303.13521>
- [106] H. W. A. Hanley and Z. Durumeric, “Twits, toxic tweets, and tribal tendencies: Trends in politically polarized posts on twitter,” 2024. [Online]. Available: <https://arxiv.org/abs/2307.10349>
- [107] Z. Cai, Z. Tan, Z. Lei, Z. Zhu, H. Wang, Q. Zheng, and M. Luo, “Lmbot: Distilling graph knowledge into language model for graph-less deployment in twitter bot detection,” 2024. [Online]. Available: <https://arxiv.org/abs/2306.17408>
- [108] R. Anderson and F. Petitcolas, “On the limits of steganography,” *IEEE Journal on Selected Areas in Communications*, vol. 16, pp. 474–481, 12 1998.
- [109] H. Wang, Z. Yang, J. Yang, C. Chen, and Y. Huang, “Linguistic steganalysis in few-shot scenario,” *IEEE Transactions on Information Forensics and Security*, vol. PP, pp. 1–1, 01 2023.
- [110] L. A. Bauer, J. K. H. IV, S. A. Markelon, V. Bindschaedler, and T. Shrimpton, “Covert message passing over public internet platforms using model-based format-transforming encryption,” *CoRR*, vol. abs/2110.07009, 2021. [Online]. Available: <https://arxiv.org/abs/2110.07009>
- [111] K. Paterson and D. Stebila, “One-time-password-authenticated key exchange,” 07 2010.
- [112] J. Rando, F. Perez-Cruz, and B. Hitaj, “Passgpt: Password modeling and (guided) generation with large language models,” 2023. [Online]. Available: <https://arxiv.org/abs/2306.01545>
- [113] S. R. Selamat, Y. Robiah, and S. Sahib, “Mapping process of digital forensic investigation framework,” vol. 8, 01 2008.
- [114] S. Paria, A. Dasgupta, and S. Bhunia, “Divas: An llm-based end-to-end framework for soc security analysis and policy-based protection,” 2023.

- [Online]. Available: <https://arxiv.org/abs/2308.06932>
- [115] M. Nair, R. Sadhukhan, and D. Mukhopadhyay, "How hardened is your hardware? guiding chatgpt to generate secure hardware resistant to cwe's," in *Cyber Security, Cryptology, and Machine Learning: 7th International Symposium, CSCML 2023, Be'er Sheva, Israel, June 29–30, 2023, Proceedings*. Berlin, Heidelberg: https://doi.org/10.1007/978-3-031-34671-2_23
- [116] B. Ahmad, S. Thakur, B. Tan, R. Karri, and H. Pearce, "On hardware security bug code fixes by prompting large language models," *IEEE Transactions on Information Forensics and Security*, vol. 19, p. 4043–4057, 2024. [Online]. Available: <http://dx.doi.org/10.1109/TIFS.2024.3374558>
- [117] I. David, L. Zhou, K. Qin, D. Song, L. Cavallaro, and A. Gervais, "Do you still need a manual smart contract audit?" 2023. [Online]. Available: <https://arxiv.org/abs/2306.12338>
- [118] M. Rodler, W. Li, G. O. Karamé, and L. Davi, "Sereum: Protecting existing smart contracts against re-entrancy attacks," in *26th Annual Network and Distributed System Security Symposium, NDSS 2019, San Diego, California, USA, February 24–27, 2019*. The Internet Society, 2019. [Online]. Available: <https://www.ndss-symposium.org/ndss-paper/sereum-protecting-existing-smart-contracts-against-re-entrancy-attacks/>
- [119] B. S. Mitchell, S. Mancoridis, and J. Kashyap, "On the automatic identification of misconfiguration errors in cloud native systems," 2024.
- [120] A. A. Pranata, O. Barais, J. Bourcier, and L. Noirie, "Misconfiguration discovery with principal component analysis for cloud-native services," in *2020 IEEE/ACM 13th International Conference on Utility and Cloud Computing (UCC)*. IEEE, 2020, pp. 269–278.
- [121] M. A. M. Ariffin, K. A. Rahman, M. Y. Darus, N. Awang, and Z. Kasiran, "Data leakage detection in cloud computing platform," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 8, no. 1, p. S1, 2019.
- [122] C. Vaidya, P. K. Khobragade, and A. A. Golghate, "Data leakage detection and security using cloud computing," 2016. [Online]. Available: <https://api.semanticscholar.org/CorpusID:46916951>
- [123] M. Henze, R. Matzutt, J. Hiller, E. Mühlmer, J. H. Ziegeldorf, J. van der Giet, and K. Wehrle, "Complying with data handling requirements in cloud storage systems," *IEEE Transactions on Cloud Computing*, vol. 10, no. 3, pp. 1661–1674, 2020.
- [124] O. Dye, J. Heo, and E. C. Cankaya, "Reflection of federal data protection standards on cloud governance," *arXiv preprint arXiv:2403.07907*, 2024.
- [125] R. Molleti, V. Goje, P. Luthra, and P. Raghavan, "Automated threat detection and response using llm agents," *World Journal of Advanced Research and Reviews*, vol. 24, pp. 079–090, 11 2024.
- [126] R. Fieblinger, M. T. Alam, and N. Rastogi, "Actionable cyber threat intelligence using knowledge graphs and large language models," in *2024 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. IEEE, 2024, pp. 100–111.
- [127] T. Ali and P. Kostakos, "Huntgpt: Integrating machine learning-based anomaly detection and explainable ai with large language models (llms)," *arXiv preprint arXiv:2309.16021*, 2023.
- [128] Y. Schwartz, L. Benshimol, D. Mimran, Y. Elovici, and A. Shabtai, "Llmccloud-hunter: Harnessing llms for automated extraction of detection rules from cloud-based cti," 2024. [Online]. Available: <https://arxiv.org/abs/2407.05194>
- [129] S. Mitra, S. Neupane, T. Chakraborty, S. Mittal, A. Pipali, M. Gaur, and S. Rahimi, "Localintel: Generating organizational threat intelligence from global and local cyber knowledge," 2024. [Online]. Available: <https://arxiv.org/abs/2401.10036>
- [130] H. Rong, Y. Duan, H. Zhang, X. Wang, H. Chen, S. Duan, and S. Wang, "Disassembling obfuscated executables with llm," 2024. [Online]. Available: <https://arxiv.org/abs/2407.08924>
- [131] H. Lu, H. Peng, G. Nan, J. Cui, C. Wang, W. Jin, S. Wang, S. Pan, and X. Tao, "Malsight: Exploring malicious source code and benign pseudocode for iterative binary malware summarization," 2024. [Online]. Available: <https://arxiv.org/abs/2406.18379>
- [132] C. Patsakis, F. Casino, and N. Lykousas, "Assessing llms in malicious code deobfuscation of real-world malware campaigns," 2024. [Online]. Available: <https://arxiv.org/abs/2404.19715>
- [133] B. Zhao, S. Ji, X. Zhang, Y. Tian, Q. Wang, Y. Pu, C. Lyu, and R. Beyah, "Uvscan: Detecting third-party component usage violations in iot firmware," 06 2023.
- [134] S. Li, G. Min *et al.*, "On-device learning based vulnerability detection in iot environment."
- [135] 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 2024 Workshop on Open-World Agents*, 2024. [Online]. Available: <https://openreview.net/forum?id=iiKOpRyLkw>
- [136] J. Su, C. Jiang, X. Jin, Y. Qiao, T. Xiao, H. Ma, R. Wei, Z. Jing, J. Xu, and J. Lin, "Large language models for forecasting and anomaly detection: A systematic literature review," 2024. [Online]. Available: <https://arxiv.org/abs/2402.10350>
- [137] X. Feng, X. Liao, X. Wang, H. Wang, Q. Li, K. Yang, H. Zhu, and L. Sun, "Understanding and securing device vulnerabilities through automated bug report analysis," in *SEC'19: Proceedings of the 28th USENIX Conference on Security*, 2019. [Online]. Available: <https://arxiv.org/abs/1907.08932>
- [138] S. Baral, S. Saha, and A. Haque, "An adaptive end-to-end iot security framework using explainable ai and llms," in *2024 IEEE 10th World Forum on Internet of Things (WF-IoT)*. IEEE, 2024, pp. 469–474.
- [139] P. N. Pathak, "How do you protect machine learning from attacks?" 2023, available online: <https://shorturl.at/AEQmP> [Accessed on January 28, 2024].
- [140] P. Liu, C. Sun, Y. Zheng, X. Feng, C. Qin, Y. Wang, Z. Xu, Z. Li, P. Di, Y. Jiang, and L. Sun, "Harnessing the power of llm to support binary taint analysis," 2024. [Online]. Available: <https://arxiv.org/abs/2310.08275>
- [141] J. Kaddour, J. Harris, M. Mozes, H. Bradley, R. Raileanu, and R. McHardy, "Challenges and applications of large language models," 2023. [Online]. Available: <https://arxiv.org/abs/2307.10169>
- [142] A. Tornede, D. Deng, T. Eimer, J. Giovanelli, A. Mohan, T. Ruhkopf, S. Segel, D. Theodoropoulos, T. Tornede, H. Wachsmuth, and M. Lindauer, "Automl in the age of large language models: Current challenges, future opportunities and risks," 2024. [Online]. Available: <https://arxiv.org/abs/2306.08107>
- [143] Computer Security Resource Center, "Information systems security (infosec)," https://csrc.nist.gov/glossary/term/information_systems_security, 2023, accessed on January 28, 2024.
- [144] CLOUDFLARE, "What is data privacy?" <https://www.cloudflare.com/learning/privacy/what-is-data-privacy/>, 2023, accessed on January 28, 2024.
- [145] D. Ganguli, L. Lovitt, J. Kernion, A. Askell, Y. Bai, S. Kadavath, B. Mann, E. Perez, N. Schieber, K. Ndousse, A. Jones, S. Bowman, A. Chen, T. Conerly, N. DasSarma, D. Drain, N. Elhage, S. E. Showk, S. Fort, Z. Hatfield-Dodds, T. Henighan, D. Hernandez, T. Hume, J. Jacobson, S. Johnston, S. Kravec, C. Olsson, S. Ringer, E. Tran-Johnson, D. Amodei, T. Brown, N. Joseph, S. McCandlish, C. Olah, J. Kaplan, and J. Clark, "Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned," *CoRR*, vol. abs/2209.07858, 2022. [Online]. Available: <https://doi.org/10.48550/arXiv.2209.07858>
- [146] P. Röttger, B. Vidgen, D. Nguyen, Z. Waseem, H. Margetts, and J. Pierrehumbert, "Hatecheck: Functional tests for hate speech detection models," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, 2021. [Online]. Available: <http://dx.doi.org/10.18653/v1/2021.acl-long.4>
- [147] F. Perez and I. Ribeiro, "Ignore previous prompt: Attack techniques for language models," 2022. [Online]. Available: <https://arxiv.org/abs/2211.09527>
- [148] D. Ganguli, L. Lovitt, J. Kernion, A. Askell, Y. Bai, S. Kadavath, B. Mann, E. Perez, N. Schieber, K. Ndousse, A. Jones, S. Bowman, A. Chen, T. Conerly, N. DasSarma, D. Drain, N. Elhage, S. E. Showk, S. Fort, Z. Hatfield-Dodds, T. Henighan, D. Hernandez, T. Hume, J. Jacobson, S. Johnston, S. Kravec, C. Olsson, S. Ringer, E. Tran-Johnson, D. Amodei, T. Brown, N. Joseph, S. McCandlish, C. Olah, J. Kaplan, and J. Clark, "Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned," 2022. [Online]. Available: <https://arxiv.org/abs/2209.07858>
- [149] G. Alon and M. Kamfonas, "Detecting language model attacks with perplexity," 2023. [Online]. Available: <https://arxiv.org/abs/2308.14132>
- [150] N. Jain, A. Schwarzschild, Y. Wen, G. Somepalli, J. Kirchenbauer, P. yeh Chiang, M. Goldblum, A. Saha, J. Geiping, and T. Goldstein, "Baseline defenses for adversarial attacks against aligned language models," 2023. [Online]. Available: <https://arxiv.org/abs/2309.00614>
- [151] Z. Cao, Y. Xu, Z. Huang, and S. Zhou, "Ml4co-kida: Knowledge inheritance in dataset aggregation," 2022. [Online]. Available: <https://arxiv.org/abs/2201.10328>
- [152] X. Qi, A. Panda, K. Lyu, X. Ma, S. Roy, A. Beirami, P. Mittal, and P. Henderson, "Safety alignment should be made more than just a few tokens deep," 2024. [Online]. Available: <https://arxiv.org/abs/2406.05946>
- [153] F. Bianchi, M. Suzgun, G. Attanasio, P. Röttger, D. Jurafsky, T. Hashimoto, and J. Zou, "Safety-tuned llamas: Lessons from improving the safety of large language models that follow instructions," in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7–11, 2024*. OpenReview.net, 2024. [Online]. Available: <https://openreview.net/forum?id=gT5hALch9z>
- [154] J. Zhao, Z. Deng, D. Madras, J. Zou, and M. Ren, "Learning and forgetting unsafe examples in large language models," in *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21–27, 2024*. OpenReview.net, 2024. [Online]. Available: <https://openreview.net/forum?id=RYnmgedvJk>
- [155] M. Wortsman, G. Ilharco, S. Y. Gadre, R. Roelofs, R. Gontijo-Lopes, A. S. Morcos, H. Namkoong, A. Farhadi, Y. Carmor, S. Kornblith, and L. Schmidt, "Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time," 2022. [Online]. Available: <https://arxiv.org/abs/2203.05482>
- [156] A. Zou, Z. Wang, N. Carlini, M. Nasr, J. Z. Kolter, and M. Fredrikson, "Universal and transferable adversarial attacks on aligned language models," 2023. [Online]. Available: <https://arxiv.org/abs/2307.15043>
- [157] S. R. Kadhe, F. Ahmed, D. Wei, N. Baracaldo, and I. Padhi, "Split, unlearn, merge: Leveraging data attributes for more effective unlearning in llms," 2024. [Online]. Available: <https://arxiv.org/abs/2406.11780>

- [158] D. Xu, S. Fan, and M. Kankanhalli, "Combating misinformation in the era of generative ai models," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 9291–9298.
- [159] A. Schwarzschild, M. Goldblum, A. Gupta, J. P. Dickerson, and T. Goldstein, "Just how toxic is data poisoning? a unified benchmark for backdoor and data poisoning attacks," 2021. [Online]. Available: <https://arxiv.org/abs/2006.12557>
- [160] K. Kurita, P. Michel, and G. Neubig, "Weight poisoning attacks on pre-trained models," 2020. [Online]. Available: <https://arxiv.org/abs/2004.06660>
- [161] V. Shah, "Machine learning algorithms for cybersecurity: Detecting and preventing threats," 12 2022.
- [162] L. Yan, Z. Zhang, G. Tao, K. Zhang, X. Chen, G. Shen, and X. Zhang, "Parafuzz: An interpretability-driven technique for detecting poisoned samples in nlp," 2023. [Online]. Available: <https://arxiv.org/abs/2308.02122>
- [163] A. Continella, Y. Fratantonio, M. Lindorfer, A. Puccetti, A. Zand, C. Krügel, and G. Vigna, "Obfuscation-resilient privacy leak detection for mobile apps through differential analysis," in *Network and Distributed System Security Symposium*, 2017. [Online]. Available: <https://api.semanticscholar.org/CorpusID:1775474>
- [164] M. Shu, J. Wang, C. Zhu, J. Geiping, C. Xiao, and T. Goldstein, "On the exploitability of instruction tuning," 2023. [Online]. Available: <https://arxiv.org/abs/2306.17194>
- [165] H. Aghakhani, W. Dai, A. Manoel, X. Fernandes, A. Kharkar, C. Kruegel, G. Vigna, D. Evans, B. Zorn, and R. Sim, "Trojanpuzzle: Covertly poisoning code-suggestion models," 2024. [Online]. Available: <https://arxiv.org/abs/2301.02344>
- [166] K. Liu, B. Dolan-Gavitt, and S. Garg, "Fine-pruning: Defending against backdooring attacks on deep neural networks," 2018. [Online]. Available: <https://arxiv.org/abs/1805.12185>
- [167] Z. Chen, Z. Xiang, C. Xiao, D. Song, and B. Li, "Agentpoison: Red-teaming llm agents via poisoning memory or knowledge bases," 2024. [Online]. Available: <https://arxiv.org/abs/2407.12784>
- [168] C. Chen, Y. Sun, X. Gong, J. Gao, and K.-Y. Lam, "Neutralizing backdoors through information conflicts for large language models," *arXiv preprint arXiv:2411.18280*, 2024.
- [169] A. G. Chowdhury, M. M. Islam, V. Kumar, F. H. Shezan, V. Jain, and A. Chadha, "Breaking down the defenses: A comparative survey of attacks on large language models," *arXiv preprint arXiv:2403.04786*, 2024.
- [170] Z. Zhang, L. Lyu, X. Ma, C. Wang, and X. Sun, "Fine-mixing: Mitigating backdoors in fine-tuned language models," 2022. [Online]. Available: <https://arxiv.org/abs/2210.09545>
- [171] G. Cui, L. Yuan, B. He, Y. Chen, Z. Liu, and M. Sun, "A unified evaluation of textual backdoor learning: Frameworks and benchmarks," 2022. [Online]. Available: <https://arxiv.org/abs/2206.08514>
- [172] Z. Xi, T. Du, C. Li, R. Pang, S. Ji, J. Chen, F. Ma, and T. Wang, "Defending pre-trained language models as few-shot learners against backdoor attacks," 2023. [Online]. Available: <https://arxiv.org/abs/2309.13256>
- [173] M. H. Wolk, "The iphone jailbreaking exemption and the issue of openness," *Cornell JL & Pub. Pol'y*, vol. 19, p. 795, 2009.
- [174] G. Mondillo, S. Colosimo, A. Perrotta, V. Frattolillo, C. Indolfi, M. M. del Giudice, and F. Rossi, "Jailbreaking large language models: navigating the crossroads of innovation, ethics, and health risks," *Journal of Medical Artificial Intelligence*, vol. 8, 2025.
- [175] A. Kumar, C. Agarwal, S. Srinivas, A. J. Li, S. Feizi, and H. Lakkaraju, "Certifying llm safety against adversarial prompting," 2024. [Online]. Available: <https://arxiv.org/abs/2309.02705>
- [176] F. Wu, Y. Xie, J. Yi, J. Shao, J. Curl, L. Lyu, Q. Chen, and X. Xie, "Defending chatgpt against jailbreak attack via self-reminder," 2023.
- [177] H. Jin, L. Hu, X. Li, P. Zhang, C. Chen, J. Zhuang, and H. Wang, "Jailbreakzoo: Survey, landscapes, and horizons in jailbreaking large language and vision-language models," *CoRR*, vol. abs/2407.01599, 2024. [Online]. Available: <https://doi.org/10.48550/arXiv.2407.01599>
- [178] A. Robey, E. Wong, H. Hassani, and G. J. Pappas, "Smoothllm: Defending large language models against jailbreaking attacks," 2024. [Online]. Available: <https://arxiv.org/abs/2310.03684>
- [179] E. Crothers, N. Japkowicz, and H. Viktor, "Machine generated text: A comprehensive survey of threat models and detection methods," 2023. [Online]. Available: <https://arxiv.org/abs/2210.07321>
- [180] B. Bababah, M. Kwiatkowski, C. Leung, C. G. Akcora et al., "Sok: Prompt hacking of large language models," *arXiv preprint arXiv:2410.13901*, 2024.
- [181] Learn Prompting, "Your guide to generative ai," <https://learnprompting.org>, 2023, accessed: January 28, 2024.
- [182] S. Schulhoff, "Instruction defense," https://learnprompting.org/docs/prompt_hacking/defensive_measures/instruction, 2024, accessed: October 11, 2024.
- [183] H. Gonen, S. Iyer, T. Blevins, N. A. Smith, and L. Zettlemoyer, "Demystifying prompts in language models via perplexity estimation," 2024. [Online]. Available: <https://arxiv.org/abs/2212.04037>
- [184] Y. Liu, Y. Jia, R. Geng, J. Jia, and N. Z. Gong, "Formalizing and benchmarking prompt injection attacks and defenses," 2024. [Online]. Available: <https://arxiv.org/abs/2310.12815>
- [185] J. Shi, Z. Yuan, Y. Liu, Y. Huang, P. Zhou, L. Sun, and N. Z. Gong, "Optimization-based prompt injection attack to llm-as-a-judge," 2024. [Online]. Available: <https://arxiv.org/abs/2403.17710>
- [186] Y. Liu, Y. Yao, J.-F. Ton, X. Zhang, R. Guo, H. Cheng, Y. Klochkov, M. F. Taufiq, and H. Li, "Trustworthy llms: a survey and guideline for evaluating large language models' alignment," 2024. [Online]. Available: <https://arxiv.org/abs/2308.05374>
- [187] M. A. K. Raiaan, M. S. H. Mukta, K. Fatema, N. M. Fahad, S. Sakib, M. M. J. Mim, J. Ahmad, M. E. Ali, and S. Azam, "A review on large language models: Architectures, applications, taxonomies, open issues and challenges," *IEEE Access*, 2024.
- [188] M. Mishra, M. Stallone, G. Zhang, Y. Shen, A. Prasad, A. M. Soria, M. Merler, P. Selvam, S. Surendran, S. Singh et al., "Granite code models: A family of open foundation models for code intelligence," *arXiv preprint arXiv:2405.04324*, 2024.
- [189] H. Zhao, H. Chen, F. Yang, N. Liu, H. Deng, H. Cai, S. Wang, D. Yin, and M. Du, "Explainability for large language models: A survey," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 2, pp. 1–38, 2024.
- [190] H. Yang, K. Xiang, M. Ge, H. Li, R. Lu, and S. Yu, "A comprehensive overview of backdoor attacks in large language models within communication networks," *IEEE Network*, 2024.
- [191] E. Moore, A. Imteaj, S. Rezapour, and M. H. Amini, "A survey on secure and private federated learning using blockchain: Theory and application in resource-constrained computing," *IEEE Internet of Things Journal*, 2023.
- [192] X. Cai, H. Xu, S. Xu, Y. Zhang et al., "Badprompt: Backdoor attacks on continuous prompts," *Advances in Neural Information Processing Systems*, vol. 35, pp. 37 068–37 080, 2022.
- [193] F. Huang, "Data cleansing," in *Encyclopedia of big data*. Springer, 2022, pp. 275–279.
- [194] A. Wan, E. Wallace, S. Shen, and D. Klein, "Poisoning language models during instruction tuning," in *International Conference on Machine Learning*. PMLR, 2023, pp. 35 413–35 425.
- [195] B. Wang, Y. Yao, S. Shan, H. Li, B. Viswanath, H. Zheng, and B. Y. Zhao, "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks," in *2019 IEEE symposium on security and privacy (SP)*. IEEE, 2019, pp. 707–723.
- [196] B. C. Das, M. H. Amini, and Y. Wu, "Security and privacy challenges of large language models: A survey," *ACM Computing Surveys*, 2024.
- [197] H. Yao, J. Lou, and Z. Qin, "Poisonprompt: Backdoor attack on prompt-based large language models," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 7745–7749.
- [198] M. Charfeddine, H. M. Kamoun, B. Hamdaoui, and M. Guizani, "Chatgpt's security risks and benefits: offensive and defensive use-cases, mitigation measures, and future implications," *IEEE Access*, 2024.