IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 4 November 2024, accepted 21 November 2024, date of publication 25 November 2024,
date of current version 4 December 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3505983

SURVEY

Application of Large Language Models in Cybersecurity: A Systematic Literature Review

ISMAYIL HASANOV[1,2], SEPPO VIRTANEN[1], (Senior Member, IEEE), ANTTI HAKKALA[1], AND JOUNI ISOAHO[1]

[1]Department of Computing, University of Turku, 20014 Turku, Finland
[2]Openfactory Nordic oy, 20500 Turku, Finland

Corresponding author: Ismayil Hasanov (ismayil.i.hasanov@utu.fi)

ABSTRACT The emergence of Large Language Models (LLMs) is currently creating a major paradigm shift in societies and businesses in the way digital technologies are used. While the disruptive effect is especially observable in the information and communication technology field, there is a clear lack of systematic studies focusing on the application and impact of LLMs in cybersecurity holistically. This article presents an exhaustive systematic literature review of 177 articles published in 2018-2024 on the application of LLMs and the use of Artificial Intelligence (AI) as a defensive measure in cybersecurity. This article contributes an analytical compendium of the recent research on the application of LLMs in offensive and defensive cybersecurity as well as in research on cyberethics, current legal frameworks, and research regarding the use of LLMs for cybersecurity governance. It also contributes a statistical summary of global research trends in the field. Of the reviewed literature, 68% was published in 2023. Nearly 30% of the articles originate from the USA and 11% from China, with other countries currently having significantly lower contributions to recent research. Most attention in recent research has been given to AI as a defensive measure, accounting for 27% of the reviewed literature. It was observed that LLMs have proven highly effective in phishing attack simulations and in managing cybersecurity administrative aspects, including defending against advanced exploits. Furthermore, LLMs show significant potential in the development of security software, further cementing their role as a powerful tool in cybersecurity innovation.

INDEX TERMS Cybersecurity, artificial intelligence, large language models, generative AI, penetration testing, cyberethics, network security, natural language processing, systematic literature review, survey.

I. INTRODUCTION

Artificial intelligence (AI) has profoundly permeated various sectors, notably through the emergence of Large Language Models (LLMs). These technologies find applications across diverse industries, including healthcare, mechanical engineering, and information technologies, with an increasing presence also in cybersecurity and with both beneficial and detrimental effects on the development of the field. Notable applications in cybersecurity include Intrusion Detection Systems (IDS) [1] and Intrusion Prevention Systems (IPS) [2], Security Information and Event Management (SIEM) systems [3], and Next Generation Firewalls (NGFW) [4]. AI assists in identifying and tracking malicious patterns, thereby substantially reducing human workload, minimizing human error, and enhancing system efficiency. Nonetheless, it is acknowledged that AI systems are not infallible, evidenced by occurrences of false positives and negatives. However, the advantages of AI are significant, most notably the capacity for round-the-clock operation without fatigue and with sustained concentration. Moreover, the ability of AI to discern subtleties that may elude human observation makes it indispensable.

Being proficient in understanding and generating human-like text, LLMs are invaluable for numerous cybersecurity tasks such as log analysis and penetration testing. LLMs are particularly beneficial for Small and Medium-sized Enterprises (SMEs) by facilitating tasks like website development

The associate editor coordinating the review of this manuscript and approving it for publication was Sedat Akleylek.

TABLE 1. Overview of related literature surveys on using LLMs in the cybersecurity context in comparison to the work presented in this article.

Reference	Year	Scope	Task focus	Time frame	Collected papers	Type
Large Language Models for Software Engineering: A Systematic Literature Review [5]	2024	LLM	Software Engineering, some aspects of cybersecurity mentioned	2017-2024	395	Journal article
LLMs for Cyber Security: New Opportunities [6]	2024	LLM	LLMs in Defensive cybersecurity	2020-2024	Not specified	Preprint
A survey on large language model (LLM) security and privacy: The Good, The Bad, and The Ugly [7]	2024	LLM	LLM in defensive and offensive cybersecurity, vulnerabilities of LLMs	2019-2024	281	Journal article
Review of Generative AI Methods in Cybersecurity [8]	2024	LLM	LLM in defensive and offensive cybersecurity, vulnerabilities of LLMs	2020-2024	Not specified	Preprint
A Survey of Large Language Models in Cybersecurity [9]	2024	LLM	LLM in offensive cybersecurity, vulnerabilities of LLMs	2021-2023	20	Preprint
Large Language Models for Cyber Security: A Systematic Literature Review [10]	2024	LLM	LLMs in defensive cybersecurity, vulnerabilities, challenges and opportunities of LLMs	2020-2024	127	Preprint
A Comprehensive Review of Large Language Models in Cyber Security [11]	2024	LLM	LLM in defensive cybersecurity	2019-2024	Not specified	Journal article
Application of Large Language Models in Cybersecurity: a Systematic Literature Review **(the work presented in this article)**	2024	LLM and AI	AI in cybersecurity, LLM in defensive and offensive cybersecurity, vulnerabilities of LLMs, ethics and legal regulations of LLMs	2018-2024	177	This article

and security audits, thereby economizing the use of resources.

The primary motivation for conducting this study is the limited number of comprehensive analyses on the implementation of AI, particularly LLMs, in cybersecurity. While there are studies that explore the defensive and offensive capabilities or ethical concerns of LLMs, they tend to focus on a single aspect, preventing a broader understanding of this emerging technology. The authors believe it is crucial to provide a holistic perspective, enabling readers to fully grasp the future potential of LLMs in the cybersecurity domain. Consequently, this literature review aims to offer a more comprehensive analysis of these technologies in this field. In this article, a thorough Systematic Literature Review (SLR) is performed to discern current methods and research directions in the application of AI and LLMs in cybersecurity. A review of other state-of-the-art surveys has also been conducted to highlight the unique contributions of this research. The findings are summarized in Table 1. As can be observed from the table, the majority of the articles concentrate on one or a few specific aspects of LLM implementation in cybersecurity. Moreover, the bulk of these works are preprints and have yet to undergo peer review. Additionally, this work does not solely focus on LLMs; it also incorporates relevant AI-related articles to offer a more comprehensive perspective than the other works summarized in Table 1.

Unlike other studies on the implementation of generative AI in cybersecurity, this work not only explores the application of LLMs but also examines the broader use of AI in this field, enabling readers to compare the effectiveness of various AI technologies in security. Furthermore, this work addresses the use of LLMs in both defensive and offensive cybersecurity, in contrast to many studies that focus solely on one aspect. Additionally, the potential implementation of LLMs in cybersecurity administration, including their role in various frameworks and policies, is discussed. Another distinguishing feature of this work is the presented analysis of cyberethics with regard to LLMs and how this reflects on proper governance of LLMs in today's digital landscape. In summary, this work makes the following key contributions:

- A review and analysis of the application of AI in cybersecurity, including the use of LLMs for both offensive and defensive purposes, while other studies typically focus on some narrower aspects (as seen in Table 1).
- A review and analysis of recent research on cyberethics in the AI context, current legal frameworks surrounding AI, and the use of LLMs for cybersecurity governance.
- A summary of statistics on global research trends in the field.

The following three research questions to which the literature review presented in this article will give elaborate answers are defined:

1) How effective are Large Language Models and Artificial Intelligence in cybersecurity applications?
2) In what ways are Large Language Models applied in cybersecurity tactics?
3) What are the challenges and limitations of Large Language Models in the context of cybersecurity?

The rest of this article is organized as follows. Section II presents the methodology used for conducting an SLR and discusses some statistical findings regarding the analyzed literature. Section III discusses key concepts and terminology

of AI, LLMs and cybersecurity to the extent needed for following the presentation of the literature analysis and the results. The analysis of the reviewed literature is presented in section IV. The order of topics presented in section IV follows the same sequence as the topical categories listed in Table 10. Following the literature analysis, the results are discussed in section V. Concluding remarks of the study are given in section VI.

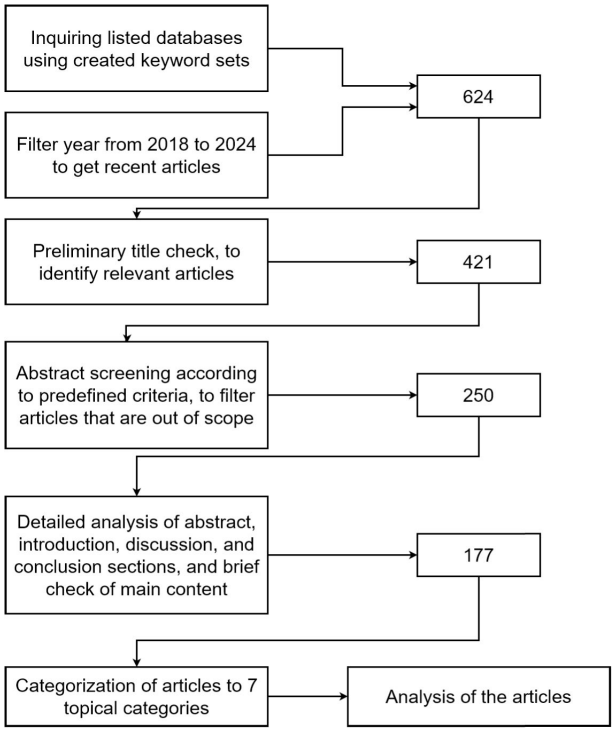

FIGURE 1. Systematic literature review process used in this work.

II. MATERIALS AND METHODS

This section describes the methodology employed for conducting the SLR and also presents statistical findings regarding the analyzed literature. The SLR methodology applied in this study is illustrated in Figure 1, which also outlines the number of articles filtered out at each stage of the analysis. The methodology draws inspiration from, and is a simplified adaptation of, two well-established approaches in the existing literature: the Kitchenham Guidelines [12] and the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) [13] framework.

The literature review entails an exhaustive analysis of several databases and research dissemination platforms, specifically: arXiv, IEEE Xplore, ResearchGate, Scopus, Google Scholar, ScienceDirect, SpringerLink and ACM Digital Library. Each database was queried using a carefully made set of keywords derived from the research questions. For every research question, seven distinct sets of keywords were developed. In addition, five sets of topic-specific keywords were crafted, resulting in a comprehensive total of 26 keyword sets used for querying the databases. The details of these keyword sets are presented in Tables 2, 3, 4 and 5.

TABLE 2. Queries for research question 1.

Queries for Research Question 1
"LLM AND Cybersecurity effectiveness"
"GPT-3.5 OR GPT-4 AND Cybersecurity applications"
"large PRE/ language PRE/ models AND cybersecurity"
"Artificial Intelligence AND Information Security effectiveness"
"generative PRE/1 ai AND cybersecurity"
"LLM in Threat PRE/ Analysis AND in Predictive PRE/ Intelligence"
"Machine Learning AND Cybersecurity comparison"

TABLE 3. Queries for research question 2.

Queries for Research Question 2
"LLM AND Cybersecurity tactics"
"GPT AND Secure Network Configuration AND Threat Detection"
"Natural Language Processing AND Cybersecurity"
"ChatGPT AND Information Security"
"LLM AND Advanced Persistent Threats OR Social Engineering"
"BERT OR FalconLLM AND Cybersecurity"
"Generative AI AND Attacks Detection in Security"

TABLE 4. Queries for research question 3.

Queries for Research Question 3
"Challenges AND LLM in Cybersecurity"
"Limitations AND GPT models in Information Security"
"AI AND ML vulnerabilities in Cybersecurity"
"LLM in Vulnerability Management AND limitations"
"Security risks AND Generative AI in Cybersecurity"
"LLaMA OR Bard AND Cybersecurity challenges"
"Ethical concerns AND AI in Cybersecurity"

TABLE 5. Topic-specific queries.

Topic-specific keyword sets
"LLM AND Phishing OR Spear Phishing detection"
"AI AND Simulating Social Engineering Attacks"
"Claude OR LLaMA OR Bard AND Infosec"
"Machine Learning AND Configuration Validation"
"Honeypots AND Generative AI AND Network Threat Detection"

For each database and corresponding keyword set, an initial screening was undertaken to compile a preliminary list of articles for subsequent analysis. Given the rapid evolution of AI and the relatively recent emergence of LLMs, the focus was on studies published in the last six years (2018–2024); older studies were filtered out. During preliminary screening, the title and abstract of each article were carefully examined. The inclusion of an article in the list was determined based on the following criteria:

- **Population/Subject**: Determine whether the article falls within the scope of the research.
- **Outcomes of Interest**: Evaluate whether the efficacy, accuracy, challenges, limitations, benefits, and applications of AI and LLMs in cybersecurity were analyzed in the article.

- **Language**: Only studies published in English were considered.

After the initial screening, 250 articles were collected and documented for further analysis. Various statistics, such as publication year, source, and keywords, were also gathered during the analysis.

The abstract, introduction, discussion, and conclusion sections of each article were analyzed in detail, while the main content was browsed cursorily. For the detailed analysis, the following criteria were applied to assess the quality of each article:

- **Study Design and Methodology**: This criterion assesses whether the paper aligns with the research scope of the SLR and evaluates the relevance and reliability of the methods employed. The authors examined whether the papers had a well-structured research design and methodology, such as empirical methods (e.g., experiments, surveys) or theoretical models (e.g., frameworks, simulations) that were relevant to the research questions. For example, in the article ''GPT-Based Malware: Unveiling Vulnerabilities and Creating a Way Forward in Digital Space,'' [14] the research design is a literature review. The authors focus on exploring the threat of GPT-based malware. The study provides a detailed overview of the threats introduced by LLMs and ways to address them. In this case, a literature review and case studies are good ways to study possible threats arising from it; therefore, this article was accepted into the SLR, providing insights into how LLMs are used to create malware. The objective of the study is to examine how threat actors are using LLMs, which aligns with the work performed. Therefore, this article meets the required ''Study Design and Methodology'' criteria.
- **Results**: This criterion assesses whether the research outcomes are aligned with the research questions presented in the article. What is more, it analyzes whether the research presents any unique or valuable information within its scope. For example, in the previously mentioned study, authors define the primary research question as identifying the vulnerabilities arising from LLMs and suggesting ways to mitigate them. The research outcomes align well with this goal since the article presents both the threats posed by LLMs and methods to address them. Moreover, the study demonstrates how GPTs could be used in the creation of polymorphic malware and includes an interesting jailbreak prompt that is particularly noteworthy.
- **Quality of the Data**: This criterion verifies whether the data employed in the training process are inherently reliable. This means the data used in the research were collected from reliable sources and are not biased or falsified. For example, in the article ''Spear Phishing Emails Detection Based on Machine Learning,'' [15] the authors perform an experiment and employ data for a Machine Learning (ML) model, which consists of 417 spear phishing emails and 13916 non-spear phishing emails, including benign and phishing emails from the companies the authors were cooperating with. The data employed for the study are considered to be of good quality since companies are usually the target of phishing; therefore, such emails are ideal for training the model. Moreover, the size of the dataset is sufficient, and the authors additionally used resampling techniques. Furthermore, the authors are transparent that the spear phishing emails were classified by an expert group, which adds extra credibility to the quality of the data.

Based on this analysis and the collected data, the articles were classified into four distinct categories: not relevant articles (73 articles), peripheral articles (10 articles), relevant articles (123 articles), and key articles (44 articles). In percentages, 18% of the collected literature comprised the key articles, while 49% of the literature constituted the relevant articles and 4% the peripheral articles. As a result of this classification, the not relevant articles (73 articles, 29%) were excluded from the remainder of the literature review either because they did not meet the selection criteria or they were not accessible, leaving 177 articles (71%) for further analysis. The distribution of these articles in terms of the databases and research dissemination platforms in which they were discovered is as follows: arXiv 44, IEEE Xplore 38, ResearchGate 33, Scopus 19, Google Scholar 18, ScienceDirect 11, SpringerLink 9 and ACM Digital Library 5. A significant portion of the articles originates from arXiv and thus has preprint status. Since the cut-off date of the SLR, many of these articles have been published in peer-reviewed forums. However, as they were still in arXiv preprint status at the time of the SLR cut-off date and during the review of articles for this study, they have been categorized as arXiv articles here. Google Scholar, ResearchGate, and Scopus contain articles from various publishers; therefore, if an article was found on one of these three services as well as on one of the specific publisher services, the article was counted for the publisher's service.

A. STATISTICS

In the next stage of the literature review, the remaining 177 articles were classified into different topical categories based on the authors' assessment of their research or methodical focus. Each paper was categorized based on a detailed analysis of its title, abstract, and content. For example, the study titled ''Getting pwn'd by AI: Penetration Testing with Large Language Models'' [16] is classified under ''LLM as an Offensive Tool'' because its main focus is on the use of LLMs in offensive cybersecurity. Similarly, ''Cyber Intrusion Detection using Natural Language Processing on Windows Event Logs'' [17] is categorized as ''NLP in Cybersecurity'' due to its emphasis on applying Natural Language Processing (NLP) technologies for cyber defense. In the rest of this article, the authors use the terms *category*

and *topical category* interchangeably when referring to how researchers classified the articles. The authors defined seven topical categories to which the articles were sorted. The defined topical categories and the number of articles for each category are shown in Table 6. While the topical categories "AI as a Defensive Measure" and "LLM in Defensive Cybersecurity" are technically related (since LLMs are a subset of AI) they were separated for the purposes of a more detailed analysis. This way a specific focus on LLMs could be provided, while the other topical category considers the broader implementation of AI in cybersecurity.

TABLE 6. Defined topical categories and article distribution by category.

Topical Category	Count
AI as a defensive measure	48
AI system vulnerabilities and mitigating them	39
LLM in defensive cybersecurity	36
LLM as an offensive tool	32
NLP in cybersecurity	11
Cyberethics and Ethical Considerations	7
LLM and law	4

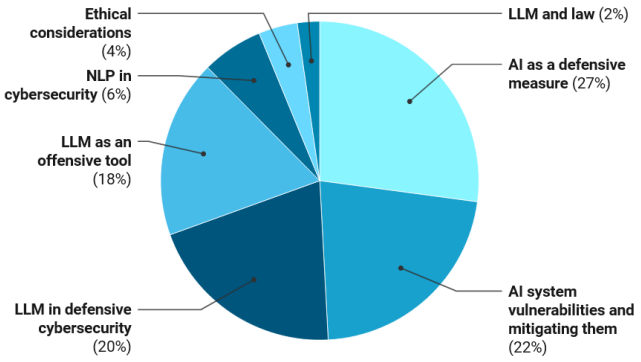

FIGURE 2. Article distribution by topical category (numbers are percentages).

The topical category with the majority of articles is "AI as a defensive measure," which includes 48 articles out of the 177 that qualified for the detailed analysis. This category is followed by "AI system vulnerabilities and mitigating them", which comprises 39 articles. The third and fourth categories pertain to the use of LLMs in defensive and offensive cybersecurity, respectively. The rest of the topical categories had significantly fewer article hits as seen in Table 6. Figure 2 illustrates the percentage distribution of the 177 articles into the topical categories, showing that nearly one-third of the analyzed literature studies the usage of AI as a defensive measure.

Table 7 shows the top five countries contributing to the analyzed literature. The country of origin for each article is determined based on the first author's affiliation as given in the article. The table also shows each country's topical focus of study based on the articles' classification. Out of the 177 articles, 52 are from the USA, followed by China with 20 articles, and India with 11 articles. These top article

TABLE 7. Article distribution by countries and the topical category most focused on in research output for each country.

Country	Count	Topical Category most focused on in articles
USA	52	LLM in defensive cybersecurity
China	20	AI system vulnerabilities and mitigating them
India	11	NLP in cybersecurity
UK	9	LLM as an offensive tool
UAE	8	AI as a defensive measure

TABLE 8. Number of articles by publication year and the topical category with most articles in each year. For 2024, only the period january-march is included in the study.

Year	Articles	Topical category with most articles
2024	20	AI as a defensive measure
2023	120	LLM in defensive cybersecurity
2022	17	AI system vulnerabilities and mitigating them
2021	7	AI as a defensive measure
2020	10	AI as a defensive measure
2019	2	Not enough data
2018	1	Not enough data

TABLE 9. Top 5 most popular keywords in analyzed literature.

Keyword	Occurrences
cybersecurity	44
chatgpt	33
machine learning	33
artificial intelligence	31
generative AI	16

counts may be explained by significant national investments the countries make in AI research. Indeed, many of the top AI companies are based in the USA. Another noteworthy observation concerns the focal areas of study across different countries. As demonstrated in Table 7, the USA primarily concentrates on LLMs and defensive security. This is unsurprising, taking into account the substantial funding the country allocates to its defense industry. In contrast, China seems to direct its resources towards system vulnerabilities and the potential applications of LLMs and AI. Meanwhile, India's research efforts seem to be predominantly centred on NLP in cybersecurity.

Another interesting statistic is demonstrated in Table 8, which shows the distribution of articles by year and the dominant focus area studied most in the articles in each year. As can be seen, almost 75% of the articles are from 2023, while 20 articles are from the first three months of 2024 (the cut-off for the review presented in this study was March 2024). Overall, the application of AI for defensive cybersecurity has dominated research efforts over the years covered in this study.

The top five keywords of the extracted articles are presented in the Table 9. The most popular keyword used is "cybersecurity," followed by "chatgpt". The most commonly used keywords within the different article categories are given in Table 10.

The general tendency is that the majority of the studies analyzed in this SLR feature either cybersecurity or AI/LLM

TABLE 10. Topical categories and the top 3 keywords used in articles for each category.

Topical Category	Keywords
Cyberethics and Ethical Considerations	ChatGPT (4)
	Computer Science (4)
	Ethics (3)
LLM and law	Artificial Intelligence (1)
	Algorithm (1)
	Law (1)
AI as a defensive measure	Machine Learning (20)
	Artificial Intelligence (14)
	Cybersecurity (10)
AI system vulnerabilities and mitigating them	Machine learning (7)
	Artificial intelligence (6)
	ChatGPT (6)
NLP in cybersecurity	Cybersecurity (8)
	Machine Learning (7)
	NLP (6)
LLM as an offensive tool	ChatGPT (11)
	Cybersecurity (9)
	Artificial Intelligence (6)
LLM in defensive cybersecurity	Cybersecurity (12)
	ChatGPT (9)
	Artificial Intelligence (7)

keywords. Looking at the co-occurrence of keywords, it can be observed that the most frequently found keyword pairs are:

- "cybersecurity" and "machine learning" (25)
- "cybersecurity" and "artificial intelligence" (22)
- "chatgpt" and "artificial intelligence" (21)
- "machine learning" and "artificial intelligence" (20)
- "cybersecurity" and "chatgpt" (19)

The most popular keyword co-occurrence combinations are the ones where a pair is formed of a security term and an AI term, aligning well and as expected with the target domain of this study.

III. FOUNDATIONAL CONCEPTS IN AI, LLMS AND CYBERSECURITY

This section introduces foundational concepts and ideas behind AI, LLMs and cybersecurity that are needed for further understanding of the topic. This section is structured as follows: first, general information about AI technologies will be provided, followed by an introduction to LLM and NLP concepts and ideas. Subsequently, the concept of jailbreaking in the context of LLMs will be defined, along with some jailbreaking techniques. This concept is crucial for understanding the logic behind offensive security in LLMs. Finally, some definitions and theoretical background for cybersecurity are presented. It must be noted, however, that the field of cybersecurity is vast, and therefore only the concepts covered in the literature analysis part of this study will be elaborated here.

A. ARTIFICIAL INTELLIGENCE

As delineated by the European Parliamentary Research Service, AI can be characterized as a system that exhibits intelligent behavior by analyzing its environment and autonomously executing actions to accomplish specific tasks [18]. This section will briefly summarize some of the most used best practices and approaches. The quintessential approach to AI is rule-based AI [19], wherein human experts develop meticulous rule-based procedures. Conversely, in ML, computers learn from data without explicit rule sets. In ML, algorithms and statistical models are used to draw inferences from extensive datasets, a procedure known as model training [20].

There are two principal techniques in ML. In **supervised learning**, the available data adhere to a specific format: $\{x_i, y_i\}_{i=1}^{n}$, where x_i denotes an input and y_i denotes the corresponding output. The primary objective is to determine the relationship between these values in order to predict unknown outputs [21].

Unsupervised learning, as the name implies, does not depend on known outcomes or outputs. Rather, it exclusively employs input data $\{x_i\}_{i=1}^{n}$ without corresponding outputs. Unlike supervised learning, which utilizes a hypothesis set comprising functions from which the data engineer extracts the solution, the objective of unsupervised learning is to discern patterns and relationships within the dataset in an exploratory fashion, without explicit instruction. Notably, unsupervised learning can also be employed to examine the data the engineer was given, to glean insights on the data (for example some tendencies), often serving as a preliminary stage in data preprocessing [21].

Deep learning (DL), a specialized subset of ML, involves artificial neural networks with multiple layers, commonly referred to as deep neural networks. These networks can tackle more intricate tasks by deconstructing them into smaller, more manageable problems across their layers [22].

Data play a pivotal role in AI, especially in ML. High-quality, expansive datasets are essential for training algorithms effectively. For instance, training an audio recognition system necessitates thousands, if not millions, of labeled samples. Sometimes there are insufficient data to equally represent each class, leading to class imbalance. In such cases, it is important to implement resampling techniques. **Oversampling** involves increasing the number of instances in the minority class to balance the dataset. **Undersampling** reduces the number of instances in the majority class to balance the dataset [23].

Another important topic is the evaluation of AI models. There are many evaluation metrics available, but the following metrics are most commonly used for classification tasks. **Accuracy** refers to the proportion of all classifications, both positive and negative, that were correctly predicted. **Recall** (also known as true positive rate) is the proportion of actual positive cases that were correctly classified as positive by the model. **Precision** measures the proportion of all positive classifications made by the model that are actually positive. The $\mathbf{F_1}$ score is the harmonic mean of precision and recall and is often used as a descriptive metric [24]. $\mathbf{F_1}$ score is especially important as it provides a more robust metric

for evaluation, making it extremely useful when working with imbalanced datasets. The **Area Under the Receiver Operating Characteristic (AUC-ROC)** score indicates how well a model can distinguish between classes, with higher values signifying better performance. The ROC curve is a plot of the true positive rate against the false positive rate at different classification thresholds [25].

In essence, AI systems analyze inputs from their environment, process this information using various algorithms, and produce outputs that range from simple decisions to complex predictions. The continuous enhancement of these systems is heavily reliant on data, computational power, and advancements in algorithm design.

B. LARGE LANGUAGE MODELS

Large Language Models are sophisticated language models encompassing billions or even trillions of parameters. These models are trained on vast datasets, often amounting to terabytes of data. Notable examples such as T5 [26], GPT, and LLaMA [27] have demonstrated remarkable proficiency in tasks such as text generation, chatbot interactions, and programming. Furthermore, it has been established that LLMs can effectively conduct penetration testing [16]. There are different types of LLMs, including general-use models and domain-specific models tailored to particular tasks. For instance, BloombergGPT [28] is designed to perform financial tasks.

The fundamental concept underpinning LLMs involves the tokenization of text, which breaks text down into smaller units known as tokens. The core principle of LLMs is to predict the next token (word) based on the context and input data. Utilizing NLP technology, LLMs can generate human-like language tailored to specific tasks, making the interpretation of their output exceedingly straightforward. One of the most prominent LLMs today is ChatGPT, which was trained using Reinforcement Learning from Human Feedback (RLHF) [29].

Traditional Reinforcement Learning (RL) involves a model learning to make decisions by interacting with its environment, receiving penalties and rewards based on its actions. These values are determined by a reward function created by engineers, making the design of this function critical. Poorly designed reward functions can lead to unintended consequences, such as agents exploiting loopholes to achieve high rewards without genuinely fulfilling the intended objectives. To address these issues, RLHF incorporates a human-in-the-loop approach. Rather than relying on a static reward function, the agent's learning objective is iteratively refined through human feedback [30].

Also, terms like **fine-tuning**, **zero-shot** and **few-shot** learning should be introduced. Fine-tuning is the process of adapting an LLM to perform better in certain domains by training it on task-oriented datasets. It helps the model to adjust its parameters, improving its ability to respond to domain-specific queries with higher accuracy and relevance. For this, techniques like Low-Rank Adaptation (LoRA) and Parameter Efficient Fine-Tuning (PEFT) are used [31]. Zero-shot learning refers to a technique where a model scores the answers without having seen any prior examples, relying entirely on its pre-existing knowledge and the provided prompt. On the other hand, few-shot learning involves giving the model a small number of examples before asking it to generate new responses, generally improving performance, especially for simpler tasks.

LLMs have built-in mechanisms designed to prevent access to illegal or offensive data. However, it is possible to bypass these safeguards through techniques known as jailbreaking [32]. Recent studies have demonstrated that cleverly crafted natural language prompts can circumvent sophisticated safety measures, leading to Prompt Injection (PI) attacks. This underscores a significant gap in current security measures, which are often reactive and narrowly focused.

Some well-known attacks have already been addressed, such as the Do Anything Now (DAN) attack, which compelled ChatGPT to provide any information requested by the user [33]. There are various types of attacks, including the Evil-Bot Prompt, ChatGPT Developer Mode v2, and Strive To Avoid Norms (STAN) Prompt. While these attacks are relatively easy to identify and rectify, more sophisticated methods involve creating adversarial prompts that exploit model weaknesses using advanced techniques like gradient-based search. This highlights a significant gap in current security measures, which are often reactive and narrowly focused [34].

C. CYBERSECURITY

In today's data-driven world, social networks, online transactions, and myriad other activities are powered by Information and Communication Technologies (ICTs). As modern technologies advance, the threat of cyberattacks escalates, making it crucial to prioritize information security and data privacy.

Cybersecurity is the practice of protecting resources, processes, and infrastructures from adversarial attacks, unauthorized access, or damage, both physical and digital. Additionally, it encompasses strategies and actions aimed at ensuring quick recovery in the event of disasters, as well as a set of policies and rules designed to safeguard these assets. This subsection will briefly discuss certain types of attacks on cybersecurity that have additional dimensions for AI and LLMs, both in attack and defense.

1) PHISHING AND SOCIAL ENGINEERING

Social engineering involves the psychological manipulation of users to obtain confidential information or prompt actions that provide access or leverage for unauthorized activities. Phishing, a form of social engineering, deceives users into disclosing sensitive information to an attacker. These attacks typically serve as an initial entry point into an infrastructure and are commonly executed via email or phone call [35].

It is believed that AI, particularly LLMs, will contribute to both attackers in crafting sophisticated phishing emails and defenders in identifying them. A more in-depth analysis of this will be conducted in section IV.

2) MALWARE

Malware encompasses all malicious software and can be defined as software the operation of which has a detrimental effect on the target computer or device. Various types of malware include viruses, worms, trojan horses, spyware, and ransomware. Like any software, malware can be created using various programming languages. Typically, hackers either develop their malware or purchase it from the darknet. However, with the advent of generative AI, it is now possible to create malware without prior programming knowledge, making malware creation accessible to almost anyone [36]. It is quite common in cybersecurity to use honeypots, which are decoy systems designed to lure attackers. Honeypots provide cybersecurity professionals with an environment to analyze attack methodologies. The data collected provide valuable insights into the behavior, evolution, and evasion techniques utilized by malware.

3) BRUTEFORCING AND ENUMERATION

Bruteforce attacks involve an attacker leveraging computer processing power, for example, to find passwords by attempting numerous username-password combinations, often sourced from dictionaries, potentially reaching millions of combinations. Generative AI can enhance this process by making educated guesses, generating passwords, and conducting adaptive bruteforce attacks [37]. Enumeration refers to systematically extracting information about a target network, system, or application, such as open network ports, running services, and website directories. This information is typically used to plan subsequent cyberattacks [38].

IV. ANALYSIS OF THE REVIEWED LITERATURE

In this section the authors present the primary findings of this study. All articles were classified into topical categories (as defined in section II) based on the central focus of the research presented in each article. In the following, the analyzed literature is discussed one topical category at a time in the following order: Cyberethics and Ethical Considerations, LLM and Law, AI as a Defensive Measure, AI System Vulnerabilities and Mitigating Them, NLP in Cybersecurity, LLM as an Offensive Tool, and LLM in Defensive Cybersecurity.

A. CYBERETHICS AND ETHICAL CONSIDERATIONS

Ethics, in its broadest sense, constitutes a branch of philosophy that addresses questions concerning what is morally right and wrong, good and bad, fair and unfair. It encompasses a system of moral principles that guide human behavior, aiding individuals and societies in determining how they ought to act in various situations. When discussing ethics in the context of AI, it pertains to the principles and guidelines that govern the development, deployment, and use of AI technologies. Given its significant role in AI, it is crucial to thoroughly analyze this topic. With the advent of ChatGPT and particularly the emergence of more powerful models, cyberethics has become a substantial concern.

1) LLMS AND PRIVACY

Wu et al. [39] assert that AI technology functions as a double-edged sword. While it is designed to enhance security, such AI systems also pose significant threats to organizations, as they can be exploited for malicious intentions. Additionally, the security of the AI system itself can lead to substantial damage if compromised. Consequently, it is important to fortify the security of these systems. Chugh [40] highlights the necessity of addressing privacy concerns and mentions the Probably Approximately Correct (PAC) learning technique as a method to safeguard sensitive corporate data when utilizing LLMs.

2) LLM AND BIAS

Wu et al. [39] emphasize that LLMs have the capacity to influence public opinion, making it crucial that the data used to train these models are free from bias and discrimination. If an LLM is trained on biased or inappropriate data, it may propagate unethical, harmful, or discriminatory practices to its vast user base, potentially reaching hundreds of millions. Another significant concern is the environmental impact of LLMs. These systems require substantial resources, including electricity, leading to increased carbon emissions and contributing to environmental pollution.

Chugh [40] underscores the importance of addressing these biases through meticulous data curation and the use of fairness-aware ML techniques. Such approaches help to mitigate biases and promote equity in AI-driven interactions. Additionally, Chugh highlights the ethical challenge of misinformation spread by LLMs. These models can inadvertently disseminate false information, amplifying societal biases and prejudices. Thus, implementing robust fact-checking mechanisms and maintaining a commitment to factual accuracy are essential. Chugh also advocates for the establishment of regulatory frameworks to safeguard against potential misuse or ethical breaches.

3) LLM AND AI GOVERNANCE

Wu et al. [39] highlight the necessity of regulating LLMs through legal frameworks, particularly concerning the copyrights of texts generated by these models. Given the popularity of LLMs, many students and researchers incorporate LLM-generated texts into their reports or homework. To address these issues, Chugh [40] recommends the use of regulatory frameworks. Another innovative approach is proposed by Gianni et al. [41], which involves a framework of democratic experimentation. This method emphasizes social inquiry and involves civil society in the governance of AI, ensuring that ethical guidelines reflect

the values and concerns of the general public. It includes public engagement and stakeholder inclusion in decision-making processes, transitioning from traditional ethical guidelines to a more inclusive and participatory model of governance.

In addressing these issues, Flaih and Jasim [42] suggest embedding ethical guidelines into chatbot AI models. Authors advocate for the development, implementation, and frequent monitoring of cyberethical frameworks and rules. Furthermore, the responsibility for the ethical use of LLMs must be shared among all stakeholders. Beyond ethical concerns, some researchers argue that modern curricula should place greater emphasis on cyberethics. According to Matei and Bertino [43], cyberethics education is insufficiently covered in cybersecurity majors. Consequently, professors often overestimate the cyberethical preparedness of students, and Matei et al. propose that cyberethical training should be integrated into curricula.

Governmental entities, such as the G7, have expressed interest in regulating LLMs. It is also crucial to ensure the proper use of LLMs like ChatGPT. As Waghmare notes [44], users have a degree of control over the data shared with and used by ChatGPT for training. Therefore, when sharing sensitive information, it is important to ensure that the data are not used for training purposes. User concerns about privacy, data security, and transparency significantly affect their loyalty to ChatGPT, as observed by Niu and Mvondo [45]. Users with strong cyberethical beliefs demand higher standards of corporate responsibility and transparency, influencing their satisfaction and loyalty to AI technologies. Thus, users are likely to seek out the most transparent and secure bots.

It is clear that while users are gradually accepting LLMs, there remains significant concern regarding data privacy and transparency. This creates a dual responsibility: users must be vigilant about the data they input into public LLMs, while developers must uphold cyberethical standards and ensure responsible data processing practices. Furthermore, existing ethical frameworks need to be updated to address the challenges posed by recent advancements in AI technologies, ensuring that privacy, bias, and governance issues are properly managed.

B. LLM AND LAW

Another crucial concept is the regulation of AI by law. As LLMs become an integral part of human life, it is essential to define what is permissible by law and what is not. For example, it needs to be defined whether it is legal to use AI or LLMs to create software. One of the notable approaches in this respect is the method proposed by Shi [46]. In the research, it was observed that all risks associated with generative AI can be categorized into two main categories. The first category encompasses all risks related to intellectual property, including issues like the ownership of AI-generated content. The second category includes all data-related risks, such as the generation of poor-quality or illegal content and data leakage. Shi criticizes the currently suggested legislation and proposes new concepts intended to strengthen enforcement. The research promotes a balanced governance strategy that focuses on both security and development. Clear regulations on copyright ownership and the enhancement of AI self-checking functions are vital. Additionally, the study suggests that international cooperation is necessary to establish unified data protection rules to address the legal challenges posed by generative AI.

LLMs like ChatGPT are not entirely new or unique; the first modern LLMs were introduced back in 2017, although a significant ''boom'' for LLMs occurred with the launch of ChatGPT. Despite widespread enthusiasm, there were notable concerns, such as Italy becoming the first country to ban ChatGPT at the governmental level. As Gualdi and Cordella [47] mention, Italy's Data Protection Authority, known as Garante, imposed a temporary ban on ChatGPT in March 2023 due to violations of the General Data Protection Regulation (GDPR). Issues like lack of transparency, data accuracy, legal basis, and age verification led to the ban. The authors critique the assumption of technological neutrality in regulations like GDPR, arguing that regulatory frameworks must evolve to address the specific characteristics of generative AI rather than applying broad, technically neutral policies.

Another important regulatory issue is that, according to Kshetri [48], LLMs like ChatGPT lower the barriers to entry for malicious actors. Techniques such as jailbreaking can force LLMs to generate malicious code or plan a cyberattack without prior expertise required from the human user. Therefore, it is crucial to regulate the use of LLMs by users.

Jeong [49] offers a comprehensive taxonomy of AI-related crimes, categorizing them into AI as a tool crime and AI as a target crime. This classification emphasizes the multifaceted nature of AI-enhanced traditional crimes, such as advanced phishing and automated hacking, and highlights the unique challenges posed by adversarial attacks targeting AI systems.

ChatGPT is highly dynamic, with datasets continuously evolving as the AI processes data and generates new outputs. As this section suggests, applying traditional regulatory approaches is insufficient for managing such systems. It is important to shift towards regulatory frameworks that incorporate a thorough understanding of AI technologies, ensuring that regulations are not only legally robust but also technologically informed.

C. AI AS A DEFENSIVE MEASURE

The advent of AI has introduced novel attack vectors, techniques, and defensive measures, while also enhancing existing defensive mechanisms. In this section, the authors assess the literature on defensive AI applications in cybersecurity and identify specific application areas.

1) AI AND MACHINE LEARNING GENERAL APPLICATIONS

The authors start this section with a discussion on the general application of AI in cybersecurity, based on the following publications: [50], [51], [52], [53], [54]. Authors note that AI in cybersecurity plays a critical role by leveraging advanced techniques such as ML, DL, and RL to detect and respond to cyber threats. A recurring theme noted in the articles is the dual role of AI as both a defender and a potential tool for cybercriminals. This underscores the necessity for robust and adaptive defense mechanisms. AI's ability to analyze vast datasets in real-time, adapt to new threats, and automate security routines enhances the speed and efficiency of defensive cybersecurity operations.

The articles also observe that AI technologies have their limitations and shortcomings. Issues such as data quality, algorithmic complexity, vulnerability to adversarial attacks, and ethical concerns related to privacy, bias, and accountability are serious concerns for AI in cybersecurity.

2) SPECIFIC AI TECHNIQUES AND APPLICATIONS

A potential strategy to apply AI techniques to cybersecurity is to embed the implementation of AI directly into the security framework. For example, Bagaa et al. [55] propose a ML-based security framework for Internet of Things (IoT) systems that leverage Software Defined Networking (SDN) and Network Function Virtualization (NFV) to provide dynamic and efficient threat detection and mitigation. AI can be utilized in Cybersecurity Named Entity Recognition (Cs-NER). Chen et al. [56] propose a novel approach to Cs-NER models that combines the Bidirectional Encoder Representations from Transformers (BERT) language representation model [57] with Long Short-Term Memory (LSTM), Bidirectional LSTM (Bi-LSTM), Iterated Dilated Convolutional Neural Networks (ID-CNNs), and Conditional Random Field (CRF) layers. This approach uses the BERT model to obtain distributed representations of words, enhancing the performance of the NER system. The study demonstrates that joint BERT models significantly outperform state-of-the-art methods in Cs-NER tasks. Another intriguing application of BERT is in malware detection, as explored by Rahali and Akhloufi [58]. Transformer-based malware detection utilizes static analysis to detect and classify malware without executing the code, making it resource-efficient and faster. Moreover, this method can not only classify software as malware or benign but also identify the type of malware. The results demonstrate satisfactory accuracy and $\mathbf{F_1}$ scores, with binary classification showing superior performance.

AI can also be applied to cryptography to enhance the security and efficiency of cryptographic algorithms such as Advanced Encryption Standard (AES), Rivest–Shamir–Adleman (RSA), and Learning with Errors (LWE) -based systems. According to Nitaj and Rachidi [59], AI can analyze and improve the security of S-boxes in symmetric ciphers and generate safe prime numbers and public/private keys in RSA.

Lastly, Hemberg and O'Reilly [60] discuss the use of a collated dataset named BRON to enhance AI applications in cybersecurity. BRON integrates multiple public threat and vulnerability sources into a graph database, supporting pattern inference, modeling, simulation, and AI planning, making it a powerful tool for cybersecurity professionals.

It is evident that AI can be applied in many specific areas of cybersecurity, making it an extremely useful tool for cybersecurity engineers. Therefore, it is important to ensure that this technology is integrated within organizations, as AI generally helps to improve security operations and strengthen overall security.

3) AI FOR INTRUSION DETECTION AND PREVENTION

Several studies have been conducted on the significant potential that AI technologies hold for intrusion detection and prevention. Park et al. [61] propose that DL models such as Convolutional Neural Networks (CNNs) and LSTM networks can be highly effective for intrusion detection. However, these models often face the challenge of insufficient data. To address this, Park et al. suggest using Generative Adversarial Networks (GANs) to generate synthetic network traffic data, thereby improving the performance of IDS by mitigating data imbalance issues.

The study by Marino et al. [62] introduces an approach to enhance the explainability of IDS models. This is achieved by identifying the minimum modifications needed to correct misclassified samples, thus providing insights into the decision boundaries of the model. The approach was tested using the NSL-KDD99 benchmark dataset, and the experiments demonstrated its effectiveness.

The integration of AI within the Open XDR framework shows promise. Pissanidis and Demertzis [63] highlight the enhanced capabilities in threat detection and response achieved by combining data from IDS, Endpoint Detection and Response (EDR), SIEM, Active Directory, and log forwarding systems. The incorporation of AI/ML enables real-time detection, reduces false positives, and enhances the overall efficiency of cybersecurity operations.

AI is also particularly useful in mitigating DDoS attacks. Ouhssini et al. [64] propose the DeepDefend framework for real-time detection and prevention of DDoS attacks in cloud computing environments. This framework leverages DL techniques and genetic algorithms and was tested on the CIDDS-001 dataset, demonstrating high accuracy in entropy forecasting and rapid, precise detection of DDoS attacks.

The use of AI in IDS, IPS, and DDoS-prevention systems has been shown to be highly efficient, contributing significantly to organizational security. This efficiency is further illustrated in the study by Latif et al. [65], where authors present an optimized IDS for Industrial IoT networks. Authors' approach utilizes Deep Transfer Learning (DTL) and Genetic Algorithms (GA), converting cybersecurity datasets into image data to enable the use of advanced CNNs for intrusion detection.

4) AI IN SPECIFIC CYBERSECURITY DOMAINS

In this section, the potential of AI across diverse cybersecurity domains will be analyzed. A particularly intriguing experiment was conducted by Veprytska and Kharchenko [66], who provide a comprehensive classification and risk analysis framework for AI-powered cyberattacks and defenses. Authors underscore the complexity of AI-against-AI scenarios, highlighting the necessity for robust AI-powered countermeasures and ongoing research to stay ahead of evolving threats. Authors' application of the IMECA method offers a systematic approach to assess risks and evaluate the effectiveness of various countermeasures. A critical consideration is adversarial ML, where the dataset is manipulated to mislead the ML algorithm's decision-making process. Vaccari et al. [67] evaluate the effectiveness of traditional ML algorithms compared to new methods designed to ensure reliability and transparency in detecting adversarial attacks. The study proposes a novel approach that utilizes explainable and reliable AI techniques to identify adversarial attacks, maximizing detection accuracy. Experimental results indicate that the proposed Reliable AI methods outperform traditional ML algorithms.

AI has also found applications in 6G networks. The study by Ferrag et al. [68] outlines various use cases of generative AI in IoT applications, including visual, audio, text-based, code-based, and IoT security applications. This implementation involves using generative AI for cyber threat-hunting in 6G-enabled IoT networks to enhance network security. The study proposes a hybrid model combining GANs and transformers for cyber threat-hunting in 6G-enabled IoT networks, with evaluations confirming the high accuracy of the proposed model in detecting various types of IoT cyber threats.

Another application in 6G is proposed by Karaçay et al. [69], focusing on the security aspects of 6G use cases, particularly the All-Senses meeting, which leverages AI and ML to enhance telepresence. The article applies the STRIDE threat modeling framework to systematically identify and classify AI/ML-specific threats in the All-Senses meeting use case, providing a detailed analysis of potential vulnerabilities and mitigation strategies.

The studies discussed above demonstrate the versatile applications of AI in addressing complex cybersecurity challenges across different domains, highlighting its role as a crucial tool in enhancing security measures and mitigating emerging threats.

5) THREAT DETECTION AND PHISHING

An innovative approach to phishing attack prevention involves combining SDN with a Deep Machine Learning technique called the CANTINA approach (DMLCA), as proposed by Raja and Ravi [70]. This methodology integrates SDN's centralized management capabilities with DMLCA to enhance network security. The DMLCA method effectively predicts phishing attacks, showing improvements in detection accuracy and suggesting that the proposed framework is a valuable tool against phishing attacks. Another method to combat phishing emails is introduced by Ding et al. [15], who combined K-means clustering with the traditional Synthetic Minority Over-sampling Technique (SMOTE), creating the KM-SMOTE algorithm. The final model achieved a recall of 95.56%, a precision of 98.85%, and an $\mathbf{F_1}$ score of 97.16%, demonstrating excellent performance. An interesting approach to phishing defense is studied by Asfour and Murillo [71]. The work involved using LLMs to effectively simulate realistic human responses to social engineering attacks. This method offers valuable insights into the susceptibility of different personality traits, enabling organizations to develop targeted cybersecurity training and awareness programs.

The method proposed by Dadvandipour and Ganie [72] employs a multi-layer strategy to mitigate spear phishing by analyzing both email content and attachments. By using Support Vector Machines (SVM) and RandomForest classifiers, sentiment analysis is applied to both email content and attachments to determine their legitimacy. The study demonstrates that this multi-layer approach, combining ML algorithms with hashing algorithms and sentiment analysis, is effective in detecting spear-phishing attacks. Generally, the SVM method is quite popular in phishing email detection [73].

A novel framework called ''Cyber Protect'' was proposed by Gawade et al. [74]. This comprehensive cybersecurity system leverages ML to detect and prevent fraudulent scams and phishing URLs. The system is trained on a large dataset of emails and messages using supervised learning. Furthermore, with the integration of NLP, the ML model achieved commendable precision, recall, and $\mathbf{F_1}$ scores, as NLP techniques significantly improved phishing detection by analyzing linguistic nuances in email content.

AI has also found significant applications in threat detection. Anande and Leeson [75] proposed generating synthetic network traffic data using GANs and classifying Advanced Persistent Threat (APT) samples using Extreme Gradient Boosting (XGBoost). As mentioned earlier, the lack of data is one of the main obstacles to AI applications in some fields of cybersecurity, and the methodology proposed by this study successfully addresses this issue. XGBoost classified synthetic samples with 99.97% accuracy, maintaining a high ROC-AUC, indicating optimal detection performance. The efficiency of Boosting was also confirmed in a study by Hasan et al. [76], in which boosting and explainable AI were used to identify APTs.

Studies by Hlatshwayo [77] and Amarasinghe et al. [78] further confirm the efficiency of AI in threat detection and response. Many benefits have been mentioned, including accelerated response times due to AI's automation of the analysis and prioritization of security alerts, high detection accuracy, and prediction accuracy. The efficiency of GAN

models was also demonstrated in a study by Ferrag et al. [79], where a two-stage intrusion detection framework was proposed, specifically designed to secure IoT environments using GAN and DL models. GANs were used to generate adversarial examples that mimic real-world attack patterns, followed by DL models to detect and classify both normal and attack traffic. The proposed framework achieved high detection accuracy, with a weighted average precision, recall, and $\mathbf{F_1}$ score of 96%, 95%, and 95%, respectively. A study by Uwagboe and Aremora [80] proposes an AI-based security analytics framework to detect and mitigate APTs in cloud environments. The framework aims to provide real-time analysis and response capabilities to minimize the impact of APTs. The proposed framework demonstrated high accuracy in detecting APTs with low false-positive rates during extensive evaluations in simulated cloud environments. A similar result could also be achieved using a Multi-Layer Protection Approach aimed at detecting and mitigating APTs. Mohamed et al. [81] designed an approach based on the MITRE ATT&CK framework, which demonstrated effective detection of APTs through Central Processing Unit (CPU) utilization monitoring. Arshad and Menon [82] explore the application of AI and ML in enhancing honeypot solutions for cybersecurity. The use of adaptive honeypots, powered by LSTM models, allows for detailed behavioral analysis of SSH attacks, achieving high detection accuracy. The research underscores the significance of feature engineering in transforming heterogeneous data into a format suitable for ML model training. Despite challenges such as data quality and occasional misclassifications, the study highlights the potential of AI to revolutionize cybersecurity by predicting attacker behaviors and improving threat detection.

AI clearly has a transformative impact on threat detection and response. However, some issues still need to be addressed, particularly ethical considerations, including privacy and bias in AI algorithms, to ensure the responsible deployment of AI in cybersecurity.

6) AI FRAMEWORKS AND THEORETICAL APPROACHES

Another intriguing application of AI is its integration into modern cybersecurity frameworks. Chomiak-Orsa et al. [83] propose embedding AI into different stages of the cyber kill chain framework. The article highlights that AI applications are particularly promising in the reconnaissance, intrusion, privilege escalation, and data exfiltration stages of the cyber kill chain. Additionally, Molina et al. [84] emphasize the power of AI in both offensive and defensive cybersecurity. Given the broad scope AI can cover, it is crucial to consider the ethical implications and potential risks associated with its use in cybersecurity.

An innovative framework, AI4CYBER, proposed by Iturbe et al. [85], leverages AI to enhance cybersecurity for critical infrastructure. This framework provides a suite of novel AI-driven services designed to manage the entire incident response lifecycle and aligns with the NIST 800-61 guidelines. Another promising AI-based tool is ''Gargoyle Guard'' [86], that improves security through continuous user authentication using Real-Time User Activity Fingerprinting (RTAF). While this technology is highly effective, it does face limitations such as user behavior variability and privacy concerns. Macas et al. [87] explore the application of DL in various cybersecurity tasks, from network intrusion detection to malware analysis and spam filtering. The survey highlights the simplicity, scalability, and reusability of DL models, emphasizing their effectiveness in automating threat detection and response. The authors present a detailed framework for deploying DL in cybersecurity, underscoring successful applications across different domains. Another interesting approach for configuration verification is explored by He et al. [88], who investigate the use of association analysis for network configuration verification in large-scale telecom networks. The proposed system leverages weak association rules to identify infrequent item sets as configuration anomalies. The framework integrates data preprocessing, model training, anomaly detection, and manual annotation to create a robust closed-loop system. Experimental results demonstrate high precision and recall rates, with the system efficiently scanning large datasets within minutes. Despite challenges in detecting frequent misconfigurations, the system's continuous improvement approach through expert feedback and additional ML techniques shows significant promise. This study provides valuable insights into enhancing network management and maintenance through advanced data mining methods.

To sum up, modern trends indicate a growing interest in the intersection of AI and cybersecurity. Increasingly, companies are incorporating AI into their Security Operations Center (SOC) operations, thereby strengthening their cybersecurity infrastructure. However, several challenges in implementation must be considered, such as the need for unified frameworks and collaborative environments [84]. The key takeaways are that numerous experts and research articles highlight the importance of AI in cybersecurity for countering contemporary threats and enhancing SOC operational workflows. Nevertheless, it is crucial to address ethical considerations and privacy concerns, ensuring that AI implementation complies with existing legislation [89], [90], [91], [92]. However, it is evident that AI can be regarded as an essential protection barrier for any organization due to its effectiveness in countering a wide range of potential attacks. Furthermore, in certain cases, AI may prove to be more efficient and cost-effective than relying solely on human resources.

D. AI SYSTEM VULNERABILITIES AND MITIGATING THEM

This subsection will delve into the potential systemic vulnerabilities inherent in AI, as well as the proposed ways of mitigating the effects of these vulnerabilities. The discussion will be organized into smaller sections: first, the vulnerabilities of AI and LLM systems are scrutinized,

followed by an exploration of potential mitigation proposals and strategies.

1) VULNERABILITIES IN AI SYSTEMS

Numerous studies address the vulnerabilities present in AI systems. Some focus on specific aspects of AI, while others adopt a broader perspective, discussing various approaches and frameworks to tackle these issues. A particularly intriguing idea was proposed by Spring et al. [93], who conducted a thought experiment on assigning Common Vulnerabilities and Exposures Identifiers (CVE-IDs) to flaws in ML systems. The article emphasizes the importance of adapting existing vulnerability management frameworks to the unique challenges posed by ML systems. By assigning CVE-IDs to ML algorithm vulnerabilities, better communication and understanding between researchers and practitioners can be fostered, ultimately leading to more secure AI/ML systems. The article also highlights the inadequacy of existing tools like the Common Vulnerability Scoring System (CVSS) in ML contexts, advocating for new frameworks that better capture the complexities of ML systems. Another noteworthy study is conducted by Grosse et al. [94]. The study examines AI vulnerabilities and the disparity between academic threat models and practical AI security by analyzing the six most common AI attacks: poisoning, backdoors, evasion, model stealing, membership inference, and property inference. The study argues that academia sometimes makes overly generous assumptions about attacker capabilities, whereas real-life attacks reveal a different scenario, characterized by more stringent access controls and limited data availability. The study underscores the need for threat models that align more closely with the day-to-day realities of AI deployment. Additionally, the frequent use of pre-trained models and the reliance on domain experts introduces unique vulnerabilities that require closer examination. Aligning research with these practical constraints can pave the way for more robust and realistic AI security measures.

Another noteworthy article addressing AI system vulnerabilities is authored by Scott-Hayward [95], who discusses the fundamental weaknesses of AI-based security systems. The study emphasizes adversarial training, a method recommended for enhancing the robustness of AI models by incorporating adversarial examples into the training dataset. However, this approach is currently implemented in an ad hoc manner. The article underscores the need for standardized adversarial robustness benchmarking, which includes agreed-upon datasets, threat models, evaluation techniques, and metrics. From data poisoning to sophisticated adversarial attacks, these threats can manipulate model outputs and compromise data integrity.

Traditional cyber risks also pose significant threats. For example, consider the potential impact of a botnet or ransomware attack on ML infrastructure. The complexity and stealth of these attacks make defending ML systems a herculean task. A robust defense strategy is essential, blending adversarial training with traditional cybersecurity practices and layered defenses [96].

The vulnerabilities within the AI/ML supply chain, particularly in commonly used libraries like TensorFlow and PyTorch, also pose serious risks [97]. Innovative approaches to counter adversarial attacks include layer-wise adversarial training and Mixed Adversarial Training (MAT), which combine multiple attack methods to improve robustness [98].

Eggers and Sample [99] explore vulnerabilities inherent in AI and ML applications within the nuclear security context. Authors illustrate how AI enhances security through applications like insider threat mitigation and autonomous perimeter defense, while simultaneously introducing new risks. The report highlights the importance of high-quality, unbiased data and robust security measures to protect AI systems from adversarial attacks. Adherence to established standards and best practices, along with continuous monitoring and human oversight, are crucial for mitigating these vulnerabilities.

An intriguing methodology proposed by Mauri and Damiani [100], known as STRIDE-AI, adapts Microsoft's STRIDE framework to address the unique security challenges of AI-ML systems. This framework, enhanced with the rigorous structure of Failure Mode and Effects Analysis (FMEA), offers a comprehensive approach for identifying and mitigating threats throughout the entire ML lifecycle. Its practical application in the TOREADOR H2020 project underscores its effectiveness in real-world scenarios, demonstrating its utility in identifying and mitigating threats in complex ML systems. As noted by Tao et al. [101], the STRIDE framework can also be applied to custom GPTs, providing a structured analysis of potential vulnerabilities and identifying 26 attack vectors with real-world implications. The study highlights the necessity of robust security measures and transparent data handling protocols.

Another notable framework for risk management is the Three Lines of Defense (3LoD) model, studied by Schuett [102]. This model, widely used in various industries, emphasizes the need for clear role assignments and coordinated efforts in managing AI risks. Schuett provides practical suggestions for implementing the model in AI companies, particularly medium-sized research labs and big tech firms. By adapting the 3LoD framework, Schuett offers a structured risk management approach that ensures comprehensive risk coverage and enhances governance. However, potential bureaucratic inefficiencies and the risk of creating a false sense of security should be considered during implementation. This approach should also be applied to securing LLMs. The study by He et al. [103] provides a thorough examination of the ethical and security challenges in LLMs, proposing advanced strategies to fortify these boundaries. By integrating sensitive vocabulary filtering, role-playing detection, and custom rule engines, the authors present a robust framework that balances high performance with stringent ethical standards. The approach addresses the

immediate risks of phishing attacks and privacy breaches while contributing to broader social equity and data protection. The necessity for security frameworks in the AI field is confirmed by numerous studies [104], underscoring the importance of guaranteeing the security of AI systems.

The increasing deployment of AI/ML hardware accelerators in critical sectors such as healthcare, aerospace, and defense has spotlighted the issue of Hardware Trojans (HTs). These covert, malicious modifications can cause significant damage, from leaking sensitive information to undermining the accuracy and reliability of ML models. The intricate and proprietary nature of these accelerators makes HT detection an exceedingly challenging task. Effective mitigation necessitates a multi-layered strategy, including design-for-trust, ML-based anomaly detection, rigorous formal verification, and side-channel analysis [105].

Poisoning attacks on ML models, particularly during the training phase, have advanced considerably. These attacks now utilize sophisticated techniques like bilevel optimization and GAN-based methods to generate highly effective poisoned data. Of particular concern are the recent developments in clean-label attacks, which employ feature collision strategies to create visually indistinguishable but malicious samples. These attacks not only degrade model performance but also evade traditional detection methods, making them exceptionally insidious [106].

AI has a profound impact on cybersecurity. Rayhan and Rayhan [107] provide a comprehensive analysis of AI's transformative role in global security, presenting a balanced view of its risks and opportunities. The study underscores AI's dual function: enhancing cybersecurity while also introducing new cyber threats, like autonomous weapons, a particularly contentious issue, which are shown to offer precision benefits but also carry significant risks of unpredictability and misuse. The most substantial potential of AI is realized when it works in collaboration with human operators, augmenting their capabilities and compensating for their weaknesses. Training and skill development for personnel working with AI are essential. By emphasizing human-AI collaboration, the authors highlight the importance of training and oversight to fully harness AI's potential. Sarker et al. [108] also emphasize the power of AI in cybersecurity, providing a valuable roadmap for integrating data science and ML into cybersecurity strategies. The work underscores the critical role of AI in identifying and mitigating cyber threats, ultimately strengthening the security infrastructure.

LLMs are currently highly popular, and it is crucial to discuss the potential vulnerabilities associated with them. A significant study by Chowdhury and Rahman [109] discusses the limitations and vulnerabilities of ChatGPT. Although recent versions have addressed many of these issues, biases in generated text and a lack of creativity remain persistent challenges. However, these can be mitigated to some extent with carefully crafted prompts. Another limitation highlighted by various studies [110], [111], [112] is the difficulty in maintaining context during extended conversations, which can lead to incoherent responses.

To protect sensitive data, especially in healthcare, robust measures such as data anonymization, encryption, and strict access controls are essential [113]. A case study by Elnawawy et al. [114] demonstrates an attack on an ML-enabled blood glucose monitoring system. By injecting adversarial data points through a known Bluetooth vulnerability [115], the ML model can be manipulated to make incorrect predictions, potentially leading to erroneous insulin dosage recommendations.

Another interesting study [116] reviews potential attacks against LLM models, including adversarial attacks, Structured Query Language (SQL) injection, DoS, and buffer overflow. These attacks exploit vulnerabilities in AI systems, emphasizing the importance of robust security measures, including regular vulnerability scanning, secure coding practices, and continuous monitoring, to mitigate these risks. Privacy and data security are additional concerns that must be addressed with LLMs [117].

Weeks et al. [118] mention toxicity injection attacks, which pose a serious threat to the integrity of open-domain chatbots. These attacks exploit the chatbot's Dialog-based Learning (DBL) framework, where the model is periodically retrained on user interactions, to inject harmful responses into the language model.

Moreover, LLMs present significant socio-economic implications, notably job displacement and widening inequalities. The automation of tasks traditionally performed by humans can lead to job losses and increased stress among workers. The deployment of AI risks exacerbating socio-economic divides, creating an 'AI divide' between those with access to AI technologies and those without [119].

Diffusion models are also vulnerable to backdoor attacks. A novel detection mechanism based on distribution discrepancy [120] achieves a high detection rate for known triggers. Additionally, the proposed evasion strategy employs end-to-end learning to minimize distribution discrepancy, maintaining high attack performance while evading detection with nearly 100% pass rates.

One of the emerging concerns in the realm of generative AI is the phenomenon of data feedback loops. These loops occur when AI-generated content is fed back into the training datasets for future models, leading to risks such as the amplification of biases, degradation of data quality, and increased vulnerability to data poisoning attacks. As AI models learn from synthetic data, the authenticity and reliability of their outputs diminish over time, a process known as ''model collapse'' [121].

An interesting attack proposed by Xu et al. [122] involves multilingual cognitive overload. By presenting harmful prompts in various languages, particularly low-resource ones, LLMs can be coerced into generating unsafe responses. Language-switching scenarios increase the effectiveness of

these attacks, with the success rate rising for languages that have greater word order distance from English. Paraphrasing harmful prompts to replace sensitive words with neutral or less common synonyms increases the likelihood of LLMs generating unsafe responses. Furthermore, LLMs can be prompted to reason backward from an effect to a cause, leading them to generate scenarios that describe how to engage in malicious behavior without facing legal consequences.

Moreover, psychological deception techniques based on persuasion principles (Reciprocation, Consistency, Social Proof, Likeability, Authority and Scarcity) can be adapted to manipulate LLMs effectively, as shown in a study by Singh et al. [123]. For instance, by using social proof and creating scenarios where the model perceives a consensus or common practice, attackers can influence LLM responses. Prompts suggesting widespread acceptance of a harmful action can lead the model to generate supportive responses.

2) MITIGATING THE AI SYSTEM VULNERABILITIES

One of the recommended ways to address risks associated with AI is leveraging frameworks like NIST's AI Risk Management and fostering a culture of continuous learning [124]. The six-dimensional framework proposed by Hu and Chen [125] offers a robust approach to dissecting these dual-edged swords, examining everything from offensive and defensive uses to the inherent vulnerabilities of the AI models themselves. Moving forward, continuous monitoring, stringent policies, and collaborative efforts are essential to harness the full potential of these systems while mitigating their risks.

It is also crucial to exercise caution with the information provided to GPT models, as highlighted in studies by Ananthachari and Singh [126] and Sieja and Wach [127]. Protecting personal information and adhering to data privacy measures is especially important when using publicly available LLMs, as the data input is often utilized for training.

A unique approach proposed by Huang et al. [128] aims to enhance AI's adaptability and robustness in the cognitive domain. Mimicry intelligence, inspired by biological systems, offers a promising method for enhancing the adaptability and resilience of AI in navigating and safeguarding the cognitive domain. By understanding and addressing the unique security challenges posed by the integration of cyber, physical, and cognitive realms, one can better prepare for the evolving landscape of AI-driven threats.

An interesting study has been carried out by Okey et al. [129], examining the cybersecurity implications of LLMs, revealing a dual narrative. On the one hand, sentiment analysis using the Valence Aware Dictionary and Sentiment Reasoner (VADER) [130] model reveals a significant proportion of positive sentiments (43.8%), suggesting an appreciation for ChatGPT's potential in enhancing threat detection and response mechanisms. Conversely, the roBERTa model highlights a considerable amount of negative sentiment (32.7%), reflecting concerns over its misuse in generating malicious code and facilitating cyber attacks. A pioneering study by Li et al. [131] introduced a semantic-preserving algorithm to generate multilingual datasets, revealing that LLMs exhibit varying degrees of vulnerability depending on the language. Notably, models like GPT-4 show enhanced defense mechanisms compared to their predecessors but still struggle with lower-resource languages. In the study, the researchers developed an advanced algorithm to create a multilingual jailbreak dataset, ensuring high semantic fidelity in translations. An empirical study was conducted on widely used open-source and commercial LLMs, including GPT-4 and LLaMa, across nine languages. The study found that certain languages make models more susceptible to jailbreak attacks. LLMs exhibit stronger defenses in English but are more vulnerable in other languages, especially those with fewer resources. Jailbreak templates significantly impacted the models' defense mechanisms, with higher success rates for attacks in low-resource languages like Swahili and Arabic. Fine-tuning techniques such as LoRA have proven effective in reducing the success rates of such attacks, underscoring the necessity for tailored, language-specific security strategies. Another noteworthy contribution is by Esmradi et al. [132], who conducted a comprehensive survey identifying significant attack methods and mitigation strategies. Authors thoroughly examined direct and indirect PIs, highlighting the vulnerabilities LLMs face throughout their lifecycle. By reviewing over 100 recent studies, authors provided insights into how attackers exploit these weaknesses to subvert AI functionalities. Additionally, authors implemented and tested various attack methods, such as self-generated attacks using upgraded DAN prompts, combination attacks, and phishing schemes utilizing LLM-generated fake websites. Authors' work underscores the urgency of addressing these threats through proactive cybersecurity measures, including RLHF, data anonymization, encryption, and advanced filtering techniques.

Jailbreaking techniques present a substantial threat to both the security and ethical utilization of LLMs. By comprehending and addressing these vulnerabilities, developers can fortify the resilience of AI models, ensuring they function within ethical confines and contribute positively to digital interactions. Future research should concentrate on devising advanced defense mechanisms that anticipate and counteract the evolving strategies of prompt-based attacks. Another critical concern is data quality. As highlighted in this section, data play a central role in model training, and in line with the data science principle ''garbage in, garbage out,'' it is essential to ensure that the data used are neither biased nor falsified. Furthermore, it is crucial to control the type of data inputted by users, especially if the data are used for model fine-tuning, in order to prevent potential attacks on the model.

E. NLP IN CYBERSECURITY

Natural Language Processing is an intriguing field that can significantly contribute to strengthening cybersecurity. NLP

is capable of performing numerous tasks within the security domain, and in this section the authors will discuss its potential applications based on recent research.

A novel application of NLP to cybersecurity is presented in the study by Li et al. [133]. The authors introduce NEDetector, an automated system designed to identify new cybersecurity terms, mainly hacking tools and groups, from hacker forums. This tool addresses the challenge of swiftly discovering and analyzing new terminology in the cybersecurity field. By employing a combination of Bi-LSTM and Random Forest algorithms, and leveraging comprehensive feature extraction from word, character, casing, and part-of-speech levels, NEDetector achieves a precision of 89.11% in detecting new cybersecurity terms. This innovative approach not only surpasses traditional neologism detection methods but also proves effective across various platforms, including Twitter, offering a robust tool for early warning and proactive defense against cyber threats.

In the ever-evolving landscape of cybersecurity, the ability to automatically analyze and understand vast amounts of related documents is invaluable. Georgescu's work [134] on an NLP model specifically designed for this purpose stands out. By developing a specialized ontology and leveraging the capabilities of IBM Watson Knowledge Studio, the model achieves impressive precision and recall rates, with an $\mathbf{F_1}$ score of 0.81 for NER and 0.58 for relation extraction.

Singh et al. [135] propose a cutting-edge solution utilizing NLP and DL models to automate software vulnerability detection. By treating source code as text and leveraging models like CodeBERT, which achieved an accuracy of 94%, the study demonstrates a significant leap in identifying and classifying vulnerabilities. This approach not only enhances detection precision but also streamlines the process through an intuitive dashboard, underscoring its potential to comprehensively fortify cybersecurity defenses.

Ukwen and Karabatak [136] prepared a comprehensive review illustrating the transformative role of NLP-based systems in digital forensics and cybersecurity. By leveraging ML and DL, these systems effectively process vast amounts of unstructured data, enhancing applications from disk and mobile forensics to intrusion detection and malware analysis. However, challenges persist in context comprehension and securing NLP systems against AI-driven attacks.

Another interesting approach is "MalVulDroid," a framework designed by Garg and Baliyan [137], to map malware to the vulnerabilities they exploit in Android systems. This mapping utilizes NLP techniques such as Bag-of-Words (BoW), n-gram probability generation, and TF-IDF, combined with ML classifiers like Multilayer Perceptron (MLP), SVM, RIDOR, and PART. This framework demonstrates exceptional accuracy, particularly with the MLP classifier achieving 98.04%. By providing a detailed many-to-many mapping matrix, MalVulDroid offers critical insights for developers and researchers to fortify applications against potential threats during the initial development phases.

Applying NLP to event logs is another promising approach. Steverson et al. [17] explore the application of transformer models and self-supervised learning for cyber intrusion detection using Windows Event Logs (WELs). Their study highlights the effectiveness of this approach in identifying and timing cyber attacks with near-perfect precision and recall. By leveraging WELs, which are widely available and capture diverse activities, this method facilitates decentralized, device-specific responses to intrusions. The high accuracy and promising results suggest a robust framework for autonomous endpoint defense systems, setting the stage for future advancements in multi-log analysis and context consideration.

Marinho and Holanda [138] present a framework for real-time identification and profiling of emerging cyber threats using NLP and ML. By harnessing dynamic data from Twitter and the comprehensive MITRE ATT&CK framework, their system achieves a 77% $\mathbf{F_1}$ score in threat profiling. This methodology exemplifies the potential of integrating open-source intelligence with structured knowledge bases to enhance situational awareness and response strategies in cybersecurity.

Andrew et al. [139] present an approach to mapping Linux shell commands to the MITRE ATT&CK framework using NLP techniques. By leveraging TF-IDF with unigram and bigram tokenization, their model achieves high recall scores, significantly aiding in the automatic identification and categorization of attacker behaviors.

Jha [140] delves into the transformative potential of integrating ML and NLP to bolster smart grid cybersecurity. By harnessing the capabilities of these technologies, smart grids can achieve enhanced anomaly detection, real-time threat response, and global threat intelligence. The fusion of ML and NLP provides a comprehensive approach to analyzing structured and unstructured data, enabling more effective incident response and adaptive security measures.

This section illustrates that NLP technologies offer significant benefits for cybersecurity in various applications. One of the most promising applications is the analysis of log messages to identify potential cyber attacks. Additionally, NLP models prove useful in analyzing textual data related to information security. In one case, NLP models were employed to identify key cybersecurity terms. In conclusion, it can be asserted that NLP technologies are highly valuable in enhancing cybersecurity efforts.

F. LLM AS AN OFFENSIVE TOOL

Generative AI also finds applications in offensive security, particularly in phishing. Schmitt and Flechais [141] present a robust framework for understanding the impact of generative AI on social engineering and phishing attacks. Authors demonstrate how AI's capabilities in realistic content creation, advanced targeting, and automated attack infrastructure significantly enhance the effectiveness of these attacks.

Hazell [142] empirically demonstrates the potential for LLMs like GPT-3.5 and GPT-4 to enhance and scale spear phishing campaigns. By creating spear phishing messages for over 600 British Members of Parliament using GPT-3.5 and GPT-4, Hazell shows that these models significantly lower the barrier to entry for cybercriminals. The approach by Seymour and Tully [143] involves training models using word vector representations of social media posts to generate spear phishing messages.

Research by Bethany et al. [144] revealed that AI-crafted emails, particularly those leveraging internal organizational information, had high success rates in eliciting responses from employees. Despite existing phishing training, many employees remained vulnerable to these sophisticated attacks, highlighting the need for ongoing and effective training programs. The study also demonstrated the effectiveness of advanced ML-based detection techniques, achieving an $\mathbf{F_1}$ score of 0.98 in identifying LLM-generated phishing emails. The study by Falade [145] demonstrates the use of generative AI models, such as FraudGPT and WormGPT, in social engineering attacks. These tools enhance the effectiveness of phishing campaigns and other malicious activities by generating highly convincing and personalized content. The study highlights the practical applications and threats posed by these technologies, including deepfake scams and voice cloning for vishing (voice phishing) attacks. FraudGPT, discovered on dark web channels, automates the creation of phishing emails, undetectable malware, and malicious websites, making sophisticated cyberattacks accessible even to less experienced attackers. WormGPT, a tool based on the GPTJ language model, is designed specifically for malicious activities, enhancing the success rate of Business Email Compromise (BEC) attacks through personalized, convincing emails.

Sharma et al. [146] investigate the comparative effectiveness of human-crafted versus GPT-3-crafted phishing emails, highlighting the significant role of cognitive biases such as authority bias in influencing susceptibility. The study reveals that while human-crafted emails are more effective, feedback mechanisms can significantly improve individuals' phishing detection skills, particularly for AI-generated emails. However, there is a need to carry out the same study with the current model of ChatGPT-4 Omni.

LLMs can be leveraged for a variety of cyberattacks beyond phishing. A study by Heim et al. [147] evaluates the potential of ChatGPT to assist in penetration testing, particularly within an educational setting using HackTheBox machines. The GPT model was employed in various stages of penetration testing, including pre-engagement interactions, intelligence gathering, threat modeling, vulnerability analysis, exploitation, and post-exploitation. The study concludes that while ChatGPT's recommendations are often valuable, they sometimes include incorrect or misleading information, underscoring the need for human oversight and verification. Furthermore, crafting effective prompts is crucial for obtaining relevant responses from ChatGPT, emphasizing the importance of prompt engineering to maximize the model's utility. A study by Feffer et al. [148] evaluated the varied practices of AI red-teaming, particularly in the context of generative AI and LLMs. Authors noted significant inconsistencies in definitions and methods, stressing the importance of iterative and inclusive approaches to effectively identify a wide range of AI vulnerabilities. The study highlighted the need for standardized reporting and transparency to enhance the utility of red-teaming results. By integrating red-teaming with other evaluation methods such as audits and model cards, a more robust framework for ensuring AI safety and trustworthiness can be established.

The study by Naito et al. [149] explored the use of ChatGPT to generate detailed attack scenarios by integrating IT asset management data and vulnerability information. This innovative approach automates the attack path mapping process, reducing reliance on traditional scanning tools. The study highlights the benefits of detailed attack scenarios that include specific CVEs, providing valuable insights for penetration testing and red-teaming. However, the practicality of these scenarios needs further validation in real-world settings.

The efficiency of generative AI in cyberattacks is also demonstrated in a study by Teichmann [150], where the author investigates how generative AI could be utilized to plan and implement ransomware attacks. The findings reveal that these AI tools significantly lower the entry barriers for non-technical criminals and enhance the sophistication of attacks by those with IT expertise. The broad availability of generative AI could lead to an increase in the number and quality of ransomware attacks. A similar conclusion can be drawn from an article by Renaud et al. [151].

The study by Yener and Gal [152] introduces the ''smart adversary'' model, where attackers employ sophisticated techniques to exploit AI/ML vulnerabilities. The study highlights the challenges of processing high-volume and high-velocity data and underscores the need for smarter, adaptive defenses. By incorporating explainable AI, smart data, and adversarial training, the research proposes a robust framework to counteract these advanced threats.

A very novel study by Deng et al. [153] introduces PENTESTGPT, a framework leveraging LLMs to automate penetration testing tasks. The study showcases the proficiency of LLMs in specific sub-tasks but identifies challenges in maintaining context. PENTESTGPT's innovative architecture, featuring Reasoning, Generation, and Parsing modules, significantly enhances the efficiency and accuracy of penetration testing. The framework's success in real-world applications underscores its superiority to the default ChatGPT model and its practical value while highlighting the need for continued research to refine AI tools for cybersecurity. A study by Happe and Cito [16] confirms the efficiency of LLMs (ChatGPT-3.5) in penetration testing. Authors demonstrate the potential of LLMs to assist in high-level task planning and low-level vulnerability hunting through

a closed-feedback loop with a vulnerable virtual machine. Additionally, ChatGPT can not only generate malware code and phishing emails but is also quite effective in SQL Injection Attacks, as demonstrated by Alawida et al. [154]. Authors also confirm AI's capability to generate polymorphic malware and craft convincing phishing emails. Tann et al. [155] explore the use of LLMs like ChatGPT, Google Bard, and Microsoft Bing in solving cybersecurity Capture The Flag (CTF) challenges and professional certification questions. Authors demonstrate that while LLMs can effectively solve factual questions and many CTF challenges, they struggle with conceptual questions and ethical safeguards can be bypassed using jailbreak prompts. In the test cases, ChatGPT solved 6 out of 7, Bard solved 2, and Bing solved only 1.

Shandilya et al. [14] delved into the emergence of GPT-based malware, emphasizing the sophistication and evasiveness of such threats. The use of LLMs like ChatGPT to create polymorphic malware presents significant challenges for traditional detection methods. The study also explores the potential misuse of AI models through jailbreak prompts, which can trick the AI into generating malicious code. To address these threats, the researchers propose advanced detection methods, improved user authentication, and regular adversarial testing.

An interesting study by McKee and Noever [156] illustrates the powerful capabilities of LLMs in generating sophisticated and varied forms of malware. The ability to automate these processes and enhance the complexity of attacks poses significant challenges for traditional cybersecurity defenses. The research highlights the urgent need for advanced AI-driven defensive strategies and real-time anomaly detection systems to counteract these emerging threats.

Beckerich et al. [157] present a proof-of-concept called "RatGPT," demonstrating how generative AI models like ChatGPT can be exploited to deploy Remote Access Trojans (RATs). By leveraging vulnerable plugins and dynamic IP generation, the study shows how attackers can establish undetected communication with victims' systems. A seemingly innocent executable is delivered to the victim, which upon execution, uses ChatGPT to generate and execute payloads. The payload connects to an attacker's command and control (C2) server, allowing remote control without direct interaction with the victim's system. The power of LLMs was also demonstrated in an article by Gupta et al. [158], where authors engaged LLMs in the creation of malware such as WannaCry, NotPetya, Ryuk, REvil, and Locky. The article demonstrates that ChatGPT can generate code snippets for ransomware, including the encryption process and ransom note generation. The AI can provide detailed steps and code to execute such attacks. Further, the power of LLMs was confirmed by Pa et al. [159] in an article where the authors explore the capabilities of generative AI technologies like ChatGPT and Auto-GPT in developing malware. Authors demonstrate that these models can generate functional malicious code within minutes, highlighting significant gaps in safety controls. In his recent study [160], Botacin confirms the strength of the GPT-3 model in malware generation but also highlights that GPT-3 struggles to generate complex malware from simple prompts. A subsequent study by Happe et al. [161] introduces Wintermute, an LLM-guided privilege escalation tool designed to automate and prototype penetration testing tasks. Wintermute utilizes prompts to guide LLMs in discovering and exploiting vulnerabilities. The empirical analysis revealed that GPT-4 outperforms other models in detecting and exploiting file-based vulnerabilities, compared to GPT-3.5-turbo and Llama2. However, several challenges were faced, such as maintaining focus during testing, coping with errors, and handling multi-step exploitation paths. For instance, LLMs often repeated enumeration commands and failed to exploit found vulnerabilities effectively.

In general, it can be seen that LLMs like ChatGPT pose a significant risk to the CIA Triad. As highlighted by Chowdhury et al. [162], there are several concerns, including the storage of sensitive user data, the generation of fake information, and the facilitation of cyberattacks. The study underscores the need for enhanced security measures, ethical guidelines, and continuous monitoring to prevent the misuse of AI. These insights are crucial for understanding the challenges associated with AI-generated content and developing effective strategies to protect information confidentiality, integrity, and availability.

Other studies also confirm the dangerous impact of LLMs on cybersecurity due to their vast range of applications in cyberattacks [163], [164], [165]. The importance of enhanced security measures, user education, and regulatory frameworks to prevent the misuse of AI technologies was highlighted in a study by Iqbal et al. [166]. Derner and Batistič [167] emphasize the need for proper protection against LLM jailbreaking, demonstrating that security measures of LLMs can be bypassed. Dash and Sharma [168] further highlight that generative AI can be exploited to create sophisticated fake content (deepfakes), posing significant threats to cybersecurity.

As demonstrated, LLMs are highly efficient in offensive cybersecurity and should be considered a valuable tool in the penetration tester's toolkit. As anticipated, LLMs excel in generating phishing messages and even producing code for malicious software. Surprisingly, they are also capable of solving various CTF challenges, further highlighting their proficiency across different attack vectors. In conclusion, while LLMs are powerful tools, their potential for harm necessitates careful and responsible usage.

G. LLM IN DEFENSIVE CYBERSECURITY

This section will examine the implementation of LLMs as a defensive tool within the cybersecurity domain. Rigaki

et al. [169] delve into the application of LLMs, such as GPT-3.5 and GPT-4, as agents in cybersecurity environments. Authors introduce NetSecGame, an innovative network security environment designed for realistic and modular testing. The study illustrates that LLM agents can surpass traditional reinforcement learning agents and human testers in planning and executing complex cybersecurity tasks. These findings underscore the potential of LLMs to revolutionize automated network security, notwithstanding challenges like cost and model instability.

An intriguing study conducted by Roy et al. [170] developed a highly accurate BERT-based detection tool designed to identify and block malicious prompts, demonstrating the tool's effectiveness across various platforms. This research underscores the critical role of proactive defense mechanisms and ethical considerations in safeguarding AI technologies.

In another study, Koide et al. [171] introduced ChatPhishDetector, a system leveraging LLMs (GPT-4V) to detect phishing sites with remarkable precision and recall. By combining textual and visual analysis the system excels in identifying suspicious domains and social engineering techniques across multiple languages. This approach not only surpasses existing detection methods but also highlights the significance of advanced prompt engineering and contextual understanding in enhancing cybersecurity defenses.

Jiang [172] delves into the usage of LLMs such as GPT-3.5 and GPT-4 for scam detection. The author outlines comprehensive steps required to build an effective scam detector, from data collection to integration into target systems. The study's preliminary evaluation demonstrated the models' proficiency in identifying scam indicators like unusual sender addresses and suspicious links. This research highlights the potential of LLMs in bolstering scam detection mechanisms and underscores the importance of continuous refinement and collaboration with cybersecurity experts to combat evolving threats.

The study by Heiding et al. [173] compared the effectiveness of phishing emails generated by LLMs like GPT-4 with manually created emails using the V-Triad framework. The authors found that combining human expertise with AI significantly improved the success rates of phishing emails, while reducing the cost and effort for attackers. On the defensive side, LLMs, particularly Claude [174], exhibited strong capabilities in detecting phishing attempts, sometimes surpassing human detection rates. The economic analysis revealed that AI-enabled phishing significantly reduces the cost and effort involved in creating sophisticated phishing attacks, thereby increasing the incentives for attackers to employ AI in phishing campaigns. Notably, emails manually created using V-Triad proved to be the most effective. The study by Trad and Chehab [175] compared the effectiveness of prompt engineering versus fine-tuning LLMs for phishing detection. Authors discovered that fine-tuning models like GPT-2 specifically for phishing URL detection significantly outperformed prompt-engineered models. Fine-tuned models achieved higher accuracy and robustness, making them better suited for real-world applications where phishing URLs are less prevalent. An interesting countermeasure to phishing was proposed by Cambiaso and Caviglione [176]. Authors engage ChatGPT to involve scammers in automated and pointless communications, with the aim of wasting scammers' time and resources. The study shows that AI successfully engaged scammers in extended email threads, with some interactions lasting up to 27 days. The AI-generated responses were effective in keeping scammers engaged and wasting their resources.

McHugh [177] proposed a method to address phishing mail generation, where an anti-expert policy model effectively reduced the generation of phishing content by GPT-3. This study demonstrates that LLMs can be controlled through policy interventions, highlighting that custom-trained policy models can significantly curb the production of phishing emails and enhance cybersecurity defenses. The research underscores the emerging threat of AI-Crime-as-a-Service and the importance of addressing ethical and legal considerations.

An intriguing invention by Kaheh et al. [178], Cyber Sentinel, is a cybersecurity dialogue system leveraging GPT-4 to streamline security tasks. This system excels in both explaining cyber threats (Explainable AI) and taking direct security actions (Actionable AI), thereby enhancing transparency and operational efficiency. Cyber Sentinel integrates multiple components, including an Indicators of Compromise (IoC) signature database, SIEM system, and LLM. The study acknowledges several limitations, such as the need for human oversight, potential privacy concerns, regulatory compliance challenges, and the resource-intensive nature of deploying and maintaining such systems, thus indicating a need for further research in the field.

Moreover, LLMs can be utilized to create Governance, Risk, and Compliance (GRC) policies aimed at mitigating ransomware attacks involving data exfiltration, as demonstrated in a study by McIntosh et al. [179]. The findings reveal that GPT-4-generated policies, when provided with tailored input prompts, can outperform traditional human-generated policies in terms of effectiveness, efficiency, and completeness. However, the study also emphasizes the critical role of human oversight to ensure accuracy and compliance with ethical and legal standards.

The approach by Lempinen et al. [180] integrates ChatGPT-3.5 and Wazuh. This chatbot analyzes security logs and performs actions such as blocking IP addresses and restarting agents. User feedback indicated that the chatbot is easy to use and effective in providing detailed security insights, particularly benefiting users with limited cybersecurity expertise. Nevertheless, technical issues and the need for further enhancements were identified. The study by Prasad et al. [181] confirms this theory and highlights the potential of ChatGPT in supporting Chief Information Security Officers (CISOs) and enhancing cybersecurity

management. ChatGPT demonstrated significant capabilities in defining the role of CISOs, creating cybersecurity frameworks, generating awareness content, and automating security operations.

The efficiency of LLMs in cybersecurity was also demonstrated in smart grid applications by Zaboli et al. [182], showcasing the broad range of applications where LLMs can be effectively deployed. LLM efficiency was also proven in a study by Ali and Kostakos [183], where researchers integrated ML-based anomaly detection with explainable AI and LLMs to enhance cybersecurity operations. The proposed system, named HuntGPT, was used to provide interpretable explanations for detected anomalies. Evaluation results indicate substantial proficiency in technical accuracy and response readability, underscoring the potential of such integrated systems to improve cybersecurity operations. However, it's not only ChatGPT that is used in defensive applications. Ferrag et al. [184] demonstrated that FalconLLM 40B is also effective in automated software vulnerability detection. Utilizing datasets such as FormAI and FalconVulnDB, SecureFalcon achieved remarkable accuracy in both binary classification (94%) and multiclassification (92%). Another study on software vulnerabilities explored the application of LLMs for detecting software vulnerabilities. Authors evaluated models like GPT-3.5-Turbo, Davinci, and CodeGen on datasets including Code Gadgets and CVEfixes. The findings indicate that while LLMs excel in recognizing subtle code patterns, they suffer from high false positive rates. Fine-tuning significantly improves performance, underscoring the importance of tailored training [185].

The BERT model also depicted decent performance in another study by Ferrag et al. [186]. Utilizing the novel Privacy-Preserving Fixed-Length Encoding (PPFLE) technique, SecurityBERT, a BERT-based architecture, achieved remarkable accuracy (98.2%) and rapid inference times (less than 0.15 seconds) on standard CPUs. This model outperformed traditional ML and DL models, demonstrating the potential of advanced LLMs in cybersecurity. The efficiency of the BERT model in cybersecurity has also been demonstrated in other studies, for example in [187] and [188]. Li and Fu, [189] explored the application of transformer-based models, specifically SecureBERT and LLAMA 2, for detecting and classifying Control Area Network (CAN) attacks in vehicular networks. The research highlights the superior performance of these models, proving that other LLMs are also efficient in cybersecurity.

Additionally, it was shown by Garza et al. [190] that LLMs are capable of generating and answering questions related to threat behaviors in the MITRE ATT&CK framework. The research by Wang [191] highlights LLMs' capabilities in identifying patterns, performing real-time analysis, and automating policy generation. These advanced AI models offer significant potential in enhancing cybersecurity resilience by generating informed and dynamic policies. Pearce et al. [192] investigated the use of LLMs like OpenAI's Codex and AI21's Jurassic J-1 for zero-shot vulnerability repair. Authors found that while LLMs can effectively repair synthetically generated and handcrafted security bugs, they struggle with real-world scenarios due to context limitations and reliability issues. Detailed and context-rich prompts significantly enhance the models' performance. However, ensuring that generated fixes are both functional and secure remains a challenge.

Cherqi et al. [193] presented ConGAN-BERT, an innovative framework integrating self-supervised contrastive learning with the GAN-BERT model to enhance cyber threat identification from Open-source Threat Intelligence Feeds (OTIFs). This approach addresses the challenge of limited annotated data by utilizing a semi-supervised learning strategy, significantly improving performance across various datasets. The framework's ability to handle complex and overlapping threat descriptions, combined with an efficient method for selecting hard negatives, leads to notable improvements in accuracy and robustness.

Another interesting idea is a novel intrusion detection framework that integrates BERT with Conditional Generative Adversarial Networks (CGAN). This approach by Li et al. [194] addresses the challenges of class imbalance and limited feature extraction capabilities in traditional IDS models. By augmenting minority attack samples and enhancing feature extraction through BERT, the proposed model achieves significant improvements in detection accuracy across multiple datasets.

The efficiency of integrating LLMs in IDS was also shown in a study by Markevych and Dawson [195], which highlights successful applications in sectors like banking and financial services, where AI-driven IDS have demonstrated high efficacy. However, challenges related to computational complexity, data privacy, and scalability remain.

Guastalla et al. [196] investigated the application of LLMs for detecting DDoS attacks in IoT networks. Utilizing datasets like CICIDS 2017 and the Urban IoT Dataset, the study demonstrated that LLMs, through few-shot learning and fine-tuning, can achieve high accuracy and provide insightful explanations for their predictions. Despite outperforming traditional neural networks, challenges such as hallucinations in fine-tuned models and the high cost of advanced models like GPT-4 were noted.

The study by Mikhalev et al. [197] reveals that GPT-4 excels in fundamental and intermediate cryptographic queries, achieving near-perfect scores. However, the model shows limitations in handling complex tasks, often propagating initial errors and making unwarranted assumptions.

Wang et al. [198] introduce SELF-GUARD, a novel methodology that equips LLMs with the capability to protect themselves against jailbreak attacks. By integrating the advantages of safety training and inherent safeguards, the SELF-GUARD method trains LLMs to scrutinize their responses and append suitable tags indicating harmful or harmless content. This two-stage training strategy not

only bolsters the LLMs' proficiency in identifying harmful material but also ensures they maintain high performance across various tasks. This approach provides robust defense against evolving attack techniques while preserving the general functionalities of the LLMs.

An intriguing application of ChatGPT is illustrated in a study by Shchavinsky et al. [199], where the authors demonstrated that the integration of AI facilitates the rapid creation of realistic and pertinent training scenarios. This significantly enhances the efficiency and effectiveness of the learning process. The study emphasizes the importance of cultivating technical and managerial competencies through practical applications and real-life situations. However, it also highlights the necessity for ongoing critical assessment and refinement of AI-generated content to ensure it aligns with legal, ethical, and contextual standards.

Marshall [200] delves into the profound influence of LLMs like ChatGPT on cybersecurity. His research underscores how LLMs can bolster cybersecurity through efficient code generation and swift threat detection. However, the study also raises pivotal concerns about the misuse of these models in generating phishing emails and malware, thereby lowering the threshold for cybercriminal activities. Real-world examples and experimental findings highlight the potential risks, emphasizing the necessity for robust safeguards and heightened awareness among cybersecurity professionals. Numerous other research articles echo these findings, demonstrating LLMs' efficacy in log analysis, threat detection, vulnerability assessment, and incident response, while also identifying issues such as job displacement and potential misuse of LLMs [201], [202], [203], [204]. Karlsen et al. [205] benchmark several LLMs for log analysis and security using the LLM4Sec pipeline. The authors' study evaluates models like BERT, RoBERTa, DistilRoBERTa, GPT-2, and GPT-Neo across six datasets, revealing that fine-tuning significantly enhances performance. Notably, DistilRoBERTa achieves near-perfect $\mathbf{F_1}$ scores, surpassing current state-of-the-art models.

Shafee et al. [206] assess the performance of various LLM chatbots, including ChatGPT and GPT4all, for Open source intelligence (OSINT) -based Cyber Threat Intelligence (CTI). The authors' findings indicate that while these chatbots excel in binary classification tasks, achieving $\mathbf{F_1}$ scores of 0.94 and 0.90 respectively, they fall short compared to specialized models in NER. This underscores the potential of LLM chatbots to enhance cyber threat awareness, but also highlights the need for targeted training and optimization to improve their NER capabilities. These findings suggest a pathway for integrating LLM chatbots into CTI tools, balancing their strengths in classification with ongoing improvements in entity recognition.

In general, researchers acknowledge the substantial potential of LLMs, yet researchers emphasize the need for robust regulatory frameworks, ethical guidelines, and continuous monitoring to ensure responsible use [207], [208].

As demonstrated in this section, LLMs hold significant potential in cyber defense. LLMs like ChatGPT have proven to be highly efficient across various industries, with particularly exceptional performance in text-related tasks such as phishing prevention and software vulnerability analysis. Notably, LLMs also exhibit strong capabilities in understanding cybersecurity frameworks and providing consultations on related matters.

V. DISCUSSION AND RESULTS

In this article, a thorough systematic literature review was conducted to discern current research directions and methods in the application of AI and LLMs in cybersecurity. The analysis of the literature was guided by the three research questions defined in the introduction of this article. In this section, the findings of the literature analysis of section IV are discussed and results from them to answer the research questions are derived.

AI has profoundly revolutionized the field of cybersecurity, for example with successful applications of ML and DL technologies to log analysis, intrusion detection and intrusion prevention. The recent emergence of LLMs has further influenced the cybersecurity landscape. AI, encompassing ML, DL, and LLMs, can be harnessed as both a defensive and an offensive instrument. The swift progression of LLMs has ushered in new possibilities for both attackers and defenders. In 2023 and 2024, several new LLMs were introduced, including ChatGPT-4, ChatGPT-4o, and LLaMa2. By the end of 2025, OpenAI is poised to unveil a new iteration of ChatGPT, anticipated to be markedly more potent than the current LLMs. The next milestone in AI development is Artificial General Intelligence (AGI) [209], foreseen as an exceedingly powerful tool. As AI advances, the security landscape is expected to evolve in tandem. The authors consider that AGI, under specific circumstances, might be capable of compromising some of the contemporary cryptographic methods. Consequently, the advent of more advanced AI will necessitate the development of more sophisticated defense mechanisms and will introduce new threat landscapes.

With the first research question, the authors set out to determine *how effective AI and LLMs are in cybersecurity applications*. Based on the literature analysis, it is possible to seamlessly integrate AI into IDS and IPS systems, where a host of ML algorithms can be tailored to address specific cybersecurity challenges. Furthermore, AI proves to be invaluable in threat detection and phishing prevention, a testament to its efficacy seen in the success of numerous spam filters. The body of research in this domain is extensive, encompassing both practical and theoretical studies. During the course of the literature analysis, it was noted that there are no existing studies discussing the potential of LLMs in network traffic analysis. However, based on this extensive review of the literature the authors hypothesize that it is feasible to train or develop an LLM specifically for network

traffic analysis, which would enable the LLM to classify certain packet streams as either legitimate or malicious. Such a model could potentially offer explanations for specific packets or frames, providing detailed insights on whether the traffic is malicious or if there are peculiarities (e.g., the use of an outdated protocol that, while not malicious, should be avoided). It is believed that such a model would make a significant contribution to the cybersecurity industry and enhance security accessibility for SMEs and users who are not highly trained.

Beyond LLMs, NLP technologies have also proven effective, notably in identifying emergent cybersecurity terminologies—such as new hacking tools and malware names from hacker forums—and in analyzing vast datasets, which is pivotal in digital forensics and cybersecurity investigations. As the second research question, the authors studied the literature to find out *in what ways are Large Language Models applied in cybersecurity tactics*. Within the cybersecurity realm, LLMs can be deployed for both offensive and defensive purposes. The authors' literature analysis underscores the formidable power of LLMs in offensive operations. Certain studies highlight the efficiency of LLMs in Capture-the-Flag (CTF) challenges, social engineering attacks, and various other domains. Hence, the regulation of this power through safety mechanisms is of paramount importance. It is essential to recognize that the risks associated with LLMs extend beyond the field of cybersecurity. LLMs can disseminate sensitive information, such as manufacturing instructions for explosives and schematics for firearms. Controlling these information sources to avoid malicious purposes is critical. Although LLM safety mechanisms aim to curb the spread of such information, jailbreak techniques can still be employed to extract sensitive data or coerce the LLM into performing illicit or unethical actions. LLMs have also expedited the creation of phishing emails and have proven effective in streamlining penetration testing processes, assisting with each phase. Furthermore, LLMs are instrumental in devising attacks against other AI systems.

While LLMs exhibit remarkable efficacy in offensive operations, they are equally potent in defense. LLMs and NLP technologies are particularly effective in countering social engineering attacks due to their capacity to analyze written and spoken language, making them invaluable tools in such contexts. Furthermore, LLMs are proficient in policy generation, code analysis, IDS/IPS systems, and other areas. Another intriguing application of LLMs is in explainable AI, where cybersecurity engineers can receive explanations or human-readable analyses for specific alerts or logs. This is viewed as having considerable potential, particularly for educational purposes. Similarly, in log analysis, explainable AI can save engineers time by providing immediate explanations for specific logs without necessitating in-depth analysis. The third research question was defined to deduce *the challenges and limitations of Large Language Models in the context of cybersecurity*. Based on the literature analysis, scholars posit that continuing research is imperative, as AI is expected to significantly transform the threat landscape in the imminent future. The majority of the analysed literature emphasized the application of AI and LLMs in defensive security. The offensive capabilities of AI were predominantly explored at the LLM level, alongside the defensive functionalities of these models. The analysis results highlight the inherent vulnerabilities in AI and LLM technologies that must be addressed during development, such as susceptibility to poisoning, backdoors, evasion, model stealing, membership inference, and property inference attacks.

The occasional "hallucinations" of LLMs necessitate consideration to mitigate potential risks. The analyzed research works advise caution against an uncritical reliance on LLMs and AI. The authors' study revealed numerous articles reporting instances where LLMs produced delusional or inaccurate outputs. This introduces significant concerns about the reliability of LLMs in scenarios where human oversight is unfeasible or the stakes are exceedingly high. For instance, entrusting LLMs with calculating rocket trajectories or developing new pharmaceuticals poses substantial risks. This caution extends to cybersecurity, questioning the prudence of relying solely on LLMs for threat identification. While findings generally indicate a high accuracy rate in LLM outputs—trustworthy in nine out of ten cases—it is imperative that LLMs employed in specific fields be trained with domain-specific data. Hence, while general-purpose LLMs demonstrate utility, they exhibit a higher error rate and are less advisable for specialized cybersecurity applications.

Establishing guidelines or ethical frameworks governing the use of LLMs is vital, and many researchers are actively engaged in this effort. Additionally, some studies focus on the legal aspects of LLMs, emphasizing the importance of determining accountability in the global deployment of LLMs, especially within the governmental sector.

The authors believe that before LLMs can be implemented at the governmental or corporate level, it is essential to establish a legal and ethical framework for their use. A good example is the policies adopted by many universities, such as the University of Turku (Finland) [210], which state that AI can be used as a tool, provided its use is disclosed. This approach offers students numerous opportunities for research and study. Similarly, the authors believe that instead of banning AI, corporations and governments should implement guidelines and limitations that regulate its use, ensuring responsible and transparent practices.

VI. CONCLUSION

This systematic literature review and analysis contributed an exhaustive examination of the current deployments and use cases of LLMs and defensive AI techniques within cybersecurity, unveiling both potential pitfalls and advantages. Additionally, it addresses cyberethics and the legal foundations for the use of LLMs.

During the literature review, it became apparent that LLMs and AI have great potential for utilization in cybersecurity.

LLMs have shown remarkable efficacy in phishing attack simulations and in cybersecurity governance, even defending against sophisticated exploits. Additionally, LLMs hold the potential for developing security software, further cementing their role as a formidable tool in cybersecurity innovation. AI and LLMs are versatile, with applications ranging from secure coding to traffic analysis. AI and LLMs substantially reduce the entry barriers for hackers while proving immensely beneficial for penetration testers. For example, LLMs are able to perform penetration testing tasks, and they are also highly efficient at generating phishing messages.

Certain limitations of LLMs were also observed during this study. For instance, in specific fields, the performance of LLMs was found to be below an adequate level. Additionally, there were instances where these models generated hallucinations or inaccurate information, underscoring the importance of thoroughly cross-checking any output provided by LLMs to ensure its reliability. The results of the literature analysis underscore the utility and power of LLM tools in data analysis and text review within the field of cybersecurity, reinforcing the argument for their value in cybersecurity applications. The results also lead to the observation that LLMs can significantly enhance the efficiency of individual workers. They can boost productivity through explainable AI and provide valuable insights into different sub-fields of cybersecurity. The usage of LLMs in cybersecurity is already extensive, including both offensive and defensive applications, but a lot of their potential is still unleashed and significant vulnerabilities and ethical concerns need to be addressed in their deployment for cybersecurity applications.

To conclude, the authors observe a lack of sufficient studies addressing the use of LLMs in network security, highlighting a potential research gap that should be explored.

ACKNOWLEDGMENT

During the preparation of this article, to enhance the clarity and coherence of their work, artificial intelligence (specifically ChatGPT-4o LLM), was used for proofreading, text refinement, and statistical analysis. After application of the AI, all text was manually reviewed by the authors to ensure the accuracy of both the content and the data presented.

REFERENCES

[1] J. Heino, C. Jalio, A. Hakkala, and S. Virtanen, "JAPPI: An unsupervised endpoint application identification methodology for improved zero trust models, risk score calculations and threat detection," *Comput. Netw.*, vol. 250, Aug. 2024, Art. no. 110606.

[2] T. Sowmya and E. A. M. Anita, "A comprehensive review of AI based intrusion detection system," *Meas., Sensors*, vol. 28, Aug. 2023, Art. no. 100827.

[3] G. Suarez-Tangil, E. Palomar, A. Ribagorda, and I. Sanz, "Providing SIEM systems with self-adaptation," *Inf. Fusion*, vol. 21, pp. 145–158, Jan. 2015.

[4] M. Patel, P. P. Amritha, V. B. Sudheer, and M. Sethumadhavan, "DDoS attack detection model using machine learning algorithm in next generation firewall," *Proc. Comput. Sci.*, vol. 233, pp. 175–183, Jan. 2024.

[5] X. Hou, Y. Zhao, Y. Liu, Z. Yang, K. Wang, L. Li, X. Luo, D. Lo, J. Grundy, and H. Wang, "Large language models for software engineering: A systematic literature review," *ACM Trans. Softw. Eng. Methodol.*, pp. 1–76, Sep. 2024.

[6] D. Mon Divakaran and S. T. Peddinti, "LLMs for cyber security: New opportunities," 2024, *arXiv:2404.11338*.

[7] Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang, "A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly," *High-Confidence Comput.*, vol. 4, no. 2, Jun. 2024, Art. no. 100211.

[8] Y. Yigit, W. J. Buchanan, M. G. Tehrani, and L. Maglaras, "Review of generative AI methods in cybersecurity," 2024, *arXiv:2403.08701*.

[9] G. de Jesus Coelho da Silva and C. B. Westphall, "A survey of large language models in cybersecurity," 2024, *arXiv:2402.16968*.

[10] 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, *arXiv:2405.04760*.

[11] M. Guven, "A comprehensive review of large language models in cyber security," *Int. J. Comput. Experim. Sci. Eng.*, vol. 10, no. 3, pp. 507–516, Sep. 2024.

[12] B. Kitchenham and S. M. Charters, "Guidelines for performing systematic literature reviews in software engineering," Keele Univ. Durham Univ., Durham, U.K.:, Tech. Rep. EBSE-2007-01, Dec. 2006. [Online]. Available: https://www.researchgate.net/publication/302924724GuidelinesforperformingSystematicLiterature ReviewsinSoftwareEngineering

[13] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hrobjartsson, M. M. Lalu, T. Li, E. W. Loder, E. Mayo-Wilson, S. McDonald, L. A. McGuinness, L. A. Stewart, J. Thomas, A. C. Tricco, V. A. Welch, P. Whiting, and D. Moher, "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *BMJ*, vol. 372, pp. 1–9, Jan. 2021.

[14] S. K. Shandilya, G. Prharsha, A. Datta, G. Choudhary, H. Park, and I. You, "GPT based malware: Unveiling vulnerabilities and creating a way forward in digital space," in *Proc. Int. Conf. Data Secur. Privacy Protection (DSPP)*, Oct. 2023, pp. 164–173.

[15] X. Ding, B. Liu, Z. Jiang, Q. Wang, and L. Xin, "Spear phishing emails detection based on machine learning," in *Proc. IEEE 24th Int. Conf. Comput. Supported Cooperat. Work Design (CSCWD)*, Dalian, China, May 2021, pp. 354–359.

[16] A. Happe and J. Cito, "Getting pwn'd by AI: Penetration testing with large language models," in *Proc. 31st ACM Joint Eur. Softw. Eng. Conf. Symp. Found. Softw. Eng.*, San Francisco, CA, USA, Nov. 2023, pp. 2082–2086.

[17] K. Steverson, C. Carlin, J. Mullin, and M. Ahiskali, "Cyber intrusion detection using natural language processing on windows event logs," in *Proc. Int. Conf. Mil. Commun. Inf. Syst. (ICMCIS)*, Hague, The Netherlands, May 2021, pp. 1–7.

[18] European Parliament. (Mar. 2024). *Artificial Intelligence Act: European Parliament Legislative Resolution*. Strasbourg. Accessed: Feb. 26, 2024. [Online]. Available: https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138EN.pdf

[19] V. Moret-Bonillo, "Emerging technologies in artificial intelligence: Quantum rule-based systems," *Prog. Artif. Intell.*, vol. 7, no. 2, pp. 155–166, Jan. 2018.

[20] I. E. Naqa and M. J. Murphy, *What is Machine Learning? In Machine Learning in Radiation Oncology: Theory and Applications*. Cham, Switzerland: Springer, 2015.

[21] Z. Zhao and H. Liu, "Spectral feature selection for supervised and unsupervised learning," in *Proc. 24th Int. Conf. Mach. Learn.*, Jun. 2007, pp. 1151–1157.

[22] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[23] R. Mohammed, J. Rawashdeh, and M. Abdullah, "Machine learning with oversampling and undersampling techniques: Overview study and experimental results," in *Proc. 11th Int. Conf. Inf. Commun. Syst. (ICICS)*, Irbid, Jordan, Apr. 2020, pp. 243–248.

[24] G. M. Foody, "Challenges in the real world use of classification accuracy metrics: From recall and precision to the Matthews correlation coefficient," *PLoS ONE*, vol. 18, no. 10, Oct. 2023, Art. no. e0291908.

[25] S. Narkhede, "Understanding AUC-ROC curve," *Towards Data Sci.*, vol. 26, no. 1, pp. 220–227, 2018.

[26] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," 2019, *arXiv:1910.10683*.

[27] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "LLaMA: Open and efficient foundation language models," 2023, *arXiv:2302.13971*.

[28] S. Wu, O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, and G. Mann, "BloombergGPT: A large language model for finance," 2023, *arXiv:2303.17564*.

[29] J. Zhou, J. Ji, J. Dai, and Y. Yang, "Sequence to sequence reward modeling: Improving RLHF by language feedback," 2024, *arXiv:2409.00162*.

[30] T. Kaufmann, P. Weng, V. Bengs, and E. Hüllermeier, "A survey of reinforcement learning from human feedback," 2023, *arXiv:2312.14925*.

[31] M. Raj J, K. VM, H. Warrier, and Y. Gupta, "Fine tuning LLM for enterprise: Practical guidelines and recommendations," 2024, *arXiv:2404.10779*.

[32] T. Mitsunaga, "Heuristic analysis for security, privacy and bias of text generative AI: GhatGPT-3.5 case as of June 2023," in *Proc. IEEE Int. Conf. Comput. (ICOCO)*, Langkawi Island, Malaysia, Oct. 2023, pp. 301–305.

[33] *Chatgpt_Dan*. Accessed: Feb. 2, 2024. [Online]. Available: https://github.com/0xk1h0/ChatGPTDAN

[34] A. Tsamados, L. Floridi, and M. Taddeo, "The cybersecurity crisis of artificial intelligence: Unrestrained adoption and natural language-based attacks," 2023, *arXiv:2311.09224*.

[35] J. A. Chaudhry, S. A. Chaudhry, and R. G. Rittenhouse, "Phishing attacks and defenses," *Int. J. Secur. Appl.*, vol. 10, no. 1, pp. 247–256, Jan. 2016.

[36] (2017). *Technical University of Denmark (DTU) Lyngby, Denmark, an Introduction to Malware*. [Online]. Available: https://orbit.dtu.dk/en/publications/an-introduction-to-malware-2

[37] L. Bosnjak, J. Sres, and B. Brumen, "Brute-force and dictionary attack on hashed real-world passwords," in *Proc. 41st Int. Conv. Inf. Commun. Technol., Electron. Microelectron. (MIPRO)*, Otavija, Croatia, May 2018, pp. 1161–1166.

[38] K. Scarfone, M. Souppaya, A. Cody, and A. Orebaugh, *Technical Guide to Information Security Testing and Assessment*, document NIST Publication SP 800-115, National Institute of Standards and Technology, U.S. Dept. Commerce, Gaithersburg, MD, USA, 2008. [Online]. Available: https://tsapps.nist.gov/publication/getpdf.cfm?pubid=152164

[39] X. Wu, R. Duan, and J. Ni, "Unveiling security, privacy, and ethical concerns of ChatGPT," 2023, *arXiv:2307.14192*.

[40] H. Chugh, "Cybersecurity in the age of generative AI: Usable security & ThreatGPT," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 12, pp. 1–11, Oct. 2023.

[41] R. Gianni, S. Lehtinen, and M. Nieminen, "Governance of responsible AI: From ethical guidelines to cooperative policies," *Frontiers Comput. Sci.*, vol. 4, pp. 1–17, May 2022.

[42] T. Flaih and Y. Jasim, "The ethical implications of ChatGPT AI chatbot: A review," *J. Modern Comput. Eng. Res.*, vol. 2023, pp. 49–57, Oct. 2023.

[43] S. A. Matei and E. Bertino, "Educating for AI cybersecurity work and research: Ethics, systems thinking, and communication requirements," 2023, *arXiv:2311.04326*.

[44] C. Waghmare, *Security and Ethical Considerations When Using ChatGPT in Unleashing the power of ChatGPT: A Real World Bus. Applications*. Berkeley, CA, USA: Apress, 2023, ch. 6, pp. 111–132.

[45] B. Niu and G. F. N. Mvondo, "I am ChatGPT, the ultimate AI chatbot! Investigating the determinants of users' loyalty and ethical usage concerns of ChatGPT," *J. Retailing Consum. Services*, vol. 76, Jan. 2024, Art. no. 103562.

[46] Y. Shi, "Study on security risks and legal regulations of generative AI," *J. Law Sci.*, vol. 2, pp. 17–23, Nov. 2023.

[47] F. Gualdi and A. Cordella, "Theorizing the regulation of generative AI: Lessons learned from Italy's ban on ChatGPT," in *Proc. Annu. Hawaii Int. Conf. Syst. Sci.* Honolulu, HI, USA: IEEE Computer Society, 2024, pp. 2023–2032.

[48] N. Kshetri, "Cybercrime and privacy threats of large language models," *IT Prof.*, vol. 25, no. 3, pp. 9–13, May 2023.

[49] D. Jeong, "Artificial intelligence security threat, crime, and forensics: Taxonomy and open issues," *IEEE Access*, vol. 8, pp. 184560–184574, 2020.

[50] M. Ozkan-Okay, E. Akin, Ö. Aslan, S. Kosunalp, T. Iliev, I. Stoyanov, and I. Beloev, "A comprehensive survey: Evaluating the efficiency of artificial intelligence and machine learning techniques on cyber security solutions," *IEEE Access*, vol. 12, pp. 12229–12256, 2024.

[51] F. Kamoun, F. Iqbal, M. A. Esseghir, and T. Baker, "AI and machine learning: A mixed blessing for cybersecurity," in *Proc. Int. Symp. Netw., Comput. Commun. (ISNCC)*, Montreal, QC, Canada, Oct. 2020, pp. 1–7.

[52] M. Macas and C. Wu, "Review: Deep learning methods for cybersecurity and intrusion detection systems," in *Proc. IEEE Latin-American Conf. Commun. (LATINCOM)*, Santo Domingo, Dominican Republic, Nov. 2020, pp. 1–6.

[53] N. Mohamed, "Current trends in AI and ML for cybersecurity: A state-of-the-art survey," *Cogent Eng.*, vol. 10, no. 2, pp. 1–30, Oct. 2023.

[54] A. Wasif, M. Hamid, and A. Abbas, "AI and cybersecurity: An ever-evolving landscape," *Int. J. Adv. Eng. Technol. Innov.*, vol. 1, no. 1, p. 5271, Jan. 2024.

[55] M. Bagaa, T. Taleb, J. B. Bernabe, and A. Skarmeta, "A machine learning security framework for iot systems," *IEEE Access*, vol. 8, pp. 114066–114077, 2020.

[56] Y. Chen, J. Ding, D. Li, and Z. Chen, "Joint BERT model based cybersecurity named entity recognition," in *Proc. 4th Int. Conf. Softw. Eng. Inf. Manage.*, New York, NY, USA, Jul. 2021, pp. 236–242.

[57] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL*, 2019, pp. 4171–4186.

[58] A. Rahali and M. A. Akhloufi, "MalBERT: Using transformers for cybersecurity and malicious software detection," 2021, *arXiv:2103.03806*.

[59] A. Nitaj and T. Rachidi, "Applications of neural network-based AI in cryptography," *Cryptography*, vol. 7, no. 3, p. 39, Aug. 2023.

[60] E. Hemberg and U.-M. O'Reilly, "Using a collated cybersecurity dataset for machine learning and artificial intelligence," 2021, *arXiv:2108.02618*.

[61] C. Park, J. Lee, Y. Kim, J.-G. Park, H. Kim, and D. Hong, "An enhanced AI-based network intrusion detection system using generative adversarial networks," *IEEE Internet Things J.*, vol. 10, no. 3, pp. 2330–2345, Feb. 2023.

[62] D. L. Marino, C. S. Wickramasinghe, and M. Manic, "An adversarial approach for explainable AI in intrusion detection systems," 2018, *arXiv:1811.11705*.

[63] D. L. Pissanidis and K. Demertzis, "Integrating AI/ML in cybersecurity: An analysis of open XDR technology and its application in intrusion detection and system log management," *Preprints*, vol. 2023, pp. 1–24, Jan. 2024.

[64] M. Ouhssini, K. Afdel, E. Agherrabi, M. Akouhar, and A. Abarda, "DeepDefend: A comprehensive framework for DDoS attack detection and prevention in cloud computing," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 36, no. 2, Feb. 2024, Art. no. 101938.

[65] S. Latif, W. Boulila, A. Koubaa, Z. Zou, and J. Ahmad, "DTL-IDS: An optimized intrusion detection framework using deep transfer learning and genetic algorithm," *J. Netw. Comput. Appl.*, vol. 221, Jan. 2024, Art. no. 103784.

[66] O. Veprytska and V. Kharchenko, "AI powered attacks against AI powered protection: Classification, scenarios and risk analysis," in *Proc. 12th Int. Conf. Dependable Syst., Services Technol. (DESSERT)*, Wulumuqi, China, Dec. 2022, pp. 1–7.

[67] I. Vaccari, A. Carlevaro, S. Narteni, E. Cambiaso, and M. Mongelli, "On the detection of adversarial attacks through reliable AI," in *Proc. IEEE INFOCOM Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, New York, NY, USA, May 2022, pp. 1–6.

[68] M. A. Ferrag, M. Debbah, and M. Al-Hawawreh, "Generative AI for cyber threat-hunting in 6G-enabled IoT networks," 2023, *arXiv:2303.11751*.

[69] L. Karaçay, Z. Laaroussi, S. Ujjwal, and E. U. Soykan, "On the security of 6G use cases: AI/ML-specific threat modeling of all-senses meeting," in *Proc. 2nd Int. Conf. 6G Netw. (6GNet)*, Oct. 2023, pp. 1–8.

[70] R. Ravi, "A performance analysis of software defined network based prevention on phishing attack in cyberspace using a deep machine learning with CANTINA approach (DMLCA)," *Comput. Commun.*, vol. 153, pp. 375–381, Mar. 2020.

[71] M. Asfour and J. C. Murillo, "Harnessing large language models to simulate realistic human responses to social engineering attacks: A case study," *Int. J. Cybersecurity Intell. Cybercrime*, vol. 6, no. 2, pp. 21–49, Aug. 2023.

[72] S. Dadvandipour and A. G. Ganie, "Analyzing and predicting spear-phishing using machine learning methods," *MultidiszciplináRis Tudományok*, vol. 10, no. 4, pp. 262–273, 2020.

[73] S. Salloum, T. Gaber, S. Vadera, and K. Shaalan, "A systematic literature review on phishing email detection using natural language processing techniques," *IEEE Access*, vol. 10, pp. 65703–65727, 2022.

[74] M. Gawade, "Cyber protect: A robust cybersecurity system for fraudulent scam and phishing detection using machine learning techniques," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 11, no. 11, pp. 2480–2487, Nov. 2023.

[75] T. Anande and M. Leeson, "Synthetic network traffic data generation and classification of advanced persistent threat samples: A case study with GANs and XGBoost," in *Proc. Int. Conf. Deep Learn. Theory Appl.*, Rome, Italy, Jul. 2023, pp. 1–18.

[76] M. M. Hasan, M. U. Islam, and J. Uddin, "Advanced persistent threat identification with boosting and explainable AI," *Social Netw. Comput. Sci.*, vol. 4, no. 3, pp. 1–9, Mar. 2023.

[77] M. Hlatshwayo, "Unleashing the power of AI: A deep dive into the integration of AI in cybersecurity for threat detection and response," *J. IoT Intell. Solutions*, vol. 1, pp. 1–25, Jan. 2024.

[78] A. M. S. N. Amarasinghe, W. A. C. H. Wijesinghe, D. L. A. Nirmana, A. Jayakody, and A. M. S. Priyankara, "AI based cyber threats and vulnerability detection, prevention and prediction system," in *Proc. Int. Conf. Advancements Comput. (ICAC)*, Malabe, Sri Lanka, Dec. 2019, pp. 363–368.

[79] M. A. Ferrag, D. Hamouda, M. Debbah, L. Maglaras, and A. Lakas, "Generative adversarial networks-driven cyber threat intelligence detection framework for securing Internet of Things," 2023, *arXiv:2304.05644*.

[80] O. Uwagboe and S. Aremora. (2023). *AI-Based Security Analytics for Cloud Infrastructure: Leveraging Machine Learning Algorithms to Detect and Mitigate Advanced Persistent Threats (APTs) in Cloud Environments*. ResearchGate preprint, Accessed: Feb. 12, 2024. [Online]. Available: https://www.researchgate.net/publication/376168059TitleAI-BasedSecurityAnalyticsforCloudInfrastructureLeveragingMachine LearningAlgorithmstoDetectandMitigateAdvancedPersistentThreats APTsinCloudEnvironments

[81] N. Mohamed, E. Alam, and G. Stubbs, "Multi-layer protection approach (MLPA) for the detection of advanced persistent threats," *J. Positive School Psychol.*, vol. 6, no. 5, pp. 1–23, Jun. 2022.

[82] T. Arshad and S. Menon, "AI-enabled honeypot," *J. Netw. Inf. Secur.*, vol. 11, no. 2, pp. 16–26, Jun. 2023.

[83] I. Chomiak-Orsa, A. Rot, and B. Blaicke, "AI in cybersecurity: The use of AI along the cyber kill chain," in *Proc. 11th Int. Conf. Comput. Collect. Intell.*, Hendaye, France, Aug. 2019, pp. 406–416.

[84] S. B. Molina, P. Nespoli, and F. G. Mármol, "Tackling cyberattacks through AI-based reactive systems: A holistic review and future vision," 2023, *arXiv:2312.06229*.

[85] E. Iturbe, E. Rios, A. Rego, and N. Toledo, "Artificial intelligence for next generation cybersecurity: The AI4CYBER framework," in *Proc. 18th Int. Conf. Availability, Rel. Secur.*, New York, NY, USA, Aug. 2023, pp. 1–8.

[86] C. R. Barone IV, M. Mekni, and M. Nassar, "Gargoyle guard: Enhancing cybersecurity with AI techniques," in *Proc. 3rd Intell. Cybersecur. Conf.*, San Antonio, TX, USA, Oct. 2023, pp. 127–132.

[87] M. Macas, C. Wu, and W. Fuertes, "A survey on deep learning for cybersecurity: Progress, challenges, and opportunities," *Comput. Netw.*, vol. 212, Jul. 2022, Art. no. 109032.

[88] X. He, S. Li, Z. He, and X. Peng, "Research on network configuration verification based on association analysis," in *Proc. 6th Int. Conf. Comput. Sci. Appl. Eng.*, Nanjing, China, Dec. 2022, pp. 1–6.

[89] G. Blanc, Y. Liu, R. Lu, T. Takahashi, and Z. Zhang, "Interactions between AI and cybersecurity to protect future networks," *Ann. Telecommun.*, vol. 77, pp. 727–729, Nov. 2022.

[90] D. Samon. (Dec. 2023). *Artificial Intelligence's Function in Cybersecurity*. Accessed: Feb. 13, 2024. [Online]. Available: https://www.researchgate.net/publication/376784670Artificial Intelligence%27sFunctioninCybersecurity

[91] G. B. Mensah and L. Acquah, *Generative AI, International Cyber-Security Infrastructure, and Geosynchronous Satellite Banking*. Berlin, Germany: ResearchGate Preprint, 2023, Accessed: Feb. 14, 2024, doi: 10.13140/RG.2.2.30417.30567.

[92] M. Ramzan and A. Abbas, "Mindful machines: Navigating the intersection of AI, ML, and cybersecurity," *J. Environ. Sci. Technol.*, vol. 2, no. 2, pp. 1–12, Jan. 2024.

[93] J. M. Spring, A. Galyardt, A. D. Householder, and N. VanHoudnos, "On managing vulnerabilities in AI/ML systems," in *Proc. New Secur. Paradigms Workshop*, Oct. 2020, pp. 111–126.

[94] K. Grosse, L. Bieringer, T. R. Besold, and A. Alahi, "Towards more practical threat models in artificial intelligence security," 2023, *arXiv:2311.09994*.

[95] S. Scott-Hayward, "Securing AI-based security systems," *Geneva Centre Secur. Policy Strategic Secur. Anal.*, vol. 25, pp. 1–25, Jun. 2022.

[96] L. N. Tidjon and F. Khomh, "Threat assessment in machine learning based systems," 2022, *arXiv:2207.00091*.

[97] D. Williams, C. Clark, R. McGahan, B. Potteiger, D. Cohen, and P. Musau, "Discovery of AI/ML supply chain vulnerabilities within automotive cyber-physical systems," in *Proc. IEEE Int. Conf. Assured Autonomy (ICAA)*, Fajardo, PR, USA, Mar. 2022, pp. 93–96.

[98] P. Bountakas, A. Zarras, A. Lekidis, and C. Xenakis, "Defense strategies for adversarial machine learning: A survey," *Comput. Sci. Rev.*, vol. 49, Aug. 2023, Art. no. 100573.

[99] S. L. Eggers and C. Sample, "Vulnerabilities in artificial intelligence and machine learning applications and data," U.S. Dept. Energy, Idaho Nat. Lab (INL), Idaho Falls, ID, USA, Tech. Rep. INL/RPT-22-66111-Rev000, Dec. 2020. [Online]. Available: https://www.osti.gov/biblio/1846969

[100] L. Mauri and E. Damiani, "Modeling threats to AI-ML systems using STRIDE," *Sensors*, vol. 22, no. 17, p. 6662, Sep. 2022.

[101] G. Tao, S. Cheng, Z. Zhang, J. Zhu, G. Shen, and X. Zhang, "Opening a Pandora's box: Things you should know in the era of custom GPTs," 2023, *arXiv:2401.00905*.

[102] J. Schuett, "Three lines of defense against risks from AI," *AI Soc.*, vol. 2023, pp. 1–24, Nov. 2023.

[103] Y. He, J. Qiu, W. Zhang, and Z. Yuan, "Fortifying ethical boundaries in AI: Advanced strategies for enhancing security in large language models," 2024, *arXiv:2402.01725*.

[104] X. Zhang, F. T. Chan, C. Yan, and I. Bose, "Towards risk-aware AI and machine learning systems: An overview," *Decis. Support Syst.*, vol. 159, pp. 1–13, Aug. 2022.

[105] K. I. Gubbi, I. Kaur, A. Hashem, S. Manoj P D, H. Homayoun, A. Sasan, and S. Salehi, "Securing AI hardware: Challenges in detecting and mitigating hardware trojans in ML accelerators," in *Proc. IEEE 66th Int. Midwest Symp. Circuits Syst. (MWSCAS)*, Aug. 2023, pp. 821–825.

[106] Z. Wang, J. Ma, X. Wang, J. Hu, Z. Qin, and K. Ren, "Threats to training: A survey of poisoning attacks and defenses on machine learning systems," *ACM Comput. Surveys*, vol. 55, no. 7, pp. 1–36, Dec. 2022.

[107] A. Rayhan and S. Rayhan. *AI and Global Security: Navigating the Risks and Opportunities*. Berlin, Germany: ResearchGate Preprint, 2023, Accessed: Feb. 12, 2024, doi: 10.13140/RG.2.2.28224.92160/1.

[108] I. H. Sarker, A. S. M. Kayes, S. Badsha, H. Alqahtani, P. Watters, and A. Ng, "Cybersecurity data science: An overview from machine learning perspective," *J. Big Data*, vol. 7, no. 1, pp. 1–29, Jul. 2020.

[109] N. Chowdhury and S. Rahman, "A brief review of ChatGPT: Limitations, challenges and ethical-social implications," Bachelor's thesis, Dept. Comput. Sci. Technol., Chongqing Univ. Posts Telecommun., Chongqing, China, Feb. 2023, doi: 10.5281/zenodo.7629888.

[110] P. P. Ray, "ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope," *Internet Things Cyber-Phys. Syst.*, vol. 3, pp. 121–154, Apr. 2023.

[111] M. Alawida, S. Mejri, A. Mehmood, B. Chikhaoui, and O. I. Abiodun, "A comprehensive study of ChatGPT: Advancements, limitations, and ethical considerations in natural language processing and cybersecurity," *Information*, vol. 14, no. 8, p. 462, Aug. 2023.

[112] K. Y. Thakkar and N. Jagdishbhai, "Exploring the capabilities and limitations of GPT and Chat GPT in natural language processing," *J. Manage. Res. Anal.*, vol. 10, no. 1, pp. 18–20, Apr. 2023.

[113] J. Li, "Security implications of AI chatbots in health care," *J. Med. Internet Res.*, vol. 25, Nov. 2023, Art. no. e47551.

[114] M. Elnawawy, M. Hallajiyan, G. Mitra, S. Iqbal, and K. Pattabiraman, "Systematically assessing the security risks of AI/ML-enabled connected healthcare systems," 2024, *arXiv:2401.17136*.

[115] D. Antonioli, N. O. Tippenhauer, K. Rasmussen, and M. Payer, "BLURtooth: Exploiting cross-transport key derivation in Bluetooth classic and Bluetooth low energy," in *Proc. ACM Asia Conf. Comput. Commun. Secur.*, Nagasaki, Japan, May 2022, pp. 196–207.

[116] A. Qammar, H. Wang, J. Ding, A. Naouri, M. Daneshmand, and H. Ning, "Chatbots to ChatGPT in a cybersecurity space: Evolution, vulnerabilities, attacks, challenges, and future recommendations," 2023, *arXiv:2306.09255*.

[117] R. Pasupuleti, R. Vadapalli, and C. Mader, "Cyber security issues and challenges related to generative AI and ChatGPT," in *Proc. 10th Int. Conf. Social Netw. Anal., Manage. Secur. (SNAMS)*, Nov. 2023, pp. 1–5.

[118] C. Weeks, A. Cheruvu, S. M. Abdullah, S. Kanchi, D. Yao, and B. Viswanath, "A first look at toxicity injection attacks on open-domain chatbots," in *Proc. Annu. Comput. Secur. Appl. Conf.*, New York, NY, USA, Dec. 2023, pp. 521–534.

[119] K. Wach, J. Ejdys, R. Kazlauskaite, P. Korzynski, G. Mazurek, J. Paliszkiewicz, E. Ziemba, and D. Duong, "The dark side of generative AI: A critical analysis of controversies and risks of ChatGPT," *Entrepreneurial Bus. Econ. Rev.*, vol. 11, no. 2, pp. 7–24, Jun. 2023.

[120] Y. Sui, H. Phan, J. Xiao, T. Zhang, Z. Tang, C. Shi, Y. Wang, Y. Chen, and B. Yuan, "DisDet: Exploring detectability of backdoor attack on diffusion models," 2024, *arXiv:2402.02739*.

[121] C. Barrett, "Identifying and mitigating the security risks of generative AI," *Found. Trends Privacy Secur.*, vol. 6, no. 1, pp. 1–52, 2023.

[122] N. Xu, F. Wang, B. Zhou, B. Zheng Li, C. Xiao, and M. Chen, "Cognitive overload: Jailbreaking large language models with overloaded logical thinking," 2023, *arXiv:2311.09827*.

[123] S. Singh, F. Abri, and A. S. Namin, "Exploiting large language models (LLMs) through deception techniques and persuasion principles," 2023, *arXiv:2311.14876*.

[124] D. R. Polaski and M. J. Brienza, "Managing AI: Risks and opportunities," *PM World*, vol. 12, no. 7, pp. 1–12, Jul. 2023.

[125] C. Hu and J. Chen, "A dimensional perspective analysis on the cybersecurity risks and opportunities of ChatGPT-like information systems," in *Proc. Int. Conf. Netw. Netw. Appl. (NaNA)*, Aug. 2023, pp. 324–331.

[126] P. Ananthachari and G. Singh, "Repercussion of ChatGPT in cybersecurity," *Int. J. Res. Publication Rev.*, vol. 4, no. 2, pp. 1429–1430, Feb. 2023.

[127] M. Sieja and K. Wach, "Revolutionary artificial intelligence or rogue technology? The promises and pitfalls of ChatGPT," *Int. Entrepreneurship Rev.*, vol. 9, no. 4, pp. 101–115, 2023.

[128] R. Huang, X. Zheng, Y. Shang, and X. Xue, "On challenges of AI to cognitive security and safety," *Secur. Saf.*, vol. 2, Jan. 2023, Art. no. 2023012.

[129] O. D. Okey, E. U. Udo, R. L. Rosa, D. Z. Rodríguez, and J. H. Kleinschmidt, "Investigating ChatGPT and cybersecurity: A perspective on topic modeling and sentiment analysis," *Comput. Secur.*, vol. 135, Dec. 2023, Art. no. 103476.

[130] C. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 8, May 2014, pp. 216–225.

[131] J. Li, Y. Liu, C. Liu, L. Shi, X. Ren, Y. Zheng, Y. Liu, and Y. Xue, "A cross-language investigation into jailbreak attacks in large language models," 2024, *arXiv:2401.16765*.

[132] A. Esmradi, D. W. Yip, and C. F. Chan, "A comprehensive survey of attack techniques, implementation, and mitigation strategies in Large Language Models," in *Proc. 3rd Int. Conf. (UbiSec)*, vol. 2034, Nov. 2024, pp. 76–95.

[133] Y. Li, J. Cheng, C. Huang, Z. Chen, and W. Niu, "NEDetector: Automatically extracting cybersecurity neologisms from hacker forums," *J. Inf. Secur. Appl.*, vol. 58, May 2021, Art. no. 102784.

[134] T.-M. Georgescu, "Natural language processing model for automatic analysis of cybersecurity-related documents," *Symmetry*, vol. 12, no. 3, p. 354, Mar. 2020.

[135] K. Singh, S. S. Grover, and R. K. Kumar, "Cyber security vulnerability detection using natural language processing," in *Proc. IEEE World AI IoT Congr. (AIIoT)*, Seattle, WA, USA, Jun. 2022, pp. 174–178.

[136] D. O. Ukwen and M. Karabatak, "Review of NLP-based systems in digital forensics and cybersecurity," in *Proc. 9th Int. Symp. Digit. Forensics Secur. (ISDFS)*, Elazig, Turkey, Jun. 2021, pp. 1–9.

[137] S. Garg and N. Baliyan, "MalVulDroid: Tracing vulnerabilities from malware in Android using natural language processing," *J. Web Eng.*, vol. 21, no. 8, pp. 2339–2361, Nov. 2022.

[138] R. Marinho and R. Holanda, "Automated emerging cyber threat identification and profiling based on natural language processing," *IEEE Access*, vol. 11, pp. 58915–58936, 2023.

[139] Y. Andrew, C. Lim, and E. Budiarto, "Mapping Linux shell commands to MITRE ATT&CK using NLP-based approach," in *Proc. Int. Conf. Electr. Eng. Informat. (ICELTICs)*, Jakarta, Indonesia, Sep. 2022, pp. 37–42.

[140] R. K. Jha, "Strengthening smart grid cybersecurity: An in-depth investigation into the fusion of machine learning and natural language processing," *J. Trends Comput. Sci. Smart Technol.*, vol. 5, no. 3, pp. 284–301, Sep. 2023.

[141] M. Schmitt and I. Flechais, "Digital deception: Generative artificial intelligence in social engineering and phishing," 2023, *arXiv:2310.13715*.

[142] J. Hazell, "Spear phishing with large language models," 2023, *arXiv:2305.06972*.

[143] J. Seymour and P. Tully, "Generative models for spear phishing posts on social media," 2018, *arXiv:1802.05196*.

[144] M. Bethany, A. Galiopoulos, E. Bethany, M. B. Karkevandi, N. Vishwamitra, and P. Najafirad, "Large language model lateral spear phishing: A comparative study in large-scale organizational settings," 2024, *arXiv:2401.09727*.

[145] P. V. Falade, "Decoding the threat landscape: ChatGPT, FraudGPT, and WormGPT in social engineering attacks," *Int. J. Sci. Res. Comput. Sci., Eng. Inf. Technol.*, vol. 9, pp. 185–198, Oct. 2023.

[146] M. Sharma, K. Singh, P. Aggarwal, and V. Dutt, "How well does GPT phish people? An investigation involving cognitive biases and feedback," in *Proc. IEEE Eur. Symp. Secur. Privacy Workshops (EuroS&PW)*, Delft, The Netherlands, Jul. 2023, pp. 451–457.

[147] M. Heim, N. Starckjohann, and M. Torgersen, "The convergence of AI and cybersecurity: An examination of ChatGPT's role in penetration testing and its ethical and legal implications," Bachelor's thesis, Dept. Comput. Technol. Inform., Norwegian Univ. Sci. Technol., Trondheim, Norway, May 2023. [Online]. Available: https://hdl.handle.net/11250/3076387

[148] M. Feffer, A. Sinha, W. H. Deng, Z. C. Lipton, and H. Heidari, "Red-teaming for generative AI: Silver bullet or security theater?" 2024, *arXiv:2401.15897*.

[149] T. Naito, R. Watanabe, and T. Mitsunaga, "LLM-based attack scenarios generator with IT asset management and vulnerability information," in *Proc. 6th Int. Conf. Signal Process. Inf. Secur. (ICSPIS)*, Nov. 2023, pp. 99–103.

[150] F. Teichmann, "Ransomware attacks in the context of generative artificial intelligence—An experimental study," *Int. Cybersecur. Law Rev.*, vol. 4, pp. 399–414, Aug. 2023.

[151] K. Renaud, M. Warkentin, and G. Westerman, "From ChatGPT to HackGPT: Meeting the cybersecurity threat of generative AI," *MIT Sloan Manage. Rev.*, vol. 64, no. 3, pp. 1–4, Aug. 2023.

[152] B. Yener and T. Gal, "Cybersecurity in the era of data science: Examining new adversarial models," *IEEE Secur. Privacy*, vol. 17, no. 6, pp. 46–53, Nov. 2019.

[153] 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," 2023, *arXiv:2308.06782*.

[154] M. Alawida, B. A. Shawar, O. I. Abiodun, A. Mehmood, A. E. Omolara, and A. K. Al Hwaitat, "Unveiling the dark side of ChatGPT: Exploring cyberattacks and enhancing user awareness," *Information*, vol. 15, no. 1, p. 27, Jan. 2024.

[155] W. Tann, Y. Liu, J. H. Sim, C. M. Seah, and E.-C. Chang, "Using large language models for cybersecurity capture-the-flag challenges and certification questions," 2023, *arXiv:2308.10443*.

[156] F. McKee and D. Noever, "The evolving landscape of cybersecurity: Red teams, large language models, and the emergence of new AI attack surfaces," *Int. J. Cryptography Inf. Secur.*, vol. 13, no. 1, pp. 1–34, Mar. 2023.

[157] M. Beckerich, L. Plein, and S. Coronado, "RatGPT: Turning online LLMs into proxies for malware attacks," 2023, *arXiv:2308.09183*.

[158] M. Gupta, C. Akiri, K. Aryal, E. Parker, and L. Praharaj, "From ChatGPT to ThreatGPT: Impact of generative AI in cybersecurity and privacy," *IEEE Access*, vol. 11, pp. 80218–80245, 2023.

[159] Y. M. P. Pa, S. Tanizaki, T. Kou, M. van Eeten, K. Yoshioka, and T. Matsumoto, "An attackers dream? exploring the capabilities of ChatGPT for developing malware," in *Proc. 16th Cybersecur. Express Test Workshop*, Marina del Rey, CA, USA, Aug. 2023, pp. 10–18.

[160] M. Botacin, "GPthreats-3: Is automatic malware generation a threat?" in *Proc. IEEE Secur. Privacy Workshops (SPW)*, May 2023, pp. 238–254.

[161] A. Happe, A. Kaplan, and J. Cito, "LLMs as hackers: Autonomous Linux privilege escalation attacks," 2023, *arXiv:2310.11409*.

[162] M. M. Chowdhury, N. Rifat, M. Ahsan, S. Latif, R. Gomes, and M. S. Rahman, "ChatGPT: A threat against the CIA triad of cyber security," in *Proc. IEEE Int. Conf. Electro Inf. Technol. (eIT)*, May 2023, pp. 1–6.

[163] M. A. Elsadig, "ChatGPT and cybersecurity: Risk knocking the door," *J. Internet Services Inf. Secur.*, vol. 14, no. 1, pp. 1–15, Dec. 2023.

[164] M. Al-Hawawreh, A. Aljuhani, and Y. Jararweh, "ChatGPT for cybersecurity: Practical applications, challenges, and future directions," *Cluster Comput.*, vol. 26, no. 6, pp. 3421–3436, Aug. 2023.

[165] P. J. Caven, "A more insecure ecosystem? ChatGPTs influence on cybersecurity," in *Proc. 51st Res. Conf. Commun., Inf., Internet Policy*, Aug. 2023, pp. 1–15.

[166] F. Iqbal, F. Samsom, F. Kamoun, and Á. MacDermott, "When ChatGPT goes rogue: Exploring the potential cybersecurity threats of AI-powered conversational chatbots," *Frontiers Commun. Netw.*, vol. 4, pp. 1–19, Sep. 2023.

[167] E. Derner and K. Batistič, "Beyond the safeguards: Exploring the security risks of ChatGPT," 2023, *arXiv:2305.08005*.

[168] B. Dash and P. Sharma, "Are ChatGPT and deepfake algorithms endangering the cybersecurity industry? A review," *Int. J. Eng. Appl. Sci.*, vol. 10, no. 1, pp. 1–5, Jan. 2023.

[169] M. Rigaki, O. Lukáš, C. Catania, and S. Garcia, "Out of the cage: How stochastic parrots win in cyber security environments," in *Proc. 16th Int. Conf. Agents Artif. Intell.*, Rome, Italy, 2024, pp. 774–781.

[170] S. S. Roy, P. Thota, K. V. Naragam, and S. Nilizadeh, "From chatbots to PhishBots?—Preventing phishing scams created using ChatGPT, Google bard and claude," 2023, *arXiv:2310.19181*.

[171] T. Koide, N. Fukushi, H. Nakano, and D. Chiba, "Detecting phishing sites using ChatGPT," 2023, *arXiv:2306.05816*.

[172] L. Jiang, "Detecting scams using large language models," 2024, *arXiv:2402.03147*.

[173] F. Heiding, B. Schneier, A. Vishwanath, J. Bernstein, and P. S. Park, "Devising and detecting phishing: Large language models vs. smaller human models," 2023, *arXiv:2308.12287*.

[174] M. Enis and M. Hopkins, "From LLM to NMT: Advancing low-resource machine translation with claude," 2024, *arXiv:2404.13813*.

[175] F. Trad and A. Chehab, "Prompt engineering or fine-tuning? A case study on phishing detection with large language models," *Mach. Learn. Knowl. Extraction*, vol. 6, no. 1, pp. 367–384, Feb. 2024.

[176] E. Cambiaso and L. Caviglione, "Scamming the scammers: Using ChatGPT to reply mails for wasting time and resources," 2023, *arXiv:2303.13521*.

[177] J. McHugh, "Defensive AI: Experimental study," Ph.D. dissertation, Dept. Cybersecurity, Marymount Univ., Arlington, VA, USA, Apr. 2023.

[178] M. Kaheh, D. K. Kholgh, and P. Kostakos, "Cyber sentinel: Exploring conversational agents in streamlining security tasks with GPT-4," 2023, *arXiv:2309.16422*.

[179] T. McIntosh, T. Liu, T. Susnjak, H. Alavizadeh, A. Ng, R. Nowrozy, and P. Watters, "Harnessing GPT-4 for generation of cybersecurity GRC policies: A focus on ransomware attack mitigation," *Comput. Secur.*, vol. 134, Nov. 2023, Art. no. 103424.

[180] M. Lempinen, A. Juntunen, and E. Pyyny, "Chatbot for assessing system security with OpenAI GPT-3.5," Bachelor's thesis, Dept. Comput. Sci. Eng., Univ. Oulu, Oulu, Finland, Jun. 2023. [Online]. Available: https://oulurepo.oulu.fi/handle/10024/42952

[181] S. G. Prasad, V. C. Sharmila, and M. K. Badrinarayanan, "Role of artificial intelligence based chat generative pre-trained transformer (ChatGPT) in cyber security," in *Proc. 2nd Int. Conf. Appl. Artif. Intell. Comput. (ICAAIC)*, Salem, India, May 2023, pp. 107–114.

[182] A. Zaboli, S. L. Choi, T.-J. Song, and J. Hong, "ChatGPT and other large language models for cybersecurity of smart grid applications," 2023, *arXiv:2311.05462*.

[183] T. Ali and P. Kostakos, "HuntGPT: Integrating machine learning-based anomaly detection and explainable AI with large language models (LLMs)," 2023, *arXiv:2309.16021*.

[184] M. A. Ferrag, A. Battah, N. Tihanyi, R. Jain, D. Maimut, F. Alwahedi, T. Lestable, N. S. Thandi, A. Mechri, M. Debbah, and L. C. Cordeiro, "SecureFalcon: Are we there yet in automated software vulnerability detection with LLMs?" 2023, *arXiv:2307.06616*.

[185] M. Das Purba, A. Ghosh, B. J. Radford, and B. Chu, "Software vulnerability detection using large language models," in *Proc. IEEE 34th Int. Symp. Softw. Rel. Eng. Workshops (ISSREW)*, Oct. 2023, pp. 112–119.

[186] M. A. Ferrag, M. Ndhlovu, N. Tihanyi, L. C. Cordeiro, M. Debbah, T. Lestable, and N. S. Thandi, "Revolutionizing cyber threat detection with large language models: A privacy-preserving BERT-based lightweight model for IoT/IIoT devices," *IEEE Access*, vol. 12, pp. 23733–23750, 2024.

[187] K. Ameri, M. Hempel, H. Sharif, J. Lopez Jr., and K. Perumalla, "CyBERT: Cybersecurity claim classification by fine-tuning the BERT language model," *J. Cybersecurity Privacy*, vol. 1, no. 4, pp. 615–637, Nov. 2021.

[188] Z. Wang, J. Li, S. Yang, X. Luo, D. Li, and S. Mahmoodi, "A lightweight IoT intrusion detection model based on improved BERT-of-theseus," *Expert Syst. Appl.*, vol. 238, Mar. 2024, Art. no. 122045.

[189] X. Li and H. Fu, "SecureBERT and LLAMA 2 empowered control area network intrusion detection and classification," 2023, *arXiv:2311.12074*.

[190] E. Garza, E. Hemberg, S. Moskal, and U. OReilly, "Assessing large language models knowledge of threat behavior in MITRE ATT&CK," in *Proc. 3rd Workshop Artif. Intell. Enabled Cybersecur. Anal. (KDD)*, Long Beach, CA, USA, Aug. 2023, pp. 1–7.

[191] F. Wang, "Using large language models to mitigate ransomware threats," *Preprints*, vol. 2023, pp. 1–12, Nov. 2023.

[192] H. Pearce, B. Tan, B. Ahmad, R. Karri, and B. Dolan-Gavitt, "Examining zero-shot vulnerability repair with large language models," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2023, pp. 2339–2356.

[193] O. Cherqi, Y. Moukafih, M. Ghogho, and H. Benbrahim, "Enhancing cyber threat identification in open-source intelligence feeds through an improved semi-supervised generative adversarial learning approach with contrastive learning," *IEEE Access*, vol. 11, pp. 84440–84452, 2023.

[194] F. Li, H. Shen, J. Mai, T. Wang, Y. Dai, and X. Miao, "Pre-trained language model-enhanced conditional generative adversarial networks for intrusion detection," *Peer Peer Netw. Appl.*, vol. 17, no. 1, pp. 227–245, Nov. 2023.

[195] M. Markevych and M. Dawson, "A review of enhancing intrusion detection systems for cybersecurity using artificial intelligence (AI)," in *Proc. Int. Conf. Knowl. Based Org.*, Sibiu, Romania, Jul. 2023, pp. 1–9.

[196] M. Guastalla, Y. Li, A. Hekmati, and B. Krishnamachari, "Application of large language models to DDoS attack detection," in *Proc. Int. Conf. Secur. Privacy Cyber-Phys. Syst. Smart Vehicles*, Oct. 2024, pp. 83–99.

[197] V. Mikhalev, N. Kopal, and B. Esslinger, "Evaluating GPT-4s proficiency in addressing cryptography examinations," *Cryptologia*, vol. 48, no. 1, pp. 1–10, Mar. 2024.

[198] Z. Wang, F. Yang, L. Wang, P. Zhao, H. Wang, L. Chen, Q. Lin, and K.-F. Wong, "Self-guard: Empower the LLM to safeguard itself," 2023, *arXiv:2310.15851*.

[199] Y. V. Shchavinsky, T. M. Muzhanova, Y. M. Yakymenko, and M. M. Zaporozhchenko, "Application of artificial intelligence for improving situational training of cybersecurity specialists," *Inf. Technol. Learn. Tools*, vol. 97, no. 5, pp. 215–226, Oct. 2023.

[200] J. Marshall, "What effects do large language models have on cybersecurity," Old Dominion Univ., Norfolk, VA, USA, Tech. Rep., May 2023. [Online]. Available: https://digitalcommons.odu.edu/covacci-undergraduateresearch/2023spring/projects/15

[201] H. J. Kam, C. Zhong, H. Liu, and A. Johnston, "The blend of human cognition and AI automation: What will ChatGPT do to the cybersecurity landscape?" in *Proc. Dewald Roode Workshop Inf. Syst. Secur. Res.*, Jun. 2023, pp. 1–22.

[202] M. A. Hadi, M. N. Abdulredha, and E. Hasan, "Introduction to ChatGPT: A new revolution of artificial intelligence with machine learning algorithms and cybersecurity," *Sci. Arch.*, vol. 4, no. 4, pp. 276–285, 2023.

[203] S. Biswas, "Role of ChatGPT in cybersecurity," *SSRN Preprint*, vol. 2023, pp. 1–3, Jan. 2023, doi: 10.2139/ssrn.4403584.

[204] M. Ayaim, "How ChatGPT can be used as a defense mechanism for cyber attacks," Old Dominion Univ., Norfolk, VA, USA, Tech. Rep., Dec. 2023. [Online]. Available: https://digitalcommons.odu.edu/covacci-undergraduateresearch/2023fall/projects/15

[205] E. Karlsen, X. Luo, N. Zincir-Heywood, and M. Heywood, "Benchmarking large language models for log analysis, security, and interpretation," 2023, *arXiv:2311.14519*.

[206] S. Shafee, A. Bessani, and P. M. Ferreira, "Evaluation of LLM chatbots for OSINT-based cyber threat awareness," 2024, *arXiv:2401.15127*.

[207] F. Okeke, "An assessment of the use of generative AI in cybersecurity: Challenges and opportunities," Bournemouth Univ., Bournemouth, U.K., Tech. Rep., Dec. 2023, doi: 10.13140/RG.2.2.20613.12001.

[208] S. Neupane, I. A. Fernandez, S. Mittal, and S. Rahimi, "Impacts and risk of generative AI technology on cyber defense," 2023, *arXiv:2306.13033*.
[209] P. Botu, "Consciousness for AGI," in *Proc. 10th Annu. Int. Conf. Biologically Inspired Cogn. Archit. (BICA)*, Seattle, WA, USA, Aug. 2019, pp. 365–372.
[210] (2024). *University of Turku Guideline on Artificial Intelligence in Teaching and Studying*. Accessed: Mar. 31, 2024. [Online]. Available: https://utuguides.fi/artificialintelligence

ISMAYIL HASANOV received the M.Sc. degree in information and communication technology from the University of Turku, Finland, in 2023. He is currently a part-time Doctoral Researcher with the University of Turku and a NOC and SOC Engineer with Openfactory Nordic oy. He has over six years of experience in networking technologies. His current research interests include network security, cybersecurity policies, the security of LLMs, and the application of AI in cybersecurity.

SEPPO VIRTANEN (Senior Member, IEEE) received the D.Sc. (Tech.) degree in communication systems from the University of Turku, Finland, in 2004. He is currently a Professor of cyber security engineering with the Department of Computing, University of Turku. His current research interests include the application of artificial intelligence and large language models to network and cyber security, security of smart environments, and cyber security in digitalization and societal processes.

ANTTI HAKKALA received the D.Sc. (Tech.) degree in communication systems from the University of Turku, Finland, in 2017. He is currently a University Teacher of communication systems and cyber security with the Department of Computing, University of Turku. He has 15 years of experience in teaching engineering students on cyber security and communication systems engineering and has supervised over 100 bachelor's and master's theses on cyber security topics. His current research interests include application of AI and LLMs to cyber and network security, digital forensics, and security and privacy in the networked information society.

JOUNI ISOAHO received the M.Sc. (Tech) degree in electrical engineering and the Lic.Tech. and Dr.Tech. degrees in information technology from Tampere University of Technology, Finland, in 1989, 1992, and 1995, respectively. Since 1999, he has been a Professor with the University of Turku, Finland. The core of his research is communication and cyber security technologies. His current research interests include security of autonomous systems and AI, human and societal cybersecurity, and smart technology and digitalization.