A Comprehensive Overview of Large Language Models (LLMs) for Cyber Defences: Opportunities and Directions

Mohammed Hassanin, Nour Moustafa

***Abstract*—The recent progression of Large Language Models (LLMs) has witnessed great success in the fields of data-cetric applications. LLMs trained on massive textual datasets showed ability to encode not only context but also ability to provide powerful comprehension to downstream tasks. Interestingly, Generative Pre-trained Transformers utilised this ability to bring AI a step closer to human being replacement in at least data-centric applications. Such power can be leveraged to identify anomalies of cyber threats, enhance incident response, and automate routine security operations. We provide an overview for the recent activities of LLMs in cyber defense sections, as well as categorization for the cyber defense sections such as threat intelligence, vulnerability assessment, network security, privacy preserving, awareness and training, automation, and ethical guidelines. Fundamental concepts of the progression of LLMs from Transformers, Pre-trained Transformers, and GPT is presented. Next, the recent works of each section is surveyed with the related strengths and weaknesses. A special section about the challenges and directions of LLMs in cyber security is provided. Finally, possible future research directions for benefiting from LLMs in cyber security is discussed.**

I. Introduction

Cyber security is an essential realm that focuses primarily on safeguarding digital systems (computer systems, networks, and data) from any wrongdoings, including unauthorized access or interruption. While technology has increasingly been integrating into every facet of life, the necessity of cyber security is growing as well. It encompasses a vast set of practices, ranging from guidelines to applications designed to protect data. With the development of digital systems including the Internet of Things (IoT) [1], Internet of Vehicles (IoV) [2], Internet of Battlefield Things (IoBT) [3], and Cyber-Physical Systems (CPS) [4], as well as technologies such as smartphones, tablets [5], autonomous vehicles, digital twins [6], cryptocurrencies [7], and smart homes [8], the scope of cyber security is expanding accordingly. The more these technologies integrate into daily life, they also widen the potential landscape for cyber threats. Therefore, they necessitate advanced security technologies to safeguard such systems from intrusions and inaccessibility.

Cyber defense is a realm of cyber security that is responsible for detecting, preventing, and responding to attacks or cyber threats [9]. It combines different advanced technologies such as deep learning (DL) [10] and large language models (LLMs), as well as maintaining set practices such as threat intelligence and incident response in order to achieve security to the digital system. The goal of cyber defence is twofold: to safeguard digital systems from cyber threats and to build robust systems able to recover from any attacks. Lately, cyber attacks has become more sophisticated, that dictates drawing intensive attention to build proactive technologies for cyber defence. To achieve that, research community need to spend more time and effort to be ahead of cyber attacks techniques.

In this digital era, the necessity of cyber defense in cannot be overstated. As every domain of the life such as economy, health, education, communities, and social welfare, the need to protect such systems from cyber threats becomes mandatory. Ranging from data breaches and ransomware to espionage, the scope of cyber threats dictates the urgent need for robust cyber defense strategies [11]. For instance, Sviatun *et al.* reported that cyber crimes cost US around 1 trillion every year and 20 billion for the whole world [12], highlighting the economic impact of lacking cyber defenses. Moreover, such development of cyber defense techniques also supports the protection of individual privacy. Therefore, cyber defense is not only an integral necessity, but it is a fundamental component of daily life to ensure a secure, stable, and fair society.

Several strategies can be used for cyber defence, however, machine learning stays on top of all the technologies. As the recent progression of machine learning and large language models (LLMs) represents a paradigm shift in AI, it impact all the science branches, particularly cyber security. Machine learning algorithms used small datasets to learn simple computer vision tasks such as plate numer recognition and other simple tasks. The invention of Deep Learning along with the increase in computational power was the game changer to machine learning paradigm, for it resolved the challenges of neural network such as that learning a networks with many layers[13]. Meanwhile, Natural Language Processing (NLP) utilised this progression of deep learning and achieved paradigm shift in text processing by the invention of Transformers, which in turn impact the whole machine learning field [14]. It achieved that success by allowing the model to weigh the importance of all parts of the input space simultaneously. This simultaneous processing helped encode long-range dependencies and thus the context [15]. Parallel to these advancements, Generative Adversarial Networks (GANs) were first proposed by Goodfellow *et al.* as a different way of training that instead based on data generation models [16]. Simply put, it composes two neural networks: a generator and

M. Hassanin is with University of South Australia, Australia, and the Faculty of Computers and Information, Fayoum University, Egypt.
N. Moustafa is with University of New South Wales, Australia.

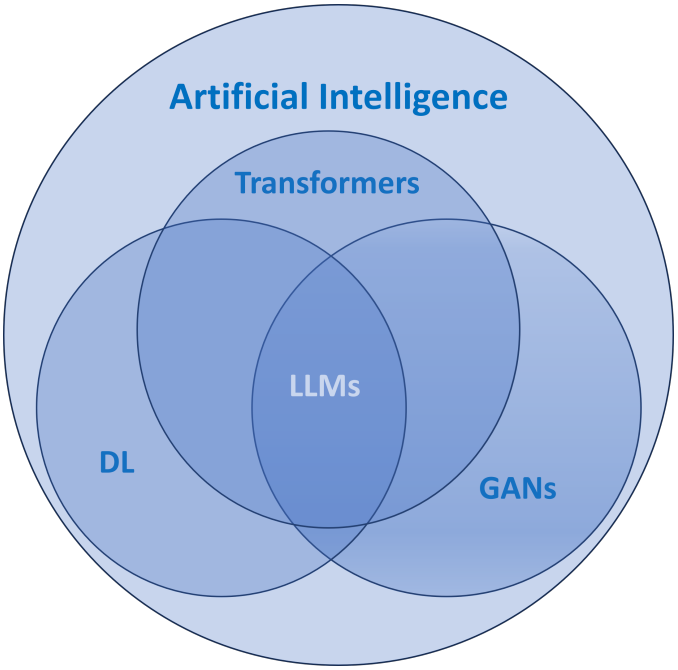

Fig. 1: The evolution of LLMs and GPT is based on deep learning, GANs, and Transformers.

a discriminator, which are trained concurrently using adversarial techniques. The generator produces synthetic data that closely resemble to the orginal sample, while the discriminator evaluates how far the generated image from the original one. They have been used in many designs and diverse fields including image synthesis, style transfer, and advancements in medical imaging techniques [17]. These three different designs of deep learning, *i.e., deep neural networks, GANs, Transformers* paved the way to LLMs evolution as Figure 1 shows.

As a new step towards the current progression of LLMs, Bidirectional Encoder Representations from Transformers (BERT) has been introduced with high rates of success. Such models are trained on massive corpora of text data and then they generate coherent. Contextually, these models generate appropriate text that can successfully perform language translation and answer questions with high accuracy [19], [20]. Recently, Generative Pre-trained Transformers (GPT) have been proposed as a hybrid evolution of Generative Neural Networks (GANs) and Masked Pre-trained Transformers. The result of this blend is that models are capable of generating coherent and contextually relevant text given a good input, called a prompt. Thus, a significant step forward towards more intelligent AI that will help automating a lot tasks, including text comprehension and visual understanding [21]. One of the beneficiaries of this progression is certainly cyber security in both sides; threats and defence [22]. Large Language Models (LLMs) like GPT (Generative Pre-trained Transformers) are becoming increasingly significant in the field of cyber defense, offering a range of capabilities that enhance both the effectiveness and efficiency of cybersecurity operations. Their extensive training on diverse datasets enables them to understand and generate human-like text (see Figure 2), which can be leveraged in several critical areas of cyber defense. The illustrated power of LLMs in analyzing massive amounts of data, including web logs, network traffic, and regular datasets, will certainly push the cyber defense realm ahead, particularly with early threat detection, which as a result allows faster and more timely response [23]. Such capability of text comprehension and contextual understanding is required for cyber security routinely tasks such as documentation and email responses, which will free security experts to focus on complicated threats. Another aspect is the benefits of LLMs in generating training scenarios and simulations which will enrich the training of algorithms as well as the staff. Even the realm of governance and guidelines, LLMs provide an amazing features to cyber security by helping in keeping regulations up to date (see Figures 3 and 5).

In this survey, we review LLM-based cyber defence techniques specific to machine learning defense strategies. We cover the most of the unique approaches in the cyber defense since the evolution of LLMs. Our survey broadly classifies LLM techniques into threat intelligence, vulnerability assessment, network traffic security, privacy preservation, personnel awareness, ethical security. We noted that some of the techniques fit within more than one category; however, we put them under the dominant of the method. Such a classification helps the researcher to draw attention to the gaps, as well as easiness to get insights. Figure 4 shows the categories of LLM-base cyber defence technologies.

We highlight that this review is focusing only on cyber defence due to vastness of cyber security fields shown in Figure 5. It is clear from Figure 3 that the number of published papers in the last year has dramatically increased compared to before as we anticipate to even exponential increase the upcoming years. Further, our article lists the main articles with unique ideas to provide the cyber defence community with keypoints to find the proper solutions for the real gaps and issues. Research gaps are provided in the current research context, as well as plausible research directions along with with future areas are projected and discusses.

Since LLMs have been used across cyber security, few surveys [24], [25] review the recent trends of LLMs in cyber security. Although LLMs provide significant accuracy, this comes at the account of a lot of issues, which limit their feasibility real-life domains. Cyber-defence strategies need consistent adaptation to fix the ever-changing attacks techniques and tricks. In [26], Yao *et al.* provide a survey on all the cyber security rather than the defense. Also, there is a recent review published on drawbacks and vulnerability of LLMs [27], which different from our scope. In contrast, our article is primarily focusing on the LLM-based cyber defence strategies and their impact on the light of the previous works.

II. Threat Intelligence

Threat intelligence is a realm of cyber defence that specialises on network traffic engineering including collection, evaluation, and analysis of any potential or current threats that harms the system [28]. It focuses primarily on understanding the cyber attacks and the environment of the cyber criminal

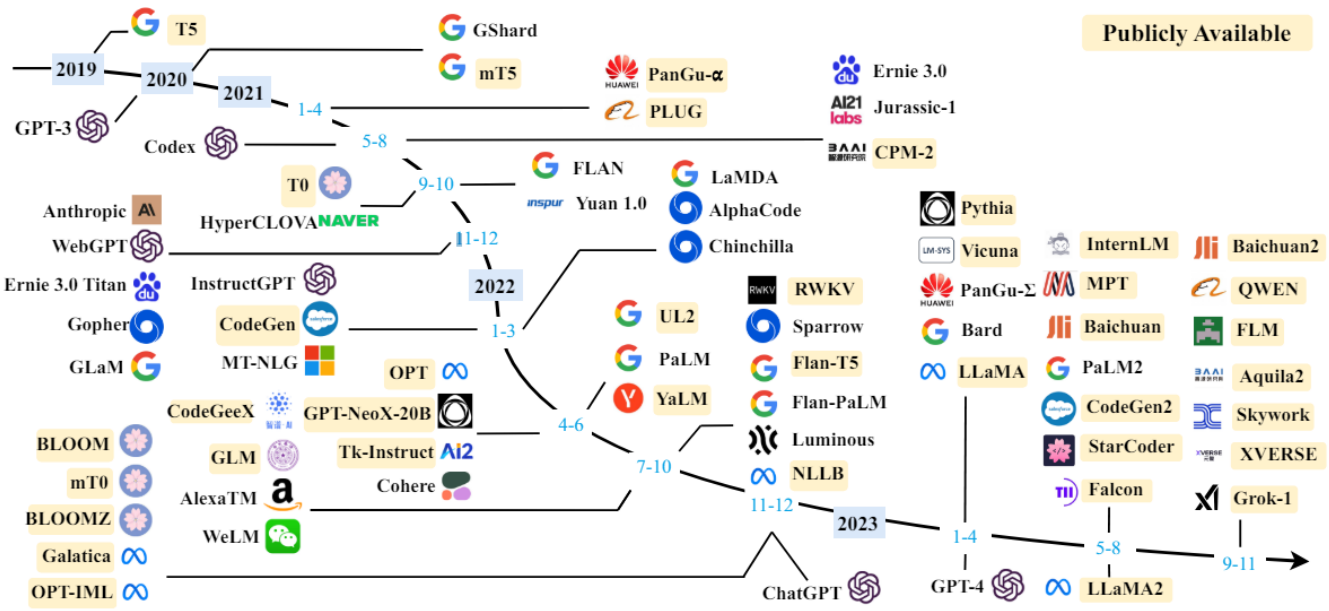

Fig. 2: The publicly available LLMs in the most existing in recent years in a timeline. The timeline is ordered based on the publishing date. Figure from [18]

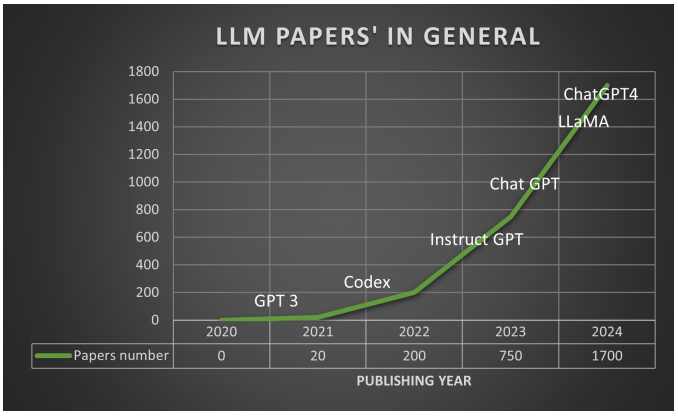

Fig. 3: Visual description of the number of LLMs usage in research in all fields in general.

or attacker. The goal is to use all the techniques to imagine the cyber criminal way of attack to respond with the best strategy and even expect it before hand. By collecting and analyzing such data from various sources, including internal technical programs, human intelligence (HUMINT), open-source intelligence (OSINT), social media, etc, cyber threats landscape can be understood and predicted. The result of this process will be used to update security strategies, help incident response teams, as well as refine threat detection capabilities, and thus improve the whole security operations.

Successful threat intelligence outputs multiple operations: collection of important data, extracting actionable insights through analysis, updating the appropriate stakeholders of the results, and finally list these insights in the security operation tasks. The outcomes of threat intelligence is helpful for organisation to have the best response to the attacks. This is the first step process of attack response, however, it is complicated due to the advancement of the attacks and the tricks, which dictates more attention from the research community of cyber defence. In this section, we are reviewing LLM-based penetration testing strategies as we provide discussion on each method, significance and drawbacks (see Figure 6).

Sufi *et al.* used GPT models to improve the capabilities of OSINT to automatically extract, analyse, and summarise important information from historical cyber incident reports [29]. They succeeded in improving the accuracy of threat intelligence and as a result ability of forecasting future cyber threats. This work emphasised LLMs capacity to conclude deep insights to cyber security professionals that help in answering efficiently to emergent risks. Siracusano *et al.* used LLMs to address the issue of extracting pertinent information from unorganised cyber threat intelligence (CTI) data, which is usually a time-consuming and labor-intensive process [30]. They presented a dataset of 204 real-world reports that adhere to the STIX ontology, which is considerably larger than existing datasets. Then, GPT-3.5 is used to enhance the effectiveness of CTI extraction. The empirical experiments demonstrate that aCTIon surpasses the previous methods with a significant margin in identifying items such as malware, threat actors, and attack patterns from cyber threat data. In [31], LOCALINTEL introduced a step forward in the process of generating organization-specific threat intelligence automation using LLMs. Hu *et al.* used LLMs to create comprehensive knowledge graphs, notably LLM-TikG extract and organising threat intelligence from extensive amounts of unorganised data [32]. This allows for more accurate anticipation, and thus, more mitigation of cyber attacks. Utilising LLMs improved the accuracy and speed of knowledge graph generation that provide a strong framework to promptly detect and address emerging threats. Sewak *et al.* presented a novel method

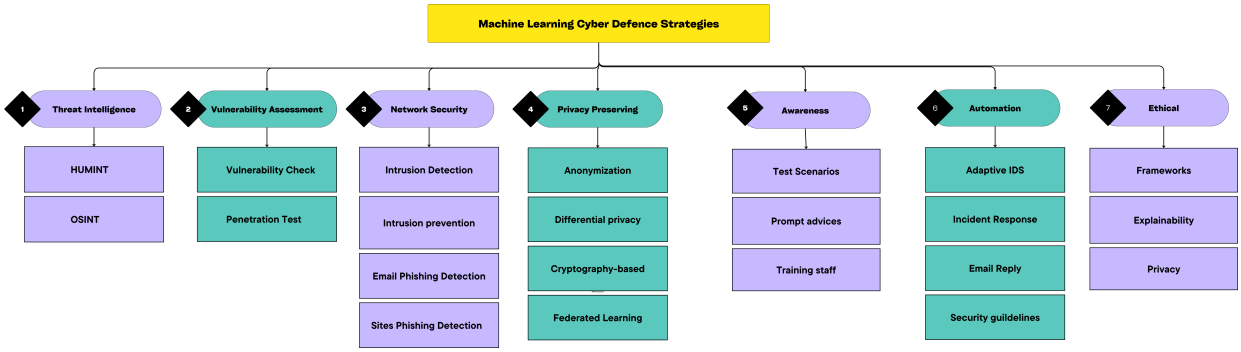

Fig. 4: A taxonomy of using LLMs in cyber defence sections. They are categorized based on the type of security. Some techniques may intersect in multiple categories; in this case, they are grouped based on the most dominant characteristic.

Fig. 5: Visual description of the number of LLMs usage in research in all fields in cyber security fields.

combining LLMs and semantic hypernetworks to improve the detection, analysis, and response to cyber threats efficiently [33]. This hybridization is proposed to enhance the identification of advanced cyber threats along with flexible solution that can be adjusted to different cyber security obstacles. Hays *et al.* incorporated LLMs into threat intelligence exchange to replicate intricate and authentic security situations which enable users to participate in decision making [34]. The development of exercises that utilise LLMs to create a wide range of threat scenarios and tactics for response have been explored. Moreover, it evaluates the consequences of LLM-enhanced TTXs on the training of security personnel, the improvement of collaborative problem-solving. It provided recommendations for LLMs in security TTXs to take into account technological and ethical considerations to optimise their effectiveness along with responsible behaviour.

III. Vulnerability Assessment

Vulnerability assessment is very crucial component of all the cyber defence strategies. It investigates the whole system for a review of security weaknesses that could be a possible exploit of threat. It provides a systematic help for the organizations to identify, quantify, and prioritize the vulnerabilities. This includes software, hardware, and network infrastructures to enable organizations to protect the digital assets. The expected outcome of this operation is to alleviate potential risks before cyber criminal exploit them. It involves scanning to the whole system with automated and manual techniques for known vulnerabilities. These vulnerabilities are checked in weak passwords, unpatched software, and insecure system configurations. The results are sort of reports vulnerabilities' details, such as severity, impact, and recommendations, which will be used to guide security teams to make informed decisions [36].

Moreover, vulnerability assessments helps organizations to comply regulations and standards of security controls. This process is recommended in a regular basis to detect new vulnerabilities that may have emerged, as well as confirming the resolution of the previously detected ones. This vulnerability assessments are integral for maintaining the security of organizational environments [37]. In this section, we review the LLM-based methods of vulnerability assessment. LLMs are used to improve the effectiveness of automatic penetration testing [38] to automate the detection weaknesses in software systems. With the help of LLMs, PentestGPT is able to anticipate security vulnerabilities and propose effective measures to address them. A substantial enhancements in accuracy is achieved compared to conventional penetration testing approaches. This achievement in penetration tests will have significant consequences for the future of cyber security defence measures. In [39], SecureFalcon used LLMs to automate cyber security measures identify, analyse, and counteract cyber threats with speed and accuracy. By incorporating real-time data analysis, SecureFalcon illustrated improved efficacy of security operations along with facilitating ongoing learning and adjustment to emerging threats. Temara *et al.* investigated the utilisation LLMs, notably ChatGPT, in the early phases of penetration testing to achieve higher performance [40]. It asserts that LLMs' comprehensive reconnaissance enhanced the results of penetration tests by providing insights about system vulnerabilities. By harnessing LLMs capacities to automate data collection and analysis, penetration testers can achieve the maximum rate of accuracy.

Happe *et al.* used LLMs to find and monitor security vulnerabilities in software systems by leveraging natural language processing skills and suggest security weaknesses [41]. First, LLMs is trained on extensive code datasets, encompassing both secure and insecure samples to identify the vulnerabil-

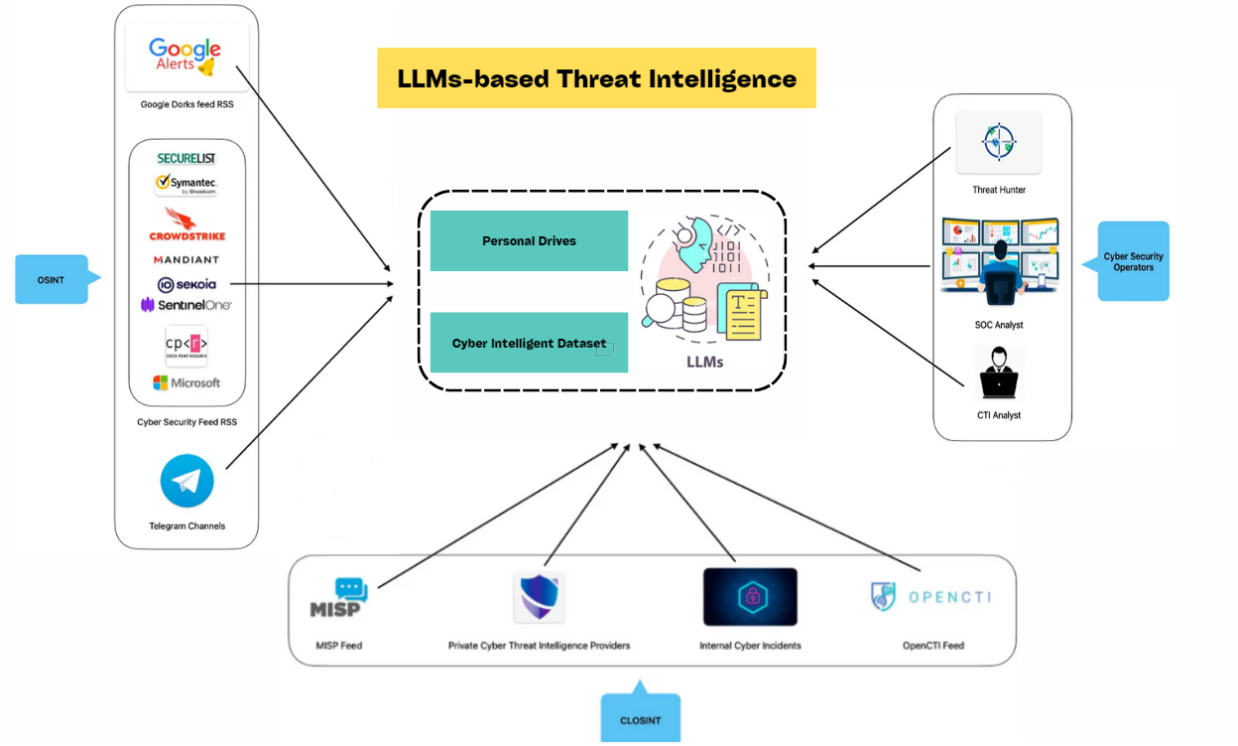

Fig. 6: Visual description of the LLMs-based threat intelligence steps. LLMs is used to analyze the possible exploits and threats.

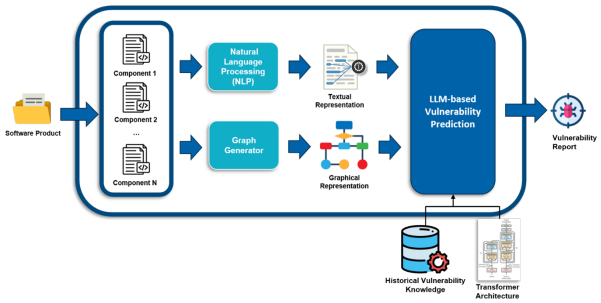

Fig. 7: Description of vulnerability prediction of LLM-based system. Image from [35]

ities. It showed that LLMs can attain superior accuracy and speed compared to conventional techniques. The use of large language models (LLMs) in penetration testing is examined. It explores the use of LLMs to model attacker behaviour, find security flaws, and evaluate the security posture of apps and systems. Penetration testers can automate a number of testing tasks, such as reconnaissance, exploitation, and reporting, by utilising LLMs' natural language understanding capabilities. The research emphasises how LLMs can improve penetration testing procedures' efficiency that will in return strengthen cyber security defences. The integration of LLMs for zero-shot vulnerability repair is investigated in [42]. Firstly, the concept of zero-shot vulnerability repair is to automatically develop patches for software vulnerabilities without having experienced these vulnerabilities before. It focused at producing patches for software flaws that are known to exist in a variety of programming languages and applications, relying on the ability of LLMs to comprehend vulnerability context and generate syntactically and semantically sound patches. The benefits and drawbacks of zero-shot vulnerability correction with LLMs are considered through testing and assessment, taking into account variables such as patch quality, vulnerability coverage, and possible hazards related to automated patch creation. The detection and monitoring of cyber security vulnerabilities is studied in [43]. LLMs as machine learning models should be used to analyse textual data, such security advisories, bug reports, and forums to find potential weaknesses in software systems. Various methods are studied to check the ability of LLMs to detect vulnerabilities in the software systems as follows: 1)native language interpretation: utilising the contextual capabilities of LLMs to comprehend and interpret natural language, and hence, recognising the vulnerabilities including error messages, code snippets, and descriptions of program behaviour. 2) pattern recognition: detecting comparable vulnerabilities in new or old software systems by comparing various timely patterns. 3) sensory analysis: it finds departures from typical behaviour that can point to the existence of a cyber threat. This work emphasises the effectiveness of LLMs in cyber security monitoring and vulnerability discovery using their natural languages' abilities.

Vulnerability assessment of software coding Ingemann *et al.* studied the software vulnerabilities using LLMs [44]. According to the results, LLMs have potential advantages over conventional techniques in the areas of functionality evaluation and vulnerability detection, including increased scalability, accuracy, and efficiency. The reason behind this is that LLMs is powered with natural language context. In this work [45], LLMs are leveraged to identify vulnerabilities in software coding systems. Since LLMs are adept at deciphering

and analysing code patterns and structures, they have shown encouraging outcomes in discovering vulnerabilities. LLMs is integrated in the identification system for vulnerability discovery, particularly to improve the efficacy and precision of security testing procedures in order to strengthen overall security posture. However, This research highlights the necessity of boosting LLMs to encounter obstacles and constraints, including problems like false positives, scaling issues, and the requirement for ongoing training and improvement to meet changing threats. As there is a an increasing concern over security flaws in online applications and in this work [46], Sakaoglu *et al.* presents an novel method, namely Kartal, to identify vulnerabilities such as command injection, cross-site scripting (XSS), and SQL injection in online applications by analysing web application code. The experiments illustrate that the accuracy of Kartal is outperforming the previous methods. Another work of Sakaoglu *et al.* [46] to focus on logical vulnerabilities such as SQL injection or cross-site scripting, which result from poor business logic rather than technical implementation flaws. It utilises refined LLMs to identify patterns suggestive of logical weaknesses in web application code. The proposed method implements an optimised LLMs to identify vulnerabilities by firstly preprocessing data, train on labelled data. Despite the existing obstacles to using LLMs, it significantly improves performance, reduces manual labor, and improves detection accuracy.

In another study of detecting vulnerabilities in smart contracts [47], *et al.* explored the use of LLMs (ChatGPT) for the purpose of detecting vulnerabilities in smart contracts on blockchain platforms. First, it analyzes the syntax and semantics of contract code to identify any security vulnerabilities. A novel mechanism is provided to optimise using a dataset that have different smart contract codes, as well as documented flaws. It concluded that ChatGPT is able to accurately identify prevalent weaknesses such as reentrancy attacks, integer overflows, and inappropriate access control.

IV. Network Security

Network security is a backbone of cyber defence strategies to safeguard network data and infrastructures against unauthorised access, misuse, change, destruction, or unlawful disclosure. In this way, the preservation of the integrity and usability of networking and data resources is guaranteed. It also refers to the range of measures and rules to prevent and monitor unauthorised access, misuse, alteration, or denial of the system's services and its resources [48].

Intrusion detection is a core element of network security since it is about the identification and response to cyber crimes that target the system or the network. Intrusion Detection Systems (IDS) are primarily used to oversee network traffic and detect abnormal patterns that might cause a network fault or system breach by an intruders. IDS has two main variants: 1) Network-based Intrusion Detection Systems (NIDS), designed to monitor and analyse any suspicious behaviour. 2) Host-based Intrusion Detection Systems (HIDS), implemented for particular hosts or devices within the network to monitor incoming and outgoing packets and promptly notifies if any abnormal activity, including system calls executed by the operating system or modifying system files. IDS has two main methods: 1) signature-based detection that relies on specific patterns of recognised threats [49], 2) anomaly-based detection, which compares activities to a baseline to identify deviations [50]. Anomaly-based detection include machine learning techniques. Strong administration is necessary to immediately respond to issues while minimising false positives that might drift the attention from true problems [51]. As complexity of cyber threats is getting worse, intrusion detection stands an essential role in network security. Continuous updates and training are necessary to the network traffic data and IDS systems to guarantee effectiveness against cyber threats. Network security is one aspect in the security plan, as it helps protect critical data from unethical access.

Since the evolution of deep learning, significant improvement has been witnessed in intrusion detection systems (IDS). Deep learning models' power to train massive datasets and learn intricate models boosted IDS to identify abnormalities in network traffic that could escape detection by conventional approaches [52]. It provides the ability to adapt to new threats without the need for explicit retraining from scratch, which help in countering zero-day exploits and advanced persistent threats (APTs) [53] 8. DL-based IDS enhances the accuracy of harmful activity identification while diminishing the occurrence of false positives, and hence, the efficiency of cyber defence is significantly improved. In this section, we are reviewing LLM-based network security strategies as we provide discussion on each method, significance, and drawbacks.

Intrusion Detection The power of LLMs in recognizing and preventing online fraud is investigated in [55]. It shows the strength of LLMs in identifying many types of online scams, such as phishing, fake emails, and misleading marketing by training LLMs on large-scale cyber security datasets. By analysing contextual textual material and using their natural language understanding capabilities through LLMs, fraudulent behaviour is recognised . The results indicate that LLMs is able to help detect and stop internet scams because of its ability to classify the infected material from the genuine ones. However, further studies are required to produce robust systems that are able to provide scam identification and mitigation, which in return will ultimately lead to a safer online environment. In this work [56], Shi *et al.* explored the possibility of creating hostile samples with language models that can avoid detection by current intrusion detectors. They show that language models are capable of producing text samples that avoid the detectors while maintaining semantic similarity with the source data. This shows the necessity of constantly improving detection techniques to keep the security safe. It implies that by spotting flaws and providing supervision for developing robust detection systems, merging red-teaming with language models is proved to increase the robustness of LLMs. Hassanin *et al.* proposed a pre-trained Large Language Model designed for anomaly detection in satellite networks [54]. They customized the network traffic data to fit Transformers input, which encode the context in cyber data (see Figure 9). Here [57], is a similar

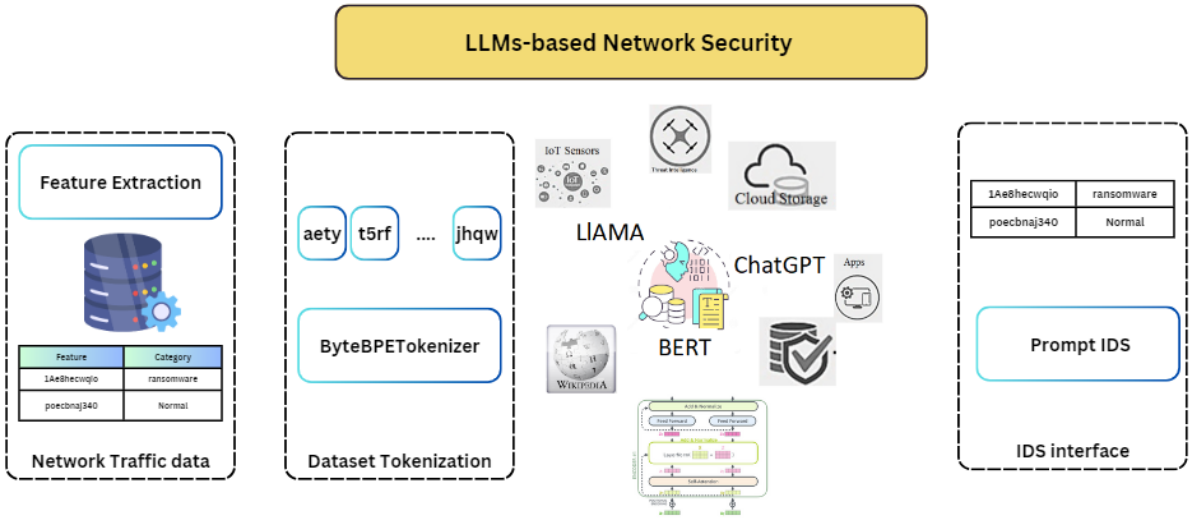

Fig. 8: Visual description of the LLMs-based network security steps. LLMs is used to predict the intrusion in the network flow. LLMs is mainly used after fine-tuning.

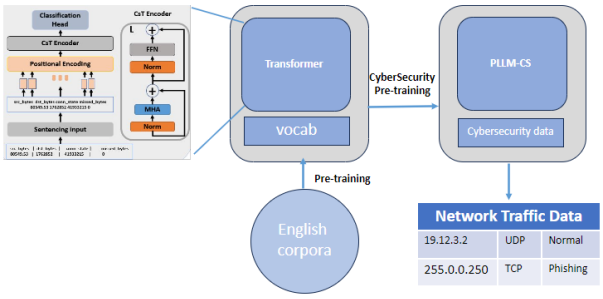

Fig. 9: Fine-tuning BERT model to predict malicious anomalies in satellite networks from [54]

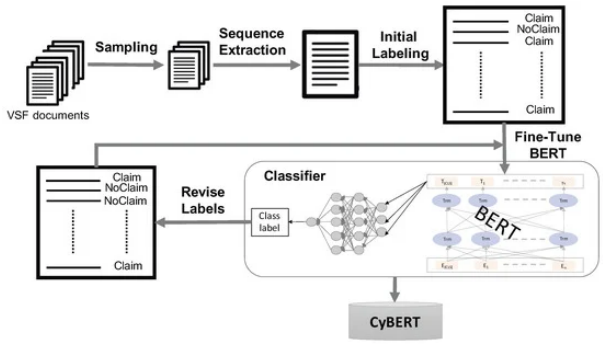

Fig. 10: Fine-tuning BERT model to predict intrusions in network data [62].

study but for network traffic data. Flowtransformer presents a framework for flow-based network intrusion detection systems (NIDS) that uses transformers as the machine learning model [58]. The suggested architecture improves network intrusion detection by capturing contextual data and temporal dependencies inside network flows. The accuracy and efficiency of identifying malicious activity in network traffic is significantly improved with the help of the Transformers [14]. However, it is still suffers from Transformers issues, including the need of brute force training with huge amount of data as well as this improvement comes on the cost of speed and resources. In this paper [59], the integration of LLMs and cyber security and its impact is investigated in various ways, including threat intelligence, security automation, adversarial defence, and cyber security awareness. Using LLMs in the landscape of cyber security offers improvement to the cyber security operations and threat detections. It summarised the benefits and risks of applying LLMs to cyber security operations. In this paper [60], Prasad *et al.* explores using the pre-trained Transformers, namely ChatGPT, is improving the cyber security tasks. Few areas in cyber security are investigated, including threat intelligence, security automation, adversarial defence, and user education. It draws attention to the model's abilities to analyse massive amounts of text data, automate security chores, identify and neutralise hostile attempts, and provide instructional materials to spread knowledge about cyber security awareness. It also highlights the significance of ChatGPT in improving cyber security principles as well as the ability of mobilization according to the changing threat environment overall. Salloum *et al.* explores the techniques used to identify the threats presented by malicious accounts on digital platforms, particularly in the context of generative AI such as ChatGPT [61]. It uses machine learning algorithms to distinguish between genuine user activities and those that are driven by malicious drives. It highlights the importance of powerful anomaly detection systems to monitor atypical activity patterns and evaluate potential risks in real-time. Moreover, it emphasises the significance of regularly updating systems and educating users to keep up with the ever-changing cyber attacks as it recommends using a multi-layered security strategy to efficiently reduce the risks of fake accounts.

In [63], Nguyen *et al.* presented an LLMs-based IDS to analyse network traffic in order to detect anomalies and abnormalities. Because they treated each flow as a word, and a series of flows is a sentence, they enabled the model to accurately predict the behaviour of network traffic. Their work

to LLMs is more than a model plugged into their defense system, rather they build a stronger network traffic flows by involving the key elements (source, destination points, protocol flags, packet numbers, etc) into the process. They used real datasets (CIDDS-001 and CIDDS-002 datasets) for model training and validation to consider the upcoming changes in the network settings, therefore, the highlighted the importance language models to tackle the issues of the network flows systems. In this paper [64], Nwafor *et al.* employed LLMs to detect intrusion detection systems (IDS) in vehicle networks. It builds the whole defensive system around Controller Area Network (CAN) bus systems [65]. It exploits the susceptibility of these networks to some cyber threats including replay, fuzzing, and DoS attacks to build a stronger model against these networks. Using LLMs in such environments achieved great success because the complexities in the language models. It shown accurate detections for abnormalities and possible risks and superior performance compared to the standard methods. In this work [66], using LLMs to detect DDoS attacks is investigated. It boosting the machine LLMs as a strong capable machine learning ability to detect DDoS early before spreading its harm to the network. Overall, it is helping the defensive systems to reduce the harm of the DDoS attacks on the network systems. Rieck *et al.* proposed LLMs model for detection of unknown attacks in the network data [67]. It primarily extracts n-grams and word-based characteristics from network payloads and these data packets and their contents are considered as language units. Then, an linguistic analysis is employed to identify irregularities in the network traffic. One of the noticeable features in this work is that it is unsupervised so it does not require any labelled datasets. In this case, not only this model achieves significant performance on the unlabelled data, it is also capable to detect any abnormality in the ever-changing environments. Also, it achieves such performance in a linear time, which make it very efficient in real-time intrusions detections. Lastly, this work addresses the issue of open vocabulary in the cyber security, which will detect any evasions alterations or obscurity by the attackers in the network traffic. In this work [68], hybrid method of LLMs and RNNs are used to identify the abnormalities in the Vocabulary Event-Level network data logs. They modeled the network logs on each individual line, utilising the tokens of the character level, This effectively helped to handle the usual diverse languages in such datasets. By modeling a collection of ASCII characters as tokens, it enables the model to depict of any log entry, irrespective of its particular details. Three models are proposed to achieve the goal of abnormalities detection, Event Model, Bidirectional Event Model, and Tiered Event Models, respectively. While traditional approaches face difficulties in handling network data, RNN showed ability to detect cyber threats. Alkhatib *et al.* proposed an LLM method, namely CAN-BERT, to detect abnormalities in Controller Area Network (CAN) systems, which are frequently employed in automotive networks [69]. The main objective of this work is to employ the LLM capabilities to analyse CAN network traffic, which is usually non-textual and highly organised. They customized their model to handle continuous data packet streams instead of individual words, dealing with CAN messages as sentences and individual signal values as words. By analysing the CAN messages, it identifies the potential intrusions or malicious activity within the network. The empirical experiments showed that both known and undiscovered attack pathways are accurately detected by CAN-BERT. Li *et al.* introduces a novel method that combines augmented with pre-trained language models (PrLMs) and Conditional Generative Adversarial Networks (CGANs) to improve intrusion detection systems (IDS) [70]. The main reason of combining these two methods is to build a sophisticated feature extraction capabilities to improve intrusion detection efficiency. CGANs are used handle the imbalance in the training datasets while LLMs are used to extract complicated pattern of the network features. Overall, this integration of LLMs with CGANs provides a potential way to address inherent issues in network intrusion detection, including the imbalanced datasets with an efficient feature extraction mechanism. It proved that a potential method to significantly enhance the accuracy of detecting abnormalities in the network data. Aghaei *et al.* present a new method for identifying anomalies such as fraudulent transactions, network intrusions, or atypical user behaviours in data by modifying the Bidirectional Encoder Representations from Transformers (BERT) [71]. This variant of BERT improves security and efficiency in detecting outliers in network datasets. Training SecureBERT across various datasets boosted its capability to detect abnormalities. In [72], Piggott *et al.* presented a chatbot system, namely Net-GPT, that is driven by LLMs and intended to act as a man-in-the-middle (MITM) agent between ground control operators and UAVs. By using LLMs, Net-GPT is able to comprehend and react to operator inputs, offering real-time support and automating certain activities. It emphasises how adding LLMs to UAV systems enhances operational effectiveness and the relationship between humans and machines in UAV. Patsakis *et al.* used LLMs to decipher hidden code that is utilised in actual malware campaigns that is frequently employed strategy by cyber criminals to evade detection by conventional security systems [73]. It suggested using LLMs to counteract different obfuscation tactics, including encryption, packing, and polymorphism. It highlights that that LLMs have the potential to simplify complex code structures, however, their usefulness is highly dependent on the complication of the obfuscation methods used as well as the specific characteristics of the virus. It stresses the necessity for additional development of these models to handle such sophisticated obfuscation tactics, as well as the need of continuous training of LLMs to combat with the ever-changing world of cyber threats. In this paper [74], Fujima *et al.* how to use ChatGPT for ransomware threat analysis and prediction. Ransomware is a kind of malware that causes serious risks to cyber security by encrypting victim's data and demands payment to unlock it. It trains LLMs on datasets containing ransomware messages to recognise the linguistic patterns and traits of ransomware threats. Since ChatGPT is powerful to model the long-range dependencies, it helps recognise common themes, strategies, and signs linked to ransomware attacks. Hence, it showed ability to forecast possible ransomware threats by looking for patterns in past data. Further, it helps

cyber security experts to spot new ransomware tendencies and issue cautionary messages. Though the inherited shortcomings of LLMs, it showed significant improvements in deciphering ransomware messages, comprehending threat environments, and forecasting prospective ransomware attacks. Sandoval *et al.* is identifying security issues that result from using LLMs in code development [75]. It shows that they are useful in enhancing programming efficiency and precision, however, they might unintentionally introduce security vulnerabilities in code generation. Pointing out security issues of LLMs in code development helps the personnel to avoide it as well as creating the awareness amongst the employees.

In [76], He *et al.* explores the capacity of LLMs to assist in software code evaluations, with a specific emphasis on identifying security vulnerabilities as well as evaluating code functionality. Firstly, this work investigates how LLMs can optimise the code review process, saves a lot of software development costs. They found that by automating certain aspects, LLMs have the ability to enhance efficiency and minimise human error. LLMs helps in vulnerability identification as a machine learning module. Though it showed the significance of LLMs in protecting, reviewing and identifying vulnerabilities for software programs, it need a training customization on this paradigm. In [77], Okey *et al.* investigates LLMs and cyber security on topic modeling and sentiment analysis. In order to provide insights into new threats and vulnerabilities, topic modelling seeks to discover popular cyber security themes and patterns within online discussions. Sentiment analysis evaluates the general attitude expressed in discussions on cyber security, which aids in determining public opinion and identifying possible warning signs of danger. The study illustrates ChatGPT's potential to improve cyber security intelligence and strategies by utilising its natural language processing capabilities. In [78], Sharma *et al.* explores the combined influence of big data analytics and LLMs on cyber security practices. It delves into how these technologies enhance threat detection, incident response, and the other security operations. Since big data enables organizations to process vast amounts of data, it helps identifying anomalies and potential threats. The main reason behind this improvement is that LLMs enhances natural language understanding and facilitates automated responses in security operations. Hence, it concludes that LLMs empower organizations to strengthen their cyber security posture, better response incidence, and effective threats detection. In [79], Lai *et al.* uses the natural language processing features, LLMs, counteract the increasing spread and sophistication of cyber threats. It uses LLM to analyse the network data and identify any abnormality in the patterns and due to the power of LLMs in pattern recognition, it achieves high performance in the field of intrusion detection. Overall, this work presents a crucial method for large-scale language model intrusion detection and hence it improves the accuracy, scalability, and flexibility in response to changing cyber attacks.

Email Phishing Labonne *et al.* proposed using T5 (Text-to-Text Transfer Transformer) in identifying email spam through few-shot learning methods [80]. A performance comparison between T5 and the other LLMs in detecting spam emails when there is a scarcity of labelled training cases is provided. The results illustrate the superiority of T5 over the other LLMs in detecting spams with less labelled data, as well as exhibiting resilience when dealing with different types of spam. Roy *et al.* discussed the increasing concern about the use of LLMs such as ChatGPT, Google Bard, and Claude in the creation of complicated phishing scams that utilise AI-driven chatbots to generate challenging phishing attempts [81]. The main purpose is to learn the cyber attacks ways to reduce their risks, such as creating AI detection tools capable of identifying phishing attempts, as well as enforcing stricter regulations on the use of generative AI technologies. In [82], Wu *et al.* compare the performance of ChatGPT and conventional spam detection algorithms, more precisely the capacity to handle the subtleties of natural language in spam text. ChatGPT surpassed traditional approaches in accuracy and speed, due to its sophisticated natural language capabilities. Further, spam techniques that frequently evaded traditional filters have been detected by ChatGPT. It also reduces the occurrence of false positives that recommend it to be plugged in email systems. Heiding *et al.* investigated the proficiency of LLMs in producing and identifying phishing emails compared to human [83]. It showed that LLMs in generating and detecting phishing emails are more efficient than humans due to their comprehension of linguistic subtleties. However, LLMs are identified as double faces for cyber security, it can be defensive assets or dangerous vulnerable sources. Li *et al.* combined LLMs with multi-modal knowledge graphs to improve the accuracy of reference-based phishing detection of phishing attempts [84]. By utilising textual and non-textual components (such as graphics and links) , KnowPhish authenticates information using a comprehensive knowledge network that encompasses verified data on authentic websites and visual identities. Unlike conventional methods that identify phishing based on one technique, KnowPhish uses system's architecture that integrates the predictive capabilities of LLMs for text analysis knowledge graphs for evaluating the credibility of emails and web pages. The performance of KnowPhish is proved to be better in identifying complex, context-based phishing schemes that use many deceptive strategies. In [85], LLMs are employed to identify SMS spam through a few-shot learning methodology; that is, using a limited number of instances to tackle the difficulty of quickly adjusting to new and changing spam strategies without retraining. Pre-training LLMs showed the ability to generalise from a limited number of instances and reliably categorise SMS messages as either spam or non spam. This shows the efficiency of LLMs on few-shot scenarios compared to conventional machine learning techniques that often necessitate bigger training datasets. Such performance indicates that LLMs not only improvise without annotated datasets but also higher performance in identifying spam. This approach has substantial ramifications for the security of mobile communication, providing a salable and flexible solution to efficiently tackle spam messages sent over SMS. Koide *et al.* used LLMs to recognise phishing emails by transforming emails into prompts to check them [86]. This makes it possible to look more closely at the content of the email in its context and spot sophisticated social engineering

techniques. Because it can look for differences between the brand name and the domain name associated with it, it spots brand impersonations effectively. They introduced a dataset of phishing emails collected from honeypots between August 2022 and October 2023. These emails have no URL in the body, which effectively removed any emails that attempted to send recipients to phishing websites. It targeted 193 brands in 19 languages—English, Portuguese, German, and Dutch. Though it is promising for identifying phishing emails, it is crucial to take into account the constraints of LLMs and the computing expenses. Patel *et al.* investigated using LLMs in identifying phishing emails, and they performed a comparative analysis between LLMs and conventional machine learning-based phishing detection systems [87]. They highlighted that LLMs has advanced learning capacities that excel detecting phishing emails, especially after training on various datasets. The results indicate that LLMs not only achieve better accuracy, but also provide resilience against new phishing techniques. Trad *et al.* provided a thorough comparison between two approaches, prompt engineering and fine-tuning, in utilising LLMs for detecting phishing attempts [88]. It is an important study because the advancement of LLMs is turning the machine learning as a blackbox. First, LLMs are given specially designed prompts or tested after fine tuning of phishing and lawful messages' training. It highlights that prompt engineering is computationally efficient and faster, however fine-tuning provides superior accuracy, as well as flexibility to new phishing strategies. Therefore, the recommendation is to use them according to the case or the problem. Nahmias *et al.* proposed LLM-based method to improve the detection of spear-phishing attacks by utilising prompted contextual vectors [89]. Because Spear-phishing is an advanced way of stealing sensitive information by sending false emails that imitates a trustworthy sender, detecting spear-phishing is very sophisticated task. The main contribution of this work is that it utilises the contextual embeddings by language models. These fine-tuned contexts are produced using specialised prompts to assist in identifying spear-phishing. This is the first study that incorporates prompted contextual vectors into current spear-phishing detection algorithms. They succeeded to differentiate between authentic messages and malicious ones by training on a large dataset. Another study [90] that presented an advanced algorithm to improve the performance of identifying hazardous emails, such as phishing attempts, spam, and genuine (ham) communications. It is mainly providing a strong training on large datasets that uses a wide range of real-life email conversations to boost the accuracy efficiency in different situations. In [91], Heiding *et al.* explored multifaceted use of large language models (LLMs) in both creating and identifying phishing emails. They generated phishing emails by LLMs that look authentic as well as they used the same way to detect phishing emails by training on large datasets. Yu *et al.* used LLMs to improve security measures through the creation of a concept called honeywords which provides a combination of false passwords and the user's real password to baffle attackers as well as identify security breaches when they try to utilise these wrong passwords [92]. Therefore, the generative capability of LLMs is used to generate these honeywords, which complicate the task of attacker to know the authentic words. Despite the effectiveness of the honeywords concept, they rely on the training data and the model's capacity to generate false yet plausible alternatives. In [93], Chataut *et al.* uses LLMs in spotting phishing attempts. Phishing attacks are serious cyber security risks that trick people into disclosing personal information via using social engineering techniques. The ability of LLMs to use contextual signals and language patterns to analyse email content with the aim of detecting phishing attempts is investigated. It uses LLM to boost the ability of their model in differentiating between phishing and legitimate emails by training them on a dataset of those emails. The results show the efficiency of LLMs to detect email phishing and secure emails. Overall, LLMs illustrate significant improvment regarding email phishing. An extensive study on log anomaly detection using Transformer-based Large Language Models (LLMs) is explored in [94]. Firstly, Transformer-based LLMs is used mainly to identify irregularities in log data by detecting unusual patterns that could be signs of system failures. Also, LLMs-based chatbots are used to facilitate verbal engagements between potential attackers and malevolent actors to acquire information that divert the attention of attackers. This study showed that LLMs is capable of overcoming the cyber attacks and boost efficiency. In this paper [95], Ferrag *et al.* proposes a novel approach to improve cyber threat detection in IoT (Intenet of Things) and IIoT (Industrial Internet of Things) systems. It presents LLM-based privacy-preserving method for sifting through device communications and data streams to look identify possible cyber attacks. With limited resources, the proposed method proved significant performance in anomaly detection, security monitoring, and intrusion detection in IoT/IIoT. Overall, it proposes an effective approach to strengthen cyber security in IoT while taking resource constraints and privacy issues into account. Ameri *et al.* proposed using a refined iteration of the BERT model for detection of abnormalities present in text [62]. It refined it by training on the specific datasets of cyber security such as as shown in Figure 10. It showed superior accuracy results compared to the conventional machine learning methods. Bayer *et al.* provided a tailored BERT model for the cyber security field to analyse the specific terminologies in cyber security writings, encompassing technical threat reports and security protocols [96]. Ranade *et al.* proposed CyBERT to improve the identification cyber security-related issues in text, including various types of malware, attack methods, and security weaknesses across reports, logs, and research articles [97]. In order to fit it to cyber security, they trained BERT on the specialised datasets, and thus, improving the performance of abnormalities detection. In [98], Wohlbach *et al.* evaluates the possible cyber security threats of NLP models, including ChatGPT and Google Bard. It discusses the ways that bad actors may use these models to commit phishing scams, disinformation campaigns, and data breaches, etc. The aim is to increase public knowledge of the possible cyber security issues of using NLP models in real-world applications by exposing their vulnerabilities and shortcomings. Furthermore, it offers valuable perspectives on approaches to mitigate cyber security risks of LLMs models, including model validation,

adversarial learning, and self-security focused model design. In [99], Zaboli *et al.* investigates improving cyber security measures using large language models (LLMs), particularly chatGPT for smart grid applications. It looks on how LLMs capabilities can be used to detect and mitigate cyber security risks in smart grid networks. It demonstrates using LLMs to evaluate and decipher intricate data patterns, find abnormalities, and offer practical insights, and thus, improve smart grid security. It provides opportunities, difficulties of integrating LLMs into smart grid cyber security plans, aiming to infrastructure resilience against cyber threats.

LLMs-based Honeypots Mckee *et al.* explored the use of LLMs to improve honeypot technology in order to counter command-line cyber-attacks [100]. Koide *et al.* explored the utilising LLMs to detect phishing websites to identify deceitful language and misleading content commonly seen on phishing websites [101]. It analyses linguistic patterns and irregularities in online website text as it asserts that incorporating LLMs into cyber security frameworks boost significantly the identification of phishing activities. Sladi *et al.* investigates using LLMs's generative power to generate dynamic, persuasive, and interactive environments that imitate authentic systems in order to deceive and engage attackers [102]. Thereby gaining valuable knowledge about attack techniques, which help in detecting the cyber attacks. Its primary way is to produce authentic network traffic, user interactions, and system logs in real-time. These LLM-based honeypots are more efficient in attracting attackers for extended durations, hence offering more profound insights into malevolent approaches.

V. Privacy Preservation

Preserving privacy is an integral part in protection of personal and organisational data from unauthorised access. Lately, it became crucial for both individuals and companies to protect sensitive information, confidentiality and integrity, particularly after the progression of AI models and the vastness of digital systems. However, it requires simultaneously to implement strong security measures to counter potential cyber threats. Diverse range of approaches precisely designed to safeguard personal data and guarantee exclusive access by authorised users. For instance, encryption is a fundamental technology in this field, employed to encode data in a manner to restrict access to only those individuals with the appropriate decryption key. Also, anonymization and data masking techniques are used to eliminate personally identifying information from data collections, thus alleviating data breaches' issues.

With the evolution of deep learning models, although their impressive analytical powers, problems regarding the protection of privacy due to their reliance on extensive datasets have risen. In order to tackle these concerns, advanced methods like differential privacy, federated learning, and homomorphic encryption are being more and more included into deep learning frameworks [103]. Stochastic perturbation to obfuscate individual data points have been used in [104]. Federated learning have been introduced to enable the training of models on distributed devices, which protects confidentiality of data by training in its original place. These strategies aid in reducing privacy issues, enabling the use of deep learning's advantages while ensuring the protection of user privacy. However, more concerns have been introduced due to the evolution of LLMs that provide more challenges to the cyber defence privacy preservation [105]. Privacy preserving still needs a lot of attention as many gaps need more investigation as Figure 11. In this section, we review the recent works on privacy preservation and LLMs.

Kim *et al.* used LLMs to address a significant concerns about the leakage of personally identifiable information (PII) while training on web-collected data, which include sensitive personal information [106]. It suggests that data subjects to test LLMs with a specific information to check whether their information will be leaked. This work handles "black-box probing" where they can check if indviduals' PII is likely to be leaked, and "white-box probing" to check if provider of LLM to improve the model's handling of private information. In [107], Raeini *et al.* provides a unique approach to large language models (LLMs) called privacy-preserving large language models (PPLLMs) to protect user privacy without compromising the efficiency. To guarantee that sensitive data stays private and secure during training and inference, these models used federated learning, differential privacy, and secure multi-party computation. It solves privacy preserving concerns including data privacy, confidentiality, and security by integrating privacy-preserving techniques with LLMs. One of the main issues of LLMs is the adversarial attacks and lack of security, therefore, PPLLM is to find a middle ground between protecting user privacy and utilising big language models for a variety of purposes, such as conversational bots, text creation, and cyber security. Though it shows improvement in the privacy preserving, it comes on the account of accuracy due to the encryption layer. Another explanatory study [108], generative AI models' impact on cyber security and privacy preserving is investigated. In threat detection and Intelligence, generative AI models are used to examine big datasets and find patterns suggestive of cyber threats, and as a result, threat detection capabilities can be improved. They also delved into the possibility of adversarial assaults directed at generative AI models. In privacy problems, they highlighted that generative AI models gives rise to privacy problems because there is a chance that they will unintentionally produce private or sensitive data, endangering people's privacy. For this, mitigation methods, such as applying privacy-preserving approaches and adding protections against adversarial attacks, are suggested to address these vulnerabilities. Exploring the ability of LLMs to dispel common misconceptions in order to provide advice on security and privacy-related issues in [105]. LLMs' capacity to respond to enquiries about security and privacy, including falsehoods like those regarding password security or data privacy procedures is investigated. The results show insights into the efficiency of LLMs as instructional resources in the fields of cyber security and privacy. However, it highlights the issues of this model that draws the attention of the research community to conduct additional research and make improvements in order to increase LLMs' capacity to a precise and trustworthy advice on security and privacy-related issues. In this paper [109], Singh *et al.* studied the range of

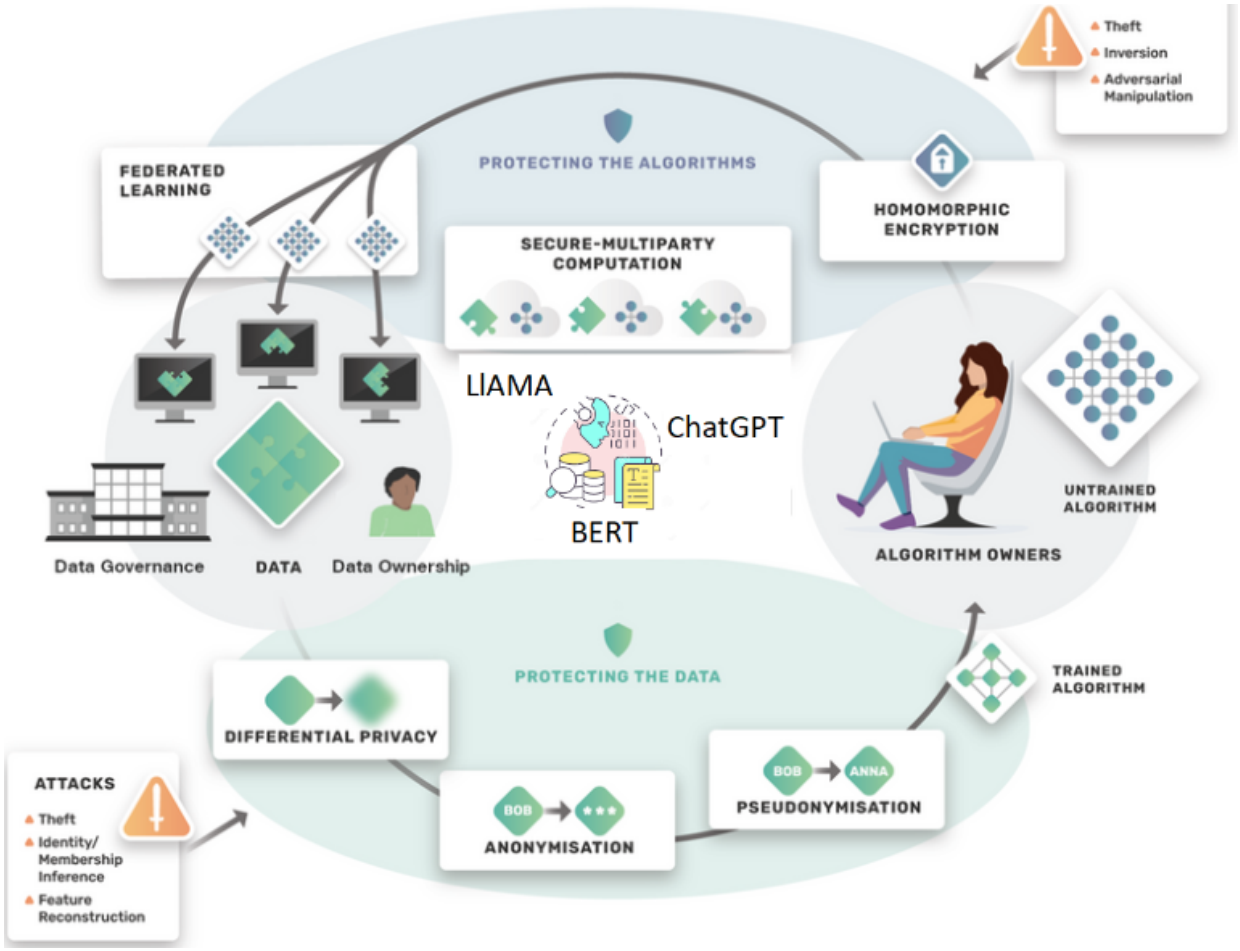

Fig. 11: Visual description of the LLMs-based privacy preservation techniques. LLMs can be integrated in all the systems of privacy preservation such as data privacy, and algorithm protection.

privacy and security issues and proposed LLMs to address them. The proposed method is simply to use zero-knowledge proofs without disclosing any particular information. They validated the integrity and accuracy of the model without revealing the underlying data or model parameters.

VI. Awareness

Awareness of the organisation staff is a crucial component of the important cyber defense techniques, strengthening the need of educating individuals and organisations about the dangers of cyber threats and how to avert them. The main objective of this is to establish a security-oriented environment where all individuals comprehend their responsibilities and roles in protecting digital resources against cyber risks [110]. Awareness programs are essential due to human mistakes continue to be a prevalent vulnerability exploited by cyber criminals. Such programs usually include several topics, including fundamental of cyber security, different types of threats, ranging from phishing, to ransomware, secure practices, personal secure networks tips. It also includes training which is frequently covering sophisticated subjects such as mobile security, and the secure use of private social media [111]. Comprehensive awareness programs not only provide staff with required knowledge to avoid threats, but also security procedures that can minimize the risk of a cyber crimes. However, consistent upgrades and courses are essential to keep up-to-date with the ever-changing cyber threats [112]. In this section, we are reviewing LLM-based network security strategies as we provide discussion on each method, significance and drawbacks. [113] offers a strategy for using ChatGPT [114] to improve information security behaviours. It identifies the issue of being individuals frequently become targets of cyber attacks because they do not pay attention to security best practices. It addresses this issue through ChatGPT's features to deliver security reminders and nudges. For instance, it remind users to reinforce secure practices, such as creating strong passwords and using two-factor authentication; provide personalised security advice. While this framework seeks to foster an awareness of security within organisations, it is a theoretical approach and lacking the evaluations. In [115], Yamin *et al.* utilised LLMs to automate the generation of authentic and intricate scenarios for cyber security training exercises that will be used for training modules to be more robust against cyber threats. These cyber security incidents and exercises can be used to train the cyber security teams to acquire knowledge and produce plausible training scenarios. It recommends incorporating LLMs into cyber security training programs that will offer ongoing learning and adjustment to

emerging cyber threats.

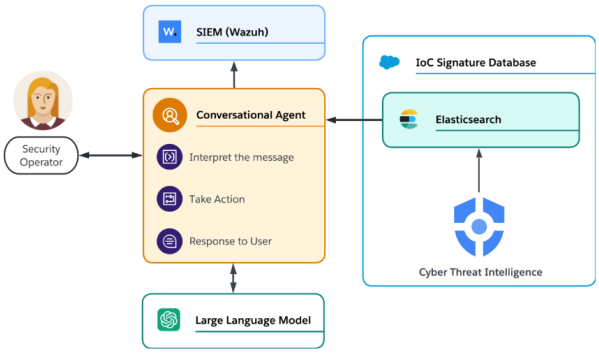

Fig. 12: The proposed LLM-based method for developing awareness of security issues [116].

This work [116] looks into the usefulness of GPT-4 [117] powered conversational agents in streamlining security-related activities. It investigates how conversational agents can help with different security activities based on superior language capabilities. It proved that it leads to the simplification of procedures like threat analysis, incident response, and security policy enforcement. The study emphasises how cyber security workflows can benefit from the use of LLMs agents to improve productivity and efficacy in security operations management (see Figure 12). The impact of large language model (LLM) chatbots in boosting cyber threat awareness through Open Source Intelligence (OSINT) is investigated in [118]. It evaluates the abilities of LLM-based chatbots to decipher and analyse OSINT data to deliver threat detection in time, spot possible cyber security threats, and instruct users on security incident response. It is mainly studies how LLM chatbots perform in providing users with precise and useful insights that in return will raising the level of awareness regarding cyber threats and intrusions.

VII. Cyber Security Operations Automation

Automating cyber security operations has become an essential requirement in defending the recent digital systems, particularly with the constant changing in cyber threats [119], [120]. with AI and LLMs, security applications can improve their capacity to efficiently identify, address, and minimise security problems without human intervention. Such operations utilise advanced technologies and platforms capable of executing duties often carried out by human analysts, such as identifying threats, responding to incidents. This automation not only speeds up the process of resolving threats, but also decreases the likelihood of human mistakes [121]. Additionally, humans can not handle massive amounts of data as they get boring quickly, however, automated cyber operations can address this issue and help fix the issue before growing up.

Implementing automation in cyber security operations also tackles the increasing problem of the cyber security skills shortage and lack of experience. Skills shortage can cause harm to the organisation due to the late response, which may cost massive loss. Automation after the last progression of LLMs, is a doable task, especially with repetitive duties, which in return will free up the security staff to concentrate with complicated tasks [119].

Security Orchestration, Automation, and Response (SOAR) technology offer a smooth integration of different security tools on automated workflows, making it easier to automate responses to incidents [122]. This helps enhancing the efficiency of security operations, particularly in the absence of skills. The last progression of AI and GPT models open the wide door for cyber security operations to rise up and close the inherent gaps. In this section, we will provide overview for the crossed steps of LLMs in the automation of cyber security operations.

Hays *et al.* investigates the application of LLMs, ChatGPT-3, in improving incident response methods to automate the crucial procedures of documentation, analysis, and review [123]. It uses historical data and predictive analytics of LLMs to produce incident reports, delivering immediate decision assistance, and proposing measures to address breaches. The main benefits LLMs is to enhance responsiveness and precision, together with the capacity to overcome emerging risks and scenarios. It also tackles obstacles such as guaranteeing the dependability of advise given by AI and the susceptibility of AI systems to manipulation by malicious individuals. It recommends guidelines to incorporate LLMs into current security operations centres (SOCs) and emphasises, particularly with continuous training to optimise the efficiency of these AI tools in incident response against ever-changing cyber attacks. In a study to use LLMs to defend the web sites from automatic hacking by executing automated cyber-attacks on them [124]. First, they utilise the sophisticated natural language processing and machine learning methods to identify weaknesses in online infrastructure, such as SQL injections and cross-site scripting problems without the need for human involvement. Then, cyber defence mechanisms will be used to predict and reduce AI-powered threats, thus protecting digital assets from automated advanced attacks.

In [125], Feffer *et al.* explored the creation of a sophisticated LLM tailored for the purpose of red teaming in order to assess and enhance its defensive capabilities. The main purpose is to automate and the cyber security operations, hence improving their resilience against a wide range of cyber threats. The results illustrated the model's capacity to detect vulnerabilities with greater efficiency. It also discusses potential security issues associated with automating red team attacks that will highlight the significance of automation of cyber security operations. plugging-in LLMs into an automated IDS is discussed in [126]. First, network data are utilised to interpret large quantities of network data and records to identify abnormal patterns. The main feature of this technique is that it promptly respond to emerging threats. Moreover, it demonstrated the superiority of the LLM-based IDS over traditional systems in identifying evolving cyber threats. Sultana *et al.* explored the use LLMs for automating cyber operations in different scenarios, including threat identification, system surveillance, and incident handling [127]. It emphasises the ability of LLMs to improve the effectiveness of LLMs in cyber operations to reduce the risks. They proposed merging LSTM with LLMs to combat the real time and ever changing cyber

threats. They highlighted the importance of involving cyber security experts in the process of LLMs to guarantee their conformity with security norms and prevent the introduction of additional vulnerabilities. One-day software vulnerabilities that are revealed but not yet fixed have been investigated [124] for detection using LLMs. Fang *et al.* trained LLMs on a large collection of security patches, vulnerability reports, and exploit code to examine publicly published weaknesses and developing appropriate exploit code. The experiments indicate that LLMs, when provided with comprehensive vulnerability knowledge, are capable of mitigate operational risks of attack scripts. Using LLMs in detection of hardware trojans that cause alterations to hardware circuits is proposed in [128]. It shows the ability of LLMs to not only improving security protocols by predicting trojan assaults in hardware, but also identifying flaws within these systems. Notable enhancement in the ability to detect hardware trojans at an early stage is achieved by LLMs compared to normal methods. This work is very important for safeguarding the hardware components as it illustrates the significance of LLMs-based cyber threats' detections. Helbling *et al.* proposed using LLMs to as a self-examination mechanism to analyse their own responses and the the prompts [129]. The methodology combines heuristic-based and machine learning techniques to enable LLMs to evaluate their operational integrity in real-time. The results showed that this self-examination of LLMs boosted the model's ability to counteract adversarial attacks. It suggests that self-aware LLMs with explainable characteristics are necessary for AI ethics and governance to guarantee interpretability. Attack trees are hierarchical structures that are used to depict possible attack scenarios and system weaknesses. In [130], Gady-atskaya *et al.* uses ChatGPT's to automatically create attack trees in order to analyse textual descriptions of cyber security risks. By streamlining the attack tree generation process, this method hopes to give cyber security experts an invaluable tool for threat analysis and threat mitigation. As a result it shows how LLMs used to automate difficult cyber security jobs as well as they are employed to boost the effectiveness of threat detection and response incidence. Shao *et al.* used LLMs to tackle one of the main offensive security tasks, Capture The Flag (CTF) challenges that target computer security issues [131]. Two distinct procedures have been investigated for solving CTF challenges utilising LLMs. The first utilises a human-in-the-loop (HITL) approach, while the other is completely automated, which explains the effectiveness of LLMs into the problem-solving process. They showed that these models can perform at least as well as, if not better than, human participants in specific situations. David *et al.* discussed automating the smart contract security by using LLMs, particularly in the presence blockchain technology [132]. This work investigated the efficacy of LLMs in identifying vulnerabilities and compares them to conventional manual auditing techniques. They concluded that although automated technologies have made great advancements and can detect numerous common vulnerabilities, they are not yet capable of completely replacing the nuanced comprehension human experts. The authors assert that manual audits continue to be a crucial element of smart contract security, especially for guaranteeing adherence to legal norms. Cambiaso *et al.* used LLMs to automate responses to fraudulent emails, aiming at assessing the ability of LLMs to interact with scammers and discourage their activities by generating contextually suitable replies [133]. By imitating ordinary human reactions, LLMs waste the scammers' time, resources, and diminishing the effectiveness. They highlighted the necessity of human supervision to prevent unforeseen outcomes, such as the continuation of detrimental content, especially after the scammers recognise they are automated. This study indicated that valuable new instrument in combating email-based scams is provided by LLMs.

VIII. Ethical LLMs

Ethical cyber defense is very crucial part in the cyber defense process in the organisation. It involves the implementation of guidelines, concepts and procedures to protect digital assets and information systems from any cyber threats. With the significant spread of artificial intelligence and its availability to cyber criminals, the focus on the ethical aspects of how these technologies have to be managed. Ethical cyber defence signifies the value of protecting user privacy, fairness, and transparency in the creation and implementation of cyber security measures. A comprehensive structure that surpasses basic adherence to laws and regulations is entailed under ethical cyber defense, as well as a dedication to making ethical decisions at all levels, ranging from technology selection to to respecting ethical limits. For instance, it may be possible to oversee all employee actions on a company network to avert data breaches, however, ethical factors should dictate the manner of such monitoring. Simply put, ethical cyber defence is the proactive obligation of organisations to implement the cyber defense process along with ensuring that no harm to others. Therefore, it encompasses the responsibility of revealing vulnerabilities, exchanging threat intelligence, and refraining from selling software with a possible exploitation. To provide machine learning-based cyber defense, the fairness of AI decisions have to be fair and just. One of the ways to provide that is to use explainable AI (XAI) that is responsible for justifying the AI conclusions and decisions. In this section, we are summarising the methods of cyber defense with large language models in the recent literature. Figure 13 shows the common design of ethical LLMs for cyber security considerations.

In a explainable LLMs study, Ali *et al.* presents a model that combines a Random Forest classifier, LLM in conjunction with XAI tools such as SHAP and Lime to detect anomalies in a high performance manner [134]. LLMs to detect abnormal patterns or anomalies in network traffic, XAI frameworks to comprehend the outcomes of the LLMs (see Figure 14). In [135], Uddin *et al.* introduced pre-trained Transformers to identify phishing emails, as well as explanations for its judgements, thereby improving the transparency and reliability. The different part here from previous methods is the incorporation of explainability mechanisms, which enable users to comprehend the underlying reasoning behind classifications. It shows that the model outperforms traditional machine learning methods in terms of accuracy and explainability.

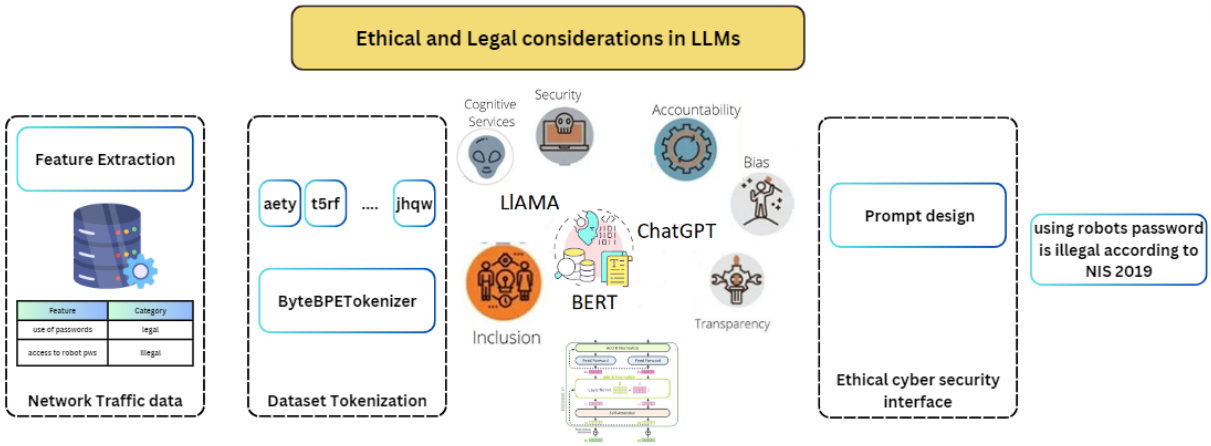

Fig. 13: Visual description of the LLMs-based ethical and legal frameworks.

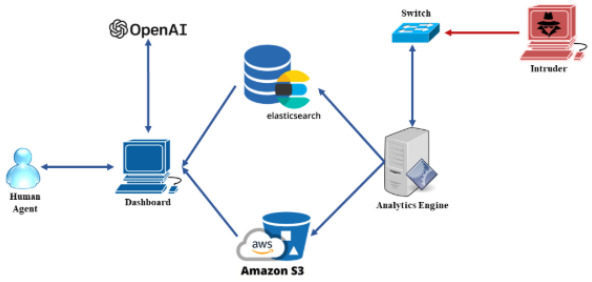

Fig. 14: The proposed LLM method to provide ethical frameworks for cyber security teams [134].

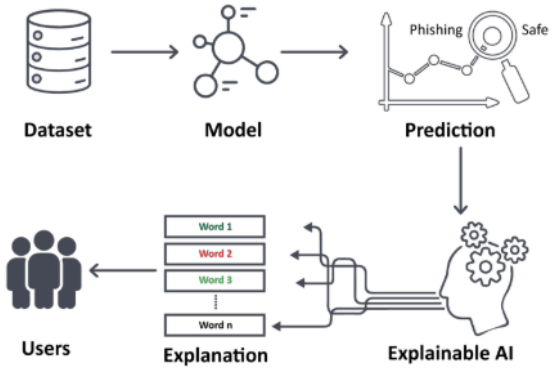

Fig. 15: The proposed LLM method of transparency and explainability [135]

In an exploratory study [136], Sebastian *et al.* investigates a number of cyber security areas, including data privacy, adversarial assault susceptibility, the possibility of disseminating false information, and social engineering strategies in the presence generative AI. It highlights the need for thorough risk assessments and security procedures to reduce potential risks. Also, it sheds light on the security issues associated with the deployment of AI chatbots. It proposes creating a more secure environment for the deployment of AI chatbots by bringing attention to handle the cyber security threats. Garvey *et al.* explored the potential use of LLMs, mainly ChatGPT and Bard, as avatars for red teaming in foresight scenarios [137]. It uses LLMs to automate the process of using an adversarial method to assess plans and tactics to check how far LLMs to replace humans. It presents a comprehensive overview of the advantages and constraints of employing LLMs in this particular context. It proved that LLMs has the ability to significantly improve the complexity of scenarios,however, under the supervision of humans.

IX. Challenges and Open Problems

Despite the noticeable success of LLMs beyond scientific research and applied sciences, they still suffer from several challenges that need to be fixed to obtain the maximum benefit from LLMs. The main impediments include susceptibility to bias, as LLMs can perpetuate existing biases present in training datasets, as well as privacy leakage concerns. Additionally, high computational resources along with massive amounts of data are needed to train from scratch or fine-tuning. Although these models can be customized to downstream tasks, their performance is not as efficient as training from scratch. There have also been some challenges in explaining and interpreting LLMs' decisions and learning. In this section, we provide an overview of these challenges and limitations, recent efforts to tackle these limitations, and discussing the open research gaps.

Customization: First, because LLMs are trained on general-purpose broad datasets, cyber security context is very limited, which limit their effectiveness in detection and respond to cyber threats. Fine-tuning techniques are proposed in the literature to address this issue, however, it is still challenging due to the scarcity and sensitivity of high-quality labeled data [138]. Additionally, making sure that these customization provide integrity to whole task is another concern. The last noticed attention from research community to this point probably provide promising gate towards fixing this issue.

Cyber attacks' ever-changing landscape: Cyber threats witnesses continuous change in complexity and frequency with rapid development of advanced attacks including as phishing, ransomware, and advanced persistent threats (APT) [139]. Such a landscape limits the ability of LLMs to provide ultimate

resilient solutions. Outdated datasets can yield an ineffective LLM against new threats, as affect all the later stages of incident response. This opens the door for new attacks and threats to be developed. As a consequence, ongoing research development is critically needed to develop resilient versions of LLMs with high ability to adapt to the changing landscapes.

Reflection of training datasets Because cyber defense is one of the most important paradigms, building models for cyber threats detection is a very sensitive task [140]. Training datasets of cyber defence techniques require specific criteria with wide range of cyber threats, cyber crimes scenarios, attacks patterns and defensive measures. On the other hand, LLMs can not train on traditional datasets of cyber security. As a consequence, there is not suitable datasets to build cyber defence specific LLMs. A need of clean dataset from any threats exploits, massive one, including cyber threats data and scenarios, providing smart attack patterns is critical.

Real-time suitability On one hand, the rapid changes of cyber threats along with the sensitivity of cyber defence require the response of threat detection in short time [141]. On the other hand, LLMs require high computational resources and massive datasets for training and retraining. Both issues make the ability of LLMs to be integrated on a real-time mode is very limited. A special attention is needed to tackle this issue. Building efficient LLMs that are able to train on small datasets might be a direction to tackle that issue.

Using LLMs to develop attacks The witnessed progression of LLMs on the machine learning paradigms presented promising solutions for all the realms of scientific research and applied sciences [142]. It also impacted the development of cyber crimes and attacks. This power of recognizing the attacks and threats is accompanied by another potential advancement of cyber attacks. As a result, the landscape of cyber threats is growing wider and complicated. Early directions for organisations and governments to bound the development of GPT with legal standards and policies should be taken into account.

Privacy Preserving LLMs has a significant concern with privacy preserving, because their need to process massive amounts of datasets that include personal information. In other words, LLMs train on general purpose datasets such as Wikipedia, Glove and other dictionaries, therefore, personal information are easily accessible [143]. These personal information may be reproduced from LLMs with simple cyber attacks. Data anonymization and implementing robust privacy-preserving techniques, such as differential privacy is crucial at this stage, however, it still challenging tasks. Using the power of LLMs with privacy preserving techniques remain an open challenge that requires special efforts from researchers and governments.

Adversarial attacks LLMs can be exploited by introducing crafted inputs designed to deceive the models and then give wrong outputs. For example, some attacks can subtly alter the input data that result at incorrect, misleading, or harmful outputs [143]. It is proved that LLMs are vulnerable to the adversarial attacks. Attackers might exploit this vulnerability to bypass security measures, manipulate automated decision-making processes. Developing robust LLMs against adversarial attacks will limit the power of LLMs to evolve, especially they do not have any processing on the input data. Adding methodology attacks to data attacks will make the challenge of LLMs more complicated. A possible direction here is to apply discriminative methodologies that are proves robustness on deep learning models such as [144] to LLMs.

Interperability and transparency The recent progression on GPT techniques paved the way to the involvement of AI in every branch in our daily life activities, including education [145], health [146], driving [147], etc. The result is that AI is about to dominate the decision making operations with a nature of "black box". With this given "black box" of LLMs, clear explanations are essential for trustworthiness and accountability. Also, security analysts should understand how the final classification of a threat is made. The lack of transparency can hinder the adoption of LLMs. Traditional XAI models to be applied for LLMs is a probable direction [148].

Overall, these are the challenges in the field of cyber security and threat detection. We focused on mentioning the related challenges to cyber security to enable the researchers to concentrate on them, looking for quick solutions. Across the cyber security, there are more challenges such as internal hallucination, etc [149].

X. Conclusions

This paper surveyed close to 80 articles of LLMs within the cyber defense paradigm. Strengths and limitations of LLMs methods in cyber security are discussed. We cover the main branches of cyber defense including threat intelligence, vulnerability assessment, network security, privacy preservation, awareness and training, and ethical guidelines. A hierarchical structure for cyber defense sections is provided.

Despite the power of the developed LLMs in overcoming a lot inherent issues in AI models, several challenges and open gaps still need to be filled, particularly when using these models in cyber defense. These challenges along research directions that are still open are listed and discussed. Although there are recent efforts proposed to address these limitations, it is still far from the acceptable solutions on such delicate field. This review will provide an overview picture for researchers on open challenges and possible directions to enable developing better LLMs suited for cyber security applications.

References

[1] N. Moustafa, I. A. Khan, M. Hassanin, D. Ormrod, D. Pi, I. Razzak, and J. Slay, "Dfsat: Deep federated learning for identifying cyber threats in iot-based satellite networks," *IEEE Transactions on Industrial Informatics*, 2022. 1

[2] J. Contreras-Castillo, S. Zeadally, and J. A. Guerrero-Ibañez, "Internet of vehicles: architecture, protocols, and security," *IEEE internet of things Journal*, vol. 5, no. 5, pp. 3701–3709, 2017. 1

[3] L. Zhu, S. Majumdar, and C. Ekenna, "An invisible warfare with the internet of battlefield things: a literature review," *Human behavior and emerging technologies*, vol. 3, no. 2, pp. 255–260, 2021. 1

[4] R. Baheti and H. Gill, "Cyber-physical systems," *The impact of control technology*, vol. 12, no. 1, pp. 161–166, 2011. 1

[5] P. Merle, S. Gearhart, C. Craig, M. Vandyke, M. E. Brooks, and M. Rahimi, "Computers, tablets, and smart phones: The truth about web-based surveys," *Survey Practice*, vol. 8, no. 6, 2015. 1

[6] M. G. Juarez, V. J. Botti, and A. S. Giret, "Digital twins: Review and challenges," *Journal of Computing and Information Science in Engineering*, vol. 21, no. 3, p. 030802, 2021. 1

[7] A. R. Sai, J. Buckley, and A. Le Gear, "Privacy and security analysis of cryptocurrency mobile applications," in *2019 fifth conference on mobile and secure services (MobiSecServ)*. IEEE, 2019, pp. 1–6. 1

[8] M. Farooq and M. Hassan, "Iot smart homes security challenges and solution," *International Journal of Security and Networks*, vol. 16, no. 4, pp. 235–243, 2021. 1

[9] J. O. Oyelami and A. M. Kassim, "Cyber security defence policies: A proposed guidelines for organisations cyber security practices," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 8, 2020. 1

[10] S. Salim, N. Moustafa, M. Hassanian, D. Ormod, and J. Slay, "Deep federated learning-based threat detection model for extreme satellite communications," *IEEE Internet of Things Journal*, 2023. 1

[11] O. v. Sviatun, O. v. Goncharuk, C. Roman, O. Kuzmenko, and I. V. Kozych, "Combating cybercrime: economic and legal aspects," *WSEAS Transactions on Business and Economics*, vol. 18, pp. 751–762, 2021. 1

[12] R. K. Shukla and A. K. Tiwari, "Security analysis of the cyber crime," in *The Ethical Frontier of AI and Data Analysis*. IGI Global, 2024, pp. 257–271. 1

[13] M. Hassanin, S. Anwar, I. Radwan, F. S. Khan, and A. Mian, "Visual attention methods in deep learning: An in-depth survey," *Information Fusion*, vol. 108, p. 102417, 2024. 1

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017. 1, 7

[15] M. Hassanin, A. Khamiss, M. Bennamoun, F. Boussaid, and I. Radwan, "Crossformer: Cross spatio-temporal transformer for 3d human pose estimation," *arXiv preprint arXiv:2203.13387*, 2022. 1

[16] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014. 1

[17] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE signal processing magazine*, vol. 35, no. 1, pp. 53–65, 2018. 2

[18] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023. 3

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018. 2

[20] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020. 2

[21] D. Saha, S. Tarek, K. Yahyaei, S. K. Saha, J. Zhou, M. Tehranipoor, and F. Farahmandi, "Llm for soc security: A paradigm shift," *arXiv preprint arXiv:2310.06046*, 2023. 2

[22] Z. Dong, Z. Zhou, C. Yang, J. Shao, and Y. Qiao, "Attacks, defenses and evaluations for llm conversation safety: A survey," *arXiv preprint arXiv:2402.09283*, 2024. 2

[23] H. Karlzen and T. Sommestad, "Automatic incident response solutions: a review of proposed solutions' input and output," in *Proceedings of the 18th International Conference on Availability, Reliability and Security*, 2023, pp. 1–9. 2

[24] J. Zhang, H. Bu, H. Wen, Y. Chen, L. Li, and H. Zhu, "When llms meet cybersecurity: A systematic literature review," *arXiv preprint arXiv:2405.03644*, 2024. 2

[25] Y. Li, H. Wen, W. Wang, X. Li, Y. Yuan, G. Liu, J. Liu, W. Xu, X. Wang, Y. Sun *et al.*, "Personal llm agents: Insights and survey about the capability, efficiency and security," *arXiv preprint arXiv:2401.05459*, 2024. 2

[26] Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang, "A survey on large language model (llm) security and privacy: The good, the bad, and the ugly," *High-Confidence Computing*, p. 100211, 2024. 2

[27] F. Wu, N. Zhang, S. Jha, P. McDaniel, and C. Xiao, "A new era in llm security: Exploring security concerns in real-world llm-based systems," *arXiv preprint arXiv:2402.18649*, 2024. 2

[28] M. Conti, T. Dargahi, and A. Dehghantanha, *Cyber threat intelligence: challenges and opportunities*. Springer, 2018. 2

[29] F. Sufi, "An innovative gpt-based open-source intelligence using historical cyber incident reports," *Natural Language Processing Journal*, p. 100074, 2024. 3

[30] G. Siracusano, D. Sanvito, R. Gonzalez, M. Srinivasan, S. Kamatchi, W. Takahashi, M. Kawakita, T. Kakumaru, and R. Bifulco, "Time for action: Automated analysis of cyber threat intelligence in the wild," *arXiv preprint arXiv:2307.10214*, 2023. 3

[31] S. Mitra, S. Neupane, T. Chakraborty, S. Mittal, A. Piplai, M. Gaur, and S. Rahimi, "Localintel: Generating organizational threat intelligence from global and local cyber knowledge," *arXiv preprint arXiv:2401.10036*, 2024. 3

[32] Y. Hu, F. Zou, J. Han, X. Sun, and Y. Wang, "Llm-tikg: Threat intelligence knowledge graph construction utilizing large language model," *Available at SSRN 4671345*, 2023. 3

[33] M. Sewak, V. Emani, and A. Naresh, "Crush: Cybersecurity research using universal llms and semantic hypernetworks," 2023. 4

[34] S. Hays and D. J. White, "Using llms for tabletop exercises within the security domain," *arXiv preprint arXiv:2403.01626*, 2024. 4

[35] M. Siavvas, "Vulnerability prediction using large language models (llms)," https://dossproject.eu/the-doss-approach-for-vulnerability-prediction-using-large-language-models-llms/, 2023, [Online; accessed 19-July-2024]. 5

[36] K. Scarfone, M. Souppaya, A. Cody, and A. Orebaugh, "Technical guide to information security testing and assessment," *NIST Special Publication*, vol. 800, no. 115, pp. 2–25, 2008. 4

[37] H. Eakin and A. L. Luers, "Assessing the vulnerability of social-environmental systems," *Annu. Rev. Environ. Resour.*, vol. 31, pp. 365–394, 2006. 4

[38] 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," *arXiv preprint arXiv:2308.06782*, 2023. 4

[39] M. A. Ferrag, A. Battah, N. Tihanyi, M. Debbah, T. Lestable, and L. C. Cordeiro, "Securefalcon: The next cyber reasoning system for cyber security," *arXiv preprint arXiv:2307.06616*, 2023. 4

[40] S. Temara, "Maximizing penetration testing success with effective reconnaissance techniques using chatgpt," *arXiv preprint arXiv:2307.06391*, 2023. 4

[41] A. Happe and J. Cito, "Getting pwn'd by ai: Penetration testing with large language models," in *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2023, pp. 2082–2086. 4

[42] H. Pearce, B. Tan, B. Ahmad, R. Karri, and B. Dolan-Gavitt, "Examining zero-shot vulnerability repair with large language models," in *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2023, pp. 2339–2356. 5

[43] V. Akuthota, R. Kasula, S. T. Sumona, M. Mohiuddin, M. T. Reza, and M. M. Rahman, "Vulnerability detection and monitoring using llm," in *2023 IEEE 9th International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE)*. IEEE, 2023, pp. 309–314. 5

[44] R. Ingemann Tuffveson Jensen, V. Tawosi, and S. Alamir, "Software vulnerability and functionality assessment using llms," *arXiv e-prints*, pp. arXiv–2403, 2024. 5

[45] N. S. Mathews, Y. Brus, Y. Aafer, M. Nagappan, and S. McIntosh, "Llbezpeky: Leveraging large language models for vulnerability detection," *arXiv preprint arXiv:2401.01269*, 2024. 5

[46] S. Sakaoglu, "Kartal: Web application vulnerability hunting using large language models: Novel method for detecting logical vulnerabilities in web applications with finetuned large language models," 2023. 6

[47] C. Chen, J. Su, J. Chen, Y. Wang, T. Bi, Y. Wang, X. Lin, T. Chen, and Z. Zheng, "When chatgpt meets smart contract vulnerability detection: How far are we?" *arXiv preprint arXiv:2309.05520*, 2023. 6

[48] A. Patel, Q. Qassim, and C. Wills, "A survey of intrusion detection and prevention systems," *Information Management & Computer Security*, vol. 18, no. 4, pp. 277–290, 2010. 6

[49] P. Ioulianou, V. Vasilakis, I. Moscholios, and M. Logothetis, "A signature-based intrusion detection system for the internet of things," *Information and Communication Technology Form*, 2018. 6

[50] V. Jyothsna, R. Prasad, and K. M. Prasad, "A review of anomaly based intrusion detection systems," *International Journal of Computer Applications*, vol. 28, no. 7, pp. 26–35, 2011. 6

[51] S. Nedelkoski, J. Cardoso, and O. Kao, "Anomaly detection and classification using distributed tracing and deep learning," in *2019 19th IEEE/ACM international symposium on cluster, cloud and grid computing (CCGRID)*. IEEE, 2019, pp. 241–250. 6

[52] M. Rabbani, Y. Wang, R. Khoshkangini, H. Jelodar, R. Zhao, S. Bagheri Baba Ahmadi, and S. Ayobi, "A review on machine learning approaches for network malicious behavior detection in emerging technologies," *Entropy*, vol. 23, no. 5, p. 529, 2021. 6

[53] N. A. S. Mirza, H. Abbas, F. A. Khan, and J. Al Muhtadi, "Anticipating advanced persistent threat (apt) countermeasures using collaborative security mechanisms," in *2014 International Symposium on Biometrics and Security Technologies (ISBAST)*. IEEE, 2014, pp. 129–132. 6

[54] M. Hassanin, M. Keshk, S. Salim, M. Alsubaie, and D. Sharma, "Pllm-cs: Pre-trained large language model (llm) for cyber threat detection in satellite networks," *arXiv preprint arXiv:2405.05469*, 2024. 6, 7

[55] L. Jiang, "Detecting scams using large language models," *arXiv preprint arXiv:2402.03147*, 2024. 6

[56] Z. Shi, Y. Wang, F. Yin, X. Chen, K.-W. Chang, and C.-J. Hsieh, "Red teaming language model detectors with language models," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 174–189, 2024. 6

[57] Y. Yang, X. Zhou, R. Mao, J. Xu, L. Yang, Y. Zhangm, H. Shen, and H. Zhang, "Dlap: A deep learning augmented large language model prompting framework for software vulnerability detection," *arXiv preprint arXiv:2405.01202*, 2024. 6

[58] L. D. Manocchio, S. Layeghy, W. W. Lo, G. K. Kulatilleke, M. Sarhan, and M. Portmann, "Flowtransformer: A transformer framework for flow-based network intrusion detection systems," *Expert Systems with Applications*, vol. 241, p. 122564, 2024. 7

[59] S. Sai, U. Yashvardhan, V. Chamola, and B. Sikdar, "Generative ai for cyber security: Analyzing the potential of chatgpt, dall-e and other models for enhancing the security space," *IEEE Access*, 2024. 7

[60] S. G. Prasad, V. C. Sharmila, and M. Badrinarayanan, "Role of artificial intelligence based chat generative pre-trained transformer (chatgpt) in cyber security," in *2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*. IEEE, 2023, pp. 107–114. 7

[61] S. A. Salloum, "Detecting malicious accounts in cyberspace: Enhancing security in chatgpt and beyond," in *Artificial Intelligence in Education: The Power and Dangers of ChatGPT in the Classroom*. Springer, 2024, pp. 653–666. 7

[62] K. Ameri, M. Hempel, H. Sharif, J. Lopez Jr, and K. Perumalla, "Cybert: Cybersecurity claim classification by fine-tuning the bert language model," *Journal of Cybersecurity and Privacy*, vol. 1, no. 4, pp. 615–637, 2021. 7, 10

[63] L. G. Nguyen and K. Watabe, "Flow-based network intrusion detection based on bert masked language model," in *Proceedings of the 3rd International CoNEXT Student Workshop*, 2022, pp. 7–8. 7

[64] E. Nwafor and H. Olufowobi, "Canbert: A language-based intrusion detection model for in-vehicle networks," in *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2022, pp. 294–299. 8

[65] S. C. HPL, "Introduction to the controller area network (can)," *Application Report SLOA101*, pp. 1–17, 2002. 8

[66] M. Guastalla, Y. Li, A. Hekmati, and B. Krishnamachari, "Application of large language models to ddos attack detection," in *International Conference on Security and Privacy in Cyber-Physical Systems and Smart Vehicles*. Springer, 2023, pp. 83–99. 8

[67] K. Rieck and P. Laskov, "Language models for detection of unknown attacks in network traffic," *Journal in Computer Virology*, vol. 2, pp. 243–256, 2007. 8

[68] A. R. Tuor, R. Baerwolf, N. Knowles, B. Hutchinson, N. Nichols, and R. Jasper, "Recurrent neural network language models for open vocabulary event-level cyber anomaly detection," in *Workshops at the thirty-second AAAI conference on artificial intelligence*, 2018. 8

[69] N. Alkhatib, M. Mushtaq, H. Ghauch, and J.-L. Danger, "Can-bert do it? controller area network intrusion detection system based on bert language model," in *2022 IEEE/ACS 19th International Conference on Computer Systems and Applications (AICCSA)*. IEEE, 2022, pp. 1–8. 8

[70] 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-to-Peer Networking and Applications*, vol. 17, no. 1, pp. 227–245, 2024. 8

[71] E. Aghaei, X. Niu, W. Shadid, and E. Al-Shaer, "Securebert: A domain-specific language model for cybersecurity," in *International Conference on Security and Privacy in Communication Systems*. Springer, 2022, pp. 39–56. 8

[72] B. Piggott, S. Patil, G. Feng, I. Odat, R. Mukherjee, B. Dharmalingam, and A. Liu, "Net-gpt: A llm-empowered man-in-the-middle chatbot for unmanned aerial vehicle," in *2023 IEEE/ACM Symposium on Edge Computing (SEC)*. IEEE, 2023, pp. 287–293. 8

[73] C. Patsakis, F. Casino, and N. Lykousas, "Assessing llms in malicious code deobfuscation of real-world malware campaigns," *arXiv preprint arXiv:2404.19715*, 2024. 8

[74] H. Fujima, T. Kumamoto, and Y. Yoshida, "Using chatgpt to analyze ransomware messages and to predict ransomware threats," 2023. 8

[75] G. Sandoval, H. Pearce, T. Nys, R. Karri, S. Garg, and B. Dolan-Gavitt, "Lost at c: A user study on the security implications of large language model code assistants," in *32nd USENIX Security Symposium (USENIX Security 23)*, 2023, pp. 2205–2222. 9

[76] J. He and M. Vechev, "Large language models for code: Security hardening and adversarial testing," in *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, 2023, pp. 1865–1879. 9

[77] 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," *Computers & Security*, vol. 135, p. 103476, 2023. 9

[78] P. Sharma and B. Dash, "Impact of big data analytics and chatgpt on cybersecurity," in *2023 4th International Conference on Computing and Communication Systems (I3CS)*. IEEE, 2023, pp. 1–6. 9

[79] H. Lai, "Intrusion detection technology based on large language models," in *2023 International Conference on Evolutionary Algorithms and Soft Computing Techniques (EASCT)*. IEEE, 2023, pp. 1–5. 9

[80] M. Labonne and S. Moran, "Spam-t5: Benchmarking large language models for few-shot email spam detection," *arXiv preprint arXiv:2304.01238*, 2023. 9

[81] 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," *arXiv preprint arXiv:2310.19181*, 2023. 9

[82] Y. Wu, S. Si, Y. Zhang, J. Gu, and J. Wosik, "Evaluating the performance of chatgpt for spam email detection," *arXiv preprint arXiv:2402.15537*, 2024. 9

[83] F. Heiding, B. Schneier, A. Vishwanath, and J. Bernstein, "Devising and detecting phishing: Large language models vs. smaller human models," *arXiv preprint arXiv:2308.12287*, 2023. 9

[84] Y. Li, C. Huang, S. Deng, M. L. Lock, T. Cao, N. Oo, B. Hooi, and H. W. Lim, "Knowphish: Large language models meet multimodal knowledge graphs for enhancing reference-based phishing detection," *arXiv preprint arXiv:2403.02253*, 2024. 9

[85] M. Hur, S. Seo, J. Hwang, H. Lim, and M. Min, "Utilizing large language models for detection of sms spam in few-shot settings," *Available at SSRN 4815382*. 9

[86] T. Koide, N. Fukushi, H. Nakano, and D. Chiba, "Chatspamdetector: Leveraging large language models for effective phishing email detection," *arXiv preprint arXiv:2402.18093*, 2024. 9

[87] H. Patel, U. Rehman, and F. Iqbal, "Large language models spot phishing emails with surprising accuracy: A comparative analysis of performance," *arXiv preprint arXiv:2404.15485*, 2024. 10

[88] F. Trad and A. Chehab, "Prompt engineering or fine-tuning? a case study on phishing detection with large language models," *Machine Learning and Knowledge Extraction*, vol. 6, no. 1, pp. 367–384, 2024. 10

[89] D. Nahmias, G. Engelberg, D. Klein, and A. Shabtai, "Prompted contextual vectors for spear-phishing detection," *arXiv preprint arXiv:2402.08309*, 2024. 10

[90] S. Jamal and H. Wimmer, "An improved transformer-based model for detecting phishing, spam, and ham: A large language model approach," *arXiv preprint arXiv:2311.04913*, 2023. 10

[91] F. Heiding, B. Schneier, A. Vishwanath, J. Bernstein, and P. S. Park, "Devising and detecting phishing emails using large language models," *IEEE Access*, 2024. 10

[92] F. Yu and M. V. Martin, "Honey, i chunked the passwords: Generating semantic honeywords resistant to targeted attacks using pre-trained language models," in *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*. Springer, 2023, pp. 89–108. 10

[93] R. Chataut, P. K. Gyawali, and Y. Usman, "Can ai keep you safe? a study of large language models for phishing detection," in *2024 IEEE 14th Annual Computing and Communication Workshop and Conference (CCWC)*. IEEE, 2024, pp. 0548–0554. 10

[94] P. Balasubramanian, J. Seby, and P. Kostakos, "Transformer-based llms in cybersecurity: An in-depth study on log anomaly detection and conversational defense mechanisms," in *2023 IEEE International Conference on Big Data (BigData)*. IEEE, 2023, pp. 3590–3599. 10

[95] 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*, 2024. 10

[96] M. Bayer, P. Kuehn, R. Shanehsaz, and C. Reuter, "Cysecbert: A domain-adapted language model for the cybersecurity domain," *ACM Transactions on Privacy and Security*, vol. 27, no. 2, pp. 1–20, 2024. 10

[97] P. Ranade, A. Piplai, A. Joshi, and T. Finin, "Cybert: Contextualized embeddings for the cybersecurity domain," in *2021 IEEE International Conference on Big Data (Big Data)*. IEEE, 2021, pp. 3334–3342. 10

[98] C. Wohlbach, M. M. Chowdhury, and S. Latif, "Evaluating cybersecurity risks in nlp models: Google bard as bard of prey and chatgpt as cyber crime aide," *Proceedings of 39th International Confer*, vol. 98, pp. 159–168, 2024. 10

[99] A. Zaboli, S. L. Choi, T.-J. Song, and J. Hong, "Chatgpt and other large language models for cybersecurity of smart grid applications," *arXiv preprint arXiv:2311.05462*, 2023. 11

[100] F. McKee and D. Noever, "Chatbots in a honeypot world," *arXiv preprint arXiv:2301.03771*, 2023. 11

[101] T. Koide, N. Fukushi, H. Nakano, and D. Chiba, "Detecting phishing sites using chatgpt," *arXiv preprint arXiv:2306.05816*, 2023. 11

[102] M. Sladić, V. Valeros, C. Catania, and S. Garcia, "Llm in the shell: Generative honeypots," *arXiv preprint arXiv:2309.00155*, 2023. 11

[103] Y. Wang, X. Liang, X. Hei, W. Ji, and L. Zhu, "Deep learning data privacy protection based on homomorphic encryption in aiot," *Mobile Information Systems*, vol. 2021, pp. 1–11, 2021. 11

[104] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, "Random-data perturbation techniques and privacy-preserving data mining," *Knowledge and Information Systems*, vol. 7, pp. 387–414, 2005. 11

[105] Y. Chen, A. Arunasalam, and Z. B. Celik, "Can large language models provide security & privacy advice? measuring the ability of llms to refute misconceptions," in *Proceedings of the 39th Annual Computer Security Applications Conference*, 2023, pp. 366–378. 11

[106] S. Kim, S. Yun, H. Lee, M. Gubri, S. Yoon, and S. J. Oh, "Propile: Probing privacy leakage in large language models," *Advances in Neural Information Processing Systems*, vol. 36, 2024. 11

[107] M. Raeini, "Privacy-preserving large language models (ppllms)," *Available at SSRN 4512071*, 2023. 11

[108] 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*, 2023. 11

[109] S. Singh, "Enhancing privacy and security in large-language models: A zero-knowledge proof approach," in *International Conference on Cyber Warfare and Security*, vol. 19, no. 1, 2024, pp. 574–582. 11

[110] D. Galinec and L. Luić, "Design of conceptual model for raising awareness of digital threats," *WSEAS transactions on environment and development*, vol. 16, pp. 493–504, 2020. 12

[111] A. Aleroud and L. Zhou, "Phishing environments, techniques, and countermeasures: A survey," *Computers & Security*, vol. 68, pp. 160–196, 2017. 12

[112] M. Lagana, "Information security in an ever-changing threat landscape," in *The Routledge Companion to Risk, Crisis and Security in Business*. Routledge, 2018, pp. 255–271. 12

[113] T. Gundu, "Chatbots: A framework for improving information security behaviours using chatgpt," in *International Symposium on Human Aspects of Information Security and Assurance*. Springer, 2023, pp. 418–431. 12

[114] K. I. Roumeliotis and N. D. Tselikas, "Chatgpt and open-ai models: A preliminary review," *Future Internet*, vol. 15, no. 6, p. 192, 2023. 12

[115] M. M. Yamin, E. Hashmi, M. Ullah, and B. Katt, "Applications of llms for generating cyber security exercise scenarios," 2024. 12

[116] M. Kaheh, D. K. Kholgh, and P. Kostakos, "Cyber sentinel: Exploring conversational agents in streamlining security tasks with gpt-4," *arXiv preprint arXiv:2309.16422*, 2023. 13

[117] K. S. Kalyan, "A survey of gpt-3 family large language models including chatgpt and gpt-4," *Natural Language Processing Journal*, p. 100048, 2023. 13

[118] S. Shafee, A. Bessani, and P. M. Ferreira, "Evaluation of llm chatbots for osint-based cyberthreat awareness," *arXiv preprint arXiv:2401.15127*, 2024. 13

[119] S. M. Mohammad and L. Surya, "Security automation in information technology," *International journal of creative research thoughts (IJCRT)–Volume*, vol. 6, 2018. 13

[120] J. Kinyua and L. Awuah, "Ai/ml in security orchestration, automation and response: Future research directions." *Intelligent Automation & Soft Computing*, vol. 28, no. 2, 2021. 13

[121] I. H. Sarker, *AI-Driven Cybersecurity andThreat Intelligence: Cyber Automation, Intelligent Decision-Making and Explainability*. Springer Nature, 2024. 13

[122] U. Bartwal, S. Mukhopadhyay, R. Negi, and S. Shukla, "Security orchestration, automation, and response engine for deployment of behavioural honeypots," in *2022 IEEE Conference on Dependable and Secure Computing (DSC)*. IEEE, 2022, pp. 1–8. 13

[123] S. Hays and D. J. White, "Employing llms for incident response planning and review," *arXiv preprint arXiv:2403.01271*, 2024. 13

[124] R. Fang, R. Bindu, A. Gupta, and D. Kang, "Llm agents can autonomously exploit one-day vulnerabilities," *arXiv preprint arXiv:2404.08144*, 2024. 13, 14

[125] M. Feffer, A. Sinha, Z. C. Lipton, and H. Heidari, "Red-teaming for generative ai: Silver bullet or security theater?" *arXiv preprint arXiv:2401.15897*, 2024. 13

[126] O. G. Lira, A. Marroquin, and M. A. To, "Harnessing the advanced capabilities of llm for adaptive intrusion detection systems," in *International Conference on Advanced Information Networking and Applications*. Springer, 2024, pp. 453–464. 13

[127] M. Sultana, A. Taylor, L. Li, and S. Majumdar, "Towards evaluation and understanding of large language models for cyber operation automation," in *2023 IEEE Conference on Communications and Network Security (CNS)*. IEEE, 2023, pp. 1–6. 13

[128] G. Kokolakis, A. Moschos, and A. D. Keromytis, "Harnessing the power of general-purpose llms in hardware trojan design," in *Proceedings of the 5th Workshop on Artificial Intelligence in Hardware Security, in conjunction with ACNS*, 2024. 14

[129] A. Helbling, M. Phute, M. Hull, and D. H. Chau, "Llm self defense: By self examination, llms know they are being tricked," *arXiv preprint arXiv:2308.07308*, 2023. 14

[130] O. Gadyatskaya and D. Papuc, "Chatgpt knows your attacks: Synthesizing attack trees using llms," in *International Conference on Data Science and Artificial Intelligence*. Springer, 2023, pp. 245–260. 14

[131] M. Shao, B. Chen, S. Jancheska, B. Dolan-Gavitt, S. Garg, R. Karri, and M. Shafique, "An empirical evaluation of llms for solving offensive security challenges," *arXiv preprint arXiv:2402.11814*, 2024. 14

[132] I. David, L. Zhou, K. Qin, D. Song, L. Cavallaro, and A. Gervais, "Do you still need a manual smart contract audit?" *arXiv preprint arXiv:2306.12338*, 2023. 14

[133] E. Cambiaso and L. Caviglione, "Scamming the scammers: Using chatgpt to reply mails for wasting time and resources," *arXiv preprint arXiv:2303.13521*, 2023. 14

[134] T. Ali and P. Kostakos, "Huntgpt: Integrating machine learning-based anomaly detection and explainable ai with large language models (llms)," *arXiv preprint arXiv:2309.16021*, 2023. 14, 15

[135] M. A. Uddin and I. H. Sarker, "An explainable transformer-based model for phishing email detection: A large language model approach," *arXiv preprint arXiv:2402.13871*, 2024. 14, 15

[136] G. Sebastian, "Do chatgpt and other ai chatbots pose a cybersecurity risk?: An exploratory study," *International Journal of Security and Privacy in Pervasive Computing (IJSPPC)*, vol. 15, no. 1, pp. 1–11, 2023. 15

[137] B. Garvey and A. Svendsen, "Can generative-ai (chatgpt and bard) be used as red team avatars in developing foresight scenarios?" *Analytic Research Consortium (ARC) August*, 2023. 15

[138] A. Kumar, A. Raghunathan, R. Jones, T. Ma, and P. Liang, "Fine-tuning can distort pretrained features and underperform out-of-distribution," *arXiv preprint arXiv:2202.10054*, 2022. 15

[139] R. Dillon, P. Lothian, S. Grewal, and D. Pereira, "Cyber security: evolving threats in an ever-changing world," in *Digital Transformation in a Post-Covid World*. CRC Press, 2021, pp. 129–154. 15

[140] A. Razzaq, A. Hur, H. F. Ahmad, and M. Masood, "Cyber security: Threats, reasons, challenges, methodologies and state of the art solutions for industrial applications," in *2013 IEEE Eleventh International Symposium on Autonomous Decentralized Systems (ISADS)*. IEEE, 2013, pp. 1–6. 16

[141] R. Brewer, "Cyber threats: reducing the time to detection and response," *Network Security*, vol. 2015, no. 5, pp. 5–8, 2015. 16

[142] B. Gordijn and H. t. Have, "Chatgpt: evolution or revolution?" *Medicine, Health Care and Philosophy*, vol. 26, no. 1, pp. 1–2, 2023. 16

[143] B. Yan, K. Li, M. Xu, Y. Dong, Y. Zhang, Z. Ren, and X. Cheng, "On protecting the data privacy of large language models (llms): A survey," *arXiv preprint arXiv:2403.05156*, 2024. 16

[144] M. Hassanin, I. Radwan, N. Moustafa, M. Tahtali, and N. Kumar, "Mitigating the impact of adversarial attacks in very deep networks," *Applied Soft Computing*, vol. 105, p. 107231, 2021. 16

[145] S. Moore, R. Tong, A. Singh, Z. Liu, X. Hu, Y. Lu, J. Liang, C. Cao, H. Khosravi, P. Denny *et al.*, "Empowering education with llms-the next-gen interface and content generation," in *International Conference on Artificial Intelligence in Education*. Springer, 2023, pp. 32–37. 16

[146] M. Sallam, "Chatgpt utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns," in *Healthcare*, vol. 11, no. 6. MDPI, 2023, p. 887. 16

[147] L. Chen, O. Sinavski, J. Hünermann, A. Karnsund, A. J. Willmott, D. Birch, D. Maund, and J. Shotton, "Driving with llms: Fusing object-level vector modality for explainable autonomous driving," *arXiv preprint arXiv:2310.01957*, 2023. 16

[148] A. Zytek, S. Pidò, and K. Veeramachaneni, "Llms for xai: Future directions for explaining explanations," *arXiv preprint arXiv:2405.06064*, 2024. 16

[149] K. Andriopoulos and J. Pouwelse, "Augmenting llms with knowledge: A survey on hallucination prevention," *arXiv preprint arXiv:2309.16459*, 2023. 16