Received: 8 June 2025 | Revised: 20 June 2025 | Accepted: 23 June 2025 | Vol. 1 | Issue 1 | June 2025
e-ISSN: 3105-2169

International Journal of Mechatronics, Robotics, and Artificial Intelligence
Review article

Harnessing Large Language Models for Enhanced Cybersecurity: A Review of Their Role in Defending Against APT and Cyber Attacks

Zainab S. Aziz*[1], **Ali A. Abed**[2],
[1] Computer Engineering Department, University of Basrah, Basrah, Iraq.
[2]Mechatronics Engineering Department, University of Basrah, Basrah, Iraq.

Correspondence
*Zainab Salim Aziz
Computer Engineering Department, University of Basrah
Email: pgs.zainab.salim@uobasrah.edu.iq

Abstract

The emergence of Large Language Models (LLMs) has opened new frontiers in artificial intelligence applications across multiple domains, including cybersecurity. This paper presents a comprehensive review of the role of LLMs in enhancing cyber defense mechanisms, with a particular focus on their effectiveness in identifying, mitigating, and responding to Advanced Persistent Threats (APTs) and other sophisticated cyber-attacks. We explore the integration of LLMs in threat intelligence, anomaly detection, automated incident response, and adversarial behavior analysis. By examining recent advancements, case studies, and state-of-the-art implementations, we highlight the strengths and limitations of current LLM-based approaches. Furthermore, we assess the challenges related to scalability, adversarial robustness, and ethical considerations inherent in deploying LLMs within cybersecurity infrastructures. The review concludes with future research directions, emphasizing the need for hybrid AI systems that combine LLMs with traditional rule-based and statistical methods to provide resilient and adaptive cybersecurity solutions in the face of evolving digital threats.

Keywords
Advanced Persistent Threat APT, LLaMA2, LLMs, Cyber-attack.

I. INTRODUCTION

The significance of data security has been highlighted on a worldwide level due to the fact that state-sponsored threat actors have compromised many multinational networks through the use of Advanced Persistent Threat (APT) attacks. Breakthrough models like ChatGPT, LLaMA, and its variants, known as large language models (LLMs), have made substantial progress in the field of artificial intelligence. These models, utilizing large datasets and sophisticated neural network structures, exhibit exceptional capabilities in comprehending and producing human language. They not only establish new standards for attaining artificial general intelligence (AGI), but also demonstrate exceptional flexibility and efficiency when working along with professionals in certain fields. This research allows for the customization of LLMs to address the unique issues in different industries, hence promoting advancement and growth in healthcare, law, education, software engineering, and other areas. Exploring LLM applications in cybersecurity can establish the basis for additional model study and deployment, while also emphasizing the potential transformational effects [1]. The escalating number of cyber-attacks presents substantial hazards to individuals, corporations, and governments, making cybersecurity a crucial concern [2]. While recent advances in Large Language Models (LLMs) have shown significant promise in various cybersecurity applications, most existing studies focus narrowly on specific tasks—such as phishing detection, malware classification, or log parsing—using individual models in isolation. There remains a lack of comprehensive reviews that systematically compare the capabilities, limitations, and suitability of different LLMs across a broad range of cybersecurity use cases. Furthermore, the literature often overlooks practical deployment considerations, such as model scalability, contextual reasoning, and real-time adaptability. To address this gap, the present study aims to provide a structured and comparative review of widely used LLMs such as BERT, RoBERTa, GPT-3, GPT-4, LLaMA2, and T5 highlighting their strengths, challenges, and potential for defending against modern cyber threats, including Advanced Persistent Threats (APTs).

II. METHODOLOGY

This study adopts a narrative and comparative review methodology to evaluate the role of Large Language Models (LLMs) in cybersecurity. The objective is to systematically explore how these models have been applied in various cybersecurity tasks, including threat intelligence extraction, phishing detection, anomaly detection, intrusion detection, and automated incident response.

A. Literature Search Strategy

The literature reviewed in this study is gathered from reputable academic databases, including IEEE Xplore, Springer Link, Science Direct, arXiv, and Google Scholar. The search is conducted using combinations of the following keywords: "Large Language Models in cybersecurity", "BERT for threat detection", "GPT-3 cyber-attacks", "LLMs APT detection", "LLM log analysis", and "transformer-based cybersecurity models". The initial search included publications from 2020 to 2025, with a few seminal earlier works included for foundational context.

B. Inclusion and Exclusion Criteria

To ensure the relevance and quality of the reviewed material, the following inclusion criteria were applied:

- Peer-reviewed journal articles, conference proceedings, and preprints
- Studies presenting applications or evaluations of LLMs in cybersecurity
- Papers written in English
- Works involving quantitative or qualitative evaluation of LLM-based models

C. Data Extraction and Comparative Analysis

Each selected study was analyzed for key characteristics, including:

- Size and training data of the LLM
- Cybersecurity use-case (e.g., phishing detection, intrusion classification)
- Performance metrics, such as accuracy, precision, recall, and F1-score
- Deployment considerations, such as scalability, open-source availability, and prompt tuning complexity

A comparative matrix (Table IV) is developed to summarize these features across popular LLMs including BERT, RoBERTa, GPT-3, GPT-4, LLaMA2, and T5. The analysis aimed to provide a balanced view of the strengths and limitations of each model in real-world cybersecurity applications.

III. NLP IN CYBERSECURITY FIELD

In the past ten years, the development of machines capable of engaging in conversation and interacting with humans in a manner resembling human behavior has become a tangible achievement. This phenomenon is observed in numerous applications that directly interact with humans, as well as in new technologies such as smart assistants, intelligent search engines, and customer support systems. Although these applications vary greatly in their respective contexts, they all share a common feature: they utilize an engine that incorporates advancements in natural language processing (NLP) techniques, leveraging breakthroughs in deep machine learning (ML) and artificial intelligence (AI). The basic goal of NLP applications is to utilize the power of ML/AI to comprehend natural languages, process and analyze them, and extract meaningful insights from these analyses, as ML/AI continue to transform the daily lives [3]. NLP is an academic discipline that investigates the interface between computers and human language. It is situated within the realms of AI and computational linguistics. Natural language processing encompasses the creation of methods and algorithms that empower computers to comprehend, scrutinize, and produce text or speech in a manner that emulates human communication. NLP has a wide range of applications and is not limited to any certain field or employment. Text classification is the categorization of text into predetermined categories or classifications. Named Entity Recognition (NER) is the process of identifying and extracting certain entities from text, such as the names of individuals, groups, locations, or dates. Sentiment analysis is the process of ascertaining the sentiment expressed in a text, whether it be positive, negative, or neutral. Language modeling involves the creation of statistical models that estimate the likelihood of a specific sequence of words appearing in a sentence or document. Machine translation is the automated process of transforming text from one language to another. Responding to queries by providing accurate and relevant answers in regular conversation, Text Summarization is the process of creating concise summaries of lengthy texts. Speech recognition is the process of transforming spoken language into written text. Natural language generation refers to the act of creating text or speech that mimics human language, using input or data as a basis. NLP techniques employ several methodologies, including rule-based systems, statistical models, machine learning algorithms, deep learning architectures, and language rule sets. Computers are now capable of reading and evaluating linguistic properties such as grammar, syntax, semantics, and pragmatics. NLP is highly influential in the subject of cybersecurity because of its several advantages. In the field of cybersecurity, NLP algorithms can identify patterns, anomalies, and indications of penetration that traditional security procedures may overlook by examining the semantic and syntactic framework of language. Integrating contextual information enhances the precision and effectiveness of threat detection and security monitoring systems. Advanced Threat Detection and Response and NLP approaches can assist in detecting suspicious behavior and potential security breaches. NLP algorithms have the capability to detect patterns that suggest fraudulent activity or social engineering attempts by examining user communications, logs, and other sources of textual data. This minimizes potential harm by enabling firms to rapidly and efficiently address security concerns. NLP enhances efficiency and automation by eliminating the requirement for

human analysts to manually evaluate and analyze textual material, as these operations are automated. Automation enables security personnel to dedicate their attention to more valuable duties, such as incident response, threat hunting, and vulnerability management, resulting in enhanced productivity. Cybersecurity systems that utilize NLP can enhance user experience by employing conversational agents, intelligent chatbots, and natural language interfaces. These interfaces enhance user interaction with security systems by providing a more intuitive experience. Users can ask questions, receive immediate responses, and access pertinent security information and guidance. NLP approaches can be utilized in the realm of cybersecurity to identify and mitigate adversary attacks. Cybercriminals may employ tactics to mislead or hide written information with the intention of evading security measures. NLP systems can aid in the detection of assaults by evaluating linguistic patterns, semantic inconsistencies, and atypical language usage[4].

IV. Large language model (LLM)

LLM is an AI system that uses deep learning and a significant amount of human language data to understand, extract, and create new linguistic material. LLMs are commonly known as generative AI. The design of these models is exclusively intended to provide content solely in the textual format. Transformer models, a deep learning architecture used in natural language processing, have quickly become an essential component of language and machine learning models such as LLMs. GPT-3 is a language model that utilizes a Generative Pre-Trained Transformer architecture. It is employed by ChatGPT, a highly regarded AI chatbot created by OpenAI. GPT-3 possesses the capability to produce persuasive information, create computer code, compose poetry in many human poetic forms, and carry out more assignments. Furthermore, GPT-3 has demonstrated remarkable efficacy as a security tool. GPT-3's recent demonstration exhibited its capacity to identify hundreds of security vulnerabilities in a solitary codebase, outperforming tens of flaws found by conventional commercial solutions offered by a well-established cybersecurity organization. Due to the growing popularity of LLMs, early studies have mostly focused on identifying the constraints, difficulties, and potential hazards related with these models in the fields of cybersecurity and privacy [5]. The advancement of LLMs has significantly bolstered the capabilities of artificial intelligence. Prominent instances of these models encompass pioneering models like ChatGPT, LLaMA, and its variations. These models exhibit remarkable expertise in producing and comprehending human language by utilizing huge datasets and cutting-edge neural network structures. They not only establish new standards for the progress of artificial general intelligence (AGI), but they also exhibit exceptional adaptability and efficiency in working together with specialists in certain domains. This study methodology allows for the tailoring of LLMs to cater to the specific requirements of many domains, hence fostering the advancement and growth of sectors such as software engineering, law, healthcare, and education. The field of modern cybersecurity has several challenges as adversaries rapidly and dynamically evolve to exploit holes and avoid detection. Conventional methods like rule-based systems and signature-based detection frequently face difficulties in keeping up with the constantly evolving threat landscape. Advancements in artificial intelligence, specifically in machine learning, have created new opportunities for improving cybersecurity. LLMs have undergone alterations to meet the requirements of the cybersecurity field. Open-source low-level languages models, such as LLaMA, have facilitated the creation of domain-specific language-level models focused on cybersecurity, such as RepairLlama and Hackmentor. However, even in the absence of specific cybersecurity knowledge, sophisticated language models such as ChatGPT can effectively tackle intricate tasks by employing swift engineering, contextual learning, and logical reasoning. These initial endeavors suggest that LLMs make a valuable contribution to cybersecurity activities, yielding favorable outcomes [6].

V. Related Works

Several studies in the literature focus on the efficacy of GenAI tools. As an illustration, Hadi et al. [7] trained GPT-3, a language model that follows a sequential pattern and has an impressive 175 billion parameters. They have extended the NLP processing, demonstrating remarkable few shot learning capabilities. This model exhibits effectiveness in several NLP tasks, such as question answering and translation, even without requiring training unique to the job. It often matches or surpasses state-of-the-art, meticulously crafted systems. Carolan et al. [8] critically analyze the potential and limitations of multi-modal LLMs, which encompass six publicly available models for text, code, picture, and video modalities, as well as proprietary models such as GPT-4 and Gemini. The study, which employed a comprehensive qualitative analysis of 230 cases, reveals a clear disparity between the performance of the GenAIs and the expectations of the public. The investigation focused on assessing generalizability. trustworthiness, and causal reasoning. These findings provide a foundation for future research to improve the transparency and reliability of GenAI in cybersecurity and other fields, which will enable the development of more complex and reliable multi-modal applications. A separate study conducts a comprehensive analysis of Google's Gemini and OpenAI's GPT models, specifically focusing on their ability to do commonsense reasoning in various multimodal tasks. This inquiry examines the advantages and disadvantages of Gemini's ability to combine practical data, focusing on areas for improvement in its competitive effectiveness in terms of temporal reasoning, social reasoning, and visual emotion detection. The text emphasizes the importance of fostering the development of commonsense reasoning in GenAI models to improve cybersecurity applications [9]. A novel classifier of cybersecurity feature claims through the refinement of a pre-trained BERT language model is presented in this publication. In this article, a claim is defined as any natural language sequence that expressly

states that a cybersecurity feature associated with a product is available. Subsequently, all other claims or sequences with no claims are marked as lacking a cybersecurity claim. CyBERT is meant to be utilized as a classification model tailored to cybersecurity that finds cybersecurity claims. The methodology comprises two primary stages: the curation of a database containing claim sequences and the refinement of BERT through fine-tuning. As far as we know, there is presently no easily accessible dataset that specifically focuses on cybersecurity and can be used to train CyBERT. Consequently, a novel dataset with labeled sequences relevant to claims was generated. The fine-tuning procedure of CyBERT begins by initializing it with the initial weights of BERT. Dense layers and dropouts are then incorporated, and all layers and parameters are further adjusted using the labeled database and architecture. The new CyBERT model improved the accuracy of the classification test by 19 percentage points compared to the original BERT text classifier [10]. In this work, a novel language model named Secure BERT, which makes use of the cutting-edge BERT NLP architecture, is introduced in order to tackle a crucial cybersecurity issue. Secure BERT processes texts with implications for cybersecurity in an efficient manner. It may be used for a wide range of cybersecurity activities, including intrusion detection, code and malware analysis, phishing detection [11], and so on, because it is sufficiently general. One of the most important components of every cybersecurity report is the pre-trained cybersecurity language model Secure BERT, which has a basic grasp of both word- and sentence-level semantics. Within this framework, an extensive collection of 1.1 billion words (with a vocabulary size of 1.6 million) from diverse cybersecurity sources such as news, reports, textbooks, publications, research papers, and videos was gathered and analyzed. In addition to the pre-trained tokenizer, a bespoke tokenization algorithm was created to maintain conventional English language while efficiently handling novel tokens related to cybersecurity. Furthermore, a pragmatic approach was employed to enhance the retraining process by introducing random perturbations to the pre-trained weights. A thorough evaluation of the proposed model was conducted by performing three distinct tasks: standard Masked Language Model (MLM), sentiment analysis, and Named Entity Recognition (NER). This was done to showcase the performance of SecureBERT in processing both cybersecurity and general English inputs. The methodology utilizes a tailored tokenizer that preserves conventional English vocabulary while also accommodating novel tokens that are pertinent to cybersecurity. This guarantees that the model can properly process both generic and domain-specific language. SecureBERT utilizes the RoBERTa architecture, which is renowned for its exceptional performance in diverse natural language processing tasks. This makes it well-suited for effectively collecting contextual relationships and semantic meanings in cybersecurity literature. The terminology used in the field of cybersecurity is often complex and intricate. SecureBERT may encounter difficulties with comprehending intricate terminologies or context-specific expressions that are inadequately represented in the training data. A proposed solution utilizes a pre-trained model to anticipate CVSS metric values. To be more precise, the XLNet model, which has been adjusted using a self-created collection of texts, is used to forecast the metric values based on the vulnerability description text. This helps to lessen the workload of the assessment process. XLNet's main component is the permutation language model. It leverages the random arrangement of the initial input order to capture bidirectional contextual information, while still maintaining the autoregressive model's one-way nature. Given a text of length n, there are n! different sorting methods, and the bidirectional contextual information can be indirectly derived by considering the complete ranking order of the text. However, computing all possible orderings would require significant computational resources. Hence, XLNet only predicts a partial sequence. The XLNet model was fine-tuned using a self-built cybersecurity corpus. Subsequently, the fine-tuned XLNet model was employed to extract semantic features from the vulnerability description language. Afterwards, the CVSS metric values were divided, transforming the multi-classification problem into a multi-label classification problem. Subsequently, multi-label classification was conducted using the extracted text features to predict vulnerability metric values. Empirical findings from testing 276,357 real vulnerabilities indicate that XLNet can attain the highest level of performance in predicting CVSS metric values [12]. A word embedding model called CySecBERT, which is based on BERT, was presented for cybersecurity text analysis. In addition to offering a model that is ideal for real-world cybersecurity use cases, the goal was to enable cutting-edge natural language processing (NLP) for the security domain and to establish a strong foundation for future research in this area. CySecBERT is trained only on a cybersecurity corpus, enabling it to comprehend and accurately describe domain-specific language and concepts. Accurate interpretation of terminology is essential, particularly when they may have divergent meanings in various contexts. For instance, the term "virus" can refer to malware instead than a biological virus. The dataset for CySecBERT was generated by carefully selecting and organizing a diverse range of data sources to construct a comprehensive cybersecurity corpus. The dataset comprises many categories of information pertaining to the field of cybersecurity, so ensuring that the model is adequately equipped to comprehend and handle technical terminology [13]. A novel strategy for task-oriented dialogue systems was employed, harnessing the power of linking GPT-4 models together and implementing efficient engineering for each sub-task. The report highlighted the crucial role of the framework in enhancing interaction and communication in cybersecurity.

It concluded that this framework improved the transparency of the system (Explainable AI) and made the decision-making process and response to threats more efficient (Actionable AI). This indicates a significant advancement in cybersecurity communication. Cyber Sentinel leverages GPT-4 as the core element of its system,

allowing it to efficiently perform cybersecurity tasks. The LLM, or Large Language Model, enhances user interactions and optimizes command processing by employing a sequential methodology that involves linking numerous GPT-4 models. Cyber Sentinel is able to process different user demands by providing the models with a sequence of chat messages that have been labeled with specific dialogue roles, such as user, assistance, and system [14]. Several studies have explored the application of advanced NLP models for spam detection. Notably, a recent study compares the performance of two top-tier models, ChatGPT-4 from OpenAI and Google Gemini, in detecting spam using the widely recognized Spam Assassin public mail corpus. The study employs a meticulous technique that encompasses dataset preparation and the utilization of recall, highlighting its potential in scenarios where identifying the greatest number of spam emails is crucial, despite a slightly higher tendency to mistakenly classify valid emails as spam. This comparison underscores the distinct strengths of each model in spam detection tasks, providing valuable insights for selecting appropriate tools based on specific requirements[15].

TABLE I. EVALUATION METRICS OF GPT-4

metric	value
accuracy	0.94
precision	0.93
recall	0.92
F1 score	0.925

TABLE II. VALUATION METRICS OF GOOGLE GEMINI

metric	value
accuracy	0.92
precision	0.89
recall	0.95
f1 score	0.92

Table III provides a comprehensive overview of the Natural Language Processing (NLP) models that have been utilized in the cybersecurity field. The integration of Large Language Models (LLMs) into cybersecurity applications has seen remarkable growth in recent years, primarily due to their superior capabilities in language understanding, contextual learning, and dynamic reasoning. These models excel at interpreting unstructured textual data, which is prevalent in cybersecurity sources such as logs, reports, alerts, and threat intelligence feeds. Among the most widely adopted models are BERT, RoBERTa, GPT-3/4, LLaMA, and T5, all of which have been applied to a range of critical tasks. These include threat intelligence extraction, automated log and alert analysis, anomaly and intrusion detection, vulnerability classification, and automated incident response. Their ability to learn from diverse textual contexts makes them valuable assets in building intelligent, adaptive, and scalable security solutions. Table IV further presents a comparative analysis of the key LLMs and their performance across cybersecurity tasks.

TABLE III. VARIETY OF MODEL WITH DATASETS AND THEIR PEPORTED PERFORMANCE METRICS.

models	dataset used	accuracy
BERT for Phishing Detection[16]	Enron Email Dataset, PhishTank Dataset	95% Precision, 90% Recall
GPT-2 for Malware Detection[17]	unlabeled 10 million assembly instructions	85.4%Accuracy
BERT for Threat Intelligence[18]	MITRE ATT&CK Framework, Open Threat Intelligence	92% F1-score
BERT for Intrusion Detection[19]	KDD Cup 1999 Dataset, NSL-KDD Dataset	99% Accuracy
RoBERTa for Vulnerability Detection[20]	CVEFixes	81% F1-score
BERT for Spam Detection[21]	SpamAssassin Public Mail Corpus, Enron Email Dataset	97% Accuracy

TABLE IV. THE LLMS IN CYBERSECURITY

model	Training Dataset	Cybersecurity Applications	Strengths
BERT [28]	BooksCorpus + Wikipedia	Log classification, threat entity recognition	High accuracy in short text, fine-tunable
ROBERTa [29]	Extended corpus (CC-News, OpenWebText)	IOC extraction, phishing detection	More robust than BERT
model	Training Dataset	Cybersecurity Applications	Strengths
BERT [28]	BooksCorpus + Wikipedia	Log classification, threat entity recognition	High accuracy in short text, fine-tunable
ROBERTa [29]	Extended corpus (CC-News, OpenWebText)	IOC extraction, phishing detection	More robust than BERT

VI. ADVANCED PERSISTENT THREAT APT

Despite advancements in workplace security measures such as firewalls and anti-virus software, fraudsters continue to develop increasingly sophisticated attack tools to detect and prevent unauthorized access. Nevertheless, studies conducted by Micro and Report have proven that this presumption is now invalid due to the rise of the advanced persistent threat (APT), which is a targeted and highly sophisticated type of assault. Irrespective of the safeguards in place, the Advanced Persistent Threat (APT) selects a target and persists in its attempts until it successfully bypasses them through various of attacks type summarized in fig 1. APTs have primarily been employed

against vital infrastructure, such power grids. APTs include the 2010 Stuxnet assault and the 2017 Triton attack [22]. Advanced persistent threats (APTs) are a significant cybersecurity issue characterized by highly financed and organized attackers who exhibit long-term persistence, execute their attacks in a secretive and evasive manner, and accurately target valuable assets [23]. Advanced persistent threats (APTs) are a significant cybersecurity issue characterized by highly financed and organized attackers who exhibit long-term persistence, execute their attacks in a secretive and evasive manner, and accurately target valuable assets. APT attacks have the potential to cause terrible harm to a number of industries and sectors, including the government, businesses, key infrastructure, and private persons. APT attacks frequently have espionage, intellectual property theft, disruption of vital information infrastructures, and theft of sensitive data as their goals. APT (such as Stuxnet, Gauss, Flame, and Duqu) accounts for about 60% of cyberattacks against governments, multinational companies, and vital infrastructures during the past two years.
A typical APT attack lifecycle comprises the following steps:

A. Understanding and Equipping Information collection, Surveillance, an alternative term for reconnaissance, is an essential phase in the pre-attack planning process.During this phase, Adversaries identify and scrutinize the targeted organization, gathering comprehensive data about its technological framework and important staff members.These data are often acquired through the use of social engineering techniques and open-source intelligence (OSINT) programs. The practice of manipulating and influencing individuals through psychological tactics and interpersonal skills. Social engineering refers to the use of psychological tactics to manipulate persons in order to attain specific goals, regardless of whether such goals are beneficial to the target or not. It is commonly used in cyberattacks to pilfer confidential information or manipulate the target into performing a certain action, such as executing malware.OSINT, short for Open Source Intelligence, refers to the collection of intelligence information from publicly available sources.Currently, it usually refers to gathering data on a topic from either free or paid online sources. OSINT enables the collection of diverse data, ranging from an employee's personal profile to the hardware and software configurations of a corporation. In addition to extracting information from the internet, attackers may utilize data mining techniques and big data analytics to automatically analyze the collected data, with the aim of generating useful insight that can be acted upon. Using the collected intelligence, advanced persistent threat (APT) actors formulate an offensive strategy and assemble the required tools. Attackers commonly equip themselves with a range of equipment to target different attack vectors, allowing them to adjust their strategies in the event of a setback [24].

B. Delivery: During this stage, attackers distribute their exploits to the specific individuals they are targeting. There are two distinct methods of delivery: direct shipping and indirect shipment. The assailants utilize diverse social engineering techniques, including as spear phishing, to directly transmit exploits to their targets for delivery. Indirect delivery is covert. In this technique, the assailants will seize control of a trusted third-party entity that is relied upon by the intended victim. They will then utilize this compromised third party to indirectly distribute and execute malicious vulnerabilities. A trustworthy third party refers to either a provider of software/hardware utilized by the specific enterprise being targeted, or a reputable website that is regularly accessed by the individuals being targeted (known as a watering hole attack). Spear phishing refers to a targeted form of cyberattack where individuals or organizations are deceived into revealing sensitive information or doing certain actions using personalized and deceptive communication. In spear phishing, individuals or organizations are specifically targeted in an effort to trick them into divulging critical information or performing certain actions through the use of tailored and misleading communication. To maximize the likelihood of success, it frequently tailors the attack to the particular target using information obtained during reconnaissance. The recipient is enticed to click on a link that leads to a fraudulent website that offers drive-by-download vulnerabilities, or to download an apparently innocent attachment that contains a vulnerability exploit. APT attacks often utilize malicious attachments instead of malicious URLs, as it is typical for files (such as reports, business documents, and resumes) to be shared by email in corporate or government environments [24].

C. Initial Intrusion Unauthorized entry or intrusion Initial intrusion: refers to the unauthorized entry obtained by a perpetrator into the computer or network of the intended victim. Social engineering can be used by attackers to get credentials, which they can subsequently use to gain access that appears legitimate. Nevertheless, the prevailing approach for intrusion involves the execution of malicious code that takes advantage of a vulnerability in the computer of the target. The attackers initially deploy malicious code during the delivery phase, then subsequently acquire unauthorized access to the target's computer during the intrusion phase, once the exploit is successfully performed. The initial intrusion is a crucial stage of an Advanced Persistent Threat (APT) attack, as it allows the APT actors to gain a strong position in the target's network [24].

D. After establishing a covert access point, Advanced Persistent Threat (APT) actors implement Command and Control (C2) mechanisms to assume remote control over compromised systems, thereby enabling deeper exploitation of the target network. To evade detection, these attackers increasingly rely on publicly available tools and well-known online services. Notably, they utilize hidden services such as social media platforms and servers configured to accept connections exclusively through the Tor network (The Onion Router). Hosting C2 infrastructure on Tor significantly complicates efforts to identify, block, or neutralize these servers due to Tor’s anonymity features. Furthermore, attackers frequently deploy Remote Access Tools (RATs) to maintain persistent control and execute

malicious activities without alerting traditional cybersecurity defenses.

E. Lateral Movement Once communication has been established between the hacked systems and the command and control (C2) servers, threat actors infiltrate the network. in order to gain more authority over the intended target, which allows them to find and gather important data. The following tasks are typically involved in lateral movement: (1) mapping the network internally to gather intelligence; (2) breaking into other systems to obtain credentials and escalated privileges; and (3) locating and gathering valuable digital assets, like trade secrets and development plans. Because the attackers seek to gather as much information as possible over an extended length of time, this stage usually lasts a long time. To avoid discovery, the activities are made to run slowly and quietly.

F. Data exfiltration refers to the unauthorized transfer or extraction of data from a computer or network. The main objective of an Advanced Persistent Threat (APT) attack (shown in Fig.1) is to illicitly acquire confidential information for the purpose of obtaining strategic advantages. Consequently, For the attackers, the data extraction process also referred to as data exfiltration is an essential step. Before the data is transferred to other locations under the attackers' control, it is typically routed to an internal staging server where it is compressed and frequently encrypted. to hide the information transmission method.

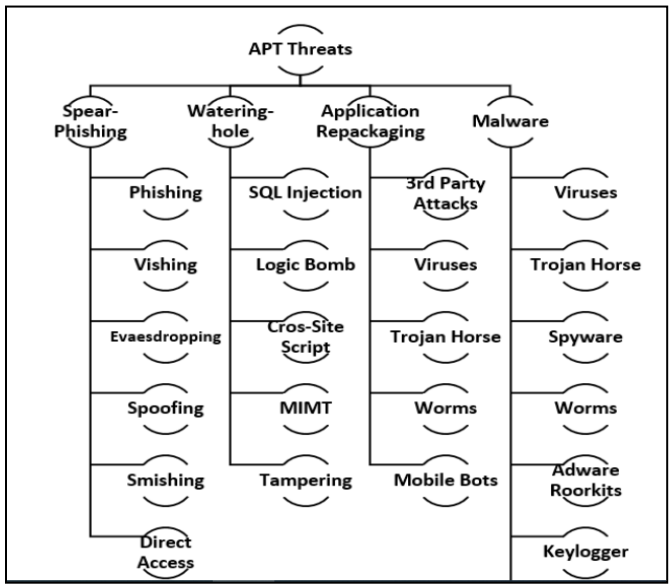

Fig. 1. APTs and their attacks

RESEARCH GAPS AND OUTLOOK

The field of NLP in the realm of cybersecurity is now encountering many obstacles and prospects for progress. Developing comprehensive datasets that encompass a wide range of cybersecurity textual materials is a vital component. These datasets are essential for a thorough evaluation and comparison of NLP approaches. The task at hand involves carefully selecting and organizing datasets that encompass a wide range of linguistic styles and settings seen in cybersecurity writings [25]. Efforts have been made to adapt LLMs (e.g., LLaMA [26]) to the cybersecurity domain. The creation of cybersecurity-enhanced domain LLMs such as RepairLlama has facilitated the resolution of distinct cybersecurity obstacles. However, advanced LLMs such as Chat GPT are able to tackle intricate tasks by utilizing prompt engineering, in-context learning, and chains-of-thought, even though they do not possess specific expertise in cybersecurity. The initial endeavors demonstrate that LLMs assist in cybersecurity duties with encouraging outcomes. Although LLMs have made early attempts in the topic of cybersecurity, there are various hurdles that the industry still encounters. Firstly, numerous studies depend on case studies that lack a rigorous methodology, which gives rise to concerns regarding scalability and reproducibility. Furthermore, the study environment seems to be divided and lacks cohesion, with a lack of interconnectedness across studies and thorough analysis. Given the significant growth in LLM research, it is essential to perform a systematic overview to guide the discipline towards a new phase of development. This phase should focus on making LLM application strategically impactful rather than experimental [27].

CONCLUSION

This study highlights the significant potential of Large Language Models (LLMs) in the domain of cybersecurity, particularly for tasks involving threat detection, incident response, and security log analysis. Through a comprehensive review, it becomes evident that LLMs offer a transformative approach to cybersecurity by enabling the automated processing of vast and complex datasets, identifying patterns indicative of malicious activities, and supporting decision-making processes through natural language interaction. LLMs demonstrate the capability to enhance situational awareness, streamline threat analysis, and provide timely insights in a field where rapid response is critical. Their integration into cybersecurity frameworks such as automated threat intelligence platforms, alert triage systems, and conversational interfaces represents a promising direction for improving the scalability, efficiency, and adaptability of defense mechanisms. However, while the benefits are substantial, the deployment of LLMs in cybersecurity also raises challenges, including data privacy concerns, the need for domain-specific fine-tuning, model explainability, and resistance to adversarial manipulation. Addressing these limitations is essential for the safe and effective use of LLMs in real-world cybersecurity operations. Overall, this review underscores the growing relevance of LLMs as powerful tools in modern cybersecurity strategies, and it calls for continued research to fully realize their potential in protecting digital infrastructure against increasingly sophisticated cyber threats.

CONFLICT OF INTEREST

One of the authors, Prof. Dr. Ali Ahmed Abed, is the Editor-in-Chief of the *International Journal of Mechatronics, Robotics, and Artificial Intelligence (IJMRAI)*. To ensure transparency and avoid any potential

conflict of interest, he was not involved in the editorial processing or peer review of this manuscript. The review process was handled independently by other qualified editors. The authors declare no other conflicts of interest related to this work.

REFERENCES

[1] L. Zheng, W. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, "Judging llm-as-a-judge with mt-bench and chatbot arena," *Advances in Neural Information Processing Systems*, *36*, pp. 46595-46623, 2023.

[2] S. Jagtap, V. Kavitkar, M. Jain, A. Ingle, J. P. Tamkhade and K. Kshirsagar, "Enhancing Cyber Security Against DDoS Attacks: A Comprehensive Review and Future Directions," *2025 1st International Conference on AIML-Applications for Engineering & Technology (ICAET)*, Pune, India, pp. 1-6, 2025, https://doi.org/10.1109/ICAET63349.2025.10932273

[3] S. Mannarswamy, & S. Roy, "Evolving AI from Research to Real Life-Some Challenges and Suggestions," *In IJCAI*, pp. 5172-5179, 2018, https://doi.org/10.24963/ijcai.2018/717

[4] N. Sabharwal and A. Agrawal, "Introduction to Natural Language Processing," *Hands-on Question Answering Systems with BERT: Applications in Neural Networks and Natural Language Processing*, pp. 1-14, 2021, https://doi.org/10.1007/978-1-4842-6664-9_1

[5] R. A. Chetwyn and L. Erdődi, "Towards Dynamic Capture-The-Flag Training Environments For Reinforcement Learning Offensive Security Agents," *2022 IEEE International Conference on Big Data (Big Data)*, Osaka, Japan, pp. 2585-2594, 2022, https://doi.org/10.1109/BigData55660.2022.10020389

[6] G. De Vito, F. Palomba and F. Ferrucci," The role of Large Language Models in addressing IoT challenges: A systematic literature review," *Future Generation Computer Systems*, vol. 171, 2025, https://doi.org/10.1016/j.future.2025.107829

[7] A. Kumar, M. Mohammed, D. Camacho and J. H. Park," A comprehensive survey on large language models for multimedia data security: challenges and solutions," *Computer Networks*, vol. 267, 2025, https://doi.org/10.1016/j.comnet.2025.111379

[8] S. Qiu, "Multi-modal Remote Sensing Visual Question Answering Algorithm Based on Large Language Model," *2024 5th International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE)*, Wenzhou, China, pp. 20-23, 2024, https://doi.org/10.1109/ICBASE63199.2024.10762433

[9] O. Perera and J. Grob, "Generative AI in Phishing Detection: Insights and Research Opportunities," *2024 Cyber Awareness and Research Symposium (CARS)*, Grand Forks, ND, USA, pp. 1-5, 2024, https://doi.org/10.1109/CARS61786.2024.10778758

[10] A. Kimia, M. Hempel, H. Sharif, J. L. 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, https://doi.org/10.3390/jcp1040031

[11] 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,* pp. 39-56, Cham: Springer Nature Switzerland, 2022, https://doi.org/10.1007/978-3-031-25538-0_3

[12] F. Shi, S. Kai, J. Zheng, and Y. Zhong, "XL Net-based prediction model for CVSS metric values," *Applied Sciences*, vol. 12, no. 18, 2022, https://doi.org/10.3390/app12188983

[13] 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, https://doi.org/10.1145/3652594

[14] J. Zhang, H. Bu, H. Wen, Y. Liu, H. Fei, R. Xi, L. Li, Y. Yang, H. Zhu and D. Meng "When LLMs meet cybersecurity: a systematic literature review," *Cybersecurity*, vol. 8, no. 55, 2025, https://doi.org/10.1186/s42400-025-00361-w

[15] T. Choudhary, "Political Bias in Large Language Models: A Comparative Analysis of ChatGPT-4, Perplexity, Google Gemini, and Claude," in *IEEE Access*, vol. 13, pp. 11341-11379, 2025, https://doi.org/10.1109/ACCESS.2024.3523764.

[16] S. Atawneh, and H. Aljehani, "Phishing email detection model using deep learning," *Electronics*, vol. 12, no. 20, 2023, https://doi.org/10.3390/electronics12204261

[17] D. Demırcı, and C. Acarturk, "Static malware detection using stacked BiLSTM and GPT-2," *IEEE Access*, vol. 10, pp. 58488-58502, 2022, https://doi.org/10.1109/ACCESS.2022.3179384

[18] R. Kaur, T. Klobučar and D. Gabrijelčič, "Harnessing the power of language models in cybersecurity: A comprehensive review," *International Journal of Information Management Data Insights*, vol. 5, no. 1, 2025, https://doi.org/10.1016/j.jjimei.2024.100315.

[19] Z. Ali, W. Tiberti, A. Marotta and D. Cassioli, "Empowering Network Security: BERT Transformer Learning Approach and MLP for Intrusion Detection in Imbalanced Network Traffic," *in IEEE Access*, vol. 12, pp. 137618-137633, 2024, https://doi.org/10.1109/ACCESS.2024.3465045

[20] T. Heričko and B. Šumak, "Commit Classification Into Software Maintenance Activities: A Systematic Literature Review," *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, Torino, Italy, pp. 1646-1651, 2023, https://doi.org/10.1109/COMPSAC57700.2023.00254

[21] I. Shen, Y. Wang, Z. Li and W. Ma, "SMS Spam Detection Using BERT and Multi-Graph Convolutional Networks," *International Journal of Intelligent Networks, vol. 6,* 2025, https://doi.org/10.1016/j.ijin.2025.06.002

[22] S. Krishnapriya, and S. Singh, "A Comprehensive Survey on Advanced Persistent Threat (APT) Detection Techniques," *Computers, Materials & Continua*, vol. 80, no.2, 2024, https://doi.org/10.32604/cmc.2024.052447

[23] H. Liu, B. An, Y. Yin, X. Huo, Z. Su, and Y. Wang, "A Trust Evaluation and Concept Drift-Based Approach for Dynamic APT Evasion Detection," *in 2024 IEEE Cyber Science and Technology Congress (CyberSciTech)*, pp. 544-547, IEEE, 2024, https://doi.org/10.1109/CyberSciTech64112.2024.00097

[24] A. Alshamrani, S. Myneni, A. Chowdhary, and D. Huang, "A survey on advanced persistent threats: Techniques, solutions, challenges, and research opportunities", *IEEE Communications Surveys & Tutorials*, vol. 21, no. 2, pp. 1851-1877, 2019, https://doi.org/10.1109/COMST.2019.2891891

[25] K. Koshekov, B. Bakirov, A. Sakhov, L. Nataliia, Y. Tanovitskiy, A. Koshekov, Y. Kurbanov, and R. Togambayev, "Cyber hygiene of the digital twin of the civil aviation occupational safety management system in the context of quantum transformation," *Radio electronic and Computer Systems*, vol. 2025, no. 1, pp. 231-247, 2025, https://doi.org/10.32620/reks.2025.1.16

[26] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," arXiv preprint arXiv:2302.13971, https://doi.org/10.48550/arXiv.2302.13971

[27] M.A.I. Mallick, and R. Nath, "Navigating the cyber security landscape: A comprehensive review of cyber-attacks, emerging trends, and recent developments," *World Scientific News*, vol. 190, no. 1, pp. 1-69, 2024.

[28] H. Guo, S. Yuan, and X. Wu, "Logbert: Log anomaly detection via BERT," *in 2021 international joint conference on neural networks (IJCNN)*, pp. 1-8. IEEE, 2021, https://doi.org/10.1109/IJCNN52387.2021.9534113

[29] M. Songailaitė, E. Kankevičiūtė, B. Zhyhun, and J. Mandravickaitė, "BERT-based models for phishing detection," *in 28th Conference on Information Society and University Studies (IVUS'2023). CEUR Workshop Proceedings,* Kaunas, Lithuania, 2023.

[30] P. Balasubramanian, J. Seby, and P. Kostakos, "Cygent: A cybersecurity conversational agent with log summarization powered by GPT-3," *in 2024 3rd International Conference on Artificial Intelligence for Internet of Things (AIIoT)*, pp. 1-6, IEEE, 2024, https://doi.org/10.1109/AIIoT58432.2024.10574658

[31] Z. Yang, and I. G. Harris,"LogLLaMA: Transformer-based log anomaly detection with LLaMA," *arXiv preprint arXiv:2503.14849*.

[32] X. Huang, K. Xue, L. Chen, J. Han, J. Li and D. S. L. Wei, "ForenSiX: Automated Network Forensics and Diagnostics for Beyond-5G and 6G Networks Using Large Language Models," in IEEE Network, vol. 39, no. 5, pp. 74-80, 2025, https://doi.org/10.1109/MNET.2025.3579925.

[33] R. Pavlich, N. Ebadi, R. Tarbell, B. Linares, A. Tan, R. Humphreys, J. Das, R. Ghandiparsi, H. Haley, J. George, R. Slavin, K. Choo, G. Dietrich and A. Rios, A. (2024). "Beyond Text-to-SQL for IoT Defense: A Comprehensive Framework for Querying and Classifying IoT Threats," *arXiv preprint arXiv:2406.17574*, https://doi.org/10.18653/v1/2025.trustnlp-main.1