World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(REVIEW ARTICLE)

Check for updates

Automated threat detection and response using LLM agents

Ramasankar Molleti [1,*], Vinod Goje [2], Puneet Luthra [3] and Prathap Raghavan [4]

[1] *Computer Science and Engineering, Jawaharlal Nehru Technological University, Hyderabad, India.*
[2] *Computer Science and Engineering, University College of Engineering Osmania University, Hyderabad, India.*
[3] *Computer Science and Engineering, Giani Zial Singh College of Engineering and Technology, Bathinda, Punjab, India.*
[4] *Computer science, Sree Sastha institute of engineering and technology, Madras university, Chennai, Tamil Nadu, India.*

World Journal of Advanced Research and Reviews, 2024, 24(02), 079–090

Publication history: Received on 22 September 2024; revised on 28 October 2024; accepted on 31 October 2024

Article DOI: https://doi.org/10.30574/wjarr.2024.24.2.3329

Abstract

The increase of cyber threats from individual cases to a worldwide problem is the reason why people have shifted their cybersecurity perspectives. Basic defense processes, originally well understood and effective, fail to match modern attacks' complexity and velocity. Taking into consideration LLMs as a recent addition to AI, this paper aims at discussing their application in integrating threat detection and response automation systems. As a result, LLMs, which have higher capabilities for natural language processing, deliver a revolutionary perspective regarding cybersecurity. Since LLM agents can review massive amounts of security data, distinguish patterns, and create contextually appropriate responses, they can bridge the gap between emerging threats and stable security systems. The paper examines the tools used by LLM agents, such as natural language processing to analyse the logs, contextual anomaly detection, pattern identification in network traffic, and the analysis of the user's behaviour. Also, it describes how LLM agents can support automated threat handling in the context of threat identification, alert prioritization, context-driven response generation, security policy enforcement, and threat handling. The integration of LLM agents into already known systems, including SIEM systems and AI-Ops platforms, is also considered, which allows for further conclusions on the opportunities to create proactive cybersecurity systems. However, open dilemmas such as adversarial attacks and interpretability are still present, the future for LLM agents in cybersecurity is still bright, and there are more possibilities in multi-modal threat analysis and quantum-safe LLM-based cryptography.

Keywords: LLM Agents; Automated Threat Detection; Cybersecurity; AI-driven Response; Contextual Analysis; Adaptive Security

1. Introduction

In today's digital environment, cyber threats have transitioned from being a rare oddity to becoming a universal concern for organizations and people. It covers all forms of threats that can be grouped, from state sponsored complex cyberattacks to greedy and unscrupulous cybercrime attacks [1]. Examples of risks are viruses, worms, spyware, Trojan horses, viruses, ransomware, hacking, DoS and DDoS attacks, inbound threats such as phishing, and others [2, see Figure 2]. The ongoing and accelerated shift to digitalization in business and the use of cloud services have widened the attacks' window, prompting traditional safeguards to become insufficient. The daily reports of cyber threats depict that the frequency, speed, and type of threats have overwhelmed the ability of personnel or software to monitor and contain them. Cyber security personnel encounter a number of disguised alerts, most of which bear no real threat, which results in alert fatigue and possibly helps them miss crucial threats [3]. The threat analysis and response done manually take a long time, are inclined to mistakes, and cannot cope with the new methods used by cybercriminals. This has brought about a gap in threats that is very much more sophisticated than the defenses put in place in organizations; thus, there is a need for automated systems that analyze threats, defend against them, and all this in real time and at machine speed

* Corresponding author: Ramasankar Molleti

to be in harmony with the nature of present day threats and attacks. Large Language Models (LLMs) are the next step in AI, mostly within the area of natural language processing. These models enable such tasks depending on the textual data upon which they have been trained; the models can comprehend, produce, and transform human-like text efficiently [4,5]. Modern LLMs, like the GPT (Generative Pre-trained Transformer) series, have proven to exceed the capability of just generating text and are capable of reasoning, problem solving and task completion in a specific area [4]. It also explains why they are appropriate for the level of analysis since they are capable of parsing context and inferring relationships, as well as churning out coherent responses. Leveraging on these ideas, this paper submits that LLM agents are revolutionary in the paradigm of agents that automate the identification and handling of threats. With enhanced natural language processing and generation capabilities, LLM agents are capable of processing large volumes of security data, detecting the tendency of threats, and responding with pertinent messages. Such agents might be able to close the gap that exists between the current advanced attacks and the otherwise rigid method of rule-based systems of security. Recent incorporation of LLM agents into the frameworks for cybersecurity seeks to deliver improved accuracy of threat identification, a shortened response time, and better means of coping with the ever-evolving threats on the cyber-front.

2. Literature review

2.1. Evolution of Threat Detection and Response Systems

Cybersecurity as a field has evolved a lot since the time it actually came into existence. Originally, threat detection was mainly conducted based on a set of features that characterized threats [6]. Such systems were highly efficient against known threats but could be beaten rather easily by new kinds of attacks. In the early nineties, stateful inspection firewalls and intrusion detection systems appeared, the latter being the programs that analyzed the traffic on the network for signs of misuse [7]. However, those systems were not able to adapt to unknown threats and raised many false alarms. Intrusion detection systems were introduced in the late 1990s, but it was in the early 2000s that intrusion prevention systems that could actively prevent discovered intrusions were developed [8].

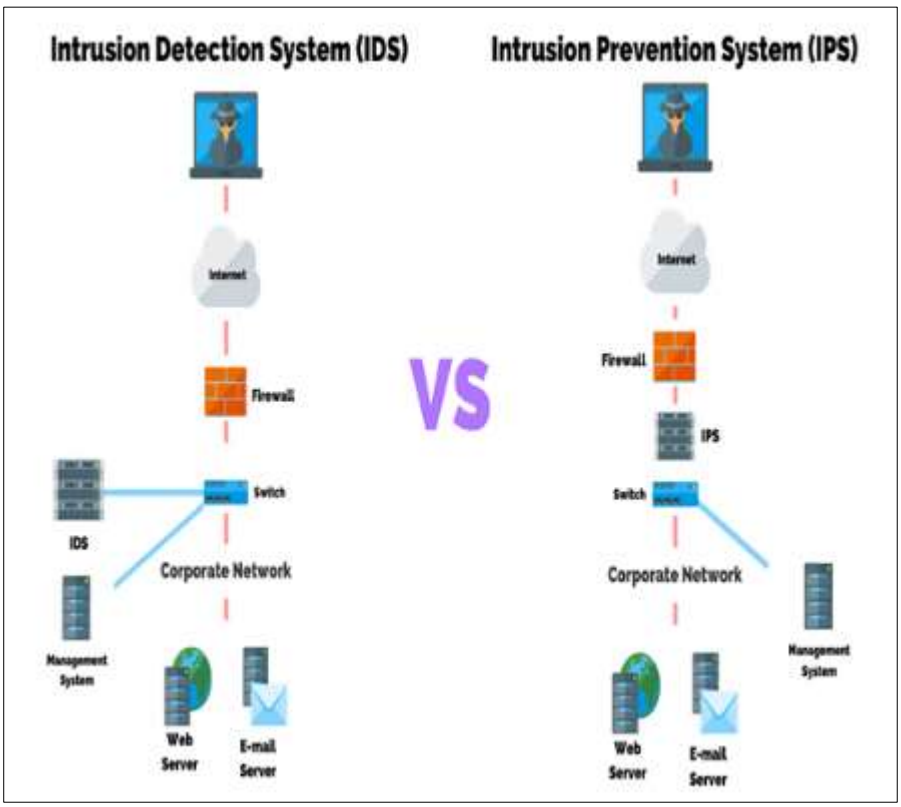

Figure 1 IDS vs IPS [9]

As shown in Figure 1., an intrusion detection system (IDS) monitors a network and delivers alerts when it detects unusual activity on the system or network. This is the primary distinction between an IDS and an IPS. In order to stop assaults from accessing targeted systems and networks, IPS responds to attacks as they happen. Although both IDS and

IPS are capable of detecting attacks, how they react to one is where they diverge most [9]. It is noteworthy that identical monitoring and detection techniques can be employed by both IDS and IPS.

Following the introduction of IPS and IDS, there was the emergence of SIEM systems that analyzed data from several sources and correlated the various inputs to give a broader picture of an organization's security needs. Still, the first-generation AI protective systems brought significant benefits, yet their main drawback was their strictly reactive approach, making an organization a target for zero-day attacks and advanced persistent threats (APTs) [10]. Sophisticated product integration more recently has led to development of Extended Detection and Response (XDR), which is a security incident detection and response system that consolidates multiple security products. Hence, this approach is meant to yield

2.2. Traditional Approaches to Cybersecurity

Conventional approaches to the protection of information systems have consisted of the "castle and moat" strategy. This comprises firewalls, anti-virus and anti-malware software, and network sub-divisions [11, 12]. While these are effective components of the old security framework, they have not been useful against current complex threats. Two more pillars of conventional information protection are worth mentioning; they are the policy of least privilege, which restricts users' access rights at the level of the minimum required for their daily activities. While this model assists with regulating the extent of the aftermath of such accounts' losses, it does not necessarily safeguard against the initial attack. Other preventive measures that have also been put into practice have also included patching and vulnerability management, where ever possible, with the aim of blocking the identified potential security weaknesses before the attackers exploit them [13]. However, the introduction of complex IT systems combined with the continuously evolving volume of vulnerabilities has made extensive patching a problem for many corporate and commercial firms. It has been remarked that user awareness training has been an optimal approach to fighting social engineering. Critically important, it is not very efficient due to the imperfection of human factors and the enhanced complexity of phishing and other social engineering schemes.

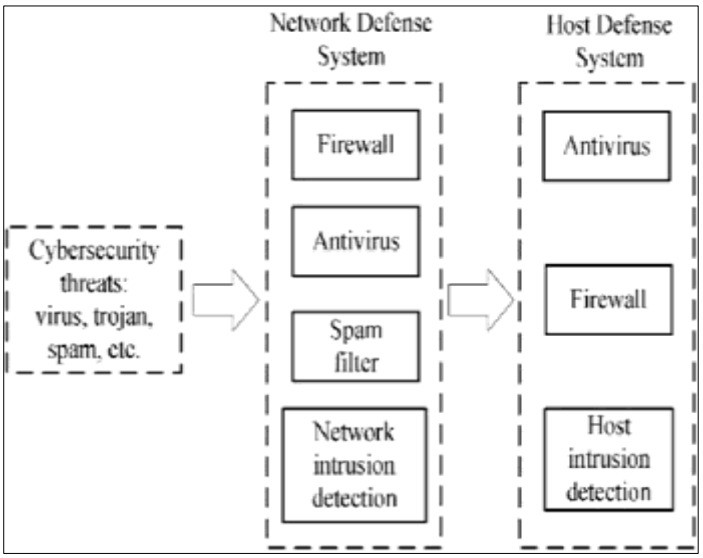

Figure 2 Common Traditional Systems [13]

Cyberdefense systems typically provide network- and host-based defense by fighting cyberthreats at the network and host levels. These two tiers of conventional cybersecurity systems are depicted in Figure 2.

2.3. The Rise of Artificial Intelligence In Cybersecurity

The failure and constraints of conventional security paradigms, the advancing flood of data, and the sophistication of threats, combined with the soaring popularity of AI and ML in the current world, have prompted cyber-security experts to incorporate and adapt AI and ML into their architecture. The above technologies present the prospects of more flexible, variant, and prognostic security mechanisms. AI's initial use case in cybersecurity was mostly the use of statistical modeling and machine learning for the identification of threats that had not been found before [14]. This approach proved to be quite effective, especially in the identification of new threats, but the early designs were crippled by high false positive ratios.

There are some very sophisticated systems now being implemented for threat identification, applying natural language processing for threat security information harvesting from different sources. All these systems enable the prediction of emerging threats and offer background information for security analysts. AI has been involved in performing repetitive safety processes like log management and alarm handling as a way of solving the more established scarcity of cybersecurity specialists [15]. Advanced patching systems provided by AI are present, and they rank vulnerabilities according to their likelihood to be exploited and their consequences.

2.4. Large Language Models: Development And Applications

Large language models are making a leap forward in artificial intelligence. These models, which are based on deep learning architectures, including the Transformer, are trained on gigantic data pieces of text and can thus emulate human writing in various fields. It can be seen that the development of LLMs can be attributed to improvements in neural network structures and the availability of larger computing systems. Some of them include the introduction of transformer models in 2017, which improved the processing of sequential data, and the introduction of pre-training techniques, which enable the generation of models with basic competency in natural language processing before they are trained to solve particular problems [16]. Several LLM capacities have been inculcated the world over, with GPT (Generative Pre-trained Transformer) models by OpenAI leading the way. Every successive release from GPT 1 to GPT 3, and the forthcoming ones, continues to exhibit improved language comprehension and generation. Some of the other well-known LLMs include Google's BERT and T5, and, of course, the recent models are PaLM and LaMDA (see Figure 3). The use of LLMs has been widely expanded in many fields and areas where they were previously not seen. In natural language processing, they have enhanced effects such as machine translation text summarization and issues relating to question answering systems. In software development, LLMs have been applied to code generation and bug detection. They have also exhibited their capability in writing, particularly fiction, content creation and even scientific works, by conducting literature searches and formulating hypotheses.

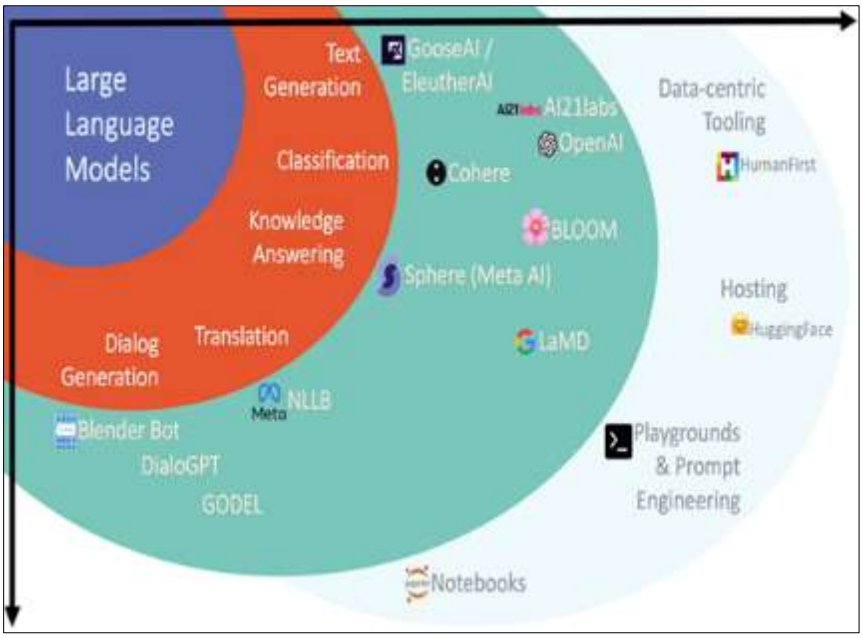

Figure 3 A Representation of Various LLMs [16]

2.5. Current State of LLM Use in Cybersecurity

Thus, cybersecurity is one of the most promising areas for the further development of application of LLMs. It is gradually getting busy with the first utilization of the possibilities provided by these models to analyze numeric and text data of any volume. There is a specific area of application that is promising, and that is threat intelligence analysis. Another is that LLMs are capable of analyzing large amounts of information from various sources, such as security blogs, forums, and social media, to define new threats and attacks' patterns [17]. They enable more specific and efficient threat hunting and better context about the environment in which the network or systems are embedded.

In the case of IR, LLMs are being subscribed to in order to handle system logs and security alerts, which in turn relay contextual information and possible remedial action for security analysts. This application plays a massive role in the faster sorting out of incidents and can even decrease the time to respond to threats. LLMs have also been seen as a

promising solution for creating and evaluating different attack plans [17]. Thus, by knowing specific details of the various attacking methods, these models can be useful for detecting weaknesses in the organization's protection and recommending improvements.

Some are working on LLMs for NL policy definition, which means security policies can be in simple English, and then LLMs convert them into IF-THEN rules for enforcement. Obviously, this approach could increase the role of security policies and make their usage and further maintenance less problematic. But the application of LLMs to deal with cybersecurity issues is not without its challenges. Risks associated with these types of models include their 'black box' nature, emerging discussion on biases possibly introduced in the data sets used while training the models, and threats of adversarial manipulations of the models, which form part of the cybersecurity discourse [17]. However, the opportunities for change in cybersecurity practices through the use of LLMs are great. As these models progress further and as researchers find ways to fix existing problems, LLMs are going to occupy a much more significant place in next generation cybersecurity platforms that are more effective, innovative, and smarter and will therefore be capable of addressing the constantly changing and dynamic nature of today's cyber threats.

3. Methodology

3.1. Automated Threat Detection Using LLM Agents

The use of large language models (LLMs) in security applications has taken a major leap in the advancement of automated threat detection systems. Typically, LLM agents possess higher levels of natural language understanding and contextual analysis to achieve what is a tremendously difficult task of finding possible threats in large data sets [18]. This section examines the main strategies used in the use of LLM agents to improve automated threat identification.

3.1.1. Natural Language Processing for log analysis

The capability of LLM agents to deal with text data makes them even more appropriate when it comes to log analysis. Most previous work on log analysis involved search algorithms such as keyword matching/common specific pattern matching to detect malicious activity. As for the compared LLM agents, the latter has an opportunity to reveal the context and semiotic appraisal of the log entries, thus implying better performance in sophisticated threat detection [18]. Such agents can be trained on huge amounts of log data and can be designed to learn not only specific signs and language expressions that might imply certain threats, but also other characteristics and patterns that can point to a security problem. For instance, an LLM agent could recognize that a particular event log in a system was an attempt at data exfiltration, although no individual event log stated the same.

3.1.2. Anomaly detection using contextual understanding

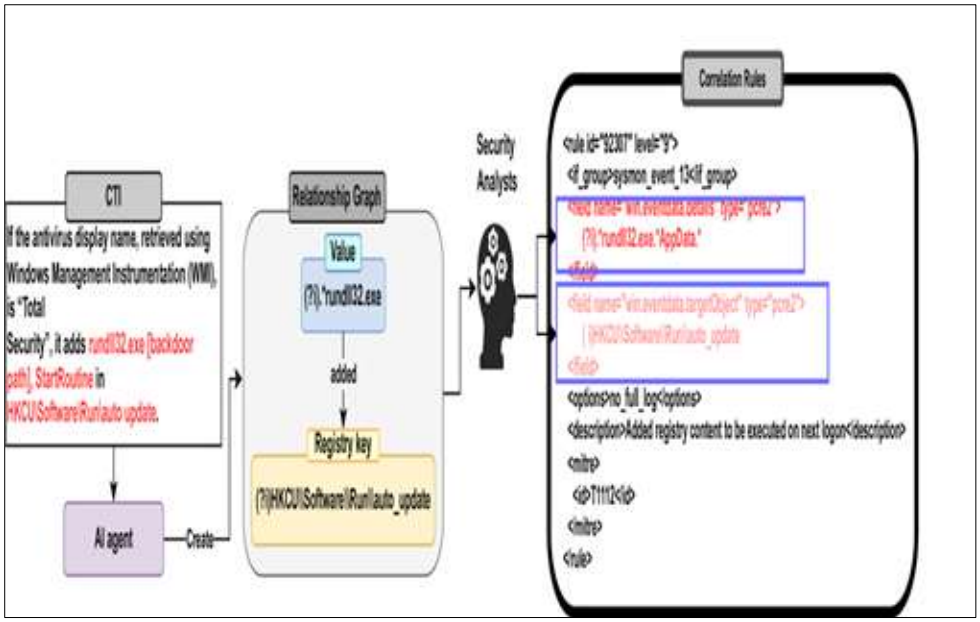

Figure 4 Example how a LLM can be used in Repetitive Anomaly Detection

Another reason that LLM agents are more effective is that they have context as well as anomaly detection. Unlike most statistical measures that tend to work on first level data, low level machine learning agents are better placed to identify an activity as anomalous while taking into consideration many factors that define that activity in its context. For instance, an LLM agent could identify the work habits of the user in the sphere of their activity, daily working time, current projects, or even world events that may influence patterns of work. This contextual understanding permits a more proper selection of the behavior that may indeed be malicious while decreasing the number of false alarms.

Figure 4 shows a repetitive task performed by security analysts along with our agent's automated version of the task. The repeated task is as follows, to start. In the SIEM event correlation rule displayed on the right side of the figure, analysts build the two red color fields based on the following paragraph from a publicly available Cyber Threat Intelligence (CTI) report [18].

The following threat pattern is specified by the SIEM rule: Once the Windows computer becomes compromised, a certain malicious registry key (such as "HKCU\Software\Run\auto_update") is created, which causes a specific malicious software (such as "rundll32.exe") to be run when the user logs in again (demonstration from Figure 4). Even though the field values cannot be predicted, this is a repeating operation because neither of the two fields belongs to a new type or name. Second, analysts can complete this tedious effort of digesting the CTI report themselves by collaborating with our AI agent. Alternatively, the suggested relationship graph, which is displayed in the figure's middle, allows them to acquire the two field values and their relationship directly. Keep in mind that the relationship graph represents our AI agent's output.

3.1.3. Pattern recognition in network traffic

Even though LLMs are aimed at natural language processing, they can be utilized in pattern recognition, specifically while diagnosing network traffic. These agents refer to network traffic data in a format that the LLMs can comprehend and subsequently analyze to identify patterns that denote malicious activity. Due to the availability of large amounts of normal and malicious network traffic, LLM agents can be trained to recognize the features of the various types of attacks that may range from simple DDoS to high-end APT [18]. For this reason, they are capable of analyzing long-range dependencies in data, making them ideal for solving problems that may involve several stages of systemic malicious attacks spanning fairly long durations.

3.1.4. User behavior analysis

Through the construction of detailed, contextual models representing normal user behavior, LLM agents can greatly aid in the analysis of user activity. These models are quite complex and are not limited to rule-based systems; they take into account factors such as the role of the user, patterns of their behaviour, the current project, and the organisational structure. By always keeping in mind what is expected from each user, LLM agents are in a better position to identify those users that might be up to no good or whose account has been compromised. For instance, an LLM agent can consider as abnormal a sequence of steps that on their own appear to be benign, but when observed in the complete context of a user’s role and current projects, they are clearly out of the ordinary.

3.1.5. Zero-day threat identification

Probably the most difficult problem for cybersecurity specialists is identifying zero-day threats, i.e., attacks that exploit new, previously unknown vulnerabilities. It should be noted that the approaches based on the signature detection method are intrinsically unsuitable for dealing with such threats. LLM agents can recognise zero-day threats, thus helping to prevent them when their computations of the normal behaviours of the system point to the fact that something new and possibly malicious is afoot [19, 20]. Data on threats that are familiar to the LLM agents includes voluminous data on known threats, normal system operation, and potential attack means. It means that, possessing this broad knowledge base, they will be able to make wise inferences about the supposed new types of threats that they have not worked with previously. For instance, an LLM agent may also discover a new attack by realizing that a certain string of activities does not correspond to any attack signature, yet they are unlike normal operations and are explained by theoretical models of attacks that were used in training the LLM agent.

Figure 5 An Architecture of an LLM Zero-day Threat Identifier (Hierarchical Planning and Task-Specific Agents) [20]

Based on Figure 5, HPTSA (Hierarchical Planner with Task-Specific Agents) is a cybersecurity framework comprising three main components: a hierarchical planner, a team manager and task-specific expert agents. The hierarchical planner describes how he interacts with the environment (for instance, a website) and then creates instructions for the team manager. Such agent instructions are chosen by the team manager after reviewing the available expert agents, and he/she coordinates the flow of information between the different agent runs [20]. Expert agents are implemented and used to target specific tasks such as SQLi or XSS. This architecture is advantageous in planned, precise vulnerability assessment because it incorporates high-level strategizing with focused specialization, which engulfs adaptive and multifaceted probing of web applications and their security.

3.2. Automated Threat Response Using LLM Agents

The described power of LLM agents does not only cover threat detection but also goes into the more essential sphere of automated threat response. As super smart language comprehension and synthesis instruments, LLM agents play a key role in boosting the threat response processes' speed, accuracy, and flexibility.

3.2.1. Incident classification and prioritization

Due to the architecture of LLM agents, it becomes possible to quickly, at the level of seconds, evaluate the nature of the security incidents and classify them based on multiple factors [21]. As opposed to rule-based systems that fail to either identify the specific type of attack or omit some significant factors in an incident, LLM agents employ their understanding of the incident's context to make accurate classifications. These agents can then rank incidents based not only on the actual level of threat that is posed, but also on the business consequences that this threat might have, the general threat environment, and the organization's risk characteristics. This kind of prioritization allows the security teams to face the threats in an ascending order of priority.

3.2.2. Automated alert triage

For security teams, alert triage is a crucial but frequently daunting duty. The management of alerts is an important, but very time-consuming process for security teams. This process can take a very long time and be a nightmare if handled manually, but through an LLM agent, the whole process of analyzing the alerts, correlating them with other parameters, and drawing possible conclusions about each alert's importance can be performed automatically [22]. These agents can extend beyond the process of alert correlation as the abilities of natural language understanding enable them to analyse information from different sources, such as security feeds, relevant blogs or even internal documents. This, in turn, enables the determination of the more comprehensive likelihood and severity of each alert.

3.2.3. Contextual response generation

The most useful aspect of LLM agents in threat response is the generation of proper responses given the situation. These agents can have access to a huge knowledge base describing the best practices in cybersecurity, certain organization's policies, and threats at the moment, which allows them to provide detailed, step-by-step action plans for each type of incident [23]. These response plans can be generalized based on the organization's needs and capacity and situations like available resources, potential business impact and regulatory factors. These plans are comprehensive, easy to understand, and accessible to both technical and non-technical stakeholders because of LLM agents' natural language-generating capabilities.

3.2.4. Adaptive security policy implementation

Incidentally or otherwise, LLM agents can be helpful in the deployment and modification of security policies every time the threats change. These agents can identify new threat intelligence, more recent incidents, and changes that have occurred to the organization's IT environment for new recommendations in security policies. Furthermore, LLM agents can assist in converting an organization's security policies into more precise and executable procedures to be followed in terms of security systems. For instance, an LLM agent could receive a natural language description of a new security policy, such as a firewall rule, and automatically derive a set of IDS signatures and access control lists for the same policy.

3.2.5. Automated threat containment strategies

During the first steps of an attack, isolation is crucial to prevent future consequences. By identifying the kind of threat and system in question, LLM agents are capable of creating containment response plans and can autogenerate containment responses (see Figure 4). Such measures could include shifting the infected systems to isolated networks, deauthorizing the used credentials, or sometimes even blocking specific traffic for a while. An understanding of the context within which LLM agents operate enables better containment actions that are not only security-oriented but also provide solutions that are relevant to the needs of the business.

3.3. Frameworks and Architectures

This article has shown that there are many factors that need to be taken into consideration in order to properly implement LLM agents for automated threat detection and response. These structures have to blend LLM capabilities with other security architectures within an environment to face issues like scaling, privacy, and real-time behavior.

3.3.1. LLM-enhanced SIEM (Security Information and Event Management)

The process of incorporating LLM agents into systems of the SIEM type is a promising way to upgrade their abilities for detecting threats and responding to them. Here, LLM agents serve as enhanced analytics for the basic SIEM features. The SIEM system keeps collecting and standardizing data from multiple sources within the organization (refer to Figure 4). Once the LLM agents receive this data, they study it in greater detail, find more intricate trends, and issue better notifications [18]. This integration enables organizations to build on what they have already invested in a SIEM while at the same time achieving a greatly heightened level of threat identification and management.

3.3.2. AI-Ops integration for cybersecurity

The addition of LLM agents to AI-Ops frameworks implies a comprehensive view of security that is systemically integrated with other IT management activities [24]. In this model, LLM agents coexist with other AI and automation solutions to deliver end-to-end protection and analysis at the security and business levels. This integration makes it easier to identify threats in relation to the rest of the IT scenario as well as to execute security responses in light of the general IT plan.

3.3.3. Multi-agent systems for distributed threat detection and response

In response to the scope and nature of contemporary IT systems, a multi-agent system solution can be used. In this approach, there are many LLM agents in the organization, and each of them addresses particular forms of data or specific aspects of security. It should also be pointed out that such agents are interactive, which means that they exchange information and act in unison. For instance, one agent may be trained to analyze log files, another may be trained to analyze traffic, and a third may work on user behavioral analysis. Thus, through the integration of these specialized agents, coordinated threat detection and response with high accuracy and adequate scalability, even in large environments, can be achieved.

3.3.4. Federated learning approaches for privacy-preserving threat intelligence

Federated learning architectures can be seen as a more viable solution for utilizing the potential of LLM agents by overcoming the problem of data privacy [25]. In this model, LLM agents are learned from each individual organization's data, which means that only the model parameters are being transferred.

This approach is quite useful for organizations as it grants a view to the threat intelligence of the combined industry without risking sensitive information. This is especially useful for industries that deal with sensitive data, like the healthcare or finance industries. Federated learning also helps to build better and more diverse LLM agents since they can be trained on more various threats, and at the same time, the participant organization's data remains secure.

4. Case studies

4.1. Effectiveness Of LLM Agents In Threat Detection

The incorporation of Large Language Model (LLM) agents into the cybersecurity defense has also shown positive impacts in the detection of threats. Such systems have demonstrated high potential for solving some of the persistent issues in the field, including issues with false positives and improving the analysis of related threats and contexts. Another important benefit of LLM agents that can be mentioned is their capability to collect and analyze large quantities of unstructured data from logs, network traffic, and other sources, as well as threat intelligence feeds. Unlike rule-based approaches, LLM agents can understand the natural language and context so that they can detect some weak clues about malicious activities that a simple rule-based system might not see. Thus, enhanced threat detection with the help of LLMs derives from their capability to take into account various contextual parameters when assessing possible threats [26]. For instance, an LLM agent can relate user actions to time of day, the user's position, and recent changes in the system to decide whether an action is malicious or just an exceptional activity that is normal at that specific time or in that certain position. It reduces the false positives by a huge margin, a major issue with other security solutions that made analysts overwhelmed by irrelevant alerts.

In addition, LLM agents perform very well in terms of scalability and real-time execution [27]. Unpredictably, the volume and velocity of data in today's IT infrastructure are increasing, so the agent's ability to promptly analyze the datasets becomes crucial. These systems can run 24/7, and we can update them with emerging patterns and threats, which means that there is a level of alertness that is impossible if analyzed only by human personnel.

4.2. Benefits of LLM agents in threat response

The applicability of LLM agents goes beyond detection to include a timely and effective reaction to threats. There is the possibility to decrease the incident response times dramatically, which is one of the most considerable advantages. Most threat agents are able to scan through threats comprehensively, sort them out in a matter of seconds, and deliver their brief to the security team within the shortest time possible. It would be important to state that using LLM agents to generate responses is noticeably more elaborate and context-sensitive than regular automatic response-providing systems [24]. These AI-driven agents are capable of evaluating different aspects of the threat, such as the threat's type, the systems exposed to the threat, potential business implications, and security policies. This leads to better means or ways of responding to different stimuli, thus minimizing the chances of overreacting or responding with very drastic measures. The last advantage is the experience of constant learning of threats due to the LLM agents that are used. Because these systems meet and study new varieties of attack, they can alter their knowledge base and regulation approaches if needed [24]. This adaptive capability is very vital, especially with today's ever-changing cyber threats, which makes it easier for organizations to address new threats that may be forming.

4.3. Integration With Existing Security Infrastructure

Despite the potential that LLM agents have in cybersecurity, their practical application can be only as good as their interaction with the rest of the security systems. These changes are not without their risks and benefits, and the process of integrating them is as follows:. On the positive side, LLM agents can improve the abilities of the current security instruments without even displacing them. For example, in cooperation with SIEM systems, LLM agents can deliver additional, more detailed analysis of the events and more precise identification of threats, which enriches rather than replaces the main SIEM objectives. However, integration also has technical issues that relate to the integration of systems. Data exchange between LLM agents and other security tools, the interaction of decision-making strategies, and the overall performance of the system are major factors that should be taken into account on purpose.

4.4. Ethical considerations and trust in AI-driven security

The use of artificial intelligence for managing cybersecurity increases numerous ethical considerations. Due to the nature of how AI is trained to make its decisions, there are concerns about possible bias in AI decisions, especially in a case where LLMs are trained on large datasets that may contain societal biases. The completeness and integrity of records, as well as trust, are the other significant matters. Because LLM agents are increasingly becoming responsible for threat identification and counteraction, it is crucial to equip security teams and organizational leaders with confidence in the assistants' decisions and suggestions [28]. This in turn poses the necessity of making AI decision-making more transparent and explainable, although the nature of LLMs might complicate this task. There are also implicit social angles. Advanced technologies such as using powerful AI systems in cybersecurity could have an even worse impact on deepening the digital divide because adopting such powerful systems could be the preserve of organizations that are well endowed with funds to acquire these systems. Sustaining optimum use of AI in the secure,

while upholding the principles of fairness and accessibility, which is likely to remain an issue within the cybersecurity fraternity.

4.5. Challenges and Limitations

While LLM agents in cybersecurity have been described above, there are several challenges and limitations that they have. Another is the issue that adversarial attacks may pose to the LLM agents themselves. As these AI systems continue to be integrated into cybersecurity, they may be locked in and purposefully selected in order to either influence and alter the systems' local decision- making and/or to provide exploitable training data. The last two challenges that can be identified as critical are interpretability and explainability. Because most LLMs are complicated, it can be troublesome for the reader to determine how the disability arrives at certain findings. This "black box" issue is potentially catastrophic in cybersecurity, primarily since the accountability and explainability of decisions are often mandatory for compliance and investigations [29]. Data privacy and regulatory compliance are another area of concern. Neither can LLM agents provide effective service if they lack access to large quantities of data; however, this factor has to take into account legislation on data protection and, of course, users' privacy. Applying LLM- based security systems in environments such as the financial and health sectors may have concealed issues of law and legal system complexities. Also, the demands of the complex LLM agents and the costs incurred when conducting computations might be high [30]. To work effectively, these systems require high-performing hardware and large storage capacity in terms of RAM and disks used, and this consumes a lot of power, which may be a hindrance to the uptake of these systems, especially among small organizations.

5. Conclusion

Large Language Model (LLM) agents' integration into automated threat detection and response systems is considered a major achievement in cybersecurity. This paper has looked at the methods utilized by LLM agents in this area, their efficiency and the difficulties that have been encountered. Such systems using AI are capable of handling complex, contextual and even relational data to an extent that has not been seen before, possibly changing the face of cybersecurity. This means that their capacity to penetrate any amount of data, perceive harasser signs, and produce incongruous responses makes them efficient weapons against emergent cyber threats. But some problems still exist, for instance, there may be adversarial attacks, interpretability, and resources. Prospects that can be expected in the near future include such issues as multi-modal threat detection, quantum-safe LLM-based cryptosystems, and human-IR symbiosis. Given the potential of LLM agents, one can only underscore the need for more research in this field as well as practical application of the discovered knowledge. The cybersecurity world must step forward to engage in these works, navigate these problems, and seek to develop these technologies into sound, moral, and efficient systems. AI can thus become a decisive factor in the development of digital security that would help people resist newfangled cyber threats in the future.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Sophos, Threat Actors Explained: Motivations and Capabilities, *SOPHOS*, 2024. https://www.sophos.com/en-us/cybersecurity-explained/threat-actors

[2] S. M, 10 Types of Cyber Attacks You Should Be Aware in [2021], *Simplilearn.com*, Nov. 11, 2022. https://www.simplilearn.com/tutorials/cyber-security-tutorial/types-of-cyber-attacks

[3] ProofPoint, What Is Alert Fatigue in Cybersecurity? - Notification Fatigue Defined | Proofpoint US, *Proofpoint*, Feb. 03, 2023. https://www.proofpoint.com/us/threat-reference/alert-fatigue#:~:text=Alert%20fatigue%20is%20a%20phenomenon (accessed Jul. 24, 2024).

[4] S. Hore, What are Large Language Models(LLMs)?, *Analytics Vidhya*, Mar. 13, 2023. https://www.analyticsvidhya.com/blog/2023/03/an-introduction-to-large-language-models-llms/#:~:text=NLP%20(Natural%20Language%20Processing)%20is (accessed Jul. 24, 2024).

[5] IBM, What Are Large Language models? | IBM, *www.ibm.com*, 2023. https://www.ibm.com/topics/large-language-models

[6] M. N. Al-Mhiqani *et al.*, A Review of Insider Threat Detection: Classification, Machine Learning Techniques, Datasets, Open Challenges, and Recommendations, *Applied Sciences*, vol. 10, no. 15, p. 5208, Jul. 2020, doi: https://doi.org/10.3390/app10155208.

[7] R. Sheldon, What is stateful inspection in networking?, *SearchNetworking*, Aug. 2021. https://www.techtarget.com/searchnetworking/definition/stateful-inspection

[8] J. Pirc, The Evolution of Intrusion Detection Prevention Then Now and the Future, *Secureworks.com*, Jul. 06, 2017. https://www.secureworks.com/blog/the-evolution-of-intrusion-detection-prevention

[9] M. Swanagan, IDS VS IPS: What's The Main Difference?, *PurpleSec*, Nov. 03, 2019. https://purplesec.us/intrusion-detection-vs-intrusion-prevention-systems/

[10] Ö. Aslan, S. S. Aktuğ, M. Ozkan-Okay, A. A. Yilmaz, and E. Akin, A Comprehensive Review of Cyber Security Vulnerabilities, Threats, Attacks, and Solutions, *Electronics*, vol. 12, no. 6, pp. 1–42, Mar. 2023, doi: https://doi.org/10.3390/electronics12061333.

[11] Checkpoint, What is Cyber Security? The Different Types of Cybersecurity, *Check Point Software*. https://www.checkpoint.com/cyber-hub/cyber-security/what-is-cybersecurity/#:~:text=The%20traditional%20security%20model%20is

[12] M. H. Langevin-Ward, Traditional Approaches to Cyber Security, *WARD IT SECURITY*, Mar. 03, 2019. https://warditsecurity.com/2019/03/traditional-approaches-cyber-security/ (accessed Jul. 24, 2024).

[13] M. A. Teixeira, M. Zolanvari, K. M. Khan, R. Jain, and N. Meskin, Flow-based intrusion detection algorithm for supervisory control and data acquisition systems: A real-time approach, *IET Cyber-Physical Systems: Theory & Applications*, May 2021, doi: https://doi.org/10.1049/cps2.12016.

[14] V. Shutenko, AI in Cyber Security: Top 6 Use Cases - TechMagic, *Blog | TechMagic*, Sep. 13, 2023. https://www.techmagic.co/blog/ai-in-cybersecurity/

[15] R. Kaur, D. Gabrijelčič, and T. Klobučar, Artificial Intelligence for Cybersecurity: Literature Review and Future Research Directions, *Information Fusion*, vol. 97, no. 101804, p. 101804, Apr. 2023, doi: https://doi.org/10.1016/j.inffus.2023.101804.

[16] Attri, Large Language Models: Basics, Working & Examples | Attri AI Blog | Attri.ai Blog, *attri.ai*, 2024. https://attri.ai/blog/introduction-to-large-language-models

[17] H. Xu *et al.*, Large Language Models for Cyber Security: A Systematic Literature Review, *arxiv.org*, 2018. https://arxiv.org/html/2405.04760v1 (accessed Jul. 24, 2024).

[18] P. Tseng, Z. Yeh, X. Dai, and P. Liu, Using LLMs to Automate Threat Intelligence Analysis Workflows in Security Operation Centers, *arxiv.org*, 2024. https://arxiv.org/html/2407.13093v1#:~:text=By%20leveraging%20the%20advanced%20capabilities (accessed Jul. 24, 2024).

[19] D. Kang, LLM Agents can Autonomously Exploit Zero-day Vulnerabilities, *Medium*, Jun. 05, 2024. https://medium.com/@danieldkang/llm-agents-can-autonomously-exploit-zero-day-vulnerabilities-e4664d7c598e (accessed Jul. 24, 2024).

[20] R. Fang, R. Bindu, A. Gupta, Q. Zhan, and D. Kang, Teams of LLM Agents can Exploit Zero-Day Vulnerabilities, *arxiv.org*, 2024. https://arxiv.org/html/2406.01637v1#:~:text=Researchers%20have%20shown%20that%20LLM (accessed Jul. 24, 2024).

[21] 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*, vol. 4, no. 2, p. 100211, Mar. 2024, doi: https://doi.org/10.1016/j.hcc.2024.100211.

[22] Dylan, Utilizing Generative AI and LLMs to Automate Detection Writing, *Medium*, May 10, 2024. https://medium.com/@dylanhwilliams/utilizing-generative-ai-and-llms-to-automate-detection-writing-5e4ea074072e (accessed Jul. 24, 2024).

[23] S. Swamy, N. Tabari, C. Chen, and R. Gangadharaiah, Contextual Dynamic Prompting for Response Generation in Task-oriented Dialog Systems, *ACLWeb*, May 01, 2023. https://aclanthology.org/2023.eacl-main.226/ (accessed Jul. 24, 2024).

[24] Personal LLM Agents: Insights and Survey about the Capability, Efficiency and Security, *arxiv.org*. https://arxiv.org/html/2401.05459v1

[25] A. Takyar, Federated learning: Unlocking the potential of secure, distributed AI, *LeewayHertz - AI Development Company*, Jan. 17, 2024. https://www.leewayhertz.com/federated-learning/#:~:text=Federated%20learning%20enables%20collaborative%20model

[26] S. M. Taghavi and F. Feyzi, Using Large Language Models to Better Detect and Handle Software Vulnerabilities and Cyber Security Threats, *Research Square*, May 21, 2024. https://www.researchsquare.com/article/rs-4387414/latest (accessed May 23, 2024).

[27] LLM Agents – Nextra, *www.promptingguide.ai*. https://www.promptingguide.ai/research/llm-agents (accessed Apr. 22, 2024).

[28] D. Schwartzer, Incorporating LLM and AI in Identity Security: Key Insights for Tech Leaders, *Medium*, Jun. 25, 2024. https://medium.com/cyberark-engineering/incorporating-llm-and-ai-in-identity-security-key-insights-for-tech-leaders-987ca2c72b3f (accessed Jul. 24, 2024).

[29] V. du Preez *et al.*, From bias to black boxes: understanding and managing the risks of AI – an actuarial perspective, *British Actuarial Journal*, vol. 29, p. e6, Jan. 2024, doi: https://doi.org/10.1017/S1357321724000060.

[30] M. Tatarek, Costs and benefits of your own LLM, *Medium*, Aug. 02, 2023. https://medium.com/@maciej.tatarek93/costs-and-benefits-of-your-own-llm-79f58c0eb47f