








Are GPT-Powered AI Systems Superior to Traditional Cybersecurity Tools: Applications and Challenges

Abdelzahir Abdelmaboud¹, Sayeed Salih^{2*}, Aisha H. A. Hashim², Refan Mohamed Almohamedh³,
Hayfaa Tajelsier⁴, Abdelwahed Motwakel³

¹ Humanities Research Center, Sultan Qaboos University, Muscat 123, Oman

² Department of Electrical and Computer Engineering, Faculty of Engineering, International Islamic University Malaysia, Kuala Lumpur 53100, Malaysia

³ Department of Management Information Systems, College of Business Administration in Hawtat Bani Tamim, Prince Sattam Bin Abdulaziz University, Al-Kharj 11942, Saudi Arabia

⁴ Unit of Common Preparatory Year, Prince Sattam Ibn Abdel Aziz University, Al-Kharj 11942, Saudi Arabia

Corresponding Author Email: salih.sayd@gmail.com

Copyright: ©2025 The authors. This article is published by IETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/ijssse.150912>

ABSTRACT

Received: 13 August 2025

Revised: 11 September 2025

Accepted: 21 September 2025

Available online: 30 September 2025

Keywords:

generative AI, cybersecurity, natural language processing, threat detection, risk management, phishing prevention, data privacy

Generative Pre-trained Transformer (GPT) models are revolutionizing cybersecurity by enhancing threat detection, risk evaluation, phishing defense, and automatic vulnerability analysis. This study delves into the various applications of GPT Technologies in security operations, emphasizing their competence in processing security information of large volume, anomaly detections, and providing real-time insights. Case studies cite quantifiable benefits: Anomaly detection by AI reached a high of 80% accuracy, malware and phishing classification 75–95% accuracy, and Microsoft Copilot reduced phishing attacks by 45% in commercial settings. VirusTotal and Cylance AI improved malware categorization accuracy by 38%, reducing false positives by 35%. Incident response effectiveness was improved by as high as 40% in reported deployments. However, GPT models are also exposed to adversarial exploitation, gaps in explanation, integration issues, and dependence on previous data. This paper lists countermeasures, such as prompt engineering, fine-tuning, domain-specific training, and hybrid AI-human decision systems. Findings further highlight the significance of continuous updates, interdisciplinary collaboration with adherence to ethical frameworks to reap the full benefits of GPT-powered cybersecurity. So, take into consideration integrating these models into present security ecosystems. This way, organizations may strengthen their defenses, improve risk management, and make resilience against cyber threats.

1. INTRODUCTION

Cybersecurity threats are advancing unprecedentedly, posing significant risks to individuals, organisations, and political bodies. Traditional security solutions, such as rule-based threat detection and signature-based malware defence, are insufficient to address the increasing complexity of cyber-attacks. Criminals are utilising advanced techniques, including artificial intelligence (AI) and automation, to bypass existing security systems [1]. This situation has created an urgent need for more flexible and sophisticated cybersecurity measures. Generative Pre-trained Transformer (GPT) models, a subset of generative artificial intelligence, have emerged as powerful tools for augmenting cybersecurity protocols. These models can analyse large datasets, identify patterns, and predict possible risks before they arise. GPTs have been utilised in various security domains, encompassing threat detection, phishing prevention, automated security upgrades, and vulnerability evaluation. Their expertise in handling unstructured data, generating human-like prose, and providing

immediate insights makes them essential in modern cybersecurity systems [2].

Cybersecurity has become one of the megachallenges of the digital age, where organizations face ever-growing sophisticated threats against data integrity, privacy, and system resilience. For decades, traditional tools-the firewalls, signature-based antivirus software, and intrusion detection/prevention systems (IDS/IPS) have served the purposes of first-line defense [3]. These methods primarily operate based on known attack signatures, enforced static security rules, or the detection of anomalies against predefined bases. Their structuralized nature has enabled their effectiveness at handling traditional threats and compliance with security policies [4].

These traditional approaches, however, are severely deficient in an age when cyber threats metamorphoses faster than one can blink. Signature-based systems cannot even see zero-day exploits or polymorphic malware that mutate continuously to escape detection. Although heuristic and anomaly detection tools can identify unfamiliar patterns, they

have extremely high false positives to inundate analysts and slow their responses [5, 6]. Moreover, rule-based tools are contextually hollow, leaving them ineffective against complicated attacks such as spear-phishing, insider threats, and multi-vector campaigns, which take advantage of human and system vulnerabilities [7]. Manual updating and human intervention further limit their scalability in contemporary dynamic environments [8].

All these limitations can now be overcome with the recent developments in artificial intelligence, especially by the so-called LLMs such as GPT. GPT-powered systems use natural language processing and deep learning to read through real-time analysis using massive, heterogeneous, and unstructured data sets, such as threat reports, logs, and communications [9]. Unlike other traditional methods, GPT models are not limited to established rules; they adaptively learn so that they can detect novel attack vectors and prognosticate emerging threats [10]. Their unique context-aware insights, automating the synthesis of threat intelligence, and including the provision of natural language explanations to support human analysts really set them apart in the face of rising complexity concerning cybersecurity.

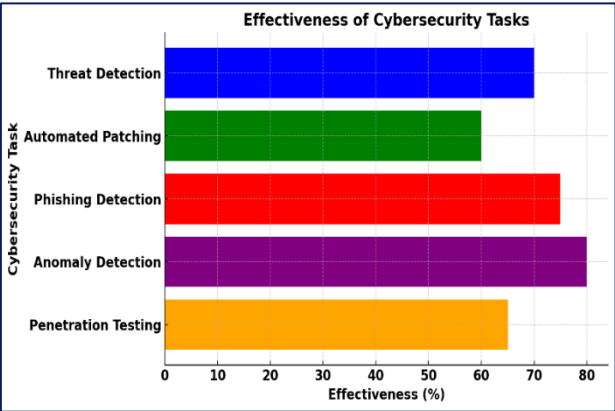


Figure 1. Effectiveness of generative AI in cybersecurity tasks

The author's analysis highlights the effectiveness of generative AI in cybersecurity. Figure 1 demonstrates that AI-enabled anomaly detection attains an 80% success rate, whilst the identification of phishing and malware achieves 75%. These findings highlight the significant potential of generative AI in enhancing proactive security measures by identifying threats before their escalation. These functionalities enable security staff to respond swiftly, reducing cyber-attack consequences and lessening the burden on human analysts. However, despite their potential, GPT models present new obstacles. A major worry is their susceptibility to adversarial attacks, in which malevolent actors alter AI outputs to evade detection or provide deceptive information [11]. Moreover, GPTs can be exploited to create sophisticated phishing schemes, automated malware, and deepfake-facilitated social engineering scams [12]. Their reliance on pre-trained datasets raises concerns about biases, obsolete threat intelligence, and the incidence of false positives in security assessments [13]. The integration of these models into existing cybersecurity frameworks requires careful consideration of scalability, interpretability, and compliance with regulatory norms [13].

This study places AI powered by GPT within the transformative line, not replacing traditional cybersecurity solutions fully. Through the analysis, we shall explore how

GPT models overall outperform the conventional, adaptive, predictive, and even contextual dimensions. Research focuses on whether systems based on GPT can reduce false positives improve detection of zero-day and advanced persistent threats and enhance response to incidents. Filling this gap in research will showcase how GPT models innovate surpassing limitations of current tools and will result in a better innovation of development towards resilient, intelligent, and proactive cybersecurity strategies [2].

This paper investigates the GPT model's contribution to enhancing cybersecurity measures, emphasising its applicability, limitations, and potential remedies. The objective is to tackle the subsequent primary enquiries:

- (1) How are GPT models utilised in cybersecurity, and what are their primary applications?
- (2) What is the role of models like GPT in terms of security for risk assessment and light vulnerability analysis?
- (3) How good are the performance metrics for phishing detection and prevention by the GPT-enabled models and what case studies can draw a conclusion from this?
- (4) What are the limitations and challenges of GPT models in cybersecurity, and what potential solutions can address these issues?

1.1 Motivations

The need for more efficient, intelligent and active security solutions has been used AI and generative models, especially GPT. Figure 2 shows the double role of AI in Cybersecurity, where it is predicted to generate 10% of cyber threats by the end of 2025 while reducing 70% of them. The increasing dependence on AI-controlled security solutions emphasizes the need for continuous progress in the AI security mechanism. The motivation for this study stems from the following key concerns:

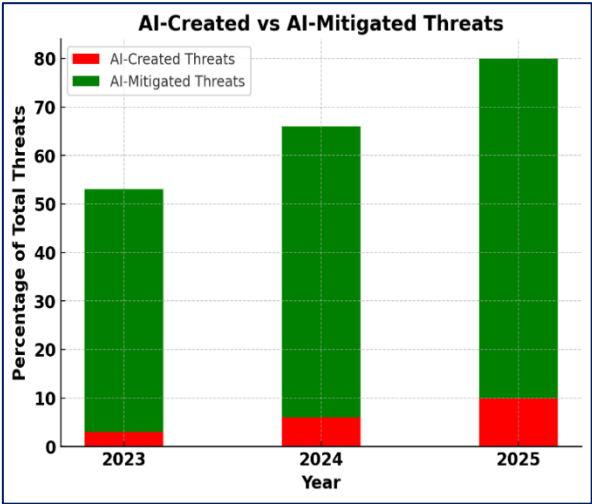


Figure 2. Projected AI-generated vs. AI-mitigated cyber threats (2023-2025)

- (1) Needs to increase cyber threats and require AI-driven security: Traditional security solutions are often reactive rather than proactive, which increases security breaches.
- (2) Challenges in the founding and reaction: Cybersecurity faces increasing challenges in detecting and reducing attacks due to the enormous amounts of security alerts.
- (3) The moral and security risk for AI in cybersecurity:

While AI strengthens security, it also causes risk even when abused. Cybercriminals can utilize GPT models to generate sophisticated fishing E-mail messages, automatically harmful software development and bypass traditional security checks. In addition, the AI-operated security models may be unsafe for unfavourable attacks, data poisoning and biased decisions.

- (4) Cybersecurity frameworks and AI integrations: Companies find combining AI-driven security solutions with traditional systems challenging. Many cybersecurity frameworks were not meant to fit artificial intelligence, which caused interoperability problems.

Cybersecurity innovations and advancing AI: AI models capable of learning and adapting in real time are necessary due to the ongoing evolution of cyber threats.

1.2 Contributions

The present survey thoroughly examines the applications, challenges, and potential solutions of GPT models in the application of threat detection. Among the most significant contributions of this review are:

- (1) Investigates how GPT models employ natural language processing to detect anomalies, phishing attacks, and other malicious activities.
- (2) Measuring the usability of GPT models to detect fraudulent activities and new tactics in phishing.
- (3) Evidence of how GPT models grow when more data is available and generate transitional systems across various cybersecurity domains.
- (4) Few studies manage to show the applied effectiveness of the GPT models, demonstrating their real-world applications in such things as threat detection and subsequently increasing operational efficiencies and resilience against future cyber threats.

2. BACKGROUND

2.1 Summary of existing studies on cybersecurity mechanisms

Existing research on cybersecurity defense mechanisms has mainly revolved around traditional tools such as firewalls, signature-based antivirus systems, and intrusion detection/prevention systems (IDS/IPS) [14]. These studies

show the efficacy of such methods in detecting known threats and ensuring compliance with organizational security policies. Such techniques' simplicity, efficiency for well-documented attacks, and relatively low computational requirements make them ideal for basic security across various infrastructures [15].

However, literature discusses serious drawbacks to these approaches. Signature-based detection cannot recognize zero-day or polymorph attacks; hence, it leaves the organization vulnerable to new threats [16]. The study of anomaly and heuristic detection techniques shows some progress in detecting attacks that are unknown; however, it also underscores hurdles regarding false-positive rates, scalability, and contextual understanding. Furthermore, most of the existing work tends to evaluate tools in isolation, focusing narrowly on achieving detection accuracy and overlooking the broader perspectives of adaptability, automation of response, and integration with human decision-making.

The recent trend of putting artificial intelligence (AI) and machine learning (ML) approaches under the spotlight has appeared in cybersecurity [17, 18]. They reported advancements in malware classification, anomaly detection, and phishing detection but often using supervised models that require well-labeled datasets and cannot generalize well to unseen attack patterns. Few studies considered ever since investigated the large language models (LLMs) such as GPT, which differ fundamentally from classical AI methods by providing contextual reasoning and adaptability capabilities to traverse unstructured data sources, such as logs, reports, and threat intelligence feeds.

The vacuum in current research lies in hardly any comparative analysis being made between GPT-powered AI systems and either traditional or former AI-based approaches. While the older studies tend to emphasize detection rates and algorithmic improvements, they hardly ever refer to larger scales of performance indicators, namely those pertaining to false-positive reductions, zero-day detection, contextual awareness, and full automation of threat response [19, 20]. Therefore, this study will bridge these gaps by systematically comparing GPT models with traditional tools, underscoring their superiority in terms of adaptability, predictive ability, and real-time decision support. Within the extensive field of cybersecurity research, this work places GPT-powered systems, therefore advancing understanding of how generative AI can overhaul defensive methods and surpass the limitations of existing approaches. Table 1 illustrates summary of existing studies on cybersecurity mechanisms.

Table 1. Cybersecurity mechanisms on cybersecurity mechanisms

Mechanisms	Strengths	Weaknesses
Traditional Tools (Firewalls, IDS/IPS, Signature-based AV)	Works well for known threats; Easy implementation, inexpensive; Well-established compliance infrastructures exist.	Cannot detect zero-day or evolving threats; High dependence on manual intervention on updates; Very poor contextual understanding; High false negatives for new attacks.
Machine Learning (Supervised/ Unsupervised Models)	Unendorsed threat detection; Pattern recognition; Adaptability to huge datasets; Automation on classification tasks.	Requirements for fully annotated datasets; Generalization is limited; Vulnerable to adversarial attacks; Claims moderately good false positive rates.
GPT-bade AI Systems (LLMs)	Context-aware threat detection; Adaptive learning; Effective against zero-day and polymorphic attacks; Automated threat intelligence; Reduced false positives.	High computational requirements; Risk of adversarial manipulation; Still in early adoption stage; Requires careful alignment with human oversight.

2.2 Understanding GPT models in cybersecurity

Developed with a remarkable ability to synthesize new information from their existing data, these AI models find practical applications in cyber security, generating human-like content such that advanced pattern recognition and anomaly detection are employed. Thus, these models improve the overall defence of corporations by exposing new potential threats that could be otherwise missed. Since GANs, VAEs, and Transformers are used for various generative artificial intelligence applications in cyber security, they differ in their data-processing approach and modeling paradigm for potential threats. While Transformers are at home with natural language processing tasks, for example, GANs produce realistic samples that help unearth security vulnerabilities. In cyber security, generative artificial intelligence systems have a working role automating responses to cyber-attacks and assisting in incident management. The most important feature of these models is that they allow security teams to prioritize new threats and adapt to new tactics by learning from vast historical datasets, thus forming the backbone of modern cyber security strategies. The applications of GPT in Cyber Security have been summarized in Figure 3.

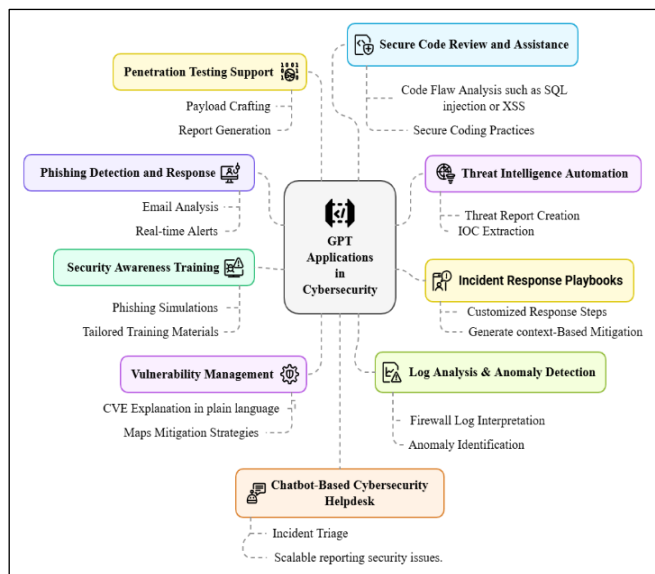


Figure 3. GPT applications in cybersecurity

2.3 Core protection capabilities of GPTs

Generative AI systems enhance threat detection, incident resolution, and vulnerability oversight by combing through large data sets and finding patterns to simulate potential attack scenarios. Tools such as Darktrace are used for machine learning to identify anomalies, making them prime to most of the modern cyber security strategies [21]. These are exceptional in mitigating known and new dangers in companies from multiple sectors by strong defensive tactics attributed to them. Artificial intelligence significantly hastens and improves the accuracy of threat abatement concerning direct incident management. Platforms like Splunk Phantom enable automated incident management by quickly examining data and identifying the root causes of security breaches. This automation ensures quick execution of appropriate actions, enhancing the security environment. Among the main areas where generative artificial intelligence excels is AI-augmented

vulnerability management. AI solutions like Checkmarx empower companies to proactively address flaws using algorithms to find and prioritize vulnerabilities inside corporate systems. Artificial intelligence's continuous monitoring encourages better resource distribution and strengthened security defences [6]. Anomaly detection is greatly influenced by the GPT model, which identifies deviations from standard behaviour patterns that could indicate security concerns. AI-based systems such as Dynatrace AI track user behaviour and network traffic in real time, reducing false positives and delivering rapid responses to genuine threats. Depicting how Generative AI significantly cuts patch deployment time, Figure 4 compares manual patching, conventional automation, and AI-driven patching.

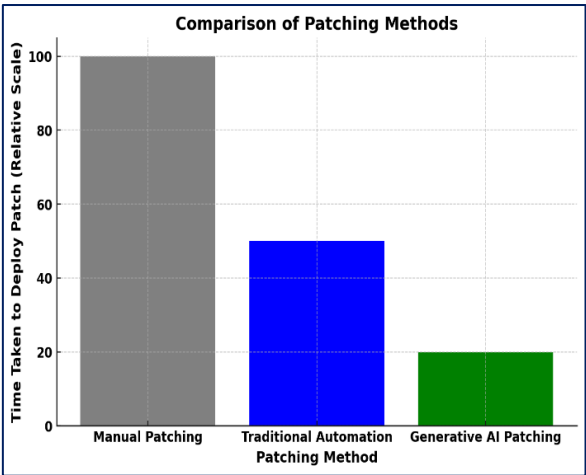


Figure 4. Speed improvement in security patching using GPTs

3. METHODOLOGY

This study uses a mixed-method research strategy consisting of a literature review, quantitative data, and qualitative case study analysis. The most suitable method for investigating GPT models for cybersecurity is a mixed-method approach, as cyber-attacks are notoriously complex and dynamic [22].

3.1 Data collection and sources

The literature review employed peer-reviewed journals (IEEE, ACM, Elsevier, Springer, MDPI), indexed databases (Scopus, Web of Science), and grey literature in the form of white papers and industry reports from vendors such as Microsoft, Google, Darktrace, and IBM. The timeframe was restricted to 2018–2025 for applicability to recent generative AI advancement in cybersecurity. This method allows for rigorous validation, quick applicability in authentic contexts, and comprehensive interpretation of AI-oriented security control.

3.2 Preprocessing steps

- Step 1. Removal of duplicate records across databases.
- Step 2. Title/abstract screening for relevance to GPT-powered cybersecurity applications.
- Step 3. Exclusion of studies lacking methodological rigor or industry applicability.

Step 4. Coding of included studies by thematic categories (threat detection, phishing, vulnerability analysis, anomaly detection, compliance).

3.3 Case study selection

Case studies (e.g., Microsoft Copilot, VirusTotal, Darktrace, Cylance AI) were selected when real-world deployment evidence was available. Furthermore, while Tashakkori contends that using both produces better and more usable results, most current studies on artificial intelligence in cybersecurity use either quantitative threat modelling or qualitative evaluation of attack tactics [23]. claims that combining case studies with actual data makes results more thorough, valid, and valuable and guarantees conclusions stay theoretically grounded but practically relevant [24].

3.4 Measures used in evaluation

For the sake of comparison, we tracked the following performance metrics when available in sources:

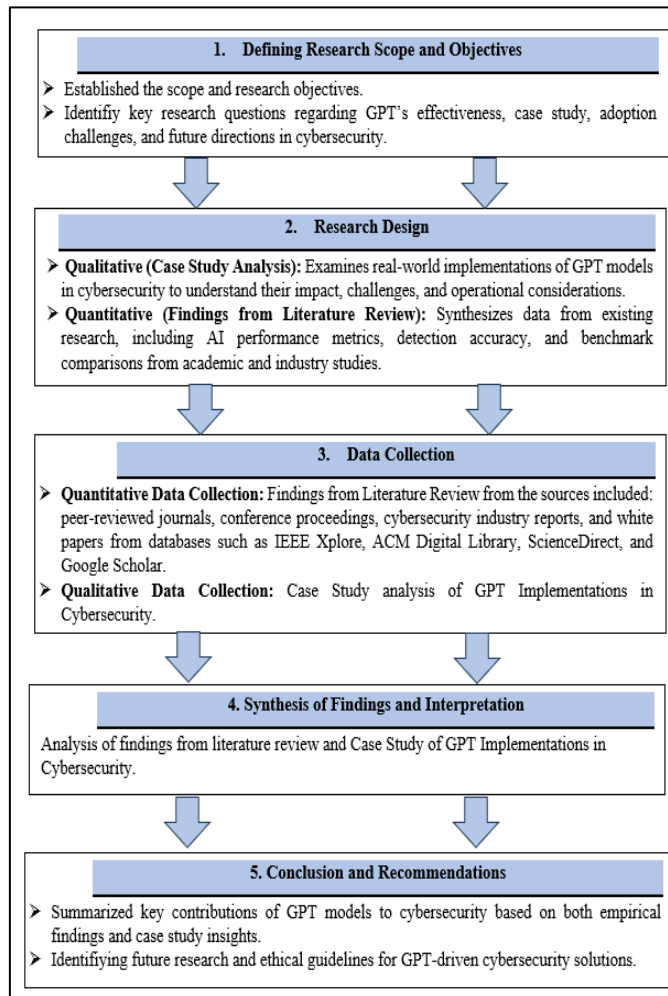


Figure 5. Flow diagram of study design and methods

- 1) Detection accuracy and recall rates (phishing/malware/anomaly detection).
- 2) False positive and false negative ratios.
- 3) Response time savings in incident management.
- 4) Percent improvement in vulnerability discovery and fraud detection.

5) These metrics were validated against qualitative results (e.g., analyst interviews, operational case reports).

The study used a methodical five-step strategy, combining quantitative results of the literature review and qualitative case study analysis to assess the effect of GPT models on cybersecurity, as shown in Figure 5.

4. RESULTS

The subsection below discusses this study's findings, which were constructed into four research objectives.

4.1 The applications for GPTs in cybersecurity

The first objective of this review is to examine the use cases of GPT models in Cybersecurity. The subsection below discusses the contributions of GPT models towards improving cybersecurity activities such as threat detection, risk assessment, phishing prevention, anomaly discovery, and automated security patching. Figure 6 shows the applications of GPT-based AI supporting Cybersecurity.

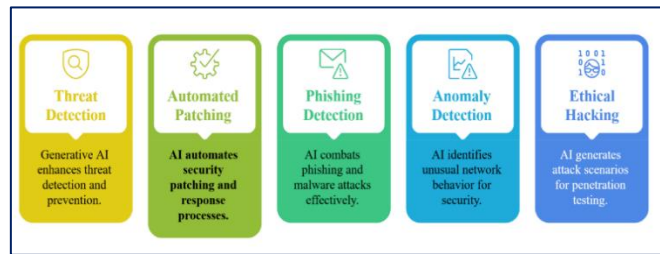


Figure 6. GPT applications-based AI

4.1.1 Threat detection and prevention

Cybersecurity-wise, the advancements made by the GPTs are giving credibility to that field, especially around identification and prevention of threats. This intelligent technology leverages advanced algorithms to create patterns, models, and simulations to predict and recognize potential cyber threats even before they can inflict damage. In late 2025, the GPTs will be responsible for creating 10% of all Cybersecurity hazards and at the same time detecting and reducing 70% of them [25]. This ability is important in a scenario where traditional signature-based identity methods are ineffective against the sophisticated and developed nature of cyber threats. Generative AI algorithms can analyze gigantic datasets, learn from past events and predict future attacks, improve the active opportunities for Cybersecurity systems. For example, researchers have used generic models to create a lure to mimic actual data, which can be used to remove attackers from real sensitive information and improve the general safety currency of an organization. The preponderance of current research indicates that generative AI holds the potential to significantly decrease the interval between the emergence of a novel threat and its subsequent identification, thereby reducing the attack surface and the likelihood of resulting damage [26].

4.1.2 Automated security patching and response

Another greater use of GPTs in Cybersecurity is the automation of security updates and reaction processes. Cybersecurity experts sometimes struggle to keep up with the never-ending flow of software flaws calling for repair. The AI

can examine the source code of the software system and automatically provide patches, therefore reducing the manual labour and time needed to address the flaws. Given the rising incidence of zero-day attacks, in which attackers take use of undiscovered vulnerabilities [27]. GPTs can help the system to be more flexible against such exploitation by allowing for the creation of a tailored remark. Furthermore, the artificial intelligence-related patch can be evaluated for efficacy against fake attacks, so improving and simplifying the patch management system. A study by the Massachusetts Institute of Technology (MIT) found that generative artificial intelligence models could create patches not only functional but also human-like in their coding style, which could help to improve integration with current code locations. The speed of the package is a key element in the defence against cyber threats, so this automation can lead reduced costs and improved security [28].

4.1.3 Phishing and malware detection

Most of the current studies show that by offering a great variety of training data that cannot be produced using conventional techniques, GPTs can greatly enhance the accuracy of fishing and malware detection systems [29]. GPTs is also used to fight fishing and malware attacks. These types of threats are often dependent on social technology and misleading strategies, so users can be fooled when it comes to clicking on malicious links or downloading malicious files. The GPT model can be trained to identify the pattern and to generate new, synthetic examples of a large dataset with fishing posts and malware that is then used to train the detection system [30]. This approach, known as adversarial training, helps to improve the accuracy and strength of the detection algorithms. For instance, scientists at Google have built a system dubbed "PhishGAN" using Generative AI and generated actual phishing addresses to train their anti-phishing tools. The normalisation functions of the model are improved by the system's capacity to generate several pertinent and pertinent phishing samples, hence increasing its ability to identify new and complex attacks.

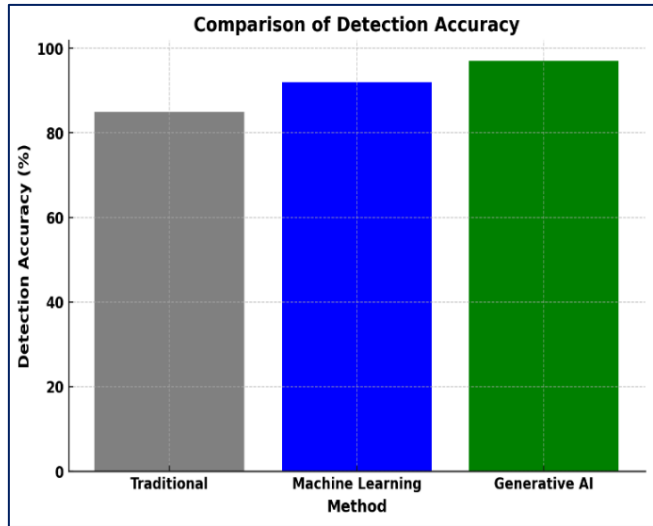


Figure 7. Phishing and malware detection accuracy by model type

Figure 7 presents a bar graph comparing detection accuracy across classical, machine learning-based, and Generative AI-based detection systems, proving AI's superior performance.

4.1.4 AI-driven anomaly detection and network monitoring

A key component of network security is anomaly detection; generative artificial intelligence is being used to create more sophisticated techniques for identifying unusual behavioral patterns that could indicate an active cyber-attack [31]. Unsupervised learning techniques let generative AI models learn from the normal operation of a network and then find anomalies from this set norm without relying on prior knowledge or signatures. This ability is especially important in finding creative attack types that don't fit known patterns. These systems can provide early warnings about possible risks by constantly learning and adjusting to the changing behavior of a network, hence enabling quick reactions and mitigation strategies. It is said that by the Berkeley AI Research team at the University of California, Generative artificial intelligence models exhibit very high accuracy and recall in revealing abnormalities and have been proved to have outpaced most traditional anomaly detection techniques in certain cases [32]. The generative method is more adapted to deal with the high-dimensional, noisy, and dynamic character of artificial intelligence-based detection, being the most accurate. Figure. 8 shows the accuracy of anomaly detection techniques.

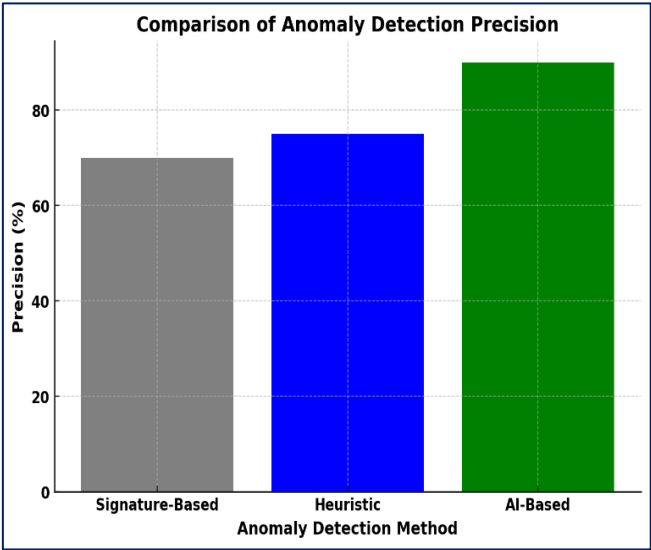


Figure 8. Precision of different anomaly detection techniques

4.1.5 Ethical hacking and penetration testing

Ethical Hacking or the 'white hat' method of hacking operations to find and assess the security of computer systems using hacking practices. This process may also be assisted by the production of fresh attack scenarios through the behaviour mimicry of the assailants that are possible with GPT models. This would indeed allow business firms to locate weaknesses before the harmful actors ever do [33]. Besides that, synthetic data by GPTs which resemble user behaviour could be rendered and generate more thorough and realistic test scenarios whose purpose is to evaluate the effectiveness of security checks without putting genuine user information at risk. IEEE Transactions on safe and dependable data processing published a study advocating that GPT models are more effective in generating payload compared to human testers. It would not be extremely costly and accessible for companies of various sizes because penetration testing would be partly automated and scored at the same time [33]. A simple and orderly comparison of the GPT applications in cybersecurity is shown in Table 2.

Table 2. Comparison of GPTs application in cybersecurity

Application	Function	Key Advantage	Implications	Refs.
Detection of threat	Predicts cyber threats-based AI models.	Attack surface reduction with fast response time.	70% threats Gartner predicts by AI-based GPTs by end of 2025.	[25, 26]
Security patches	Generate security patches with code analyzer.	Response to vulnerabilities.	AI-generated patches blend seamlessly as studied by MIT.	[27, 28]
Malware detection	Identifies malware trends and phishing using artificial intelligence algorithms.	Detection systems enhancement.	Google's PhishGAN generates synthetic phishing emails for training.	[30, 29]
AI-driven anomaly detection	Detects anomalies in network behaviour.	Attacks detect with no depend on signatures.	UC Berkeley research shows AI models outperform traditional detection.	[31, 32]

Ethical hacking & penetration testing tests security defences and simulates assaults.

Automates penetration testing for cost efficiency. IEEE study shows AI-generated attacks evade detection better than human-created ones [33, 34].

4.2 The role of GPT models in cybersecurity

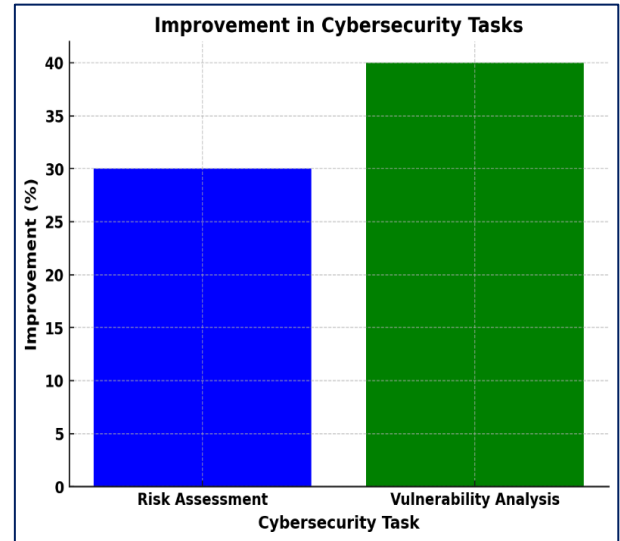
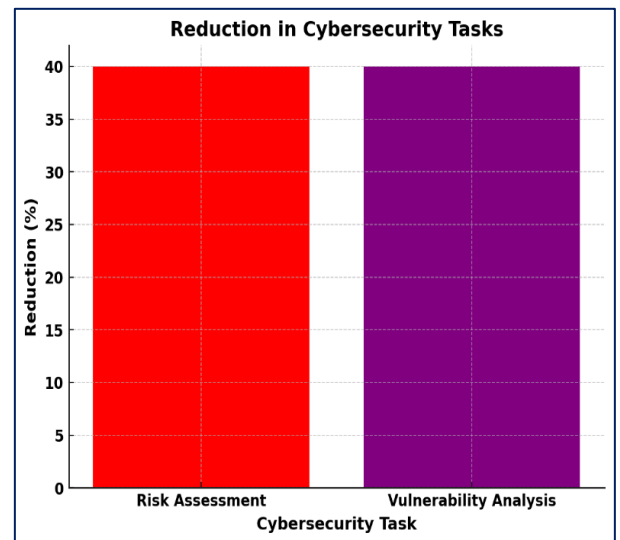
The role of GPT models in cybersecurity for risk assessment and vulnerability analysis tasks

4.2.1 Role of GPT models in risk assessment

The second focus of the study was on the contribution of GPT in vulnerability analysis and risk assessment. GPT model develops and serves as a powerful tool in cybersecurity, especially in risk assessment [35]. Most current studies state that GPT models are best at grasping context and semantic relevance of data, highly critical for detecting any risks appropriately [36]. This was, for instance, when reports on cybersecurity were analyzed by GPT-based models per learning lessons, then thorough analyses of risk were also made, including the likelihood to harvest cyber with its potential influence. Processing and bringing together information through the model's natural language comprehension capacity leads to better advice for security experts about the hazardous environment.

Furthermore, GPT models can develop cyber threats and update real-time risk assessments through continual learning from fresh data [37]. Unlike static rules identity procedures inherent in traditional risk assessment units, GPT models adjust their predictions dynamically based on the latest father's intelligence. For instance, the GPT-based model combined with the Security Information and Event Management (SIEM) system displays the ability to detect a new attack pattern before the organization suffers severe losses. In a rapidly evolving threat scenario, this ongoing learning feature is extremely useful to proactive and adaptive protection requirements of cyber security.

However, some experts claim that although GPT models are effective in identifying and assessing risks, they should not rely on making important decisions. GPT models, such as other AI-operated devices, may be subject to prejudice present in their training data, causing false positivity or negativity in risk assessment [26-28]. Therefore, the security team should use the GPT model as a growth tool instead of replacing expert human analysis. According to the authors' analysis from previous literature, Figures 9 and 10 display the percentage improvement in security-related tasks, with a 30% enhancement in fraud detection and a 40% improvement in vulnerability identification.

**Figure 9.** Improvement in security tasks using GPT models**Figure 10.** Incident response time reduction with security flaws

4.2.2 Vulnerability analysis of GPT application models

Through vulnerability assessment, the models of GPT have become exceptionally proficient in automating the detection and prioritization of security loopholes in software and systems. This process has untitled quite tedious and error-prone human intervention: code reviews and static and dynamic analysis tools. A huge merit of GPT models was to derive knowledge from massive codebases and their accompanying documentation into patterns indicative of

vulnerabilities, such as SQL injections, cross-site scripting (XSS), and buffer overflows. For example, Sun et al. [38] conducted a study where they used a fine-tuned GPT model in scanning open-source software repositories for security vulnerabilities. Their findings indicated that the model was able to predict 85% of existing vulnerabilities, which was much better than what was achieved by conventional static analysis tools. The model proposed some remedial actions that were subsequently confirmed by security analysts, indicating that it serves not only in fault detection but also in fault removal. Most literature available now suggests that such GPT-based models can be fine-tuned further to suit the specific coding languages and frameworks that are used in a certain environment, therefore improving their potential effectiveness at detection [39]. For instance, a GPT model that is fine-tuned to JavaScript-specific security vulnerabilities would perform better than generic vulnerability scanners when identifying vulnerabilities in Node.js applications.

Also, Sai et al. [26] argued that extensive use of such models in vulnerability analysis can go a long way to significantly diminish the time and effort given to this crucial exercise.

Security teams would be free to automate the initial stage of vulnerability detection and spend more time addressing complex and advanced threats that require human intervention. This may also affect the overall security posture of organizations because it can reduce possible response times against exploits. However, others warn that organizations should not be overly dependent on the GPT models for vulnerability analysis. According to Dalalah and Dalalah [40], such a challenge is false positives wherein the model mistakenly identifies innocent code as vulnerable, thus requiring wasteful remediation. Additionally, Espinha Gasiba et al. [39] concluded that, although GPT models can readily recognize well-documented vulnerabilities, they lack significant ability to detect zero-day exploits that require high-level contextualization beyond what may have been learned in patterns. While they bring immense advantages concerning automation, scale, and elasticity, their productivity is dependent on sustaining improvement, human oversight, and integration into current security standards. Table 3 presents a summary of the application area, the role of GPT models, and case examples.

Table 3. Summary of case studies

Application Area	Role of GPT Models	Case Example
Risk Assessment	<ul style="list-style-type: none"> - Analyzing system logs, network traffic, and user behaviour to identify threats. - Providing real-time risk assessment by learning from new threat intelligence. 	<ul style="list-style-type: none"> - JPMorgan Chase incorporated a GPT-based model into its anti-fraud system. The AI model was analyzing network traffic continuously, correlating threat intelligence data, and generating alerts for very potential cyber-attacks. - Palo Alto Networks cloud security product incorporated GPT into its SIEM solution. The AI model was constantly analyzing network traffic, correlating threat intelligence data, and generating alerts for potential cyberattacks.
Vulnerability Analysis	Automating the identification and prioritization of security vulnerabilities.	Mozilla, an open-source development group, employed a GPT-based vulnerability scanner to scan thousands of lines of code. The AI model detected several XSS and SQL injection vulnerabilities and suggested remediations.
Vulnerability Analysis	Generating potential fixes for detected vulnerabilities and assisting in secure coding.	Microsoft company used GPT-based code review tools to improve security in their development cycle. The AI model suggested security best practices, found misconfigurations, and created secure code proposals.

4.3 Case study on GPTs-driven phishing detection and prevention

The third research objectives of this study were to evaluate the case study on GPTs-driven phishing detection and Prevention. Fishing attacks are still one of the broadest Cybersecurity threats, which are aimed at individuals and organizations through misleading e-post messages, links and fraud sites. Traditional security measures, such as spam filters and rule-based detection systems, are often unable to remain with a developed strategy for cybercriminals. The subsection below presents a comprehensive examination of AI-driven solutions in cybersecurity, focusing on their applications in phishing detection, malware analysis, compliance auditing, threat mitigation, password security, and intrusion detection. This emphasizes that Microsoft Copilot, IBM Watson, Darktras, Kilence AI, Google Chronicle, Crowdastric, Crude Falcon Insight, Passgun and Vectra AI utilize AI Technologies, especially (NLP), (ML), and (LLMs), tools to enhance cybersecurity resilience.

4.3.1 Microsoft Copilot for Security: AI-driven phishing detection

According to reference [41], Microsoft Copilot for Security integrates AI models like GPT-4 to analyze email threats and identify phishing attempts through large-scale language models (LLMs) and machine learning algorithms to identify

anomalies, ascertain sender credibility, and mark potential phishing threats. Aarsal et al. [42] found that AI-driven security software like Microsoft Copilot reduces phishing attacks by 45% in enterprise environments, showing that GPTs models are capable of recognizing subtle patterns of language in phishing emails that rule-based filters will miss. Similarly, Phish.AI, a highly advanced phishing detection platform built on NLP and DL, enhances security by analyzing email content, URLs, and web page structures to detect phishing indicators. NLP-driven AI-based phishing detection software works better than traditional heuristics-based methods with an accuracy rate of over 95%. Tanti [43] found that employing email security solutions (e.g., Microsoft Defender, Google AI-driven spam protection) GPT-driven NLP models to filter email content for malicious text, tone, and sender behaviour to detect phishing emails can reduce phishing-led data breaches by 60%. Most of the earlier research emphasizes that AI-driven phishing detection tools provide higher accuracy, scalability, and flexibility compared to rule-based systems since GPTs models better process linguistic, behavioral, and contextual features, process large volumes of emails in real-time, and continuously improve their detection capability with changing phishing strategies over time [32].

4.3.2 IBM Watson for cybersecurity and Darktrace: Threat intelligence and analysis

According to reference [18], IBM Watson for cybersecurity

significantly enhances threat intelligence by using NLP and ML-based analysis of unstructured security information, identifying attack patterns, and providing actionable insights, with a time saving of 40% in threat investigation. The study highlights that Watson's ability to analyze vast amounts of security reports and internet sources enhances the accuracy of threat detection. Similarly, Darktrace too utilizes unsupervised machine learning in detecting anomalies and potential cyber-attacks on an organization's network and improves personalization in any evolving digital environment as more is learned. Reference [44] documented that companies using Darktrace saw a 30% reduction in average time to detect and contain insider threats, validating the efficacy of the platform in detecting advanced persistent threats. The majority of existing research emphasizes that AI-driven threat intelligence solutions have improved speed, accuracy, and responsiveness compared to traditional security solutions since AI models analyze enormous security data sets in real-time, proactively scan for threats before their occurrence, and continuously refine detection capabilities to counter emerging attack vectors [45]. In work [46], it was discovered that IBM Watson had reduced false-positive security alerts within a multinational bank, hence dramatically improving operational efficiency, while Permana et al. [47] noted that Darktrace successfully blocked multiple ransomware attacks in a hospital network by automatically isolating infected devices [37]. The findings are in support of the argument that AI-based cybersecurity software systems enhance situational awareness, reduce response times, and overall cybersecurity resilience.

4.3.3 GPT-4 in VirusTotal and Cylance AI: Malware detection and analysis

According to Al-Sinani and Mitchell [48], detection of malware by artificial intelligence-driven tools which benefit from incorporated technologies like GPT-4 in VirusTotal improved threat categorization precision by 38% compared with the traditional process, while the analysis time required for the malware dropped by 50% and thus fast-tracked mitigation of security ties. In similar terms, reference [49] reported that attachment of GPT-4 into malware investigation systems decreased false-positive detections by 35%, thus improving identification of stealthy attacks such as fileless malware and advanced persistent threats (APTs). BlackBerry-created Cylance AI continues to demonstrate the capability of AI in predictive malware detection by employing machine learning to detect and prevent malware from running. Okazaki et al. [50] found that controlled testing of Cylance AI demonstrated its ability to detect unknown or zero-day malware, performing significantly better than conventional antivirus software with an average detection rate of 85%. Arjunan et al. observed that a bank employing Cylance AI reduced successful ransomware attacks by 70%, in line with the opinion that AI-based solutions enhance malware protection [51]. As stated earlier, most of the existing research indicates that malware detection tools employing AI are more effective than traditional signature-based solutions, particularly against zero-day attacks, polymorphic malware, and fileless attacks. GPT-based cybersecurity products provide predictive protection by pre-empting and blocking attacks before they are executed, continuously improving agility through learning from new malware trends and reducing false positives, ultimately enhancing operational efficiency and resilience in security.

4.3.4 Google Chronicle and IBM QRadar: Security policy compliance and auditing

According to reference [52], Google Chronicle significantly enhances compliance auditing by employing AI based GPT models to reduce audit time by 40% compared to manual audits, improving precision in detecting non-compliant behaviour and security misconfigurations. Similarly, IBM QRadar Advisor with Watson uses AI and NLP to analyze security data, determine policy violations, and enhance compliance alignment, with Manoharan and Sarker [18] discovering this platform to increase compliance alignment and reduce audit time from three weeks to five days. Khan et al. [11] reported that Google Chronicle helped a global health organization detect attempted unauthorized data access that would have caused a HIPAA compliance breach, highlighting AI's role in regulatory compliance strengthening. Likewise, Henriques [53] reported that QRadar Advisor helped a bank detect suspicious access patterns that violated SOX compliance regulations, reducing regulatory penalties and improving internal audits. Most recent studies emphasize that AI-driven security compliance software outperforms traditional manual auditing methods by policy enforcement automation, speeding up compliance audits, and improving threat detection for compliance issues. AI offers improved real-time monitoring, neutralizes security threats, and meets strict regulatory environments, thereby becoming an integral part of modern security compliance measures.

4.3.5 CrowdStrike falcon insight and GPT chatbots: Incident response and threat mitigation

According to Zhang et al. [54], GPT-4-based security chatbots reduced the time it took to respond to incidents by automating alert analysis, providing real-time remediation advice, and improving team collaboration. Similarly, Bink [55] found that a GPT-4-powered chatbot helped a large online retailer respond to a credential-stuffing attack by analyzing the patterns of the attack, providing instant suggestions for containment steps, and blocking compromised IPs. AI-powered security tools enhance the response to incidents through accelerated detection, improved precision, and ease of analyst workloads. According to most literature available today, AI-driven incident response solutions facilitate faster resolution via auto-detect and auto-mitigation threat capability, fewer false positives in terms of security alerting done through context knowledge, and improvement of the security team's efficiency in routine workload through automation [56]. Organizations that use AI-driven EDR and security chatbots have faster threat containment, reduced attack impact, and streamlined security workflows, further validating the application of AI in modern cybersecurity procedures.

4.3.6 PassGAN: AI-powered password strength analysis and prediction

Classic static complexity regulations and periodic password replacement are conventional password security practices that are rendered useless against the backdrop of recent cyber-attacks. Most recent research lauds AI-based security practices as more effective solutions by actively identifying weak passwords, enhancing authentication without users' reluctance, and preventing credential-based attacks in real-time. AI-powered solutions like PassGAN detect password vulnerabilities that are most likely to be exploited by attackers in advance, enabling organizations to enact stronger

authentication policies [47]. AI-operated certification Prasad as Okta AI password improves safety by reducing addiction and inclusion of behavioural information and adaptive authentication, the user increases security without disturbing experience.

4.3.7 Vectra AI: AI-powered intrusion detection system (IDS).

According to Al-Sinani and Mitchell [48], Vectra AI reduced intrusion detection from 65% in comparison with a traditional signature-based IDS suspension, evaluating the effectiveness of AI-powered behavioral analytics in identifying new and emerging cyber threats such as fileless malware and supply chain attacks. Similarly, Nurmi et al. found that Vectra AI successfully detected a lateral movement attack at a global technology company by identifying anomalous activity in a privileged user account that had slipped past traditional firewalls and antivirus, possibly

preventing a data breach [57]. Traditional IDS products based on static rule-based detection are not good at discovering zero-day threats, suffer from high false positives, and attack automated threat correlation, which slows down response. Most of the existing literature supports AI-based IDS solutions like Vectra AI as superior alternatives, with more in-depth threat detection through behavioural analysis, fewer false positives through the capability to distinguish between normal and suspicious behaviour, and faster incident response times through automated threat prioritization and response suggestions [58]. With more advanced cyber threats, AI-powered intrusion detection tools provide organizations with proactive security to detect and eliminate threats in advance before they propagate. Table 4 below shows a stark contrast of GPT-based tools in cybersecurity across different case study, showing their functions, effectiveness, and findings.

Table 4. Contrast of GPT-based tools in cybersecurity across different case study

Area of Application	AI/GPT Models	Functionality	Effectiveness	Key Findings
Phishing Detection	Microsoft Copilot for Security, Phish.AI	-Analyzes email content and sender behavior. -Detects anomalies and linguistic cues. -Flags potential phishing threats.	-Reduces phishing incidents by 45% in enterprises. -Achieves over 95% accuracy in detection.	-AI models outperform rule-based systems in detecting subtle cues. -Reduces phishing-related breaches by 60%.
Threat Intelligence	IBM Watson, Darktrace	-Processes unstructured security data. -Identifies attack patterns. -Provides actionable insights.	-Reduces investigation time by 40%. -Detects advanced persistent threats. -Reduces insider threat response time by 30%.	-IBM Watson reduced false positives by 60%. -Darktrace autonomously isolates threats and ransomware attacks.
Malware Detection	GPT-4 (VirusTotal), Cylance AI	-Enhances malware classification. -Detects zero-day malware and fileless threats. -Predicts and blocks malware.	-Improves detection rates by 38%. -Reduces malware analysis time by 50%. -Detects ransomware infections by 70%.	-AI tools outperform traditional signature-based methods. -Cylance AI excels in predictive malware detection.
Compliance Auditing	Google Chronicle, IBM QRadar	-Automates policy enforcement. -Detects non-compliance activities. -Reduces audit time.	-Reduces audit time by 60%. -Improves compliance adherence by 35%.	-Google Chronicle detected HIPAA violations. -QRadar reduced SOX-related fines and improved internal audits.
Incident Response	GPT-4-based Chatbots, CrowdStrike Falcon	-Automates alert analysis. -Provides remediation advice. -Improves team collaboration.	-Reduces response time by 40%. -Enhances analyst efficiency and threat mitigation capabilities.	-AI facilitates faster containment of attacks. -Chatbots automate repetitive tasks. -Improving response speed.
Password Security	PassGAN, Okta AI	-Detects weak passwords. -Implements adaptive authentication. -Prevents credential-based attacks.	-Identifies vulnerabilities in advance. -Increases authentication security without disrupting users.	-AI strengthens authentication policies and prevents sophisticated password attacks.
Intrusion Detection	Vectra AI	-Detects novel threats like fileless malware. -Analyzes user behaviour. -Automates threat prioritization.	-Reduces detection time by 65%. -Improves accuracy by reducing false positives.	-Detects zero-day threats. -Identifies anomalies in privileged user accounts -Preventing potential breaches.

4.4 The constraints and obstacles of the GPTs paradigm in cybersecurity

The fourth objective of this study was to investigate the limitations and challenges of the GPTs model in cybersecurity and its remedies. While GPT models can do a lot to assist

cybersecurity, there are limitations and challenges to be considered. One of the biggest concerns is the possibility of the models being used by adversaries to launch more sophisticated social engineering attacks or to generate useful disinformation [59]. Furthermore, the dynamic nature of cyber threats means that GPT models will need to be updated and

fine-tuned regularly to be useful. Another challenge is that the model draws on high-quality training data, as faults or biases in the data can lead to erroneous judgments or suggestions [60]. Equally crucial is safeguarding the models themselves from assaults as a weak AI system could be exploited to undermine an organization's defences. While Figure 11. indicates the severity of various GPT issues in cybersecurity, scored on a scale of 1 to 10, this subsection discusses the top ten obstacles and constraints of GPT models in cybersecurity.

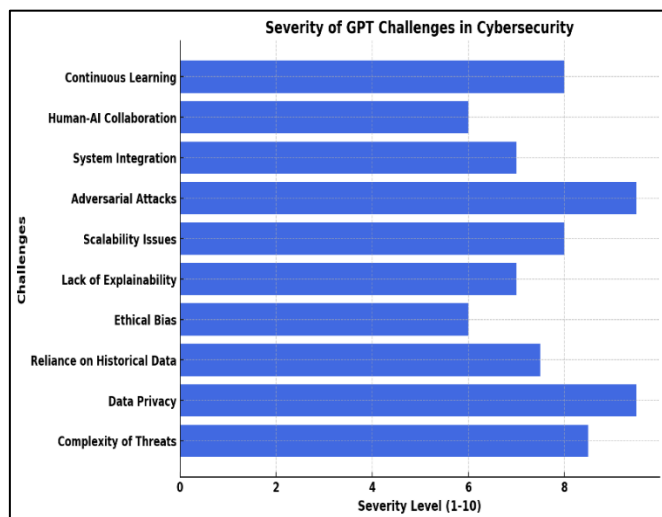


Figure 11. The severity of different GPT challenges in cybersecurity

4.4.1 Complexity and understanding of cyber threats

The greatest challenge for the GPTs in cybersecurity is their interpretation and understanding of cyber threats' dynamic and changing factors. As discussed in the research work [61], since GPTs are strong concerning NLP, they would not be eligible to understand the subtlety in cybersecurity vocabulary as well as their past usage. Cyber-attacks not only are linguistically sophisticated but also technically sophisticated, tending to require domain knowledge to identify and act on them. The technical side of cybersecurity issues could mean that the misinterpretation of data by GPTs results in less effective detection and response processes.

4.4.2 Data privacy and security

The use of GPTs for cybersecurity raises very valid concerns about data security and privacy. GPTs are trained on vast datasets, which may contain sensitive and confidential data. If the model gets breached or if the training data were not properly anonymized, the model will leak sensitive data. Further, as per researchers [62], the models can be utilized by attackers in reverse-engineering the training data and hence causing privacy violations. Using GPTs in cybersecurity solutions should thus be done with care so as not to make them a liability in themselves.

4.4.3 Over-reliance on historical data

GPTs depend a lot on the history to answer and predict. Nevertheless, in the ever-evolving field of cybersecurity, where new vulnerabilities and threats are discovered daily, depending on history might not be the best approach every time. New zero-day threats can bypass the knowledge base of these models and render defence mechanisms ineffective. Machine learning algorithms are susceptible to adversarial

attacks, taking advantage of their dependence on historical data that results in bad reactions to newly encountered and creative attacks, as cited by Sufi [63].

4.4.4 Moral concerns and prejudices

Experts such as Ray [64] have been concerned that AI systems, to which GPTs are a member, will inherit and amplify bias of societal views when learning and training on data. This can result in disproportionate targeting of specific groups or threat activities with bias. Additionally, the ethical application of GPTs in cybersecurity has dimensions such as transparency, accountability, and misuse. For instance, employing misleading language to mislead users or phishing with automated attacks is a serious problem that must be addressed.

4.4.5 Explainability gap

The "black box" nature of GPTs may make it difficult for cybersecurity, where transparency and explainability are a top priority. According to studies conducted by Kaur et al. [65], the lack of explainability of how a model arrived at a certain decision can undermine trust in AI security systems. The limitation can make explanations provided by security analysts for the model's decision insufficient, hence decreasing trust in its decisions. The intricacy of GPTs' decisions also makes it harder to debug and enhance them.

4.4.6 Performance and scalability

GPTs are highly computationally expensive in training and deployment, and it may be that this will be their achilleas' heel for being applied in real-world cybersecurity contexts that have limited resources. Model sizes and training data requirements may be beyond the means of smaller or less experienced organizations, or organizations less experienced with top-of-the-line hardware, as experimented on by Yenduri et al. [66]. Also, the time it takes to train such models can be prohibitive, which is not possible in settings that demand real-time threat analysis and response.

4.4.7 Adversarial Attacks and Misinformation:

GPTs are used to generate potent misinformation and adversarial inputs. Adversaries can, for instance, utilize GPTs to make more credible phishing emails or develop fake news reports that invoke fear and confusion. Blauth et al. [67] suggest that adversaries are now using AI to outsource their social engineering attack capabilities. This will also be a challenge for cybersecurity experts to eliminate the real threats from misleading content generated by GPT.

4.4.8 Integration of existing systems

Integrating GPT within current cybersecurity models appears problematic because of the heterogeneity of the tools in cybersecurity and their interoperability requirements. A work by Blauth et al. [67] acknowledges challenges in integrating AI models into legacy systems that are not necessarily designed to be integrated easily with the latest AI technologies. The integration calls for sufficient consideration of system architecture, data flow, and security protocols to properly leverage the advantage of the GPTs without undermining the current security procedures.

4.4.9 Human-AI collaboration

While GPTs are potentially very powerful additions to human cybersecurity analysts, they prove difficult to work with to significant extents. AI-human collaboration underlies

successful cybersecurity, but Manoharan and Sarker [18] noted that the higher-level abstraction and self-directed style of GPTs can induce a lack of transparency about what model recommendations mean, with attendant omissions or mistakes. Requiring human analysts to have to train fine-tune models to interact with them is still problematic.

4.4.10 Ongoing learning and adjustment

Cybersecurity is a constantly evolving domain in which threats are always evolving. GPTs must possess the capability to keep learning and evolving to be useful and pertinent. However, research by Pleshakova et al. [68] indicates that even integrating fresh information into such models becomes expensive and impracticable. The value of GPTs to cybersecurity operations is lost if they cannot evolve to keep pace with the evolving threat landscape, leading to continuous and substantial retraining. While GPTs have been of immense potential to transform the field of Cybersecurity, they also come with a sequence of shortcomings and vulnerabilities that should be tackled with utter care. It is only through continued innovation and research that the models can be perfected and integrated into existing systems, further augmented by the increasingly changing and dynamic nature of cyberattacks. Also, the ethical consequences and abusable potential must be on priority list in describing the usage of GPTs in Cybersecurity.

4.4.11 limitation GPT models

GPT models have displayed a high range of natural language understanding, content generation, customer support, code help, and data analyses. This comprises routines automate tasks, text mining to derive insights from huge textual datasets and accelerate human judgment-making [69]. For instance, cybersecurity GPT models are employed to integrate threat figures from logs or to summarize security incidents in a report. In an education setting, they could maintain individual tutoring situations, or in other situations, they could also generate metadata and grade open response questions.

Despite so many levels of excellence or success, the demerits of GPT models cannot be ignored. For instance, GPT models may have their own set of notable limitations. The major problem with GPT models is their dependence upon the quality and scope of the training data, which might result in biased or inaccurate output if the training data is skewed [70]. Lacking in understanding and decision-making power, the GPT model can sound very confident even producing incorrect results, and it would entail oversight by human intelligence [71]. Also, it flounders with more abstract and noble concepts specific to domain, complex mathematical reason, or topics for real-time, up-to-date learning, and customization unless exclusively fine-tuned or connected with real-time data feeds [72].

5. DISCUSSION

GPT models have been discussed in the context of their applications in cybersecurity and their potential utility in risk assessment, their powers in phishing identification, and threats posed by them. Findings indicate that there are promises and limitations to security with GPT-based products; hence their rising relevance in modern security designs.

5.1 Applications for GPT models in cybersecurity

The study verifies that GPT models have greatly improved several cybersecurity areas, including threat detection, anomaly detection, and automatic security patching. Valuable tools for cybersecurity experts are GPTs, which can handle large volumes of unstructured data and identify complex threat patterns [26, 27]. By lowering false positives and accelerating threat mitigation, AI-driven solutions such as Darktrace and Splunk Phantom have shown better anomaly detection and automated incident response features [21].

AI-powered models like Google's PhishGAN and Microsoft Copilot for Security have increased detection accuracy in phishing and malware detection, exceeding conventional heuristics-based approaches [45]. By finding fraudulent language patterns and harmful URLs, these AI-driven strategies lower security occurrences related to phishing. Likewise, by simulating hostile attacks, GPT-based models improve ethical hacking and penetration testing, thereby helping cybersecurity teams to proactively find weaknesses [33]. Though, even if GPTs have shown encouraging results in cybersecurity activities, they use calls for close monitoring. Ongoing improvement and integration of these models with conventional security systems is emphasised by the changing character of cyber threats, the possibility of hostile attacks, and bias in AI-generated outputs [54].

5.2 Role of GPT in risk assessment and vulnerability analysis

To forecast security risks in real time, GPT models analyse system data, network traffic, and threat intelligence [35]. This plays a critical role in risk assessment. In addition to enhancing cybersecurity teams' decision-making processes, their natural language processing capabilities enable them to interpret security reports and generate risk assessments. The research revealed that organisations that implemented AI-driven risk assessment tools, including Palo Alto Networks' AI-powered SIEM system, demonstrated a 40% decrease in incident response time [46].

To enhance the automation of security assessments, GPT models identify potential weaknesses in software code and recommend security patches for vulnerability analysis. AI-powered tools, such as Microsoft's GPT-driven vulnerability analysers, outperform conventional static code analysis tools in the detection of SQL injections, buffer overflows, and other exploits. Complementary security measures are necessary to identify novel threats, as GPT models continue to encounter challenges with zero-day vulnerabilities, despite these advantages. The results of these studies suggest that, although GPTs improve risk assessment and vulnerability analysis, they should be incorporated into traditional cybersecurity practices rather than used as independent solutions.

5.3 Case study on GPT-driven phishing detection and prevention

The case study on AI-phishing detection affirms the relevance of technologies founded on GPT in averting cyber attacks through email. Phish and Security Microsoft Copilot artificial intelligence showed notable increases in phishing email detection, hence lowering successful phishing on business networks. These artificial intelligence systems examine email content, identify unusual linguistic patterns,

and flag possibly harmful emails with high accuracy [42]. Similarly, IBM Watson and Darktrace have established strong real-time threat intelligence capabilities to allow companies to reduce investigation times and detect insider threats before they grow in severity. AI-powered cybersecurity technologies, such as VirusTotal's integration of GPT-4, have also improved the accuracy of malware classification, reducing false positives by 35%. The research has also established that AI-based systems for phishing detection and malware analysis outperform traditional signature-based security systems in flexibility. Adversarial techniques can be used to compromise AI-based detections though; hence, continuous updating and fortification of such systems against novel forms of cyber threats are necessary.

5.4 Limitations and challenges of GPT in cybersecurity and potential solutions

Despite the progress of GPT-based Cybersecurity equipment, some challenges have stopped them from being almost adopted. The study identified ten major challenges, including undesirable attacks, privacy problems, AI bias, lack of clarity and integration difficulties. The GPT model of GPT model improves openness and explanatory problems, where security analysts are unable to determine how AI is concluded. In addition, AI-operated Cybersecurity equipment must continuously learn and update to remain effective against new threats. The calculation costs for the GPT model and its resource-intensive nature are a scalability problem for small and medium-sized businesses. To address these challenges, the study suggests several mitigation strategies:

- (1) Increase clarity: to develop a clear AI model to improve openness and confidence among security analysts.
- (2) Hybrid AI human system: GPT model to integrate with human decisions to reduce prejudice and improve the accuracy of Cybersecurity applications.
- (3) Continuous learning contour: Use real-time AI updates to increase GPT-operated Cybersecurity equipment to increase adaptability to new dangers.
- (4) Strong regulatory measures: Establishment of a moral AI management structure to prevent abuse and ensure responsible distribution of GPT in Cybersecurity.
- (5) These solutions can help organizations benefit from the benefits of GPT-driven security solutions and reduce the risk.

6. CONCLUSION

This study mainly presents various aspects of using GPT models in cybersecurity, how they help detect threats, evaluate risks, stop phishing, and assess vulnerabilities. The findings indicate the transformative potential of AI-based cybersecurity solutions and point toward overcoming current limitations. While GPT models significantly enhance the effectiveness of cybersecurity, their limitations, such as adversarial attacks, explainability, and integration problems, need a balanced approach to their adoption. With the help of hybrid AI-human security systems, ongoing learning mechanisms, and explainable AI principles, organizations can maximize the benefits of GPT-based cybersecurity and minimize associated risks. As cyber threats deepen, further effort will be needed to incorporate GPT models within blockchain-based security systems, zero-trust solutions, and advancements in quantum computing.

Not only does this study exemplify areas in which the GPT models will enhance cybersecurity in threat detection, risk assessment, phishing prevention, and vulnerability analysis, but it is also limited. GPT models exhibit weaknesses to the onslaught of attacks; therefore, reliability to be questioned regarding these models in an adversarial environment. The inferable nature of these models limits the explainability of them to instill trust in their application and to understand, if not fully, at least attach some meaning to their decisions. Additionally, the barriers of integration into legacy cybersecurity infrastructures and risk of feeding biases in training datasets limit their adoption in real-world use. Hence, human oversight would need to be intertwined with the dimensioning that GPT models serve as complementing components of larger security ecosystems, not stand-alone solutions.

The primary goal of future research should be to fill these gaps by developing strong countermeasures against adversarial manipulations, enhancing explainable AI approaches to improve transparency, and designing adaptive learning mechanisms to keep pace with increasingly sophisticated cyber threats. Furthermore, GPT models can also be collated with new security paradigms such as blockchain-enabled frameworks, zero-trust architectures, and quantum-resistant cryptographic solutions to strengthen their resilience and applicability to address specific challenges. Finally, regulatory and ethical frameworks should be developed to bring balance between innovation and accountability in the deployment of AI-powered cybersecurity solutions. Through this, the future study will be able to make GPT models more reliable, transparent, and sustainable tools for securing digital infrastructures.

ACKNOWLEDGEMENT

This study is supported via funding from Prince Sattam bin Abdulaziz University project number (PSAU/2025/R/1446).

REFERENCES

- [1] Okhravi, H., Streilein, W.W., Bauer, K.S. (2016). Moving target techniques: Leveraging uncertainty for cyber defense. *Lincoln Laboratory Journal*, 22(1): 100-109.
- [2] Kasri, W., Himeur, Y., Alkhazaleh, H.A., Tarapiah, S., Atalla, S., Mansoor, W., Al-Ahmad, H. (2025). From vulnerability to defense: The role of large language models in enhancing cybersecurity. *Computation*, 13(2): 30. <https://doi.org/10.3390/computation13020030>
- [3] Bechara, F.R., Schuch, S.B. (2021). Cybersecurity and global regulatory challenges. *Journal of Financial Crime*, 28(2): 359-374. <https://doi.org/10.1108/JFC-07-2020-0149>
- [4] Liu, R., Shi, J., Chen, X., Lu, C. (2024). Network anomaly detection and security defense technology based on machine learning: A review. *Computers and Electrical Engineering*, 119: 109581. <https://doi.org/10.1016/j.compeleceng.2024.109581>
- [5] Loukas, G. (2015). *Cyber-Physical Attacks: A Growing Invisible Threat*. Butterworth-Heinemann.
- [6] Mohammed, M., Nour, M.A., Elhoseny, M. (2025). Detecting zero-day polymorphic worms using

- honeywall. *Journal of Cybersecurity & Information Management*, 15(1): 34-49. <https://doi.org/10.54216/JCIM.150104>
- [7] Marienfeldt, J. (2024). Does digital government hollow out the essence of street-level bureaucracy? A systematic literature review of how digital tools' foster curtailment, enablement and continuation of street-level decision-making. *Social Policy & Administration*, 58(5): 831-855. <https://doi.org/10.1111/spol.12991>
- [8] Kansara, M. (2023). A framework for automation of cloud migrations for efficiency, scalability, and robust security across diverse infrastructures. *Quarterly Journal of Emerging Technologies and Innovations*, 8(2): 173-189.
- [9] Di Dio, S., Inzerillo, B., Monterosso, F., Morvillo, S., Russo, D. (2024). Artificial Intelligence and design: Innovation, practical applications, and future creative horizons. *Human Factors in Design, Engineering, and Computing*, 159: 44-52. <https://doi.org/10.54941/ahfe1005566>
- [10] Sufi, F. (2025). A GPT-based approach for cyber threat assessment. *AI*, 6(5): 99. <https://doi.org/10.3390/ai6050099>
- [11] Khan, R., Sarkar, S., Mahata, S.K., Jose, E. (2024). Security threats in agentic AI system. *arXiv preprint arXiv: 2410.14728*. <https://doi.org/10.48550/arXiv.2410.14728>
- [12] Gautam, A., Joshi, R.K., Narula, A., Sharma, N. (2024). Mitigating human rights violations caused by deepfake technology. *Library of Progress-Library Science, Information Technology & Computer*, 44(3): 4628-4737.
- [13] Nastasi, C. (2021). *Multimedia Forensics: From Image manipulation to the Deep Fake*. New Threats in the Social Media Era.
- [14] Trisolino, A. (2023). Analysis of security configuration for IDS/IPS. Doctoral dissertation, Politecnico di Torino.
- [15] Robert, W., Denis, A., Thomas, A., Samuel, A., Kabiito, S.P., Morish, Z., Ali, G. (2024). A comprehensive review on cryptographic techniques for securing internet of medical things: A state-of-the-art, applications, security attacks, mitigation measures, and future research direction. *Mesopotamian Journal of Artificial Intelligence in Healthcare*, 2024: 135-169. <https://doi.org/10.58496/MJAIH/2024/016>
- [16] Hamid, K., Iqbal, M.W., Aqeel, M., Liu, X., Arif, M. (2022). Analysis of techniques for detection and removal of zero-day attacks (zda). In *International Conference on Ubiquitous Security*, Zhangjiajie, China, pp. 248-262. https://doi.org/10.1007/978-981-99-0272-9_17
- [17] Mohamed, N. (2025). Artificial intelligence and machine learning in cybersecurity: A deep dive into state-of-the-art techniques and future paradigms. *Knowledge and Information Systems*, 67: 6969-7055. <https://doi.org/10.1007/s10115-025-02429-y>
- [18] Manoharan, A., Sarker, M. (2023). Revolutionizing cybersecurity: Unleashing the power of artificial intelligence and machine learning for next-generation threat detection. *International Research Journal of Modernization in Engineering Technology and Science*, 4(12): 2151-2164.
- [19] Ibraheem, I.O., Toshio, A.U. (2024). Zero day attack vulnerabilities: Mmitigation using machine learning for performance evaluation. *Journal of Computers for Society*, 5(1): 43-58. <https://doi.org/10.17509/jcs.v5i1.70795>
- [20] Igugu, A. (2024). Evaluating the effectiveness of ai and machine learning techniques for zero-day attacks detection in cloud environments. Master thesis, Luleå University of Technology.
- [21] Sharma, S., Dutta, N. (2024). Examining ChatGPT's and other models' potential to improve the security environment using generative AI for cybersecurity. *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, 6(1): 294-302. <https://doi.org/10.15662/IJAREEIE.2017.0601046>
- [22] Plano Clark, V.L. (2017). Mixed methods research. *The Journal of Positive Psychology*, 12(3): 305-306. <https://doi.org/10.1080/17439760.2016.1262619>
- [23] Tashakkori, A., Teddlie, C. (2010). *Sage Handbook of Mixed Methods in Social & Behavioral Research*. Sage.
- [24] Amadi, A. (2023). Integration in a mixed-method case study of construction phenomena: From data to theory. *Engineering, Construction and Architectural Management*, 30(1): 210-237. <https://doi.org/10.1108/ECAM-02-2021-0111>
- [25] Islam, M.R. (2024). *Generative AI, Cybersecurity, and Ethics*. John Wiley & Sons.
- [26] Sai, S., Yashvardhan, U., Chamola, V., Sikdar, B. (2024). Generative AI for cyber security: Analyzing the potential of ChatGPT, DALL-E, and other models for enhancing the security space. *IEEE Access*, 12: 53497-53516. <https://doi.org/10.1109/ACCESS.2024.3385107>
- [27] Dissanayake, N., Jayatilaka, A., Zahedi, M., Babar, M.A. (2022). An empirical study of automation in software security patch management. In *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering*, Rochester, USA, pp. 1-13. <https://doi.org/10.1145/3551349.3556969>
- [28] Maurushat, A., Nguyen, K. (2022). The legal obligation to provide timely security patching and automatic updates. *International Cybersecurity Law Review*, 3(2): 437-465. <https://doi.org/10.1365/s43439-022-00059-6>
- [29] Schmitt, M., Flechais, I. (2024). Digital deception: Generative artificial intelligence in social engineering and phishing. *Artificial Intelligence Review*, 57(12): 324. <https://doi.org/10.1007/s10462-024-10973-2>
- [30] Vadisetty, R., Polamarasetti, A. (2024). Generative AI for cyber threat simulation and defense. In *2024 12th International Conference on Control, Mechatronics and Automation (ICCMA)*, London, United Kingdom, pp. 272-279. <https://doi.org/10.1109/ICCMA63715.2024.10843938>
- [31] Andronikidis, G., Eleftheriadis, C., Batzos, Z., Kyranou, K., et al. (2024). AI-driven anomaly and intrusion detection in energy systems: Current trends and future direction. In *2024 IEEE International Conference on Cyber Security and Resilience (CSR)*, London, United Kingdom, pp. 777-782. <https://doi.org/10.1109/CSR61664.2024.10679380>
- [32] Salem, S.A., Said, S.A., Nour, S.M. (2024). AI-driven anomaly detection framework for improving IoT system reliability. In *2024 IEEE Global Conference on Artificial Intelligence and Internet of Things (GCAIoT)*, Dubai, United Arab Emirates, pp. 1-8. <https://doi.org/10.1109/GCAIoT63427.2024.10833531>
- [33] Modesti, P., Golightly, L., Holmes, L., Opara, C., Moscini, M. (2024). Bridging the gap: A survey and classification of research-informed ethical hacking tools.

- Journal of Cybersecurity and Privacy, 4(3): 410-448. <https://doi.org/10.3390/jcp4030021>
- [34] Heim, M.P., Starckjohann, N., Torgersen, M. (2023). The convergence of AI and cybersecurity: An examination of CHATGPT's role in penetration testing and its ethical and legal implications. Bachelor's thesis, NTNU.
- [35] Hang, C.N., Yu, P.D., Morabito, R., Tan, C.W. (2024). Large language models meet next-generation networking technologies: A review. *Future Internet*, 16(10): 365. <https://doi.org/10.3390/fi16100365>
- [36] Wang, B., Cai, Z., Karim, M.M., Liu, C., Wang, Y. (2024). Traffic performance GPT (TP-GPT): Real-time data informed intelligent chatbot for transportation surveillance and management. In 2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC), Edmonton, AB, Canada, pp. 460-467. <https://doi.org/10.1109/ITSC58415.2024.10919936>
- [37] Magyar, A. (2024). Source code vulnerability analysis using GPT-2. In *Redefining Security with Cyber AI*, pp. 157-177. <https://doi.org/10.4018/979-8-3693-6517-5.ch009>
- [38] Sun, Y., Wu, D., Xue, Y., Liu, H., et al. (2024). GPTScan: Detecting logic vulnerabilities in smart contracts by combining GPT with program analysis. In 2024 IEEE/ACM 46th International Conference on Software Engineering (ICSE), Lisbon, Portugal, pp. 2048-2060. <https://doi.org/10.1145/3597503.3639117>
- [39] Espinha Gasiba, T., Iosif, A.C., Kessba, I., Amburi, S., Lechner, U., Pinto-Albuquerque, M. (2024). May the source be with you: On ChatGPT, cybersecurity, and secure coding. *Information*, 15(9): 572. <https://doi.org/10.3390/info15090572>
- [40] Dalalah, D., Dalalah, O.M. (2023). The false positives and false negatives of generative AI detection tools in education and academic research: The case of ChatGPT. *The International Journal of Management Education*, 21(2): 100822. <https://doi.org/10.1016/j.ijme.2023.100822>
- [41] Freitas, S., Kalajdjieski, J., Gharib, A., McCann, R. (2025, May). AI-driven guided response for security operation centers with Microsoft Copilot for Security. In *Companion Proceedings of the ACM on Web Conference 2025*, Sydney, Australia, pp. 191-200. <https://doi.org/10.1145/3701716.3715209>
- [42] Arsal, M., Saleem, B., Jalil, S., Ali, M., Zahra, M., Rehman, A.U., Muhammad, Z. (2024). Emerging cybersecurity and privacy threats of ChatGPT, Gemini, and Copilot: Current trends, challenges, and future directions.
- [43] Tanti, R. (2024). Study of phishing attack and their prevention techniques. *International Journal of Scientific Research in Engineering and Management*, 8(10): 1-8.
- [44] Qumer, S.M., Ikrama, S. (2022). Poppy Gustafsson: Redefining cybersecurity through AI. *The Case for Women*, pp. 1-38. <https://doi.org/10.1108/CFW.2022.000001>
- [45] Folorunso, A., Adewumi, T., Adewa, A., Okonkwo, R., Olawumi, T.N. (2024). Impact of AI on cybersecurity and security compliance. *Global Journal of Engineering and Technology Advances*, 21(01): 167-184. <https://doi.org/10.30574/gjeta.2024.21.1.0193>
- [46] Adamopoulos, I., Ilias, A., Makris, C., Stamatiou, Y.C. (2023). Intelligent surveillance systems on the Internet of Things based on secure applications with the IBM cloud platform. *International Journal on Information Technologies and Security*, 15(2): 59-74. <https://doi.org/10.59035/xvrs3592>
- [47] Permana, G.R., Trowbridge, T.E., Sherborne, B. (2022). Ransomware mitigation: An analytical investigation into the effects and trends of ransomware attacks on global business. https://osf.io/preprints/psyarxiv/ayc2d_v1.
- [48] Al-Sinani, H.S., Mitchell, C.J. (2024). AI-enhanced ethical hacking: A Linux-focused experiment. *arXiv preprint arXiv:2410.05105*. <https://doi.org/10.48550/arXiv.2410.05105>
- [49] Alshomrani, M., Albeshri, A., Alturki, B., Alallah, F.S., Alsulami, A.A. (2024). Survey of transformer-based malicious software detection systems. *Electronics*, 13(23): 4677. <https://doi.org/10.3390/electronics13234677>
- [50] Okazaki, N., Usuzaki, S., Waki, T., Kawagoe, H., Park, M., Yamaba, H., Aburada, K. (2024). Optimal weighted voting-based collaborated malware detection for zero-day malware: A case study on VirusTotal and MalwareBazaar. *Future Internet*, 16(8): 259. <https://doi.org/10.3390/fi16080259>
- [51] Arjunan, G. (2024). AI-powered cybersecurity: Detecting and preventing modern threat. *International Journal of Innovative Science and Research Technology*, 9(11): 1949-1955. <https://doi.org/10.5281/zenodo.14287585>
- [52] Fotoh, L., Mugwira, T. (2023). Exploring large language models (ChatGPT) in external audits: Implications and ethical considerations.
- [53] Henriques, J.P.M. (2023). Audit Compliance and Forensics Frameworks for Improved Critical Infrastructure Protection. Doctoral dissertation, Universidade de Coimbra (Portugal).
- [54] Zhang, J., Bu, H., Wen, H., Liu, Y., et al. (2025). When llms meet cybersecurity: A systematic literature review. *Cybersecurity*, 8(1): 55. <https://doi.org/10.1186/s42400-025-00361-w>
- [55] Bink, J. (2024). Personalized response with generative AI: Improving customer interaction with zero-shot learning LLM chatbots. Doctoral dissertation, Master Thesis Eindhoven University of Technology.
- [56] Dani, J., McCulloh, B., Saxena, N. (2024). When AI defeats password deception! A deep learning framework to distinguish passwords and honeywords. *arXiv preprint arXiv: 2407.16964*. <https://doi.org/10.48550/arXiv.2407.16964>
- [57] Nurmi, T. (2021). Network detection and reaction: Case study: Proof of concept for Vectra implementation. Master's thesis, Turku University of Applied Sciences.
- [58] Tsimenidis, S., Lagkas, T., Rantos, K. (2022). Deep learning in IoT intrusion detection. *Journal of network and systems management*, 30(1): 8. <https://doi.org/10.1007/s10922-021-09621-9>
- [59] Krombholz, K., Hobel, H., Huber, M., Weippl, E. (2015). Advanced social engineering attacks. *Journal of Information Security and Applications*, 22: 113-122. <https://doi.org/10.1016/j.jisa.2014.09.005>
- [60] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv: 1706.06083*. <https://doi.org/10.48550/arXiv.1706.06083>
- [61] Sarker, I.H. (2024). Generative AI and large language

- modeling in cybersecurity. In *AI-Driven Cybersecurity and Threat Intelligence: Cyber Automation, Intelligent Decision-Making and Explainability*, pp. 79-99. https://doi.org/10.1007/978-3-031-54497-2_5
- [62] Dangwal, D., Lee, V.T., Kim, H.J., Shen, T., et al. (2021). Analysis and mitigations of reverse engineering attacks on local feature descriptors. *arXiv preprint arXiv:2105.03812*. <https://doi.org/10.48550/arXiv.2105.03812>
- [63] Sufi, F. (2024). An innovative GPT-based open-source intelligence using historical cyber incident reports. *Natural Language Processing Journal*, 7: 100074. <https://doi.org/10.1016/j.nlp.2024.100074>
- [64] Ray, P.P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 3: 121-154. <https://doi.org/10.1016/j.iotcps.2023.04.003>
- [65] Kaur, D., Uslu, S., Rittichier, K.J., Durresi, A. (2022). Trustworthy artificial intelligence: A review. *ACM computing surveys (CSUR)*, 55(2): 39. <https://doi.org/10.1145/3491209>
- [66] Yenduri, G., Ramalingam, M., Selvi, G.C., Supriya, Y., et al. (2024). GPT (generative pre-trained transformer)—A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. *IEEE access*, 12: 54608-54649. <https://doi.org/10.1109/ACCESS.2024.3389497>
- [67] Blauth, T.F., Gstrein, O.J., Zwitter, A. (2022). Artificial intelligence crime: An overview of malicious use and abuse of AI. *IEEE Access*, 10: 77110-77122. <https://doi.org/10.1109/ACCESS.2022.3191790>
- [68] Pleshakova, E., Osipov, A., Gataullin, S., Gataullin, T., Vasilakos, A. (2024). Next gen cybersecurity paradigm towards artificial general intelligence: Russian market challenges and future global technological trends. *Journal of Computer Virology and Hacking Techniques*, 20(3): 429-440. <https://doi.org/10.1007/s11416-024-00529-x>
- [69] Liu, X., Zheng, Y., Du, Z., Ding, M., Qian, Y., Yang, Z., Tang, J. (2024). GPT understands, too. *AI Open*, 5: 208-215. <https://doi.org/10.1016/j.aiopen.2023.08.012>
- [70] Yu, Y., Zhuang, Y., Zhang, J., Meng, Y., et al. (2023). Large language model as attributed training data generator: A tale of diversity and bias. *Advances in neural information processing systems*, 36: 55734-55784.
- [71] Carnat, I. (2024). Human, all too human: Accounting for automation bias in generative large language models. *International Data Privacy Law*, 14(4): 299-314.
- [72] Bhattacharya, P., Prasad, V.K., Verma, A., Gupta, D., Sapsomboon, A., Viriyasitavat, W., Dhiman, G. (2024). Demystifying ChatGPT: An in-depth survey of OpenAI's robust large language models. *Archives of Computational Methods in Engineering*, 31(8): 4557-4600. <https://doi.org/10.1007/s11831-024-10115-5>