



Generative AI revolution in cybersecurity: a comprehensive review of threat intelligence and operations

Mueen Uddin¹ · Muhammad Saad Irshad² · Irfan Ali Kandhro³ · Fuhid Alanazi⁴ · Fahad Ahmed² · Muhammad Maaz² · Saddam Hussain⁵ · Syed Sajid Ullah⁶

Accepted: 29 March 2025 / Published online: 7 May 2025
© The Author(s) 2025

Abstract

Cyber threats are increasingly frequent in today's world, posing challenges for organizations and individuals to protect their data from cybercriminals. On the other hand, Generative Artificial Intelligence (GAI) technology offers an efficient way to automatically address these issues with the help of AI models and algorithms. It can work on more critical security aspects where human intervention is required and handle everyday threat situations autonomously. This research paper explores GAI in enhancing cybersecurity by leveraging AI Models and algorithms. GAI can autonomously address common security issues, detect novel threats, and augment human intervention in critical security aspects. Moreover, this research study also highlights autonomous security enhancements, improved security posture against emerging threats, anomaly detection, and threat response. Besides this, we have discussed the GAI limitations, such as occasional incorrect results, expensive training, and the potential for misuse by malicious actors for illegal activities. This research study also provides valuable insights into the balanced adoption of GAI in cybersecurity, ensuring effective threat migration without compromising system integrity.

Keywords Generative AI · Security · Artificial intelligence · Machine learning · Natural language processing · Learning systems · LMM

1 Introduction

The idea of building robots that could mimic human intelligence was first explored by researchers in the middle of the twentieth century, making the official beginning of the history of Artificial Intelligence (AI). The field of AI formally started with the coining of the term “artificial intelligence” in 1956 during a conference held at Dartmouth College. Early attempts were centered on deciphering natural language and logical puzzles. Due to a lack of computing capacity, progress was gradual, but by the 1980s, developments in computing and algorithms had produced expert systems that could carry out tasks. With the introduction of huge data, potent processors, and sophisticated machine learning techniques—intense learning—the field of AI underwent a boom in the twenty-first century,

leading to advances in autonomous systems, gameplay, picture and speech recognition, and more. AI is now integrated into various aspects of daily life, from virtual assistants to medical diagnostics. However, the idea of creating machines capable of intelligent behavior can be traced back much (Grzybowski et al. 2024).

“Generative AI” describes computer methods that use training data to produce new, meaningful output, such as text, images, or audio. The way we work and communicate is changing due to technologies like GPT-4, Copilot, and DALL-E 2. GAI has far-reaching applications across various domains, including practical tasks like question-answering and creative pursuits like writing, drawing and music composition. It has already demonstrated value in routine tasks like recipe generation, health advice and IT support. Industry estimates suggest that GAI may replace 300 million knowledge-based jobs and boost the global economy by 7% in 2022 (Zhou and Lee 2024; <https://www.goldmansachs.com/insights/articles/generative-ai-could-raise-global-gdp-by-7-percent>). However, this technology also represents significant opportunities and challenges for responsible and sustainable use, particularly in business, information systems, arts and design (Kim and Choi 2024). In the telecommunications sector, generative AI can optimize network efficiency and predict path loss (Vu et al. 2024). It can potentially improve patient care, research and diagnosis (Sai et al. 2024a). In education, it can provide personalized learning experiences (Kharrufa and Johnson 2024). However, it also poses additional threats to protected health data and cybersecurity. As our reliance on digital technology grows, cybersecurity is becoming increasingly crucial.

Effective cybersecurity measures include firewalls, encryption, multi-factor authentication and frequent software updates. The use of GAI in cybersecurity is gaining attention, particularly with the rise of ChatGPT. Overall, GAI can potentially revolutionize various industries and aspects of our lives. However, addressing the challenges and risks associated with this technology is essential to ensure responsible and sustainable use. Some view GAI as beneficial and detrimental as threat actors begin utilizing it in cyber-attacks (Sai et al. 2024b). The cyber security domain is one of the vital areas of the IT industry where GAI is emerging to show a more significant positive impact (Kam et al. 2024). The potential of AI being exploited maliciously in cybersecurity is growing as AI and machine learning use expands quickly, particularly with the emergence of GAI (Wang 2024; Zhang et al. 2024). Moreover, GAI methods allow security products to detect phishing attacks with high precision and recall (Kaushik et al. 2024). The current security methods are based on GAI, and there is a strong knowledge of defensive attacks (Vu et al. 2024; Takale et al. 2024). The use of GAI in networking fields such as network segmentation, firewalls, and network protection help many tasks, from preparing to mechanized pressure testing, enabling rapid visibility and potential data vector loss. providing an at-a-glance solution (Sriram 2024).

GAI can simulate real-world scenarios, facilitating security system testing, performance assessment, fault identification and overall system readiness enhancement (Sabherwal and Grover 2024). Moreover, GAI exhibits threat intelligence capabilities, enabling efficient identification and response to incoming threats (Vu et al. 2024). While GAI still requires advancement in robustness and predictive accuracy, its potential to significantly aid various industries, particularly cybersecurity, is undeniable. GAI is emerging as a promising tool, poised to revolutionize cybersecurity. This work highlights the need for cybersecurity systems to consider the limitations and flaws that exist within such systems. The survey provides limitations of GAI, such as occasional incorrect results, expensive training, and

the potential for misuse by malicious actors for illegal activities. This survey also intends to provide current and future researchers, policymakers, and organizational leaders with an understanding of the opportunities and risks associated with the development of GAI. It also further defines the long-term vision of developing cybersecurity, specifically focusing on the effective and meaningful role of innovation and sustainability in implementing cybersecurity programs. Table 1 shows the comparison of the proposed survey with state-of-the-art surveys. Finally, Fig. 1 provides the organization of the survey.

2 An overview of GAI

GAI is endowed with creative potential as a family of artificial intelligence models. Using the training data, these models can create new content like text, images, music, or even code (Ooi et al. 2023). They generate entirely new instances, which are similar to the training data but are unique in their way, confirming their capacity to generate original pieces. This is accomplished using the patterns and structure extracted from the training data, as suggested in Bandi et al. (2023). The power of GAI is found in their ability to combine and manipulate large, superior datasets and derivatively generate content from that data. This technique employs high forms of neural network architectures and algorithms as constant optimization (Kumar et al. 2023). The output that the system produces can sometimes be perceived as quite natural and was criticized for having creative aspects, owing to the substantial amount of data that goes into training GAI's algorithm and the arbitrary way the algorithm produces the output (Babcock and Bali 2021). Advanced deep-learning techniques GAI-enhanced architectures such as Types of GAI: Generative Adversarial Networks, Variation autoencoders, Transformers and others can use the database by extracting patterns and structures in them (Gupta et al. 2024; Zhu et al. 2011). After a model is trained, it is expected to generate new content by drawing samples from the learned distribution.

The significant promotion brought by GAI can be attributed to ChatGPT and DALLE (Dhoni and Kumar 2023). ChatGPT, a free Chabot developed by Open AI, can provide answers to most of the queries it asks. Its ability to generate unique content, such as computer codes, essays, and poems, among other stuff, quite quickly has contributed to its growing popularity. DALLE, another GAI tool, is admired for its capacity to create realistic images and art from natural language inputs. These GAI tools, including ChatGPT and DALLE, are predicted to revolutionize the nature of work in many areas permanently. However, the scale of that impact and the involved threats remain unclear. In the context of security, the evolution of GAI from early experimentation with Gas and VAEs to breakthroughs in deep-fakes and AI-generated malware, rapid progress in AI-powered security threats, widespread adoption of GAI in security, and growing concerns about AI-driven security risks and ethical implications as discussed in Fig. 2, is a significant development.

3 Applications of GAI in cybersecurity

There are several applications of GAI in the field of security (See Figs. 3 and 4). In this section, we present a detailed study of the same.

Table 1 Summary of related surveys on GAI security

References	Year	Method	Limitation	Dataset	Focus	Contributions
Wang et al. (2023)	2023	AIGC	Balancing privacy with data utility	CIFAR-10	Security and Privacy concerns on GAI contents	Evaluate the security and privacy concerns on GAI contents Analyze GAI contents from the perspective of information security properties
Zhao et al. (2024)	2024	GANs, AEs, VAEs and DMs	Managing unclean and irrelevant sensor data	MNIST	GAI for securing Physical Layer Communication	Offer a comprehensive review on the uses of GAI in physical layer security Cover GAI architecture, classification, and basic ideas Introduce several security characteristics including confidentiality, authenticity, availability, and integrity
Colda et al. (2024)	2024	GANs and RNN	Hallucinations and fabrications	Institution data	Security and Privacy concerns in GAI	Presents a comprehensive overview of GAI Evaluate the security and privacy concerns of GAI from perspective of user, institutional, regulatory, ethical, and technical factors
Yigit et al. (2024)	2024	Google Gemini, ChatGPT and DALL-E	Hallucinations significant concern	HumanEval	GAI in cybersecurity	Presents a detail overview of various applications of GAI in cybercrimes Reviewed current state of the art deployment of GAI
Chen and Esmailzadeh (2024)	2024	GAI	regulatory and policy challenges	EHRs	GAI in medical field from security and privacy viewpoint	Explore various applications of GAI in the medical field identified potential threats to security and privacy within each phase of the life cycle of such systems
Proposed	2025	GAI	–	–	GAI in cybersecurity	Analyzed the current state of GAI in the healthcare industry Highlights the need for cybersecurity systems to consider the limitations and flaws that exist within such systems Provides limitations of GAI, such as occasional incorrect results, expensive training, and the potential for misuse by malicious actors for illegal activities Outlines a strategic vision for the future of cybersecurity, emphasizing the crucial role of innovation and resilience in protecting our digital infrastructure

3.1 Password guessing attacks

Guessing password attacks presents a new frontier in cybersecurity in the GAI. Traditional methods are brute force attacks and rainbow table attacks. GAI can create robust methods to identify patterns and structures. Thus, enabling it to produce new passwords or cracking process password guessing and aiding password security assessments. A phishing attack is one of the use cases in this context, which is a significant threat to cyber security as hackers are always on the lookout for new strategies to make people reveal their personal details through email. The authors in Eisenberg (2011) have also attempted to analyze the performance of LLMs in the context of identifying phishing emails. The authors experimented on three LLMs, namely GPT-3.5, GPT-4, and customized ChatGPT, with phishing and genuine email datasets. The study indicated that the LLMs perform well in the recognition of phishing emails, but their scores are not consistent. The authors have evaluated the performances of models and the possibility of improving email security.

3.2 Generating better passwords

It can also capture and analyze various dynamic patterns involving the user's passwords, including login and authentication failure patterns (Khan, et al. 2024). This allows GAI to recognize any suspicious activity that might be related to weak passwords or unauthorized access to the system, hence reducing security infringements. Furthermore, GAI can create much stronger and more secure passwords since it is constructed by analyzing prior patterns of passwords and avoiding predictable patterns, thus reducing the chances of getting a password cracked.

3.3 Detecting GAI text in attacks

LLMs like ChatGPT, together with Google lambda, enable the detection and watermarking of AI-generated text, which helps identify both phishing emails and polymorphic programs (Bryce et al. 2024). The analysis of suspicious email senders, together with their domains and text links to potentially dangerous websites, enables LLMs to identify and prevent cyber threats effectively.

3.4 Provide illustrations of adversarial attacks

The purpose of adversarial attacks is to feed incorrect information to machine learning systems to deceive them. Image recognition models, along with NLP models, represent some of the targets where these attacks are considered appropriate (Motlagh et al. 2024).

3.5 Simulated attacks

The assessment of machine learning model security and resilience through simulated attacks presents itself as a strong evaluation method. Through simulated attacks on the system, researchers discover weaknesses that lead to improved security recommendations for future system development (Agrawal et al. 2024). The evaluation practices benefit from

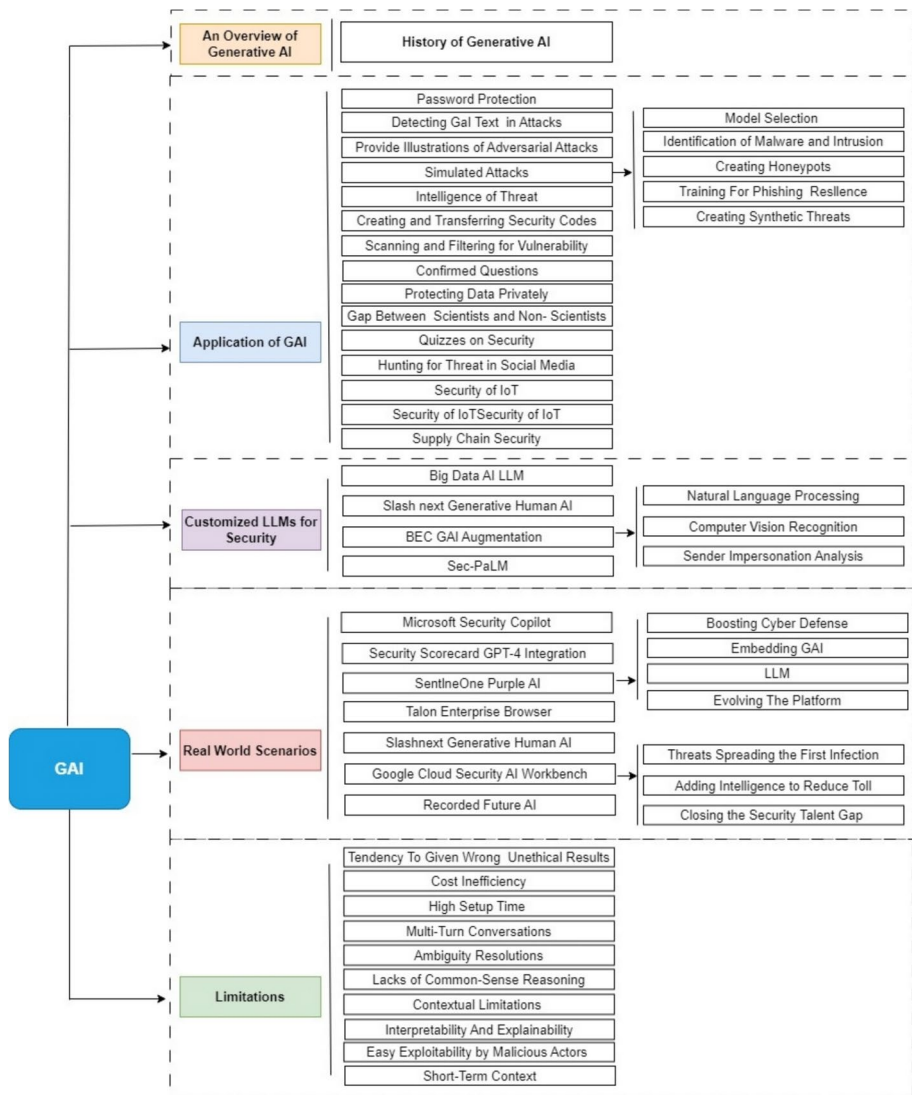


Fig. 1 Applications of GAI, customized LLMs and real-world applications

this method, which enables complete assessments of machine learning systems so security-enhancing practices can be deployed.

3.5.1 Model selection

GAI models should be trained on known attack vectors or use various attack approaches to create realistic attack cases for conducting red teaming activities (Novelli et al. 2024). These simulations helped the red team determine security flows in origination and measure the security measures the red side had in place. Furthermore, GAI models can create various

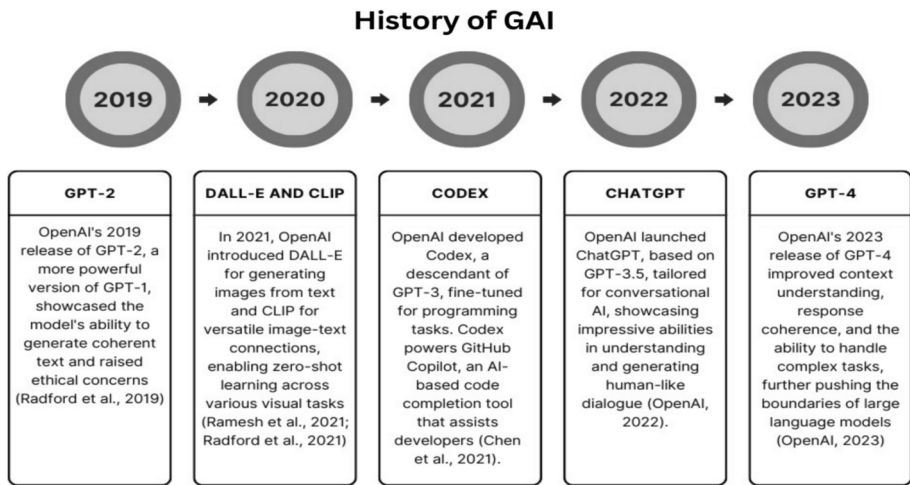


Fig. 2 Timeline of generative AI with years

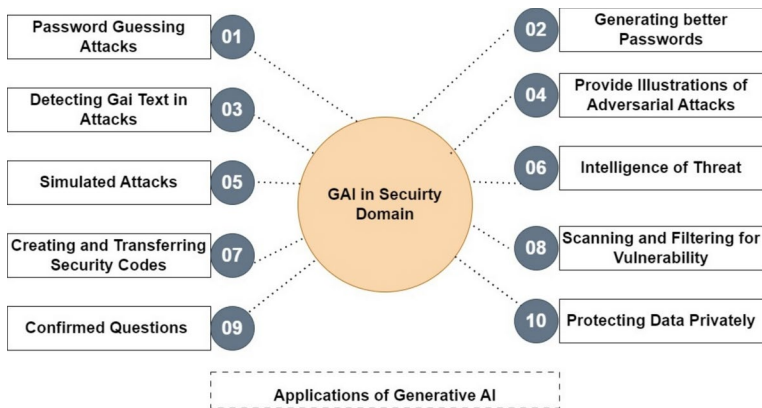


Fig. 3 Application of GAI in the security domain

attacks such as emails, social engineering situations, and exploitation (Gibert 2024). The comprehensive detection of vulnerabilities through GAI models in the red team exercise is useful in enabling organizations to effectively advance their security loop whole and enhance the defense system. Additionally, by doing this, the security systems may shift from reactive to proactive, allowing them to anticipate dangers before they arise. This may result in even more robust security. With GAI, one can create virtual spaces cyber ranges that can Give specialists in cyber security practical training (Eibeck et al. 2024); a talent intelligence platform that uses GAI to assist businesses in training their staff in new cyber security responsibilities and abilities is one of the practical use cases. Within a cyber range, GAI can generate realistic network topologies, traffic patterns, and attack scenarios (Agrawal et al. 2024). This can help security personnel try various approaches, hone their skills in a safe setting, and obtain real-world experience dealing with cyber-attacks.

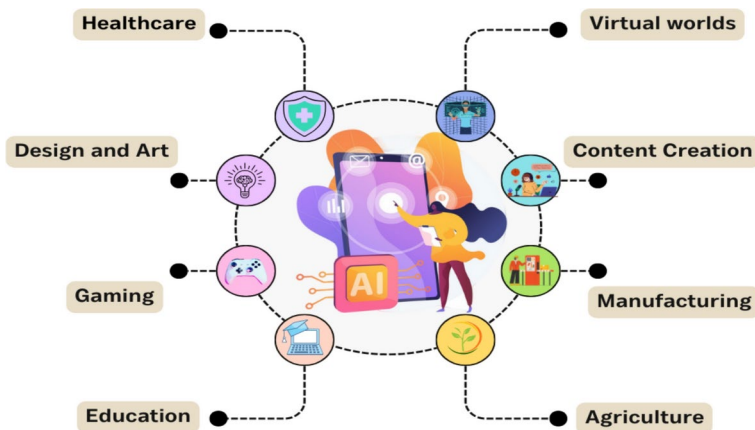


Fig. 4 Applications of GAI

3.5.2 Identification of malware and intrusion

By learning from an enormous collection of malware data, one can use generative models to produce realistic malware representations. GAI for Cyber Security: An Analysis of the Potential of Security (Yao et al. 2024). Administrators can leverage synthetic samples generated by GAI to test and enhance their malware detection system, improving their ability to recognize and manage diverse malware stains. Read-world examples include sentinel One Purple AI and Google Cloud Security AI Workbench, which neutralize GAI methodology to offer effective threat-hunting capabilities. Moreover, GAI can identify new or altered malware variants by discovering fundamental characteristics and patterns shared by multiple malware families (Yosifova 2024). This enables the GAI model to identify and categorize unknown malware based on their resemblance to recognized pattern three by strengthening the security system resilience. Similarly, security officials can train GAI algorithms on standard network traffic data to generate a representation of normal network behavior, enabling the detection of anomalies and potential security breaches (Yigit et al. 2024).

3.5.3 Creating honeypots

GAI can be used to entice attackers to build attractive decoy networks or systems, such as phony websites and applications. GAI can produce misleading content in addition to honey pots. Chatbots that support natural language understanding (GAI), such as ChatGPT and Meta LLaMA (Falade 2024), can converse with attackers humanly and obtain meaningful insights into their actions. Then, by using this private information about the attackers, essential insights into their methods of operation can be obtained. The GAI models can create dynamic honeypots that reflect the most recent attack trends by training on real-time threat intelligence and attack datasets. Therefore, the current honeypot can be automatically created and adjusted using GAI to the current threat environment, keeping the honeypots current and ready to respond to new attacks. Following an attacker's interaction with the system, organizations can observe the attacker's behavior and gain important information

about the type of attack and the tools and strategies the attacker is using. Later, the company can utilize this information to improve its security measures (Mavikumbure et al. 2024).

3.5.4 Training for phishing resilience

LLMs that create simulated phishing messages, such as ChatGPT, LLaMA, and others, can offer increased frequency and usefulness of phishing resistance training for staff members inside a company (Morreel et al. 2024). This can serve as an excellent substitute for the antiquated cyber threat awareness training courses that are currently in use. LLMs can be used to construct phishing emails. The employees must identify any suspicious indicators, like typos or strange email addresses, and report their conclusions regarding whether the email is phishing. Alternately, a baiting attack scenario should be set up, and staff members should be asked to identify whether the information or activity being presented to them is accurate (by email); it is unclear whether this is a baiting attempt. An easy and affordable phishing resilience training program can be swiftly created to educate an organization's personnel by putting such messages produced by ChatGPT (or any other successful LLM) into an email marketing tool (Xia et al. 2024), as can be seen in Fig. 2.

3.5.5 Creating synthetic threats

Fake threat environments can be generated using GAI models to test and assess a system's security. By learning from the real-life dataset, GAI algorithms can generate synthetic scenarios that closely resemble real-world tracks. They have the potential to generate artificial malware samples, which can be utilized to educate a GAN about the traits and designs of malevolent code. This process of creating new malware variants is crucial as it allows cyber security experts to test their systems against emerging threats. By using this knowledge to create new malware variants with comparable characteristics, experts can gain important insights on how to strengthen the security of their systems (Gupta et al. 2024).

3.6 Intelligence of threat

GAI trains its learning algorithm on a huge dataset, which allows it to recognize patterns and indicators of compromise and recognize and neutralize threats before they can breach the front-line defense. In certain circumstances, further cyber security technologies that could be needed to increase the security of the current system can also be predicted by GAI. This part of GAI differs from the proactive approach in that it generally looks at threats before identifying a particular system's pertinent characteristics. The proactive approach involves anticipating threats based on general characteristics, while the organization-specific methodology focuses on anticipating internal dangers based on the unique characteristics of a particular system. Beforehand, Sentinel One Purple AI, Slash Next Generative Human AI, Google Cloud AI Workbench, and other real-world security systems can display threat intelligence and respond more proactively to ever-emerging threats. GAI's capacity to analyze threats from both a wider and a closer perspective makes it a more effective defense tool against current and potential dangers (Sai et al. 2024c). Traditional cyber security methods rely on signature-based detection, manual analysis, and education, which are resource-intensive, limited by human cognition, and prone to human error. In contrast,

AI-powered approaches utilize pattern recognition, predictive analytics, automation, and behavioral analytics to detect and respond to threats more efficiently and effectively, as discussed in Table 2.

3.7 Creating and transferring security codes

Code that increases system security can be easily generated and transferred using LLMs such as ChatGPT (Surameery et al. 2023). Let us say, for instance, that a phishing attempt effectively revealed numerous employees' credentials.

The employees who clicked on the phishing email are known throughout the business, but it is unclear whether they unintentionally executed the malware intended to steal their login information. A Microsoft 365 Defender Advanced Hunting query can be used to determine the ten most recent logins made by the recipients of phishing emails within 30 min of receiving the malicious emails to investigate this case (Alvarez 2013). Subsequently, these inquiries can be employed to detect any login activity linked to the compromised credentials. Here, you can utilize ChatGPT to ask Microsoft 365 Defender Advanced hunting to look for attempts at access to the compromised email accounts (see Fig. 5). This can assist in locating and thwarting attackers and provide users with information on whether they should update their login credentials. As a result, there may be a quicker reaction time to the cyber-attack situation. Suppose the system on which the Microsoft 365 Defender hunting query is to be run does not run the KQL programming language. In that case, ChatGPT's feature using its underlying Codex model to take a piece of code written in one programming language and convert it to another can be used to do a programming language style transfer (Sindiramutty 2023).







3.8 Scanning and filtering for vulnerability

By training on datasets including false positives, the GAI models can acquire the ability to provide GAI and assist in lowering false positives in vulnerability scanning by using filters or rules that differentiate between real vulnerabilities and benign ones. By using this GAI capability, security teams can spend less time looking at false positive alarms and more time addressing actual vulnerabilities. Additionally, GAI models can be trained to consider context while filtering vulnerabilities by adding factors like system settings, network topology, user access privileges, asset criticality, etc., to training data. This can help GAI rank the vulnerabilities according to how likely they are to cause harm and how easily they can be exploited in a particular situation (Hu et al. 2024).

3.9 Confirmed questions

LLMs such as ChatGPT, Meta LLaMA, and Google LaMDA can be utilized to generate queries for threat hunting, like those for tools for researching and detecting malware, such as YARA (Ding et al. 2020), improving the system's efficiency and response time. This can free up cyber security personnel to concentrate more on other crucial security facets while GAI quietly and quickly detects and neutralizes possible threats. Because of this LLM's utility, security systems are more resilient in an environment where hostile activity is constantly changing. Moreover, using historical data, GAI models can be taught to recognize

Table 2 The AI factor in cyber security: enhancing detection, response, and prevention

References	Threat(S)	Old-approach	New-AI approach	Startups leveraging the AI approach
Agrawal et al. (2024)	Malware	Anti-virus etc. flag attacks Covers known vulnerabilities	Uses signature-based detection to new attacks Can cover “zero-day” exploits	 SentinelOne
Yu et al. (2021)	DDoS (Distributed Denial Of Service)	Analysts monitor network traffic to sport an ongoing DDoS attack Resource intensive, limited by human cognition, reactive	Algorithms auto-detect abnormal network resource allocation Efficient analyst resources, automated, faster response	 Vectra  Zenedge
Vu et al. (2024)	Iota & Endpoints	Manual device-level security updates through the cloud Ad-hoc security, ineffective at scale	Network-level behavior-analytics and entity-anomaly-detection Real-time security, effective at scale	 Cujo
Zhang et al. (2024)	Social Engineering	Education on digital hygiene and counteracting hackers’ tactics Prone to human error	Education + social-biometrics and user-anomaly-detection Less prone to human error	 Tanium  BehaviorSec

the familiar pattern of typical system behavior and network traffic. These models can then be used to create queries that find the differences and irregularities from the anticipated behavior, which may be utilized to identify any security lapses, efforts to exfiltrate data, or unapproved access to the system or network (Jiang et al. 2021). Through the process of training GAI models on real-time threat data, security warnings, or incident reports, the system can acquire the ability to dynamically modify its generated queries in response to changes in the threat landscape (Sun et al. 2018a). Thus, these models' queries can consider the most recent indicators or developing threat trends. As a result, the cyber security teams can detect new threats more rapidly by creating pertinent questions that might not have been possible with static query generation. Some of the most recent threat-hunting technologies that use an LLM to increase security analyst productivity include Sentinel One Purple AI, Google Cloud Security Workbench, and Microsoft Security Copilot.

3.10 Protecting data privately

Organizations can use shared synthetic data generated by training GAI models for various objectives, such as training machine learning models to detect online fraud (Mozolevskyi and AlShikh 2024), recommending tailored products to predict loan eligibility, and more. Due to privacy concerns and data protection rules, this may limit the exchange of customer data, which may improve data privacy protection. Moreover, privacy-preserving machine learning models can be created with GAI, for instance, can be used to enhance federated learning methods to produce artificial intelligence-generated data on edge devices that reflect regional patterns and traits. By doing this, users' privacy can be protected, and collaborative learning models can be trained without revealing sensitive data to the central server (Gadre et al. 2024). The visual components and interaction patterns of interfaces can be taught to GAI models through training on a collection of user interface designs. Using strategies like hiding or masking sensitive data fields and offering privacy-focused options when sharing data, it is possible to reduce exposure to sensitive information by developing privacy-aware user interfaces (Sannon and Forte 2022; Chukwurah 2024). Users may benefit from improved data privacy as a result. Talon is one of the more recent instances of a corporate browser that has integrated Microsoft Azure OpenAI into its architecture to improve organization security (Zou et al. 2024).

3.11 Closing the gap between scientists and non-scientists

The capacity to express one's thought process allows LLMs such as OpenAI ChatGPT, DeepMind Chinchilla AI Meta LLaMA, and others to delve into and convert the functionality of a variety of technical files, such as source code and configuration files into plain language (Zoubi and Mohammed 2024). This can make it possible for those with little technical expertise to comprehend the structure, goal, and possible repercussions of these files and their inner workings. With this knowledge, they can make more informed technological decisions and avoid unintentionally causing cyber security problems. This capability of GAI may also assist an organization in bridging the knowledge gap between non-experts and professionals in cyber security by offering easily understandable explanations for various cyber-related concerns.

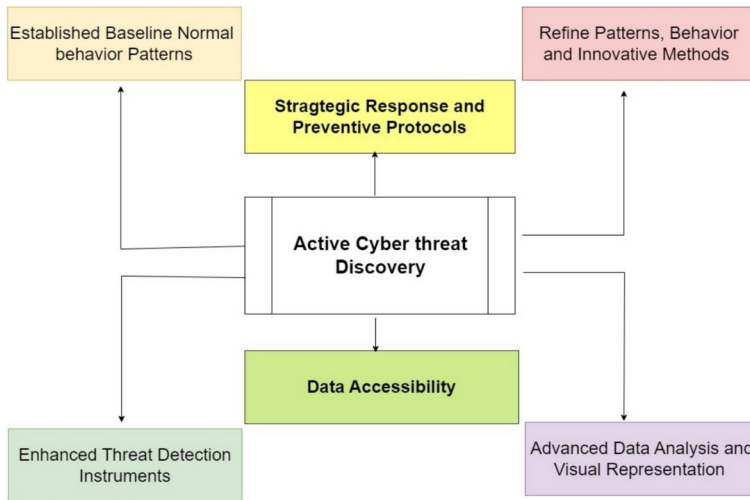


Fig. 5 Threat hunting in cybersecurity

3.12 Quizzes on security

The cyber security posture is significantly impacted by the danger posed by third parties (almost 60% of data exposures). Are the consequence of hacked third-party vendors) (Alawida et al. 2024). Using security questionnaires is one of the most effective techniques for identifying security risks across our vendor network. One can use LLMs like ChatGPT to draft these surveys rather than creating them from scratch. A sizable dataset of inquiries and answers about security can be used to train GAI models. As a result, the model can produce queries about different facets of risk management, cyber security, and compliance (Li et al. 2024). As a result, security personnel may save time as they will only need to spend some time modifying these questionnaires to increase their accuracy. Furthermore, by training the GAI model on a dataset of current security information, including actual examples of security events, they can use GAI models to modify and enhance the questionnaires they generate by the present danger scenarios. By doing this, a company can maintain a strong security posture and avoid possible security risks.

3.13 Hunting for threats in social media

Social media threat hunting collects information from social media platforms to find weaknesses and possible risks. Social media channels are searched for particular keywords to identify prospective phishing attacks or potential exposure points of sensitive data based on the organization's view of such data for social media threat hunting; LLMs such as ChatGPT, LLaMA, Chinchilla AI, or LaMDA can be employed (Pendzel et al. 2024). Social media data from various channels can be gathered, transferred to large language Models (LLMs) for analysis, and then used to generate intelligent prompts to obtain relevant information (Kineber 2024). GAI model can be trained to analyze social media post's text, images, and user interactions, understanding the context, tone, and purpose to identify potential threats or harmful actions, such as hate speech, extremist content, or unlawful

activities (Erbati 2024). This allows proactive risk monitoring and mitigation. Additionally, GAI models can learn from databases of known harmful or trustworthy users, allowing them to analyze user reputation and trustworthiness on social media platforms. This enables the generation of social media account credibility indicators, facilitating the identification and reporting of questionable accounts. Furthermore, GAI has potential applications in various fields to enhance efficiency and save time.

3.14 Security of IoT

The IoT settings and gadgets can be made more secure with the help of GAI. The Internet of Things is a network of networked gadgets that can talk to each other and share information online. These gadgets can include industrial sensors, medical equipment, and smart household appliances (Alwahedi et al. 2024). There are now more serious security problems because of the quick spread of IoT devices. Due to their lack of strong security features, many IoT devices are open to hackers. Using anomaly detection and behavioral analysis, GAI may aid in IoT security. It can create models of typical behavior for various IoT devices and settings. Generative models can identify deviations from standard patterns in data from IoT sensors and devices (Grover et al. 2021). These deviations indicate possible security breaches or unauthorized access, enabling the organization to promptly counteract the cyber threat and bolster security, as mentioned in Fig. 5.

Additionally, GAI can be very helpful in identifying and thwarting Internet of Things threats. It can simulate different attack scenarios to find potential weaknesses in IoT systems, enabling security experts to create more effective plans to stop and lessen cyberattacks that target Internet of Things devices (Mitra et al. 2024). Moreover, it may help in searching for the possible loopholes that may exist in the embedded hardware and software of the Internet of Things. It can generate a fake code to simulate how the device responds to inputs and potential means of exploitation. Therefore, it can be concluded that the security of devices' firmware and software can be improved to meet the company's security standards. To evaluate the effectiveness of access control measures, GAI can assist in the authentication process to enable safe authentication methods for IoT devices, such as biometric or multi-factor authentication (Zhou et al. 2023). Besides, it can also assist in the wonders of collecting and masking IoT data to protect users' privacy and create fake data that is realistic and aids in the privacy of sensitive data. It is also seen that organizations can enhance IoT security by triggering GAI, the dependability of IoT devices on digital attacks, and correspondingly, the aptitude to identify and address security breaches and deliver buyers and their networked devices a safer prospect (Khoo et al. 2022).

3.15 Deepfake detection and prevention

Deepfakes are synthetic images, sounds, or videos created with sophisticated computer tools. They can depict unreal events. They can be good, for instance, as in a movie, but they are negative in cases such as falsely disseminating information or invading people's privacy. This is something that worries people a lot since they could be exposed to fake news, cheating, or privacy infringement (Xu et al. 2022). This kind of fraudulent activity can be detected and avoided by using intelligent computer tools. For instance, algorithms can then recognize fails in structure-like facial characteristics or lighting patterns in forged artificial

material through the study of large volumes of authentic and fake information (Cox 1999). They can also develop methods that can determine if such photographs or videos are real or tampered with. Additionally, they can create pictures and videos showing the fake work to spread awareness. However, with the constantly improving fake creation methods, AI computer technologies should be up to date to prevent the creation of unauthentic content on the internet (Hassija et al. 2021). In addition, it noted that these tools can inform people on how to identify it (Bansal et al. 2019). Particularly, as nearly all information is increasingly likely to be processed with the help of artificial intelligence in the moments that matter most, we must have corresponding intelligent tools to allow us to trust the information we encounter.

3.16 Supply chain security

Supply chains refer to the multiple processes of delivering products from the basic materials to transportation to consumers. Maintaining the integrity and confidentiality of information through this process is very important to avoid interception or tampering. Smart computer tools such as GAI can be used to produce codes or holograms for each product to facilitate their recognition of the genuineness (Jackson et al. 2024). Furthermore, GAI can design tamper-evident packaging and supply chain mapping that will make it easier to monitor the processes to identify any susceptibility to tampering (Fosso Wamba et al. 2023). Also, by using historical data, GAI can predict future risks and, consequently, make appropriate adjustments. It is important to retain supply chain security to increase consumer trust and assurance of standards and quality. It has the capabilities of eradicating fake information cognitively avoiding 'fake news' and fake data and authenticating data at every chain of supply (Nguyen et al. 2024a) (Fig. 6).

3.17 Blockchain security

Blockchain on its own is a secure and decentralized system, with the help of practices such as GAI, the security can be even more strengthened. Through the application of GAI, smart

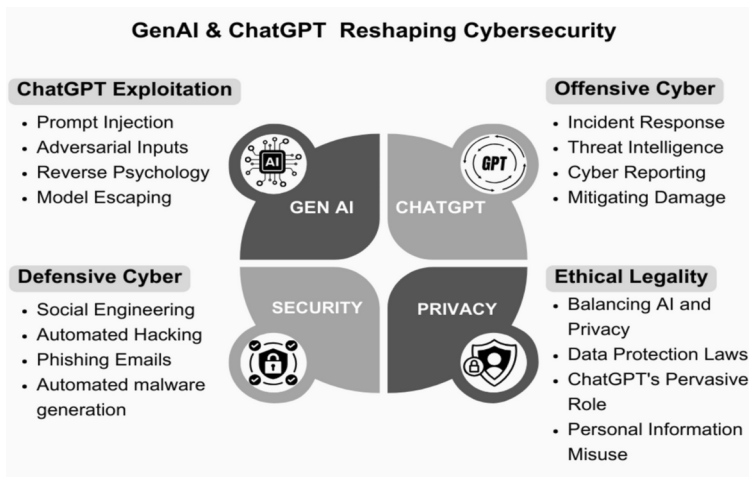


Fig. 6 Gen AI and ChatGPT reshaping cybersecurity

contracts can be analyzed before deployment to check for weaknesses and test whether the codes are free from errors. In addition, GAI can also create strong and diversified private keys for blockchain users, informing them on security aspects (Rabieinejad et al. 2023; Sun et al. 2018b). GAI is also able to supervise the blocks, identify deviations from the norm, inform the users of possible illegal actions, forecast possible negative consequences, and act correspondingly. Furthermore, GAI can help to make blockchain more secure or accurate by adding codes or signatures that will verify it. Table 3 also lists several stack technologies integrated with GAI based on the knowledge of their enterprise ROI, risk, and security aspects. Moreover, each layer as a concept has its advantages and disadvantages. While being rated as moderate threats within the interface layer, due to result in timely injections as well as hallucinations, the proposed solutions offer a high potential of ROI through boosting customer interactions.

While the application layer has great potential for ROI, there are significant threats inherent to adversarial attacks and decision biases, which can undermine model trustworthiness. The data layer, essential for optimal model performance, offers significant returns on investment in areas like fraud detection. However, it is vulnerable to issues such as data poisoning and leakage. To maximize the platform's potential while mitigating hazards, tailored security measures addressing the distinct risks of each layer are crucial. By leveraging GAI, blockchain technology can become even more secure and reliable. Additionally, it verifies that nothing has been altered without authorization by comparing the data across the blockchain (Wang et al. 2019). Not only is it safer when GAI is used with blockchain, but it also gains greater credibility. They are also more inclined to adopt blockchain for other purposes when they know that GAI contributes to security measures. This facilitates the application of blockchain's advantages and problem-solving skills by organizations and other groups (Sharma et al. 2021).

Table 3 Securing the GAI technology stack

References	Layer	Platform components	Enterprise Roi potential	Enterprise risk potential	Sample security product categories
Zhao et al. (2024)	Interface layer	User interfaces Developer portals	Significant (e.g. Customer Services Chatbots)	Moderate (e.g. Prompt injections, Hallucinations)	Data loss protection LLM firewalls
Erbati (2024); Zhao et al. (2024)	Application layer	ML orchestration ML deployment Experimentation testing	Moderate (e.g. Business Intelligence, Analytics)	Significant (e.g. Adversarial Attacks, Decision Bias Injection)	Treat intelligence Model testing Supply chain security Observability & explainability
Hoofnagle (2019); Liu et al. (2022); Zhao et al. (2024)	Data layer	Data ingest Data pipelines Data management	Significant (e.g. Fraud Models, Risk Scoring)	Moderate (e.g. Data Leakage, Inference Attacks, Data Poisoning, Model Mentoring)	Data security Posture management Data quality Data access Privacy & governance LA backup & recovery

4 Customized LLMs for security

4.1 Big data AI LLM

This section discusses Big Data's unique computer application designed to support enterprise data security. It examines and arranges all an organization's data to ensure that it is secure and easily accessible. Whether the data is kept in the company's systems or on the internet, it can function with various kinds of data. One neat feature of Big Data is its ability to protect individuals' private information (Ma et al. 2023). This implies it can carry out its duties without disclosing private information to servers or other programs. Additionally, it features an assistant named Big Chat that responds to inquiries and assists in adhering to data security regulations such as GDPR and CCPA (Hoofnagle 2019). Sometimes, issues arise if this algorithm is trained with incorrect data. For instance, if trained on confidential data, it can unintentionally divulge that data or facilitate hacker access to it. However, businesses can utilize big data to categorize their data according to criteria and level of sensitivity, ensuring that the program is trained exclusively on safe data. They can ensure the program functions properly in this way without jeopardizing anyone's privacy (Pardau 2018). The core of this service is delving deeply into data to improve understanding. It makes data more accessible to find, comprehend, and safeguard by utilizing intelligent computer technologies.

4.2 Slash next generative human AI

A unique form of artificial intelligence known as Slash Next Generative Human AI combats complex email scams, supply chain attacks, and other forms of fraud (Jia et al. 2024). Slash NeXT's new tool complements its current Human AI capabilities. By utilizing computer vision, machine learning, language comprehension, contextual awareness, and contextual understanding, Human AI performs tasks like those of actual human security professionals. It assists in thwarting intricate attacks that originate from multiple sources. It generates several versions of potential threats and improves its ability to identify them to prepare for these attacks (Zhao et al. 2024).

4.3 BEC GAI augmentation

To aid in thwarting future attacks, Human AI can create thousands of new email threat variations (Arachchige 2023).

4.3.1 Relationship graphs & contextual analysis

Human AI adheres to a standard language and writing style that benefits each employee and their partner businesses. This makes identifying odd or distinctive speech or writing patterns easier (<https://slashnext.com/press-release/slashnext-launches-industrys-firstgenerative-ai-solution-for-email-security/>). At the same time, combined relationship graphs with contextual analysis provide comprehensive attacks, such as adaptive learning cross-referencing communication patterns and context and automated scoring. In adaptive learning, GAI systems continuously learn from relationship and context data. However, in cross-referencing

communication graphs, relationships may designate CEO communication with the CEO as usual, and contextual analysis might disclose irregularities in language. Together, these tools assign risk scores emails based on the likelihood of a BEC attack (Zoubi and Mohammed 2024; Zhou et al. 2023).

4.3.2 Natural language processing

Human AI can analyze email content and associated files to determine the sender's subject matter, tone, potential emotional reactions, and malicious intent (e.g., deceiving someone into doing something). (Zhang et al. 2024). The NLP identifies the main phrases associated with BEC attacks, such as payment method, confidential information, and urgent transfer; NLP models go behind keywords to infer the intent behind the email. Even when attackers use nuanced social engineering techniques, they can discern whether the email is attempting to obtain sensitive information or requires immediate financial action. NLP algorithms are trained on massive phishing email datasets to recognize suspicious email patterns. This allows them to recognize emails that appear authentic at first glance but may conceal dangerous attachments (Ma et al. 2023; Chowdhery et al. 2022).

4.3.3 Computer vision recognition

Human AI employs a tool called Slash NeXT's LiveScan to instantly scan online URLs for any variations in their appearance, such as images or page layout, to identify phony websites that attempt to steal credentials (Liu et al. 2022). For example, Human AI can detect minute variations in spoof Microsoft 365 login pages using advanced computer vision, preventing users from accessing them. To stop ransomware attacks, Human AI may also examine the strategies employed in malicious software and email attachments.

4.3.4 Sender impersonation analysis

To stop impersonation attacks, Human AI uses email subject lines to determine whether an email is from the person it claims to be. Approximately 700,000 new threats are found and analyzed daily using a large SlashNext database (Yu et al. 2021). This database also aids Human AI in identifying risks that could induce fear or a sense of urgency in humans, leading them to act improperly. For instance, Human AI can detect whether an email attempts to influence someone's emotions to get them to act quickly if it contains the Words "Urgent!" or "Hurry up!".

4.4 Sec-PaLM

Sec-PaLM is comparable to Google's intelligent software, PaLM 2 (Shao et al. 2017), which is superior to PaLM, the program's previous version. It is quite sophisticated and can comprehend numerous languages, solve puzzles, and even write code (Anil et al. 2023). A customized version of PaLM 2 called Sec-PaLM has been specially trained to comprehend and handle cybersecurity issues. It is comparable to a significant advancement in protecting data and computers from malicious actors. It is available via Google Cloud, and it employs arti-

ficial intelligence to analyze and interpret the behavior of computer programs, particularly those that might be attempting to harm.

5 Real world scenarios

Several practical cybersecurity technologies use a technology known as GAI to increase their strength and security, as mentioned in Fig. 5. Here are a few instances: Sec-PaLM is a modified version of Google's intelligent software PaLM 2, used to analyze cybersecurity threats (Chowdhery et al. 2022). HumanAI: This program analyzes emails and looks for possible phishing attempts and other cyber threats using cutting-edge algorithms (Google 2023). Among other technologies, these use GAI's power to enhance cybersecurity, as in Table 2.

5.1 SentinelOne purple AI

On its Singularity Skylight platform, cybersecurity firm SentinelOne has developed a new tool dubbed Purple AI (Purple 2023) that is intended to assist in identifying and thwarting cyber-attacks (Singularity Skylight 2023). This technology makes threat hunting easier for security experts by utilizing a learning machine model (LLM) to identify and thwart threats quickly. GPT-4, an LLM produced by OpenAI, is among the LLMs that Sentinel One employs for Purple AI (<https://openai.com/gpt-4>). Additionally, they have trained and adjusted their LLMs utilizing their data to ensure optimal cybersecurity performance. Purple AI is capable of the following things:

5.2 Boosting cyber defense

Conventional cybersecurity tools can be extremely complex, requiring specialized training to operate correctly. However, things are different with Purple AI (Xu et al. 2022a). Simple inquiries like "Are there any bad guys here?" can be asked of Purple AI by security specialists, and it will provide them with precise responses. This greatly facilitates and speeds up the process for analysts to identify any security holes in their systems. After that, they can devote their time to other crucial activities (Ma and Hu 2022a). Additionally, Purple AI may provide an overview of its findings, which helps analysts get more done in less time.

5.3 Embedding GAI

One additional feature added to the Singularity Skylight platform is Purple AI. It is conveniently located within the platform's UI and is simple. This gives users flexibility and facilitates their gradual acclimatization to the GAI tool (Generative and AI|Google Cloud Blog 2023a). The learning curve for users is lowered by integrating GAI into platform estimation features, encouraging users to gradually become accustomed to and incorporate GAI technologies into daily tasks because the purple is conveniently located within the user interface so that users may experiment. Also, by initiating possibilities for more extensive applications across diverse departments and user groups, GAI becomes more accessible for both technical and non-technical users. Additionally, the integration allows for ongoing system

improvements to the user's requirements, gradually improving procedures like automation, context creation, and decision-making support (Vu et al. 2024).

5.4 LLM

SentinelOne combines many LLMs, both publicly available and proprietary, to increase the power of their model. To improve the performance of their cybersecurity tool, they start with a pre-made model, such as OpenAI's GPT-4, and adjust. Security teams may search for threats more efficiently, thanks to the SentinelOne Purple AI technology. Purple AI is an intuitive platform, unlike others, that requires sophisticated knowledge. It helps firms increase security without hiring additional professionals because even non-techies can use it.

5.5 Google cloud security AI workbench

Sec-PaLM is a new security-specific Large Language Model (LLM) used in Google's Cloud Security AI Workbench (<https://openai.com/gpt-4>). This strategy uses the security information provided by Google Cloud, utilizing Mandiant's well-known threat intelligence on vulnerabilities, malware, threat actors, and threat indicators and Google's broad visibility into threat data. The Workbench can manage an abundance of threat data, navigate a plethora of intricate security tools, and close the talent gap in the sector.

5.6 Containing threats from spreading beyond the first infection

To stop possible waves of adversarial attacks using AI/ML systems, Google combines extremely effective threat intelligence with real-time event analysis and creative AI-driven detections and analytics with the Cloud Security AI Workbench.

Among the google cloud security AI workbench tools with these features are the following:

- Virus Total Code Insight uses Google's Sec-PaLM to examine and explain the actions of scripts that can be harmful.
- Using Google Cloud and Mandiant's Threat Intelligence, Mandiant Breach Analytics for Chronicle automatically notifies users of ongoing environment breaches. To provide context and facilitate an efficient response to such important concerns (Xu et al. 2022b).

5.7 Adding intelligence to reduce toil

Organizations may simplify their security toolkits and enable their systems to automatically improve their security posture with the help of the Google Cloud Security AI Workbench. Such capabilities are provided by a few products in the Google Cloud Security AI Workbench, such as:

- Assured OSS leverages Large Language Models (LLMs) to help Google extend its security solution for open Open-Source-Software (OSS) by guaranteeing the inclusion of carefully selected and vulnerability-tested packages comparable to those used by

Google.

- Sec-PaLM is used by Mandiant Threat Intelligence AI to quickly discover, compile, and address threats pertinent to the enterprise by utilizing Mandiant's enormous threat graph (Ma and Hu [2022b](#)).

5.8 Transforming how practitioners do security to close the talent gap

Google's Cloud Security AI Workbench is intended to demystify and streamline security so that even those without specialist security training may understand it. For this goal, the Cloud Security AI Workbench offers two solutions:

- Chronicle AI: Without the need for in-depth cybersecurity knowledge, Chronicle AI allows users to quickly generate detections, search through enormous volumes of security events, and engage with real-time findings (Generative and AI|Google Cloud Blog [2023b](#)).
- Security Command Center AI: This program can display intricate attack graphs in simple formats. It advises on mitigating the damage and delivers insights into the impacted assets. It also provides succinct summaries of privacy, security, and compliance findings unique to Google Cloud.

5.9 Microsoft security copilot

Microsoft created Microsoft Security Copilot; an AI assistant explicitly designed for cybersecurity professionals. It helps defenders find vulnerabilities, examine large amounts of data, and get essential insights from everyday tasks (Security and Copilot|Microsoft Security [2023a](#)). It stays current on the newest vulnerabilities by referencing resources like Microsoft's threat intelligence repository, the National Institute of Standards and Technology's vulnerability database, and the Cybersecurity and Infrastructure Security Agency. Driven by Microsoft's proprietary security models and OpenAI's GPT-4 GAI, Microsoft Security Copilot has an easy-to-use interface with a straightforward prompt box. Security experts can use this interface to extract information from files, URLs, code snippets, and incident data from other security tools to help with security investigations and summarize events for reporting (Code and Insight: Empowering Threat Analysis with Generative AI [2023](#)). Every interaction, including prompts and responses with Copilot, is recorded for auditing purposes. Copilot uses Microsoft's extensive threat intelligence resources to detect threats effectively in the background, processing 65 trillion signals daily. However, rather than taking the position of security experts, its main objective is to support them. Copilot has a feature that lets users pin transaction outcomes in a shared workspace, which makes it easier for security teams to work together on threat analysis and investigation. Prompt book functionality is a noteworthy feature of Copilot that lets security experts organize a series of procedures or automation into easily actionable prompts. Teams can, for example, provide shared prompts for reverse engineering scripts so that analysis can proceed without consulting an expert. Additionally, Copilot makes creating PowerPoint slides that show attack paths and occurrences easier. Additionally, the system incorporates a feedback mechanism that allows users to comment if Copilot produces inaccurate findings, thereby improving the system's replies and lowering errors.

5.10 Talon enterprise browser

Talon Enterprise Browser and Microsoft Azure OpenAI Service have been integrated by Talon Cyber Security (Analytics and for Chronicle|Active Breach Detection 2023) to provide enterprise access to sophisticated AI text-generating tools like ChatGPT. By maintaining ChatGPT interactions inside a safe environment and restricting data transfers to third-party services, this integration allows customers to take advantage of already existing Azure resources while guaranteeing data protection. Additionally, the browser allows administrators to prevent users from submitting sensitive material into ChatGPT, such as source codes and payment card numbers. The Talon Enterprise Browser also has productivity tools like AI-powered message summarization and email answer generation. The option to ban public ChatGPT extensions and query logs significantly simplifies compliance reporting.

5.11 Slashnext generative human AI

Slashnext has made significant strides in the market with its Generative Human AI, which uses GAI to fight financial fraud, corporate email compromise (BEC), executive impersonation, and sophisticated supply chain threats (Source and Software|Google Cloud 2023). This creative approach improves upon SlashNext's current Human AI capabilities, which use a combination of computer vision, machine learning, and natural language processing to mimic human threat researchers. It successfully thwarts sophisticated multi-channel message attacks using relationship graphs and rich contextualization. Human AI uses cloning and AI data augmentation techniques to predict a wide range of possible BEC threats. To self-train the system on various scenarios, it assesses the main hazards and creates several versions. The following are the features of Human AI: Enhancement of BEC GAI By creating thousands of new BEC threat variations based on existing ones, human AI can proactively prevent future intrusions (Threat Intelligence|Cyber Threat Intelligence Platform 2023). Relationship Charts and Contextual Interpretation To detect abnormal communication patterns and conversation styles, Human AI creates a core framework of accepted communication standards and writing styles for all workers and suppliers. Natural Language Processing: To identify tone, emotion, themes, manipulation triggers, and intent related to social engineering techniques, Human AI thoroughly examines email bodies and attachments (Introducing AI-powered Investigation in Chronicle Security Operations|Google Cloud Blog 2023).

1. Visual Accuracy: Using SlashNext's Live Scan, Human AI analyzes URLs in real-time, looking for visual differences like altered layouts and images to spot legitimate phishing websites. For example, it uses computer vision to identify minute variations in phony Microsoft 365 login sites and block access (Command and Center|Google Cloud 2023).
2. Inspection of Attachments: By analyzing the social engineering characteristics of attachments and malicious code, Human AI can stop ransomware attacks.
3. Analysis of Sender Impersonation: Human AI assesses email authentication results and headline information to prevent impersonation attempts. Slash Next uses a large database to examine about 700,000 new threats daily, including ones that aim to exploit weaknesses in people. In addition, Human AI mimics these actions and emotions during

its detection process to identify how threat actors use human emotions, such as creating dread through urgent requests.

5.12 Recorded future AI

With Recorded Future's integration of OpenAI's GPT transformer model into its Intelligence Cloud, artificial intelligence is poised to transform the intelligence sector. The company's threat research subsidiary, Insikt Group, provides a massive archive of threat analysis data that spans more than ten years, which is used to train the Recorded Future AI model (Security and Copilot[Microsoft Security 2023b]). The Recorded Future Intelligence Graph provides more insights that improve this model. Using NLP and ML approaches, it automatically collects, organizes, and maps information about adversaries and victims from various sources, including text, photos, and technical data. In real-time, it analyzes and maps insights across billions of items. The Recorded Future AI provides Real-time threat landscape analysis (Liu et al. 2024). It enables prompt action across the whole internet.

Additionally, it increases analyst productivity, making up for a need for more skills. It also provides intelligence-driven insights that enable firms to make proactive decisions that protect them from threats from competition. To process and analyze more than 100 terabytes of various data kinds and turn them into actionable insight, Recorded Future has built sophisticated algorithms and analytics over the last ten years. Using this data, firms can proactively mitigate against such attacks by identifying risks and weaknesses. This expertise is significant when dealing with complex cyber and physical threat scenarios, where more than standard methods might be required. The world's most advanced intelligence repository, the Recorded Future Intelligence Cloud, was made possible by the Recorded Future AI. Recorded Future AI quickly prioritizes essential threats and weaknesses by integrating AI and ML throughout the intelligence cycle (<https://talon-sec.com/>), from analysis and production to distribution, providing analysts with real-time actionable insight. It automates labor-intensive operations related to threat research with AI, freeing analysts to focus on more advanced strategic initiatives.

5.13 Security scorecard GPT-4 integration

OpenAI's GPT-4 is integrated into Security Scorecard, a platform for security evaluation. Through natural language processing capabilities, this interface enables cyber security professionals to obtain quick responses and mitigation advice for high-priority cyber hazards. This makes it possible for security specialists to elicit more information about their exposure to cyber threats, quickly spot security holes, and increase overall cyber resilience by posing questions (Web and Browser—Talon Cyber Security 2023). The innovation division of Security Scorecard, ScorecardX, is responsible for developing the solution. Its mission is to build and provide technological solutions that meet significant client concerns. Customers can use this solution to ask open-ended questions about their business ecosystem, like "Tell me which of my critical vendors had breaches in the last year" or "Name my ten lowest-rated vendors," and get timely answers that help them make better risk management decisions. The AI-driven search feature works for every company that the Security Scorecard monitors, saving executives much time and labor by lowering the manual labor needed for data analysis. In addition, the search function keeps learning and improving to serve

users better (<https://www.prnewswire.com/news-releases/slashnext-launches-industry-first-generative-ai-solution-for-email-security-301757649.html>) (Fig. 7).

6 Transforming cybersecurity GAI impact (2023–2024)

GAI transformed cybersecurity between 2023 and 2024 by introducing new threats and improving defensive capabilities. Important advancements include zero-day threat identification, robust defense mechanisms, and GAI-driven threat detection and prevention using sophisticated anomaly detection systems. To train incident response teams and intrusion detection systems (IDS), Generative Adversarial Networks (GANs) imitated cyberattacks. Adversarial AI made complex attacks possible, which led to the development of adversarial training and counter-GAI strategies. Phishing detection, malware detection, and automated incident response were all enhanced by GAI. Cybercriminals, however, used GAI to create malware, phishing, and deepfakes. The regularity system is defined as a mechanism of structural arrangement in which specialists in a particular field are trained. The GAI system poses legal and regulatory risks governed by ethical standards and propriety of usage. The LLM challenges understanding vulnerability and threat intelligence identification in cybersecurity (Zhao et al. 2024; Golda et al. 2024).

6.1 Generative adversarial networks (GANs) for attack simulation

GAI, mainly through Generative Adversarial Networks (GANs), has become instrumental in simulating cyberattacks. GANs can generate realistic attack scenarios, helping cybersecurity teams anticipate and counter future threats. This has been especially useful in training Intrusion Detection Systems (IDS) and Incident Response Teams, allowing them to experience complex, real-world attack scenarios and prepare proactive defenses (Babcock and Bali 2021; Golda et al. 2024).

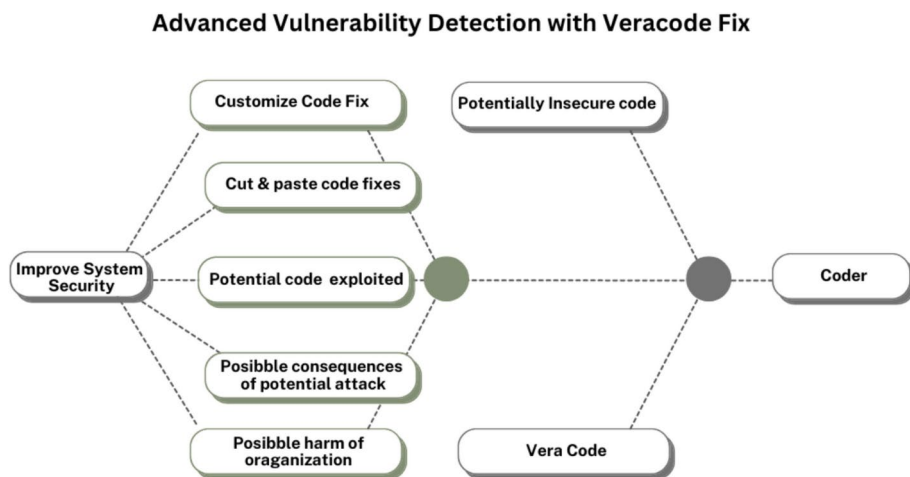


Fig. 7 Advance vulnerability detection

6.2 Adversarial AI and countermeasures

Cybercriminals have also begun using GAI to bypass security systems by generating sophisticated adversarial attacks. These attacks use GAI to manipulate data or exploit vulnerabilities in machine learning models to deceive detection systems. In response, cybersecurity experts have developed counter-GAI techniques, where generative models are used to test systems against adversarial examples and improve their resilience. This has led to the development of adversarial training, where security models are hardened against GAI-driven attacks by being exposed to various generative manipulations.

6.3 Phishing detection and prevention

Phishing attacks are becoming increasingly sophisticated, making it harder for businesses to spot them. GAI has been employed for offensive and phishing purposes. The cybercriminal uses GAI to identify social engineering and phishing emails. Dynamically generating content that evades traditional detection mechanisms. On the defensive side, GAI has significantly developed automated phishing detection methods, where models are trained to identify subtle patterns in email content and structure, flagging potentially harmful messages even when they exhibit new tactics. These models indemnify the red flags of phishing attacks (such as misspellings, attempts to coerce the recipient, and URL structure and targets) (Yigit et al. 2024; Xu et al. 2024).

6.4 Automation of incident response

Automation of GAI has enabled the development of automated cybersecurity incident response systems. GAI helps speed the response to any cyber incident by creating predicted responses to detected cyber threats. GAI algorithms can plan optimal replies to attack types, such as automatically disconnecting hacked computers, applying security updates, and preventing hostile network activity. This has been particularly useful in managing devices in cloud security and the Internet of Things, as it would only be possible to respond manually due to the number and variety of devices.

6.5 GAI for malware detection and generation

GAI has been instrumental in developing advanced malware detection systems by creating polymorphic malware, which constantly alters its signature to avoid conventional detection. Artificial intelligence. Tools can analyze what features such evolving malware might possess and identify it in its infant stages. On the other hand, cybercriminals have benefited from GAI when generating malware variants, which makes it difficult for static security systems to detect and remediate. The arms race between GAI-based offensive and GAI-based defense technologies has accelerated the development of behavior-based malware detection techniques.

7 Limitations

Although General Artificial Intelligence (GAI) has much to offer the security industry, its application has several obstacles (Zheng et al. 2024). The following are some disadvantages of using GAI in the security industry:

7.1 Tendency to give wrong/unethical results

The long-term effects of GAI still need to be discovered due to their novelty. This implies that their use carries inherent dangers related to known and unknown factors. For example, while communicating information, Large Language Models (LLMs) like ChatGPT can sound quite convincing even though the information may occasionally be wrong. Furthermore, ChatGPT is vulnerable to social biases that could be used to support immoral or illegal activity (Wazid et al. 2022).

7.2 Cost inefficiency

Advanced artificial intelligence (AI) security systems can be costly. Only businesses with sufficient resources and experience can afford to set up and maintain these systems. This implies their data and resources will be protected to the highest standard. Others may be forced to employ fewer safe techniques, which could put them at greater risk from emerging dangers. These problems may lead to certain groups receiving assistance with the cost of AI technologies. Nonprofits, for instance, may be eligible for subsidies to support their efforts to protect individuals' data.

7.3 High setup time

It can take several weeks or even months to train GAI models. Organizations that strive for rapid progress may find it slowing down due to its lengthy setup time.

7.4 Easy exploitability by malicious actors

One of the main issues with GAI systems is that malicious actors can use them to discover ways to attack other systems. They might use that information to access those systems by figuring out where they can get into without authorization. They might also create malicious programs, viruses, and convincingly false emails or communications to deceive others (Singh et al. 2022).

7.5 Interpretability and explainability

The inner workings of GAI models are like a mystery box since they are quite complex. This can challenge critical security systems when it is critical to know what is happening because it makes it difficult to interpret or explain the findings they provide.

7.6 Contextual limitations

In some circumstances, GAI systems may struggle to understand the context and provide logical responses. This could lead to strange or ineffective responses, which would be confusing and might be a security risk. People need context to comprehend and react appropriately in talks. One must consider previous remarks, the subject at hand, and any pertinent background information to provide appropriate answers. However, one drawback of GAI models is that they could better address all aspects of context (Future and AI: AI-driven Intelligence to Elevate Your Security Defenses [2023](#)).

7.7 Short-term context

GAI models need help to keep up with lengthy conversations or large amounts of data.

7.8 Lack of common-sense reasoning

GAI models can learn and remember new information but frequently need more common sense in everyday situations.

7.9 Ambiguity resolution

In most cases, GAI models cannot consider broader contexts to interpret data accurately. This can lead to their providing responses that, although appearing correct, need to match the circumstances as they misinterpreted the inquiry.

7.10 Multi-turn conversations

Sometimes, back-and-forth conversations are too lengthy for GAI models to record. This may cause them to provide illogical or inconsistent responses. These problems may lead to misunderstandings between the GAI security team and increase system vulnerability (An Inside Look at How Insikt Group Produces Leading Threat Research|Recorded Future [2023](#)).

7.11 Difficulty with long-range dependencies

Long sequences may cause GAI models to lose coherence and logic, leading to fragmented or nonsensical outputs (<https://www.recordedfuture>). This problem is caused by several variables, such as the short-term memory limitations of GAI models, their dependence on token sequences with fixed lengths, and the vanishing gradient problem during training. These restrictions may allow for system weaknesses, jeopardizing its effectiveness and security.

7.12 Data-related concerns

GAI tools can potentially jeopardize data privacy in several ways, which is a major cyber security concern. Here are a few of them:

7.13 Data breaches

If we use GAI tools carelessly, there's a chance that our personal information will be accessed without authorization or shared inappropriately. This could result in issues with privacy and someone exploiting our personal information in ways we do not desire (Openai and GPT-4|Securityscorecard 2023).

7.13.1 Inadequate anonymization

GAI models require private or sensitive data to learn or provide results. This data might reveal people's identities if identity-concealing methods are not used, harming their privacy (Scorecardx|Securityscorecard 2023).

7.13.2 Biases and discrimination

GAI models can unintentionally retain biases from the training set. Imagine if the information utilized to instruct them contains biased or unjust material. The results of the GAI models may then demonstrate and exacerbate these biases, resulting in further unfair treatment or discrimination against groups of people. The results of human safety inspections are seen by comparing the findings of human safety tests comparing LLaMA 2-Chat to MPT Vicuna and Falcon (Blease et al. 2024).

7.13.3 Lack of consent and transparency

Users' privacy rights may be violated, and their trust may be damaged if we gather, use, and share their data without obtaining their consent or providing them with clear information about this (Fui-Hoon Nah et al. 2023).

7.13.4 Inadequate data retention and deletion practices

A GAI model may facilitate unauthorized access to user data or allow unintended uses if it retains user data for longer than necessary or fails to remove it when requested or after the necessary (Tyson 2023).

7.14 Lack of control

Users have little control over the output of GAI models. This is particularly evident when the models create material independently of user input. In GPT-3.5, users provide a prompt or context, and the model reacts accordingly. Although the initial prompt allows users to modify the output, they have no control over all the finer points or subtleties of the result. Because of this lack of control, cybersecurity professionals may find it challenging to identify and address subtle system vulnerabilities that require close inspection, as summarized in Table 5 (Zhou et al. 2023).

7.14.1 Need for empirical evaluation

The efficacy of various GAI models and commercially available security devices is not standardized and cannot be measured or tested. This makes selecting the ideal model or product to get the desired outcomes challenging (Azaria 2022).

Table 5 highlights the proposed classification schema. It establishes a broad context within which the social consequences of AI-enabled cyber warfare are analyzed. It extends current cyber security dynamics to incorporate various issues, including disruption of businesses and physical threats to people's lives. In addition, the schema also brings forth the inherent need for inter- and multi-disciplinary approaches that call for the synergy of politics, economics and public health perspectives in the analysis of cyber security issues. Such provides impetus for formulating new, integrated theoretical models that could embrace the complexities of AI-influenced cyberattacks. Also, the schema highlights such categories as 'Human and Public Safety' and 'Political and Social Stability', therefore depicting the need for theoretical frameworks that encompass human and societal issues in the technological-based area of cybersecurity. However, today, it is common knowledge that technological means for this purpose are not powerful enough to address cyber risks, and AI-powered ones are not even mentioned. This makes a strong argument for changing and expanding the focus of theoretical innovation when analyzing AI-powered cyberattacks, their counteractions, sources, and consequences on society. According to the classification scheme devised in Tables 4 and 5, it can be confidently argued that governments and organizations can use it to formulate strong strategies to address AI-related cyber warfare. Its scope is broad; therefore, it would be useful in expending resources on areas where people can channel more income and efforts toward addressing the worst societal impacts.

Further, the schema in discussion also advocates for public education as an important defensive mechanism against the strategically diverse threats that such attacks encompass. Legal institutions would also benefit from it as it can provide a deep understanding that could guide the formulation of legal arrangements, thereby increasing the precision and effectiveness of remedies and enforcement actions. All these aspects take us back to the importance of the schema in enforcing a structural and coordinated approach to one of the consequences of AI use, that is, its use for cyberattacks. Having analyzed the various facets of AI-driven cyberattacks, including their types of approaches to how they can be mitigated, their causes, and their potential impact on society, we will now be highlighting how these

Table 5 GAI in cyber security: a thorough foundation for comprehending hostile and antagonistic AI

References	Privacy and human safety	Privacy and personal security	Systemic risk and infrastructure	Economic implications	Data and information security
Cox (1999)	Potentially fatal outcomes	Theft of identity	Systemic shortcomings	Disturbances to the economy	Effect on the security of data
Morreel et al. (2024)	Misdiagnoses in medicine	Effect on personal security and privacy	Dangers to the environment	Financial expenses	Effect on the property of intellectual property
Chen and Esmailzadeh (2024)	Inaccurate medical diagnosis	Money and personal information are being compromised	Harm to the infrastructure	Disruptions to finances	Assure authorized individuals may access data
Zhang et al. (2024)	Media mistakes	Occurrences of data breaches	Disruptions to operations	Loss of money for commercial entities	To avoid unwanted access

Table 4 GAI in cyber security: applications, benefits, and challenges

References	Applications	Description	Potential Benefits	Challenges	Examples
Purple (2023); Xu et al. (2024)	Automated hacking	Using Gen AI to automate the process of finding the vulnerability in systems	Fast and more efficient. Ability to test systems at scale	Potential for misuse by attackers. Difficulty in detecting Gen AI generated attacks	Automated penetration on testing tools. Vulnerability scanning bots
Yigit et al. (2024); Xu et al. (2024)	Phishing and Social Engineering	Using Gen AI to create more convincing and targeted phishing emails and social engineering attacks	Ability to personalize attacks at scale. Improve ability to bypass security measures	Potential for large-scale attacks. Difficulty in detecting Gen AI-generated attacks	Personalized Phishing emails. Fake social media profiles
Dhoni and Kumar (2023); Xu et al. (2024)	Reverse Cryptography	Using Gen AI to break encryption and decode decrypted data	Ability to test the strength of encryption algorithms. Potential for decrypting sensitive data	Potential for misuse by attackers. Difficulty in detecting Gen AI-based attacks	Cryptanalysis tools Decryption bots
Bryce et al. (2024); Xu et al. (2024)	Malware Generation	Using Gen AI to create new and more sophisticated malware	Ability to create polymorphic malware. Potential for bypassing security measures	Potential for large-scale attacks. Difficulty in detecting Gen AI-generated malware	Poly-morphic viruses. Targeted ransomware

multiple aspects can be integrated to address the fifth objective of the study. The following section, therefore, seeks to develop a model that addresses an overall framework integrating components mentioned in the first four research objectives, offering a holistic understanding of the dynamics and implications of AI-driven cyberattacks. This framework aims to serve as a cornerstone for both academic research and practical interventions in the cybersecurity domain (Vu et al. 2024; Eibeck et al. 2024; Zheng et al. 2024).

8 Conclusion

GAI holds great potential to enhance cybersecurity capabilities significantly, providing a formidable defense against increasingly sophisticated threats. By leveraging GAI models, we can improve threat detection, simulate diverse attack scenarios for analysis, identify system anomalies, crack passwords, detect phishing attempts and malware, and automate security responses. Leading technology companies are already pioneering innovative products that harness GAI's capabilities to deliver cutting-edge security solutions. However, while the opportunities are vast, it is equally important to recognize the risks associated with GAI. Malicious actors could exploit these advanced techniques to orchestrate highly sophisticated attacks, such as deep fakes or hyper-realistic phishing campaigns. To address these challenges, it is critical to establish ethical guidelines, promote responsible use of GAI, and advance cybersecurity measures in parallel with the development of GAI technologies. Only through a balanced approach combining innovation with robust safeguards can we mitigate the risks and prevent misuse. Ultimately, GAI presents an exciting opportunity to

strengthen cybersecurity defenses, but its development must be guided by careful ethical considerations and proactive measures to ensure its responsible deployment.

The survey highlights the transformative potential of LLMs in revolutionizing cybersecurity. By tapping into the power of LLMs, we can design more resilient and advanced cybersecurity systems capable of addressing the dynamic and evolving threat landscape. This paper outlines a strategic vision for the future of cybersecurity, emphasizing the pivotal role of innovation and resilience in safeguarding our digital infrastructure. We strongly encourage further research and implementation efforts to unlock the capabilities of LLMs fully, enabling us to stay ahead of emerging threats and build a safer digital ecosystem.

9 Opportunities and future direction for GAI models in cybersecurity

GAI has become an essential norm in the ever-evolving digital world for potentially impacting cybersecurity by automating processes. With the adept adoption of GAI practices, organizations could effectively leverage their cybersecurity standards excellently. Apart from leveraging the quintessential security standards, organizations must proactively adopt robust incident response plans to counter, negate, and nullify the risky forces of cyberattacks and malicious threats. The following are some of the opportunities for GAI Models in cybersecurity.

9.1 Automating cybersecurity with GAI

GAI has become a new standard in the continuously developing world, as it holds massive potential as a tool that can revolutionize cybersecurity by automating many operations. When it comes to implementing GAI practices in an organization, the organization would be able to enhance its cybersecurity to the next level. The implementation of GAI in the current systems is useful in the detection and surveillance of vulnerabilities, as well as the mobilization of counteractions in the shortest time possible. Apart from increasing security measures, there is a dire need to develop and implement an effective incident response mechanism that will help organizations mitigate, neutralize, or eliminate cyber threats. Furthermore, GAI has the potential to enhance the visibility of threats and enhance the security envelope of the organization (Agrawal et al. 2024; Mavikumbure et al. 2024). One of the areas where GAI has performed immensely well is in the identification of incidents and the ability to manage and control them during response, as well as searching for temporal patterns using security tools.

9.2 Threat intelligence with big data

Cyber threats are complex and evolving day by day, while GAI models can make use of huge datasets and compute sophisticated patterns to create intelligence to counter these threats. It is worth noting here that GAI depends heavily on large amounts of data, and it is evident that the kinds of models that are seen to be most intelligent are the ones closest to human-level intelligence. These systems develop heuristic purposes that address human interests, design artificial constructs like chatbots and interactive interfaces, and contribute towards the advancement of cooperative cyber defence systems. Hence, through the sharing

of information relating to cybersecurity across industries, companies, organizations, and governments, GAI has fortified the general approach towards enhancing cybersecurity. The processing of big data leads to the early identification of risks, which provides a primary advantage to GAI in the sphere of cybersecurity.

9.3 Simulating attacks for real-world scenarios

Organizations can use GAI to develop attack simulations of adversarial attacks and phishing attempts to prepare against real-world threats in their environment (Sai et al. 2024). Organizations can teach cybersecurity personnel to handle real-time attacks properly by running training simulations that minimize downtime and reduce damage. Through its ability to simulate attack scenarios, GAI builds organizational readiness and develops ongoing learning about cybersecurity practice improvement. An organization can effectively leverage this capability to address threats in the current volatile cybersecurity environment since attackers frequently adjust their methods.

9.4 Improving biometric security and privacy

GAI has a great impact on enhancing features of biometric authentication, homomorphic encryption and privacy-preserving techniques (Ahmad et al. 2025). This means that the latest biometric systems powered by GAI make it possible to verify users in a much more accurate and secure manner than what is delivered through password-based security systems, which are most of the time easily hacked. Furthermore, there is an attempt to work on new encryption techniques. For example, homomorphic encryption enables computations to be made on data that is encrypted without having to decrypt the information. Privacy preservation measures are undertaken to ensure the security of private data, despite enabling their use in analysis and decision-making.

9.5 Fraud detection

Through GAI technology, institutions can address transaction history monitoring to identify abnormal behavior as well as fraudulent records. Financial institutions, together with e-commerce platforms, can stop fraud from escalating through the real-time identification of suspicious activities using GAI (Shafik 2024). Transaction analysis through GAI systems enables the detection of security threats by revealing unauthorized access patterns that signify potential malicious intent, thus protecting business operations and client safety (Yosi-fova 2024; Yigit et al. 2024; Falade 2024; Mavikumbure et al. 2024).

9.6 Ensuring data transparency

Blockchain technology can use GAI to protect smart contracts and maintain both data transparency and immutability about contributions (Nguyen et al. 2024b). Smart contracts can experience coding flaws because they need no intermediary parties to operate. The audit of smart contracts by GAI helps identify all vulnerabilities to confirm their proper execution. GAI can protect blockchain system integrity because it monitors consistently to uphold transparent transaction processes with no tampering possible.

9.7 Futureproofing cybersecurity defenses

Moving forward, GAI will enhance threat intelligence and defense mechanisms and strengthen overall defenses for better protection against cybersecurity threats (Golda et al. 2024; Yigit et al. 2024; Chen and Esmailzadeh 2024; Fetaji 2024; Sai et al. 2024c; Admass 2024; Zheng et al. 2024; Ozkan-Ozay, et al. 2024; Abdi et al. 2024). Incorporating GAI into existing cybersecurity frameworks will improve the effectiveness of security threat countermeasures. The prospect of cybersecurity is about developing intelligent computing systems based on GAI to build dynamic security systems that can successfully match modern threats.

Acknowledgements This open-access research is supported by the Qatar National Library QNL.

Author contributions Muhammad Saad Irshad and Mueen Uddin: Conceptualization, Methodology, Writing- Original draft preparation, Irfan Ali Kandhro and Fuhid Alanazi: Data curation and Visualization: Fahad Ahmed, Muhammad Maaz and Saddam Hussain: Writing-Reviewing and Editing, Fuhid Alanazi, Syed Sajid Ullah: Investigation, Software. Mueen Uddin: Muhammad Saad Irshad and Irfan Ali Kandhro: Validation, Data curation, Formal analysis. Fahad Ahmed, Muhammad Maaz, Syed Sajid Ullah: resources, and analysis, All authors reviewed the manuscript.

Funding Open Access funding provided by the Qatar National Library.

Data availability The data will be made available upon reasonable request.

Declarations

Competing interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Abdi N, Albaseer A, Abdallah M (2024) The role of deep learning in advancing proactive cybersecurity measures for smart grid networks: a survey. *IEEE Internet Things J* 11:16398
- Admass WS, Munaye YY, Diro AA (2024) Cyber security: state of the art, challenges and future directions. *Cyber Secur Appl* 2:100031
- Agrawal G, Kaur A, Myneni S (2024) A review of generative models in generating synthetic attack data for cybersecurity. *Electronics* 13(2):322
- Ahmad I, Nasim F, Khawaja MF, Naqvi SA, Khan H (2025) Enhancing IoT security and services based on generative artificial intelligence techniques: a systematic analysis based on emerging threats, challenges and future directions. *Spectr Eng Sci* 3(2):1–25
- Alawida M et al (2024) Unveiling the dark side of ChatGPT: exploring cyberattacks and enhancing user awareness. *Information* 15(1):27
- Alvarez VM (2013) YARA: the pattern matching Swiss knife for malware researchers (and everyone else)

- Alwahedi F et al (2024) Machine learning techniques for IoT security: current research and future vision with generative AI and large language models. *Internet Things Cyber-Phys Syst* 4:167
- An Inside Look At How Insikt Group Produces Leading Threat Research|Recorded Future. Accessed: Jul. 30, 2023. [Online]. Available: <https://www.recordedfuture.com/leading-threat-research>
- Anil R et al. (2023) PaLM 2 technical report. *arXiv:2305.10403*.
- Assured Open Source Software|Google Cloud. Accessed: Jun. 22, 2023. [Online]. Available: <https://cloud.google.com/assured-open-sourcesoftware>
- Azaria A (2022) ChatGPT usage and limitations. [Online]. Available: <https://hal.science/hal-03913837>
- Babcock J, Bali R (2021) Generative AI with python and TensorFlow 2: create images, text, and music with VAEs, GANs, LSTMs, and transformer models. Packt Publishing Ltd. ChatGPT. Accessed: Jun. 22, 2023
- Bandi A, Adapa PVSR, Kuchi YEVPK (2023) The power of generative AI: a review of requirements, models, input-output formats, evaluation metrics, and challenges. *Future Internet* 15(8):260
- G. Bansal, V. Hasija, V. Chamola, N. Kumar, and M. Guizani, "Smart stock exchange market: A secure predictive decentralized model," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2019, pp. 1–6.
- Blease C, Torous J, Mcmillan B, Hägglund M, Mandl KD (2024) Generative language models and open notes: exploring the promise and limitations. *JMIR Med Educ* 10:e51183. <https://doi.org/10.2196/51183>
- Breach Analytics for Chronicle|Active Breach Detection. Accessed: Jun. 22, 2023. [Online]. Available: <https://www.mandiant.com/advantage/breach-analytics>
- Bryce C et al. (2024) Trends in large language models: actors, applications, and impact on cybersecurity.
- Chen Y, Esmailzadeh P (2024) Generative AI in medical practice: in-depth exploration of privacy and security challenges. *J Med Internet Res* 8(26):e53008
- Chowdhery A et al (2022) 'Palm: Scaling language modeling with pathways.' *J Mach Learn Res* 24(240):1–13
- Chukwurah EG (2024) Proactive privacy: advanced risk management strategies for product development in the US. *Comput Sci IT Res J* 5(4):878–891
- Cox A (1999) Power, value, and supply chain management. *Supply Chain Manag Int J* 4(4):167–175. <https://doi.org/10.1108/13598549910284480>
- Dhoni P, Kumar R (2023) Synergizing generative AI and cybersecurity: roles of generative AI entities, companies, agencies, and government in enhancing cybersecurity. *Authorea Preprints* 14:1–11
- Ding Y et al (2020) FraudTrip: taxi fraudulent trip detection from corresponding trajectories. *IEEE Internet Things J* 8(16):12505–12517
- Eibeck A et al (2024) "Research data supporting" a simple and efficient approach to unsupervised instance matching and its application to linked data of power plants. *J Web Seman.* <https://doi.org/10.1016/j.websem.2024.100815>
- Eisenberg I (2011) Lead-user research for breakthrough innovation. *Res Technol Manag* 54(1):50–58
- Erbati S (2024) Towards the optimal orchestration of service function chains to enable ultra-reliable low latency communication in an NFV-enabled network. Diss. Dissertation, Duisburg, Essen, Universität Duisburg-Essen
- Falade PV (2024) Deciphering ChatGPT's impact: exploring its role in cybercrime and cybersecurity. *Int J Sci Res Comput Sci Eng* 12(2):15
- Fetaji B (2024) Analyses of leveraging generative AI (GAI) for proactive cybersecurity.
- Fosso Wamba S, Queiroz MM, Chiappetta Jabbour CJ, Shi C (2023) Are both generative AI and ChatGPT game changers for 21st-century operations and supply chain excellence? *Int J Prod Econ* 265:109015
- Fui-Hoon Nah F, Zheng R, Cai J, Siau K, Chen L (2023) Generative AI and ChatGPT: applications, challenges, and AI-human collaboration. *J Inf Technol Case Appl Res* 25(3):277–304. <https://doi.org/10.1080/15228053.2023.2233814>
- Gadre SY et al. (2024) Language models scale reliably with over-training and on downstream tasks. *arXiv preprint arXiv:2403.08540*.
- Godakanda Arachchige PGB (2023) Detecting business email compromise and classifying for countermeasures.
- How Google Cloud Plans To Supercharge Security With Generative AI|Google Cloud Blog. Accessed: Jun. 22, 2023. [Online]. Available: <https://cloud.google.com/blog/products/identity-security/rsa>
- How Google Cloud Plans To Supercharge Security With Generative AI|Google Cloud Blog. Accessed: Jun. 22, 2023. [Online]. Available: <https://cloud.google.com/blog/products/identity-security/rsgoogle-cloud-security-ai-workbench-generative-ai>
- Gibert D (2024) Machine learning for windows malware detection and classification: methods, challenges, and ongoing research. *arXiv preprint arXiv:2404.18541*.
- Golda A, Mekonen K, Pandey A, Singh A, Hassija V, Chamola V, Sikdar B (2024) Privacy and security concerns in generative AI: a comprehensive survey. *IEEE Access* 12:48126

- Google AI: what to know about the palm 2 large language model. Accessed: Jul. 29, 2023. [Online]. Available: <https://blog.google/technology/ai/google-palm-2-ai-large-languagemodel/>
- GPT-4. Accessed: Jun. 22, 2023. [Online]. Available: <https://openai.com/gpt-4>
- Grover H et al (2021) Edge computing and deep learning enabled secure multitier network for the internet of vehicles. *IEEE Internet Things J* 8(19):14787–14796
- Grzybowski A, Pawlikowska-Łagód K, Lambert WC (2024) A history of artificial intelligence. *Clin Dermatol* 42:221
- Gupta P et al (2024) Generative AI: a systematic review using topic modeling techniques. *Data Inf Manag* 8:100066
- Gupta P, Ding B, Guan C, Ding D (2024) Generative AI: a systematic review using topic modelling techniques. *Data Inf Manag* 15:100066
- Hassija V, Chamola V, Gupta V, Jain S, Guizani N (2021) A survey on supply chain security: application areas, security threats, and solution architectures. *IEEE Internet Things J* 8(8):6222–6246
- Hoofnagle CJ, Van Der Sloot B, Borgesius FZ (2019) The European Union general data protection regulation: what it is and what it means. *Inf Commun Technol Law* 28(1):65
- <https://www.goldmansachs.com/insights/articles/generative-ai-could-raise-global-gdp-by-7-percent>
- Hu Z et al (2024) Dynamically retrieving knowledge via query generation for informative dialogue generation. *Neurocomputing* 569:127036
- Introducing AI-powered Investigation in Chronicle Security Operations|Google Cloud Blog. Accessed: Jun. 23, 2023. [Online]. Available: <https://cloud.google.com/blog/products/identity-security/rsaintroducing-ai-powered-investigation-chronicle-security-operations>
- Introducing Recorded Future AI: AI-driven Intelligence to Elevate Your Security Defenses. Accessed: Jul. 29, 2023. [Online]. Available: <https://www.recordedfuture.com/introducing-recorded-future-ai>
- Introducing virustotal code insight: empowering threat analysis with generative AI. Accessed: Jun. 22, 2023. [Online]. Available: <https://blog.virustotal.com/2023/04/introducing-virustotal-codeinsight.html>
- Jackson I, Ivanov D, Dolgui A, Namdar J (2024) Generative artificial intelligence in supply chain and operations management: a capability-based framework for analysis and implementation. *Int J Prod Res*. <http://doi.org/10.1080/00207543.2024.2309309>
- Jia Z et al. (2024) LangSuite: planning, controlling and interacting with large language models in embodied text environments. *arXiv preprint arXiv:2406.16294*.
- Jiang H et al (2021) A utility-aware general framework with quantifiable privacy preservation for destination prediction in LBSs. *Ieee/acm Trans Netw* 29(5):2228–2241
- Kam HJ, Zhong C, Johnston A (2024) The impacts of generative AI on the cybersecurity landscape.
- Kaushik K et al (2024) Ethical considerations in AI-based cybersecurity. *Next-generation cybersecurity: AI, ML, and blockchain*. Springer, Singapore, pp 437–470
- Khan FB et al (2024) Design and performance analysis of an anti-malware system based on generative adversarial network framework. *IEEE Access* 12:27683
- Kharrufa A, Johnson IG (2024) The potential and implications of generative AI on HCI education. *arXiv preprint arXiv:2405.05154*.
- Khoob B, Phan RC, Lim CH (2022) Deepfake attribution: on the source identification of artificially generated images. *Wiley Interdiscip Rev: Data Min Know Discovery* 12(3):e1438
- Kim J, Choi G (2024) Assessing the impact of generative artificial intelligence on customer engagement in business-to-customer scenarios. *Asia-Pacific J Conver Res Interchange*. <https://doi.org/10.47116/apjcri.2024.02.09>
- Kineber AF (2024) Identifying the internet of things (IoT) implementation benefits for a sustainable construction project. *HBRC J* 20(1):700–766
- Kumar S, Musharaf D, Musharaf S, Sagar AK (2023) A comprehensive review of the latest advancements in large generative AI models. *Communications in computer and information science*. Springer, Cham, pp 90–103
- Li LY et al (2024) HOT” ChatGPT: the promise of ChatGPT in detecting and discriminating hateful, offensive, and toxic comments on social media. *ACM Trans Web* 18(2):1–36
- Liu D, Cao Z, Jiang H, Zhou S, Xiao Z, Zeng F (2022) Concurrent low-power listening: a new design paradigm for duty-cycling communication. *ACM Trans Sensor Netw* 19(1):1–24. <https://doi.org/10.1145/3517013>
- Liu T, Xu H, Zhang L, Han Z (2024) Source selection and resource allocation in wireless-powered relay networks: An adaptive dynamic programming-based approach. *IEEE Internet Things J* 11(5):8973–8988
- Ma P et al. (2023) InsightPilot: an LLM-empowered automated data exploration system. In: *Proceedings of the 2023 conference on empirical methods in natural language processing: system demonstrations*.
- Ma J, Hu J (2022a) Safe consensus control of cooperative-competitive multi-agent systems via differential privacy. *Kybernetika* 13:426–439. <https://doi.org/10.14736/kyb-2022-3-0426>

- Ma J, Hu J (2022b) Safe consensus control of cooperative-competitive multi-agent systems via differential privacy. *Kybernetika* 13:426–439. <https://doi.org/10.14736/kb-2022-3-0426>
- Mavikumbure HS et al. (2024) Generative AI in cyber security of cyber physical systems: benefits and threats. Microsoft Security Copilot|Microsoft Security. Accessed: Jun. 22, 2023. [Online]. Available: <https://www.microsoft.com/enin/security/business/ai-machine-learning/microsoft-security-copilot> Accessed: Jun. 22, 2023. [Online]. Available: <https://cloud.google.com/vertex-ai>
- Microsoft Security Copilot|Microsoft Security. Accessed: Jun. 22, 2023. [Online]. Available: <https://www.microsoft.com/enin/security/business/ai-machine-learning/microsoft-security-copilot>
- Mitra A, Mohanty SP, Kougianos E (2024) The world of generative AI: deepfakes and large language models. arXiv preprint [arXiv:2402.04373](https://arxiv.org/abs/2402.04373).
- Morreel S, Verhoeven V, Mathysen D (2024) Microsoft Bing outperforms five other generative artificial intelligence ChatBots in the Antwerp University multiple-choice medical license exam. *PLOS Digital Health* 3(2):e0000349
- Motlagh FN et al. (2024) Large language models in cybersecurity: state-of-the-art. arXiv preprint [arXiv:2402.00891](https://arxiv.org/abs/2402.00891).
- Mozolevskiy D, AlShikh W (2024) Comparative analysis of retrieval systems in the real world. arXiv preprint [arXiv:2405.02048](https://arxiv.org/abs/2405.02048).
- Nguyen CT, Liu Y, Du H, Hoang DT, Niyato D, Nguyen DN, Mao S (2024a) Generative AI-enabled blockchain networks: fundamentals, applications, and case study. [arXiv:2401.15625](https://arxiv.org/abs/2401.15625).
- Nguyen CT, Liu Y, Du H, Hoang DT, Niyato D, Nguyen DN, Mao S (2024b) Generative AI-enabled blockchain networks: fundamentals, applications, and case study. *IEEE Netw* 39:232
- Novelli C et al. (2024) Generative AI in EU law: liability, privacy, intellectual property, and cybersecurity. arXiv preprint [arXiv:2401.07348](https://arxiv.org/abs/2401.07348).
- Ooi K-B et al (2023) The potential of generative artificial intelligence across disciplines: perspectives and future directions. *J Comput Inf Syst* 65:1–32
- Ozkan-Ozay M et al (2024) A comprehensive survey: evaluating the efficiency of artificial intelligence and machine learning techniques on cyber security solutions. *IEEE Access* 12:12229
- Pardau SL (2018) The California consumer privacy act: towards a European-style privacy regime in the United States. *J Tech L Poly* 23:68
- Pendzel S et al (2024) Generative AI for hate speech detection: evaluation and findings. Regulating hate speech created by generative AI. Auerbach Publications, New York, pp 54–76
- Purple AI|Empowering Cybersecurity Analysts With AI-driven Threat Hunting, Analysis & Response—Sentinelone. Accessed: Jul. 22, 2023. [Online]. Available: <https://www.sentinelone.com/blog/purple-ai-empowering-cybersecurity-analysts-with-ai-driven-threat-hunting-analysis-response/>
- Rabieinejad E, Yazdinejad A, Parizi RM, Dehghantanha A (2023) ‘Generative adversarial networks for cyber threat hunting in Ethereum blockchain.’ *Distrib Ledger Technol Res Pract* 2(2):1–19. <https://doi.org/10.1145/3584666>
- Sabherwal R, Grover V (2024) The societal impacts of generative artificial intelligence: a balanced perspective. *J Assoc Inf Syst* 25(1):13–22
- 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
- Sai S et al (2024a) Generative AI for transformative healthcare: a comprehensive study of emerging models, applications, case studies and limitations. *IEEE Access* 12:31078
- Sai S et al (2024b) Generative AI for cyber security: analyzing the potential of chatgpt, dall-e, and other models for enhancing the security space. *IEEE Access* 12:53497
- Sai S et al (2024c) Generative AI for cyber security: analyzing the potential of chatgpt, dall-e and other models for enhancing the security space. *IEEE Access* 12:53497
- Sannon S, Forte A (2022) Privacy research with marginalized groups: what we know, what’s needed, and what’s next. *Proc ACM Human-Comput Interact* 6(CSCW2):1–33
- Scorecard Inteates With Openais GPT-4|Securityscorecard. Accessed: Jul. 30, 2023. [Online]. Available: <https://securityscorecard.com/blog/scorecards-integrates-with-open/>
- Scorecardx|Securityscorecard. Accessed: Jul. 30, 2023. [Online]. Available: <https://securityscorecard.com/company/scorecardx/>
- Secure Enterprise Web Browser—Talon Cyber Security. Accessed: Jun. 22, 2023. [Online]. Available: <https://talon-sec.com/secureenterprise-browser/>
- Security Command Center|Google Cloud. Accessed: Jun. 22, 2023. [Online]. Available: <https://cloud.google.com/security-command-center>
- Shafik W (2024) The role of generative artificial intelligence in e-commerce fraud detection and prevention. Strategies for E-commerce data security: cloud, blockchain, AI, and machine learning. IGI Global, NY, pp 430–469

- Shao S et al. (2017) Real-time IRC threat detection framework. In: 2017 IEEE 2nd international workshops on foundations and applications of self* systems (FAS* W). IEEE.
- Sharma P, Jain S, Gupta S, Chamola V (2021) Role of machine learning and deep learning in securing 5G-driven industrial IoT applications. *Ad Hoc Netw* 123:102685
- Sindiramutty SR (2023) autonomous threat hunting: a future paradigm for AI-driven threat intelligence. arXiv preprint [arXiv:2401.00286](https://arxiv.org/abs/2401.00286)
- Singh S, Sulthana R, Shewale T, Chamola V, Benslimane A, Sikdar B (2022) Machine-learning-assisted security and privacy provisioning for edge computing: a survey. *IEEE Internet Things J* 9(1):236–260
- Singularity Skylight. Accessed: Jun. 22, 2023. [Online]. Available: <https://assets.sentinelone.com/skylight/singularity-skylight>
- Slashnext Launches Industry First Generative AI Solution for Email Security|Slashnext. Accessed: Jul. 29, 2023. [Online]. Available: <https://slashnext.com/press-release/slashnext-launches-industrys-firstgenerative-ai-solution-for-email-security/>
- Slashnext Launches Industry's First Generative AI Solution for Email Security. Accessed: Feb. 19, 2024. [Online]. Available: <https://www.prnewswire.com/news-releases/slashnext-launches-industry-first-generative-ai-solution-for-email-security-301757649.html>
- Sriram S (2024) Cyber security control systems for operational technology. *Industrial control systems*. Wiley, Hoboken, pp 285–302
- Sun G et al (2018a) Service function chain orchestration across multiple domains: a full mesh aggregation approach. *IEEE Trans Netw Serv Manag* 15(3):1175–1191
- Sun G et al (2018b) Cost-efficient service function chain orchestration for low-latency applications in NFV networks. *IEEE Syst J* 13(4):3877–3888
- Surameery NMS, Shakor MY (2023) Use chat GPT to solve programming bugs. *Int J Inf Technol Comput Eng* 31:17–22
- Takale DG, Mahalle PN, Sule B (2024) Cyber security challenges in generative AI technology. *J Netw Secur Comput Netw* 10(1):28–34
- Talon Cyber Security. Accessed: Jul. 12, 2023. [Online]. Available: <https://talon-sec.com/>
- The Recorded Future Intelligence Graph. Accessed: Jul. 30, 2023. [Online]. Available: <https://www.recordedfuture.com/platform/intelligence-graph>
- Threat Intelligence|Cyber Threat Intelligence Platform. Accessed: Jun. 22, 2023. [Online]. Available: <https://www.mandiant.com/advantage/threat-intelligence>
- Tyson J (2023) Shortcomings of ChatGPT. *J Chem Educ* 100(8):3098–3101. <https://doi.org/10.1021/acs.jchemeduc.3c00361>
- Vu T-H, et al. (2024) Applications of generative AI (GAI) for mobile and wireless networking: a survey. arXiv preprint [arXiv:2405.20024](https://arxiv.org/abs/2405.20024).
- Wang M (2024) Generative AI: a new challenge for cybersecurity. *J Comput Sci Technol Stud* 6(2):13–18
- Wang K, Dong J, Wang Y, Yin H (2019) Securing data with blockchain and AI. *IEEE Access* 7:77981–77989
- Wang T, Zhang Y, Qi S, Zhao R, Xia Z, Weng J (2023) Security and privacy on generative data in aigc: a survey. arXiv preprint [arXiv:2309.09435](https://arxiv.org/abs/2309.09435).
- Wazid M, Das AK, Chamola V, Park Y (2022) Uniting cyber security and machine learning: advantages, challenges and future research. *ICT Exp* 8(3):313–321
- Xia Y et al. (2024) AICoderEval: improving AI domain code generation of large language models. arXiv preprint [arXiv:2406.04712](https://arxiv.org/abs/2406.04712)
- Xu X, Liu W, Lean Yu (2022) Trajectory prediction for heterogeneous traffic agents using knowledge correction data-driven model. *Inf Sci* 608:375–391
- Xu Y, Wang E, Yang Y, Chang Y (2022a) A unified collaborative representation learning for neural-network-based recommender systems. *IEEE Trans Knowl Data Eng* 34(11):5126–5139
- Xu Y, Wang E, Yang Y, Chang Y (2022b) A unified collaborative representation learning for neural-network-based recommender systems. *IEEE Trans Knowl Data Eng* 34(11):5126–5139. <https://doi.org/10.1109/TKDE.2021.3054782>
- Xu H et al. (2024) Large language models for cyber security: a systematic literature review. arXiv preprint [arXiv:2405.04760](https://arxiv.org/abs/2405.04760)
- Yao Y et al (2024) A survey on large language model (LLM) security and privacy: the good, the bad, and the ugly. *High-Conf Comput* 4:100211
- Yigit Y et al. (2024) Critical infrastructure protection: generative ai, challenges, and opportunities. arXiv preprint [arXiv:2405.04874](https://arxiv.org/abs/2405.04874).
- Yigit Y, Buchanan WJ, Tehrani MG, Maglaras L (2024) Review of generative AI methods in cybersecurity. arXiv preprint [arXiv:2403.08701](https://arxiv.org/abs/2403.08701).
- Yosifova V (2024) application of open-source large language model (LLM) for simulation of a vulnerable IoT system and cybersecurity best practices assistance.

- Yu J, Lu L, Chen Y, Zhu Y, Kong L (2021) An indirect eavesdropping attack of keystrokes on a touch screen through acoustic sensing. *IEEE Trans Mobile Comput* 20(2):337–351
- Zhang X, Deng H, Xiong Z, Liu Y, Rao Y, Lyu Y, Li Y, Hou D, Li Y (2024) Secure routing strategy based on attribute-based trust access control in social-aware networks. *J Signal Process Syst* 96:1–18
- Zhang J et al. (2024) When llms meet cybersecurity: a systematic literature review.” *arXiv preprint arXiv:2405.03644*.
- Zhao L, Qu S, Xu H, Wei Z, Zhang C (2024) Energy-efficient trajectory design for secure SWIPT systems assisted by UAV-IRS. *Veh Commun* 45:100725
- Zhao C, Du H, Niyato D, Kang J, Xiong Z, Kim DI, Letaief KB (2024) Generative AI for secure physical layer communications: a survey. *arXiv preprint arXiv:2402.13553*.
- Zheng Y et al (2024) An overview of trustworthy AI: advances in IP protection, privacy-preserving federated learning, security verification, and GAI safety alignment. *IEEE J Emerg Select Top Circ Syst* 14:582
- Zheng W, Lu S, Cai Z, Wang R, Wang L, Yin L (2024) PAL-BERT: an improved question answering model. *Comput Model Eng Sci* 139(3):2729–2745. <https://doi.org/10.32604/cmes.2023.046692>
- Zhou E, Lee D (2024) Generative artificial intelligence, human creativity, and art. *PNAS Nexus*. <https://doi.org/10.1093/pnasnexus/pgae052>
- Zhou T et al (2023) In pursuit of beauty: aesthetic-aware and context-adaptive photo selection in crowdsensing. *IEEE Trans Know Data Eng* 35(9):9364–9377
- Zhou J, Ke P, Qiu X, Huang M, Zhang J (2023) ChatGPT: potential, prospects, and limitations. *Front Inf Technol Electron Eng* 2023:1–6
- Zhu F et al (2011) Mining top-k large structural patterns in a massive network. *Proc VLDB Endowment* 4(11):807–818
- Zou Z et al (2024) A pilot study of measuring emotional response and perception of LLM-generated questionnaire and human-generated questionnaires. *Sci Rep* 14(1):2781
- Al Zoubi AM (2024) Spam reviews detection models in multilingual contexts applying sentiment analysis, metaheuristics, and advanced word embedding

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Mueen Uddin¹ · Muhammad Saad Irshad² · Irfan Ali Kandhro³ · Fuhid Alanazi⁴ · Fahad Ahmed² · Muhammad Maaz² · Saddam Hussain⁵ · Syed Sajid Ullah⁶

✉ Mueen Uddin
mueen.uddin@udst.edu.qa

✉ Syed Sajid Ullah
syed.s.ullah@uia.no

Muhammad Saad Irshad
saadburney123@gmail.com

Irfan Ali Kandhro
Irfan@smiu.edu.pk

Fuhid Alanazi
alanazi@iu.edu.sa

Fahad Ahmed
fbnashfaq@gmail.com

Muhammad Maaz
Mazmazz733@gmail.com

Saddam Hussain
21h8564@ubd.edu.bn

¹ College of Computing and IT, University of Doha for Science and Technology, 24449 Doha, Qatar

- ² Department of Software Engineering, Sindh Madressatul Islam University, Karachi, Pakistan
- ³ Department of Computer Science, Sindh Madressatul Islam University, Karachi, Pakistan
- ⁴ Faculty of Computer and Information Systems, Islamic University of Madinah, Madinah 42351, Saudi Arabia
- ⁵ School of Digital Science, Universiti Brunei Darussalam, Jalan Tungku Link BE1410, Brunei Darussalam
- ⁶ Department of Information and Communication Technology, University of Agder, (UiA), N-4898 Grimstad, Norway