IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 9 March 2024, accepted 31 March 2024, date of publication 4 April 2024, date of current version 19 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3385107

SURVEY

Generative AI for Cyber Security: Analyzing the Potential of ChatGPT, DALL-E, and Other Models for Enhancing the Security Space

SIVA SAI[1], UTKARSH YASHVARDHAN[2], VINAY CHAMOLA[1,3], (Senior Member, IEEE), AND BIPLAB SIKDAR[4], (Senior Member, IEEE)

[1]Department of Electrical and Electronics Engineering, Birla Institute of Technology and Science, Pilani (BITS Pilani), Pilani Campus, Pilani 333031, India
[2]Department of Computer Science and Information Systems, Birla Institute of Technology and Science, Pilani (BITS Pilani), Pilani Campus, Pilani 333031, India
[3]Anuradha and Prashanth Palakurthi Centre for Artificial Intelligence Research (APPCAIR), Birla Institute of Technology and Science, Pilani (BITS Pilani), Pilani Campus, Pilani 333031, India
[4]Department of Electrical and Computer Engineering, National University of Singapore, Singapore 119077

Corresponding author: Vinay Chamola (vinay.chamola@pilani.bits-pilani.ac.in)

ABSTRACT This research paper intends to provide real-life applications of Generative AI (GAI) in the cybersecurity domain. The frequency, sophistication and impact of cyber threats have continued to rise in today's world. This ever-evolving threat landscape poses challenges for organizations and security professionals who continue looking for better solutions to tackle these threats. GAI technology provides an effective way for them to address these issues in an automated manner with increasing efficiency. It enables them to work on more critical security aspects which require human intervention, while GAI systems deal with general threat situations. Further, GAI systems can better detect novel malware and threatening situations than humans. This feature of GAI, when leveraged, can lead to higher robustness of the security system. Many tech giants like Google, Microsoft etc., are motivated by this idea and are incorporating elements of GAI in their cybersecurity systems to make them more efficient in dealing with ever-evolving threats. Many cybersecurity tools like Google Cloud Security AI Workbench, Microsoft Security Copilot, SentinelOne Purple AI etc., have come into the picture, which leverage GAI to develop more straightforward and robust ways to deal with emerging cybersecurity perils. With the advent of GAI in the cybersecurity domain, one also needs to take into account the limitations and drawbacks that such systems have. This paper also provides some of the limitations of GAI, like periodically giving wrong results, costly training, the potential of GAI being used by malicious actors for illicit activities etc.

INDEX TERMS Security, Artificial Intelligence, machine learning, natural language processing, learning systems.

I. INTRODUCTION

GAI has made a remarkable impact in a large number of fields throughout the world. GAI, prophecized to have a broad applicability by various influencial tech leaders has already started showing its influence over lots of different areas of the world be it (1) creative industries like art, music and literature [1] industry where GAI generated work is being deemed as great piece of art and is acting as a source of inspiration to the artists or in (2) content generation [2] where GAI is being used to streamline and automate content creation process or in (3) personalization and recommendation systems [3] where GAI is being used to generate personalized recommendations for various products and improve customer experience or in (4) healthcare industries or medical research [4] where GAI is being used to simulate biological processes, generate synthetic data lfor research, improve medical imaging analysis etc. or in (5) Virtual reality (VR) and augmented

The associate editor coordinating the review of this manuscript and approving it for publication was Tai-Hoon Kim.

reality (AR) [5] where it is being used in developing highly realistic and immersive AR/VR experiences or in (6) Natural language processing [6] and chatbots to engage in more polished human-like conversations or (7) to provide greater access to digital content to the people with disabilities and hence reduce communication gap etc.

One of the vital areas of the IT industry where GAI is emerging to show a more significant positive impact is the cybersecurity domain [7], [8]. GAI has the capability to allow the existing security products to detect advanced phishing attacks with greater accuracy [9], scan the network for malicious activity and automatically respond back to ongoing attacks with higher chances of success against them [8]. Current security products using AI take a defensive approach against an attack, that is, they start taking countermeasures once an attack has been made [10] while using GAI with these products will let them take an offensive approach against these threats, that is, it will allow the security team to always remain some steps ahead of the attack vendors by knowing in advance what is the most probable practice they might adopt to disrupt the security of their systems, thus enabling them to be prepared for such attacks and handling them more efficiently [11]. Also, the presence of large amounts of data in the cybersecurity field [12] allows for better training of the GAI model, which helps the model perform better to provide security to various systems [13].

The use of GAI, along with network protections like firewalls, network segmentation etc, can help us perform many tasks, from preparing to mechanized pressure testing. Using firewalls with GAI can offer us a quick at-a-glance solution to gain visibility into control shadow-IT usage and potential data loss vectors [14]. GAI can help us tackle lateral movement in ransomware attacks in traditional network-based segmentation by implementing a simplified zero-trust approach wherein continuous authentication and authorization are done to secure specific groups of applications and resources [15]. GAI has the capability to create conditions that mirror true situations, which can help us test security systems, assess their performance, identify faults in them and improve the overall readiness and security of the system [16]. After getting trained over a huge volume of domain-specific structured and unstructured data, such as malicious URLs, indicators of compromise (IOCs), malware samples etc [17], GAI becomes capable of exhibiting threat intelligence, which allows it to identify and respond to incoming threats to an institution efficiently [18].

Although GAI still has a long way to go when it comes to robustness and accuracy of prediction, there is no doubt that GAI is emerging as a promising tool that has the potential to significantly aid various industries, particularly the cybersecurity realm [19].

II. AN OVERVIEW OF GAI

The natural question that arises when we observe the GAI's influence over so many fields of importance is what is the secret to the power of GAI? The answer to that would be that the output of GAI is carefully calibrated combinations of the data fed to it to train its algorithm [20]. Due to the massive amount of data which goes into training the algorithm of GAI and the randomness with which it selects the elements of its output, its output appears to have creative and lifelike elements to it [21].

GAI models use advanced deep-learning techniques like Generative Artificial Networks, Variational Autoencoders, Transformers etc. [22], to learn patterns and structures from large datasets. Once trained, GAI models generate fresh content by sampling from this learned distribution.

Most of the hype generated by GAI is due to the significant attention being grabbed by ChatGPT [23] and DALLE [24]. ChatGPT is a free chatbot which can generate answers to almost any question asked of it. Developed by OpenAI, ChatGPT has gained much popularity for its ability to quickly and efficiently produce unique content, be it computer code, college-level essays, poems, etc. DALLE, which is a GAI tool with the ability to generate realistic images and art from natural language, is also gaining much admiration from users. Clearly, the GAI tools like ChatGPT and DALLE have the potential to change how various jobs are carried out for good, but the full extent of that effect is still unknown, as are the risks involved.

III. APPLICATIONS

There are several applications of GAI in the field of security (See figure 1). In this section, we present a detailed study on the same.

A. PASSWORD PROTECTION

By training the GAI over large password datasets, one can make its algorithm capable of identifying structures and patterns in commonly used passwords, thus enabling it to produce new passwords or prioritize specific combinations during the cracking process hence improving the efficiency of password guessing and aiding in password security assessments. One of the use cases in this context is PassGAN [25] which uses a Generative Adversarial Network to automatically learn the distribution of real passwords from actual password leaks, thus eliminating the requirement for manual password analysis. Training GAI models over large password datasets also gives it the capability to generate strong and unique passwords that adhere to the pattern of secure passwords learnt by its model. This can enable users to create passwords that are difficult to guess or crack, thus reducing the chances of unauthorized access to their system [26]. GAI can also be used to model and analyze the user's behaviour patterns related to password usage, such as login patterns, password changes, authentication failures, devices used to access the system, etc. This can allow GAI to detect unusual or abnormal behaviour that may point to compromised passwords or unauthorized access to the system, thus aiding in mitigating potential security violations.

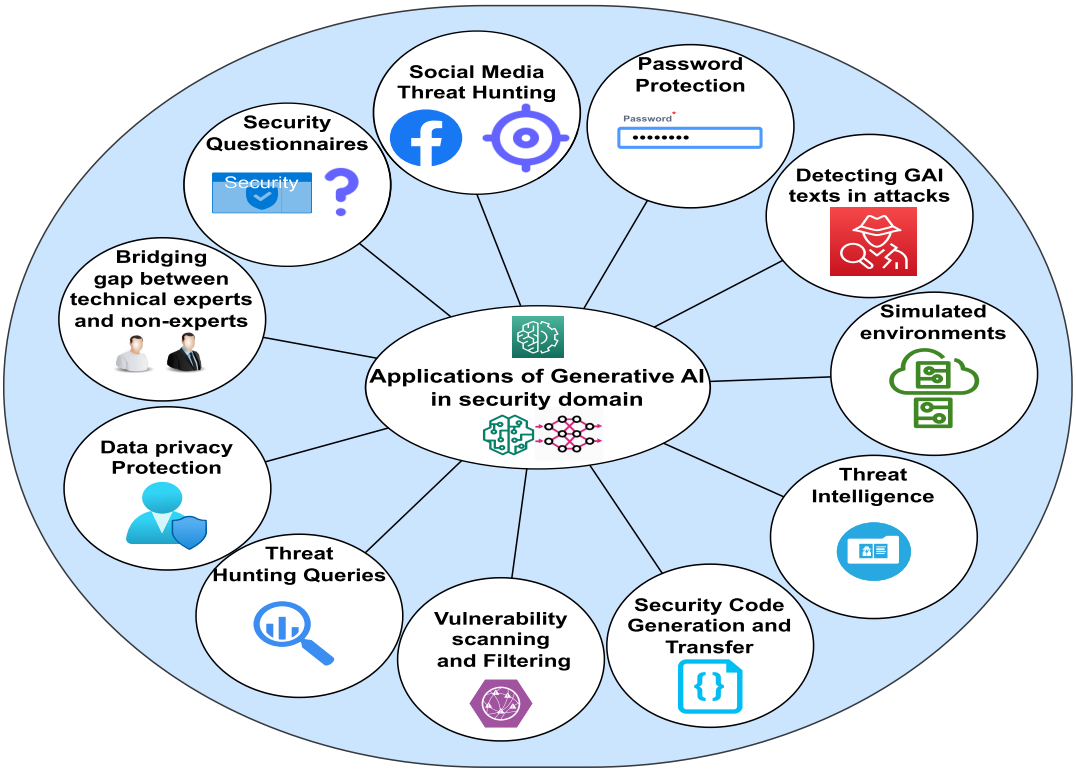


FIGURE 1. Applications of GAI in cyber security.

B. DETECTING GAI TEXT IN ATTACKS

Large Language Models (LLM) like Google LaMDA [27] and ChatGPT can be utilized to detect and watermark AI-generated text. This utility of LLMs can be used to identify phishing emails and polymorphic codes by detecting uncharacteristic email address senders or their respective domains or links in the text, which could lead to malicious sites.

C. GENERATE EXAMPLES OF ADVERSARIAL ATTACKS

GAI can be used to generate adversarial attacks specifically intended to expose the vulnerabilities of GAI text models. This can be achieved by creating carefully crafted input texts to trick the model into revealing its limitations and biases [28]. These adversarial examples can be used to understand the shortcomings of GAI models and hence develop more robust defences against future attacks [29].

D. SIMULATED ENVIRONMENTS

Using GAI, one can simulate threatening environments to mirror real-world situations, which can help security officials get trained to recognize and handle such threats and test and evaluate the effectiveness of their security systems. Intelligent and automated testing of such kind can significantly improve the overall posture of the security systems.

1) SIMULATED ATTACKS

By training over known attacks and techniques, GAI models can become capable of generating realistic attack scenarios and tactics to carry out red team simulations [30]. Red teams simulate real-world adversaries to detect vulnerabilities in an organization's security posture. GAI models can be used to produce attack patterns, phishing emails, social engineering scenarios or exploitative techniques that can be used to challenge an organization's security defences and test the effectiveness of their responses [31]. This can also enable the security systems to move from a defensive stance to a more proactive stance, where they will be able to predict the threats in advance. This can lead to further hardened security.

GAI can be used to generate simulated environments (cyber ranges) that can provide hands-on training to cyber security professionals. One of the real-life use cases is Draup [32] which is a talent intelligence platform that leverages GAI to help companies train their employees in emerging cyber security skills and roles. GAI can create realistic network topologies, traffic patterns and attack scenarios within a cyber range [33]. This can allow security officials to practice their abilities in a controlled environment, test different strategies and gain practical experience handling cyber attacks.

2) MALWARE AND INTRUSION DETECTION

Using Generative models, one can create realistic malware representations by learning from a huge set of malware data

[34], [35]. Using these synthetic samples, security officials can test and evaluate their malware detection systems, thus increasing the efficiency of their system in identifying and handling various malware variants. Some real-life use cases in this regard are SentinelOne Purple AI and Google Cloud Security AI Workbench, which provide efficient threat-hunting capabilities by leveraging GAI techniques. Further, GAI can be used to detect new or mutated variants of malware by learning underlying features and patterns shared among different malware families [36]. This can allow GAI models to detect and classify previously unseen malware variants based on their similarity to known patterns, thus increasing the robustness of the security system. Similarly, by training the GAI algorithm over standard network traffic data, one can make GAI generate a representation of normal network behaviour which can then be used to compare and detect anomalies in normal patterns of the network, which can enable security officials to detect potential security breaches or suspicious intrusion activities.

3) CREATING HONEYPOTS

GAI can be used to create persuasive decoy systems like fake websites and applications or persuasive decoy networks, which can be used to attract attackers. GAI can also be engaged in generating deceptive content along with honey pots. For example, GAI-enabled chatbots like ChatGPT, Meta LLaMA [37] etc., can be used to engage in human-like interactions with attackers and gain crucial information about their behaviour. This sensitive information about attackers can then be utilized to provide valuable insights into their modus operandi. The GAI models can be trained on real-time threat intelligence and attack datasets, enabling them to generate dynamic honeypots that mirror the latest attack trends. Thus, GAI can be utilized to automatically create and adapt the current honeypot to the present threat scenario, thus keeping the honeypots up-to-date and responsive to emerging threats. After the attackers engage with the system, the organizations can monitor the behaviour of the attackers and obtain valuable insights about the nature of their attacks and the methods and techniques used by them. The organization can later use this data to enhance their security system [38].

4) PHISHING RESILIENCE TRAINING

LLMs like ChatGPT, LLaMA etc., can be used to generate simulated phishing messages, which can be used to provide more frequent and practical phishing resilience training to the employees of an organization [39]. This can act as an excellent replacement for the current out-of-date cyber threat awareness training programs. LLMs can be leveraged to create phishing emails in which the employees are asked to recognize any suspicious signs of it being a phishing mail, like typos or unusual email addresses, and report back their decisions if it is a phishing email or not. Or simulate a baiting attack scenario where the employees are asked to recognize if the data or action being put forward in front of them (in the form of email etc.) is too good to be accurate and hence is a potential baiting attack or not.

By feeding such messages generated by ChatGPT (or any other efficient LLM) in an email marketing solution and sending them to the organization's employees, one can quickly create an easy and cost-effective phishing resilience training program for their organization [40] (Refer figure 4).

5) SYNTHETIC THREAT GENERATION

GAI models can be utilized to create artificial threat environments to test and evaluate the security of systems. GAI algorithms can learn from the real-life dataset to create synthetic scenarios that closely mirror actual attacks. They can be used to create synthetic malware samples, which can be used to train a GAN to learn the characteristics and patterns of malicious code. This knowledge can then be utilized to produce new malware variants with similar traits, which can be used by cyber security officials to test their systems against growing threats hence giving them valuable information as to how to increase the security of their systems. Reference [41] GAI can also be used for tasks like phishing campaign simulation, network traffic simulation, adversarial attack generation etc., which can allow security teams to assess their systems' vulnerability, evaluate their security measures' effectiveness and hence better develop appropriate defences against potential threats.

E. THREAT INTELLIGENCE

Using the massive dataset that GAI uses to train its learning algorithm, it becomes capable of identifying patterns and indicators of compromise hence giving it the strength to detect and handle threats in real-time before they have made it through the front-line defence. In some cases, GAI can also predict some other cyber security technology that may be required to better the security of the existing system. This aspect of GAI is different from the proactive stance mentioned above in the sense that GAI is looking at a threat in general and then determining the relevant features corresponding to a given system. In contrast, the proactive stance mentioned above uses an approach more limited to the organization to predict threats within it in advance. Real-world security tools like Google Cloud AI Workbench, SentinelOne Purple AI, SlashNext Generative Human AI etc., are capable of exhibiting threat intelligence to deal with ever-emerging threats more proactively. The ability of GAI to interpret the threats using both a broader and narrower approach allows it to become a better defensive weapon against existing and future threats [41].

F. SECURITY CODE GENERATION AND TRANSFER

LLMs like ChatGPT can be readily used to generate and transfer code meant to make a system more secure [41]. For example, suppose there was a phishing attack that successfully exposed the credentials of many employees

TABLE 1. A record of content types that LLMs could examine at present and possible new editions in the future.

Content Type	Potential applications	Availability
Document/PDF analysis	Document summary creation and research paper analysis	No
Image analysis	Content and image context detection	In 2023
Sound	Sound-based interactions and voice-to-text applications	No
Text/Conversations	Text analysis, chatbots, documentation	Yes
Video	Video content analysis and deepfakes creation	No

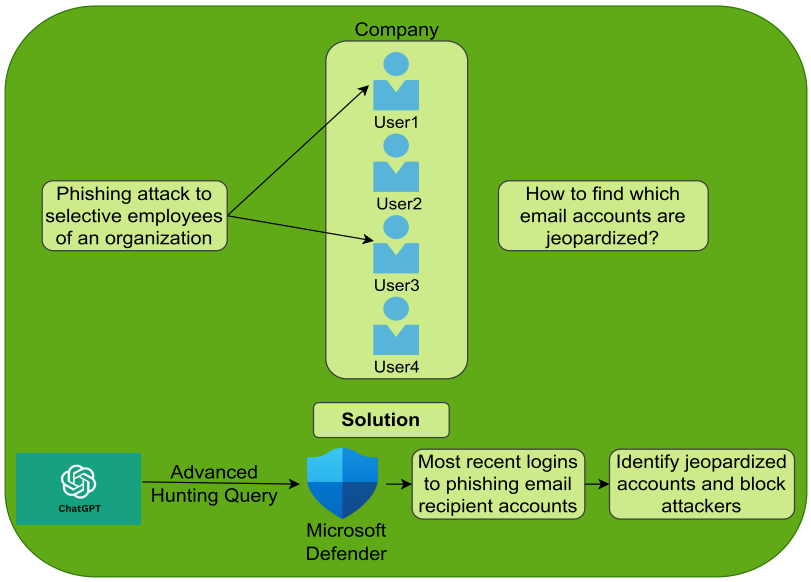


FIGURE 2. Security code generation using OpenAI's ChatGPT.

within the company. It is known which employees accessed the phishing email, but if they accidentally ran the code meant to steal their credentials - this needs to be clarified. To enquire into this situation, a Microsoft 365 Defender Advanced Hunting [42] query can be used to find the 10 most recent logins performed by the phishing email recipients within 30 minutes of them getting the malicious emails. These queries can then be used to identify login activity in relation to the compromised credentials.

Here ChatGPT can be used to give Microsoft 365 Defender Advanced Hunting query to check for login attempts to the jeopardized email accounts (Refer figure 2). This can help identify and block attackers and give users clarification on whether they need to change their login passwords or not. This can, in turn, help reduce response time to the cyber attack incident. The feature of ChatGPT—to use its underlying Codex [43] model to take a piece of code written in one programming language and convert it to another programming language can be used in this situation to do a programming language style transfer if the system on which the Microsoft 365 Defender hunting query is to be run does not operate with the KQL [44] programming language. This can simplify the end user's query process and help him perform queries smoothly.

G. VULNERABILITY SCANNING AND FILTERING

The GAI models can be trained on datasets comprising false positives, which can allow them to learn to generate filters or rules that distinguish actual vulnerabilities from benign ones hence allowing GAI to aid in reducing false positives in vulnerability scanning. This feature of GAI can be utilized to help security teams prioritize their efforts to address real vulnerabilities while reducing the time spent by them investigating false positive alerts. Also, GAI models can be trained to take context into consideration while filtering vulnerabilities by incorporating into training data features like system configurations, network topology, user access privileges, asset criticality etc. This can allow GAI to order the vulnerabilities based on their potential impact and exploitability within a specific setting. This capability of GAI can be utilized to help security officials to focus on vulnerabilities that pose the most significant risk to the organization [45].

GAI models can also be used to effectively scan various programming languages for vulnerability (Refer figure 3). They can be used to detect insecure code samples in numerous languages, thus aiding developers in addressing potential susceptibilities before they become significant security issues. These GAI models can also give valuable context

to the existing code, like customized code fix suggestions and cut-and-paste code fixes, along with the potential for the code to be exploited by malicious actors, the possible consequences of the attack and the potential harm to the organization, which can be leveraged to improve the security of the system further [46]. Veracode Fix is a specialized tool which leverages GAI technology to accomplish vulnerability detection and remediation in codes.

H. THREAT-HUNTING QUERIES

LLMs like ChatGPT, Meta LLaMA and Google LaMDA can be leveraged to create threat-hunting queries like queries for malware research and detection tools like YARA [47], thus enhancing the response time and efficiency of the system. This can allow cyber security officials to focus their efforts more on other critical aspects of security while GAI is working in the background to identify and handle potential threats swiftly. This utility of LLMs helps improve the robustness of security systems in a landscape of ever-evolving malicious activities. Further, GAI models can be trained to learn the standard pattern of typical system behaviour and network traffic from historical data. These models can then be leveraged to generate queries to identify the deviations and anomalies from the expected behaviour, which can be used to detect potential security breaches, data exfiltration attempts or unauthorized access within the system or network [48].

By training GAI models over real-time threat data, security alerts or incident reports, GAI can learn to dynamically adapt the queries it generates based on the evolving threat landscape [49]. The queries thus generated by such models can reflect the latest indicators or emerging threat trends. This can allow the cyber security teams to proactively hunt for threats by making appropriate queries to quickly identify newly arising threats which may not have been possible with static query generation. SentinelOne Purple AI, Google Cloud Security Workbench and Microsoft Security Copilot are some of the newly announced threat-hunting tools which utilize an LLM to improve the productivity of security analysts.

I. DATA PRIVACY PROTECTION

GAI models can be trained to produce shareable synthetic data, which can be used by organizations for various purposes like training ML models for detecting online frauds [50], customized product recommendations, loan eligibility predictions etc. This can reduce the sharing of customer data due to privacy concerns and data protection laws, which could lead to better data privacy protection [20]. Further, GAI can be used to develop privacy-preserving machine learning models. For example, GAI can be used to improve federated learning techniques to generate synthetic data on edge devices reflecting local patterns and characteristics. This can ensure the training of collaborative learning models without exposing sensitive data to the central server, thus preserving users' privacy [51].

The GAI models can be trained over a dataset of user interface designs, allowing it to learn visual elements and interaction patterns of interfaces. This knowledge can be employed to create privacy-aware user interfaces [52] that can lead to a reduction of exposure to sensitive information by using techniques such as masking/obscuring sensitive data fields, providing privacy focussed options during data sharing etc [53]. This could lead to better data privacy for users. Talon enterprise browser is one of the emerging examples which has embedded Microsoft Azure OpenAI [54] service within its framework to ensure better security for the organizations.

J. BRIDGING GAP BETWEEN TECHNICAL EXPERTS AND NON-EXPERTS

LLMs like OpenAI ChatGPT, DeepMind Chinchilla AI [55], Meta LLaMA etc. have the ability to articulate their thought process, which enables them to probe in and translate the functionality of various technical files, including source code, configuration files etc., in plain language. This can allow users with limited technical knowledge to understand the inner workings of these files and to comprehend their purpose, structure and potential consequences. This understanding can help them in making better technical decisions, thus preventing them from creating accidental cyber security issues. This ability of GAI can also help an organization bridge the gap between cyber security experts and non-experts by providing readily understood explanations for various cyber issues and hence helping them collaborate together better to ensure greater security of the system.

K. SECURITY QUESTIONNAIRES

The threat posed by third parties has a notable impact on the cyber security posture (nearly 60% of data exposures are caused by compromised third-party vendors [56]). One of the most efficient ways to recognize security hazards throughout our vendor network is by making use of security questionnaires. Instead of designing these questionnaires from scratch, one can use LLMs like ChatGPT to draft them [57]. GAI models can be trained on a large dataset comprising security-related questions and responses. This can enable the model to learn to generate questions covering various aspects of compliance, cyber security and risk management. This can lead to time-saving for security officials who just have to spend time now adjusting these questionnaires to improve their accuracy instead of designing the complete questionnaires from scratch. Further, GAI models can be used to adapt and improve the questionnaires generated by them according to the ongoing threat scenarios by training the GAI model over a dataset consisting of up-to-date security information, including real-life examples of security incidents. This can help an organization maintain a robust security posture and avoid potential security hazards.

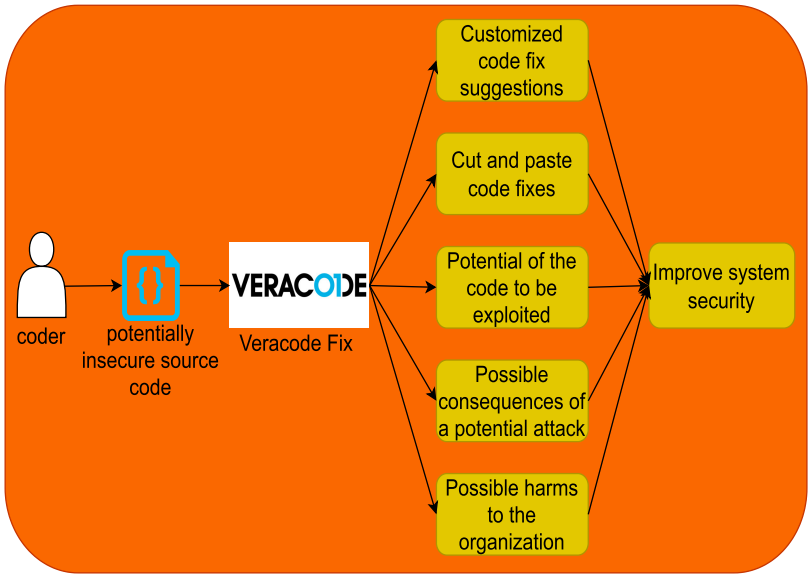


FIGURE 3. Vulnerability scanning using Veracode Fix.

L. SOCIAL MEDIA THREAT HUNTING

Social media threat hunting involves the analysis of data gathered from social media to identify vulnerabilities and potential hazards. Social media channels are scoured for specific keywords (based on the organization's perception of sensitive data) to find the possible exposure points of sensitive data or potential phishing attacks [58]. LLMs like ChatGPT, LLaMA, Chinchilla AI or LaMDA can be used for social media threat hunting by first collecting the data from various social media channels, feeding it to those LLMs for analysis and then making intelligent prompts to get desired information [59].

GAI models can be trained to examine the context of various social media posts, including the texts, images and user interactions. Hence by understanding the context, sentiment and intent of the social media content ChatGPT (or any other efficient LLM model) can be used to detect potential threats or harmful activities such as hate speech, extremist content or illegal activities. This can help in proactive monitoring and mitigation of risks [60]. Further, GAI models can be trained on datasets comprising known trustworthy or malicious users, which can help them learn to analyze user reputation or trustworthiness on social media platforms. This can allow the GAI models to produce credibility metrics for social media accounts which can help an organization in detecting and flagging suspicious or potentially malicious accounts for further examination [61].

There are several other fields where there is a potential for leveraging GAI for getting higher efficiency as well as time-saving. Refer Figure 2.

M. IoT SECURITY

GAI can be used to enhance the security of devices and environments within the Internet of Things. IoT refers to an interconnected network of devices that can communicate and exchange information over the Internet. These devices can range from smart home appliances to medical devices and industrial sensors [62]. The rapid proliferation of IoT devices has led to an increase in the number of significant security concerns. Many IoT devices lack robust security measures, thus making them vulnerable to cyberattacks.

GAI can help in IoT security by performing behavioural analysis and anomaly detection. It can be utilized to develop models of normal behaviour for different types of IoT environments and devices. By analyzing data from IoT sensors and devices, generative models can recognize divergence from standard patterns [63], signalling potential security breaches or unauthorized access, thus allowing the organization to take quick actions against the cyber threat hence promoting robustness in security. GAI can also play a significant role in IoT threat detection and prevention. It can be used to simulate various attack scenarios on IoT systems to identify their potential vulnerabilities, thus allowing security professionals to develop better strategies to prevent and mitigate cyberattacks targeting IoT devices [64]. It can also help in analyzing the firmware and software of IoT devices for potential security flaws or vulnerabilities. It can generate synthetic code to test the device's response to different inputs and potential attack vectors. Based on the results, devices' firmware and software security measures can be improved to meet the security demands of the company. GAI can aid in designing and testing secure authentication methods for IoT devices, including biometric or multi-factor authentication, thus, assisting in generating synthetic authentication data to evaluate the effectiveness of access control mechanisms [65]. It can also aid in the anonymization and aggregation of IoT data to protect user privacy while still allowing for meaningful analysis.

It can generate synthetic data that resembles actual data to help preserve the confidentiality of sensitive information. By leveraging GAI in IoT security, organizations can enhance the resilience of IoT devices against cyber threats, proactively identify and address security risks and provide a safer and more secure environment for users and their interconnected devices [66].

N. DEEPFAKE DETECTION AND PREVENTION

Deepfakes are highly realistic artificial media developed using GAI techniques, often involving the manipulation of images, videos, or audio to depict events or situations that did not happen in real life. While deepfakes have potential positive applications, such as in entertainment and visual effects, they also raise significant concerns related to misinformation, fraud, and privacy violations [67]. GAI tools can be employed to detect and mitigate the adverse impacts of deepfakes.

It can be used to develop algorithms to analyze visual and audio cues to recognize inconsistencies in deepfake content. These cues may include unnatural facial expressions, incorrect lighting, mismatched lip movements etc. They can also be trained on a diverse dataset of both real and deepfake content to enable them to learn to differentiate between authentic and manipulated media [68]. GAI can also help develop forensic tools to trace the origin and editing history of digital media, helping one verify the authenticity of the content. It can generate visualizations that highlight alterations and anomalies in deepfake content, aiding in identifying tampered elements. It can also aid in the development of techniques to counteract the effects of deepfakes, such as watermarking, content authentication, media fingerprinting/hashing and blockchain-based verification [69].

GAI can create simulated deepfake examples for educational purposes, helping individuals understand the potential risks of deepfakes and recognize the signs of manipulated media. Significant challenges to media authenticity and trust are posed due to the rapid advancement of deepfake technology [70]. GAI plays a crucial role in staying ahead of these challenges by enabling the development of sophisticated detection methods and mitigation strategies. As the arms race between deepfake creation and detection continues, GAI will be pivotal in ensuring the accuracy and reliability of digital media in an increasingly interconnected world.

O. SUPPLY CHAIN SECURITY

Supply chains involve the movement of goods, information, and resources across various stages, from raw material acquisition to production and distribution. Ensuring the integrity and authenticity of products within the supply chain is crucial to prevent counterfeiting, tampering, and other security threats [71]. GAI can play a significant role in enhancing supply chain security through various methods. It can be used to design unique identifiers, such as QR codes or holograms, for each product to ensure authenticity and provide consumers with a means to verify products [72]. GAI can create holographic labels or seals with intricate patterns that are difficult to replicate, enhancing the security of product packaging. GAI can also perform optimization of package designs to make tampering or unauthorized access more evident, deterring potential security breaches [73]. It can also assist in creating digital maps of supply chain routes, locations, and intermediaries, helping to track the journey of goods and identify potential vulnerabilities [74].

GAI can analyze historical supply chain data to predict potential security risks, enabling proactive risk mitigation and response planning. Supply chain security is crucial to maintaining consumer trust, product quality, and regulatory compliance. GAI offers innovative solutions to detect counterfeits, prevent tampering, and ensure the authenticity of products as they move through complex supply chain networks. Organizations can build a more secure and transparent supply chain ecosystem by applying GAI techniques [75].

P. BLOCKCHAIN SECURITY

GAI can have several valuable applications in enhancing the security of blockchain technology. Blockchain is known for its decentralized and secure nature, but GAI can further strengthen its capabilities and address certain challenges [76].

One way is it can help in the analysis of smart contracts to identify potential vulnerabilities and security risks before they are deployed on the blockchain. It can simulate various scenarios to test the behaviour of smart contracts and ensure they function as intended without exposing security loopholes. GAI can also assist in generating highly secure and random private keys for users, enhancing the cryptographic strength of blockchain accounts. It can provide help to the users to securely manage their private keys by providing recommendations for key storage and backup methods [77].

GAI can monitor blockchain network activities and transactions to detect unusual patterns or behaviours that may indicate security breaches or attacks [78]. It can generate predictive models that recognize potential threats and trigger proactive security measures. Further, it can assist in the integrity of blockchain data by creating digital signatures or hashes, ensuring its integrity and authenticity throughout the entire chain. It can also assist in verifying data consistency across distributed nodes to detect any discrepancies or unauthorized changes [79].

Incorporating GAI into various aspects of blockchain security not only strengthens the technology's resilience against attacks and vulnerabilities but also fosters greater trust in the adoption of blockchain solutions across industries [80]. These applications empower organizations to overcome security challenges and fully leverage the benefits that blockchain technology offers.

IV. CUSTOMIZED LLMs FOR SECURITY

This section discusses three LLMs that are customized for security.

A. BigID BigAI LLM

BigID's BigAI [81] is a LLM designed to analyze and categorize organizations' data to optimize their security and boost their risk management endeavours. It allows enterprises to scan structured and unstructured data, which can be stored in either cloud or on-premises and suggest titles and descriptions for data tables, columns and clusters using a combination of ML-driven classification and GAI techniques so that they are easier to locate via search.

One of the notable features of BigAI is that it approaches the issue while keeping consumer data private, thus enabling the LLM to work on the enterprise's data with the vendor's private servers and models without revealing any sensitive information to public models. BigID uses a virtual personal assistant called BigChat, which was created with the goal of acting as a copilot, allowing the management of compliance challenges. This acts as a critical discriminator of BigID's BigAI from services like Google Cloud Security AI, which focuses mainly on threat identification and management use cases. BigChat utilizes BigID's documentation, guides, articles, forum posts and PDFs to answer users' queries. For example, a user could query to detect and categorize sensitive data to ensure adherence to privacy regulations like GDPR [82] and CCPA [83].

Suppose one trains an LLM on regulated customer data (essentially data outside the defined purpose) [84]. In that case, it can lead to a violation of user privacy and can accelerate risk on data one knows and the data one does not know. Even when training LLMs on confidential intellectual property, it likely elevates the danger that the private information will be breached, leaked or hacked. BigID BigAI can allow one to categorize, designate and tag data by regulation, type sensitivity and even intention of use - across structured and unstructured data and everything within the range [85]. That makes it simple to categorize and designate sensitive data related to customers, privacy, regulated intellectual property etc. By doing so, organizations can choose appropriate data sets that are more low-risk and relevant to train LLMs more efficiently to get more accurate results. One can choose, for example, to leave out sensitive human resources data and avoid violating collected and tagged employee data. As an alternative, one can use public, non-confidential data to train LLMs, thus ensuring that they are trained on data that will not compromise security or privacy.

BigID's BigAI can help organizations identify, filter and administer both structured data for rational AI and unstructured data for novel conversational AI. BigID has the capability to enable consumers to extend data administration and security to modern conversational AI and LLMs, thus making sure that they are acting in a responsible manner. It is a service that is built on profound data exploration. It has the capability to leverage GAI and make data simpler to identify, analyze and ultimately secure.

B. SLASHNEXT GENERATIVE HUMAN AI

SlashNext Generative Human AI is the industry's first Artificial Intelligence solution that uses GAI to defend against advanced business email compromise (BEC) [86], supply chain attacks, executive impersonation and financial fraud [87].

This new solution created by SlashNext joins its existing HumanAI capabilities, which mimic human threat researchers by amalgamating natural language processing, computer vision and machine learning with relationship graphs and deep contextualization to thwart sophisticated multi-channel messaging attacks. It anticipates vast numbers of potential AI-generated BEC threats by using techniques such as AI data augmentation and cloning to assess core threats and then generate thousands of other versions of that same core threat, which enables the system to train itself on possible variations. The following are the features of HumanAI:

1) BEC GAI AUGMENTATION

HumanAI has the capability to generate thousands of new BEC variants from current threats to stop attacks in the future [88].

2) RELATIONSHIP GRAPHS & CONTEXTUAL ANALYSIS

HumanAI provides a baseline of known good communication patterns and writing styles for each employee and supplier to detect unusual communication and conversation types [89].

3) NATURAL LANGUAGE PROCESSING

HumanAI can perform analysis over text in the email body and attachments for the topic, tone, emotion, intent and manipulation triggers associated with social engineering strategies [90].

4) COMPUTER VISION RECOGNITION

HumanAI leverages SlashNext's LiveScan [91] to inspect URLs in real-time for any visual deviations, such as images and layouts, to detect credential phishing webpages. For example, HumanAI uses computer vision to detect highly subtle divergence from imposter Microsoft 365 log-in pages and block access.

5) FILE ATTACHMENT INSPECTION

HumanAI also has the capability to break down social engineering traits of attachments and malicious codes to stop ransomware.

6) SENDER IMPERSONATION ANALYSIS

HumanAI also evaluates headline details and email authentication results to stop impersonation attacks. There is a

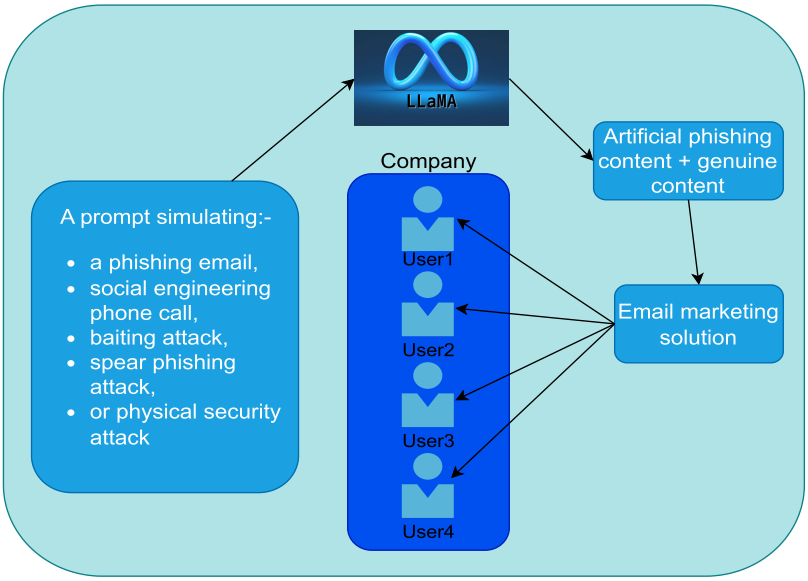


FIGURE 4. Phishing resilience training using Meta LLaMA.

massive database that SlashNext uses as the source for zero-hour detections to analyze around 700,000 new threats per day [87]. It sources threats for human compromises with any security vendor's most virtual sandbox crawlers. So SlashNext HumanAI is highly effective. HumanAI also has the unique ability to spot how threat actors play off human emotions, such as sending fear-generating requests that ask users to take wrong actions quickly. These include requests like "Urgent!" or "Hurry up!" etc. HumanAI simulates those same emotions and behaviours in its detection process.

C. Sec-PaLM

Sec-PaLM is based on PaLM 2 [92], a next-generation LLM built by Google over its previous work PaLM [93]. It is a cutting-edge model with enhanced multilingual, reasoning and coding abilities [94]. A comparison between GPT-4 and PaLM 2 is given in table 2.

Sec-PaLM is a specialized version of PaLM 2, which has been trained on datasets with security use cases. It is a potential breakthrough in cyber security analysis. It is accessible through Google Cloud, and it uses AI to help study and explain the behaviour of scripts that have the potential to be malicious and to better detect the scripts that pose risks to people and companies in extraordinary times.

V. REAL WORLD SCENARIOS

Various real-world cyber security products are leveraging the power of GAI with the aim to strengthen their security fronts (Refer figure 5). Some of them are listed below:

A. SENTINELONE PURPLE AI

SentinelOne Purple AI [95] is a threat-hunting tool released by SentinelOne for its Singularity Skylight [96] platform. The tool utilizes an LLM intending to make threat hunting more accessible hence leading to a significant increase in the productive capacity of security analysts. One of the LLMs that SentinelOne utilizes to implement Purple AI is OpenAI's GPT-4 [97]. Moreover, SentinelOne has trained LLM models on its own data, doing much fine-tuning on the way to customize them to give high-quality results in the security domain. Some of the features of SentinelOne Purple AI are as follows:

1) BOOSTING CYBER DEFENSE

The traditional threat-hunting tools are often too sophisticated and tedious to use and require dedicated training to operate on them for professional purposes. Using Purple AI, security analysts will be able to query the tool for potential threats in the environment with simple statements like "Is there any threat actor (of a particular kind) present in this (or any other) environment?" and the Purple AI will be able to give accurate results to them based on their query. This will help security analysts detect and identify potential threats with greater ease and speed, thus allowing them to save time which can be utilized to deal with more critical tasks at hand [98]. Purple AI is also capable of summarizing the results of its action, which helps alleviate tedium around doing actual analysis.

2) EMBEDDING GAI

Purple AI tool is provided as an add-on to the Singularity Skylight platform, and the interface is made available directly within the platform. The add-on feature of the tool can give users the flexibility to use the app as per their comfort, thus allowing them to get accustomed to the GAI tool gradually [99].

3) LLM

SentinelOne utilizes the capabilities of several LLMs, both open-source and proprietary, to increase the efficiency of their model. SentinelOne primarily uses third-party models like OpenAI's GPT-4 as a pre-trained model and then fine-tunes it to make their threat-hunting tool suitable for working efficiently in the security domain.

4) EVOLVING THE PLATFORM

The purple AI threat-hunting tool allows the security operations centre team to increase their efficiency and scale up their threat-hunting activities due to the ease of its use. A typical threat-hunting platform requires a high level of skill for its usage. The SentinelOne Purple AI platform allows a user to interact with it without needing a high level of technical skill, thus making threat hunting simple and feasible for non-technical people. This allows organizations to improve their security without investing in additional human capital.

B. GOOGLE CLOUD SECURITY AI WORKBENCH

Google's Cloud Security AI Workbench [100] is powered by a new security-specific LLM called Sec-PaLM. The model makes use of Google Cloud's security intelligence via Google's broad visibility into threat data and Mandiant's acclaimed threat intelligence around susceptibilities, malware, threat actors and threat indicators. The Workbench aims to reduce overload from threat data, many complex security tools in use and the talent gap. The customers can feed their private data to the platform at inference time. They can also control their data with enterprise-grade abilities like data isolation, protection, sovereignty, and compliance support. This is possible due to Security AI Workbench being built on Google Cloud's Vertex AI [101] infrastructure. Some feature offerings of Google Cloud Security AI Workbench are as follows:

1) CONTAINING THREATS FROM SPREADING BEYOND THE FIRST INFECTION

With Cloud Security AI Workbench in the picture, Google is able to achieve the combination of highly efficient threat intelligence with point-in-time incident analysis and novel AI-based detections and analytics to help counter a potential cascade in adversarial attacks using AI/ML systems.

Some Google Cloud Security AI Workbench tools which possess such capabilities are:

- *VirusTotal Code Insight* [102], which uses Google's Sec-PaLM to examine and explain the behaviour of potentially malicious scripts, and
- *Mandiant Breach Analytics for Chronicle* [103], which utilizes Google Cloud and Mandiant's Threat Intelligence to automatically alert us in case of the presence of active breaches in our environment. It also uses Sec-PaLM to provide context and effectively respond to these crucial issues.

2) ADDING INTELLIGENCE TO REDUCE TOIL

Google Cloud Security AI Workbench can enable organizations to simplify their security tools and controls and empower their systems to secure themselves. Some Google Cloud Security AI Workbench tools which can provide organizations with such capability are:

- *Assured OSS* [104], which utilizes LLMs, can be used to help Google include more OSS packages in their OSS vulnerability solution, which offers the same curated and vulnerability-tested packages that are used at Google.
- *Mandiant Threat Intelligence AI* [105], which is based on Mandiant's huge threat graph, can be used to leverage Sec-PaLM to quickly find, summarize, and act on threats pertinent to the company.

3) TRANSFORMING HOW PRACTITIONERS DO SECURITY TO CLOSE THE TALENT GAP

Cloud Security AI Workbench offered by Google can be utilized to make security more approachable and understandable even for those who are not security specialists by training. Two of the solutions provided by Google in its Cloud Security AI Workbench in this regard are:

- *Chronicle AI* [106]: Chronicle AI allows its users to search billions of security events, have real-time interaction with the results, and quickly generate detections without possessing high cyber security skills.
- *Security Command Center AI* [107]: The Security Command Centre has the capability to present complex attack graphs in easily understandable explanations. It can also provide information about impacted assets and recommendations for mitigation. Further, it can summarize security, compliance and privacy findings for Google Cloud in an effective manner.

C. MICROSOFT SECURITY COPILOT

Microsoft Security Copilot [108] is an AI assistant for cyber security professionals released by Microsoft. It can help defenders identify breaches and analyze and derive essential results from the vast amount of data available to them daily. It uses information from the cyber security and Infrastructure Security Agency, the National Institute of Standards and Technology's vulnerability database, and Microsoft's own threat intelligence database to get information on the latest vulnerabilities.

Microsoft Security Copilot is powered by OpenAI's GPT-4 GAI and Microsoft's own security-specific models. Its interface has a simple prompt box using which security professionals can gain help in security investigations, summarizing events for reporting, getting a summary of feed in files, URL content, code snippets for analysis or incident information from other security tools [109]. All the prompts to and the responses generated by the Copilot are saved in a database so there is a full inspection trail for investigators. Behind the scenes, the Copilot makes use of 65 trillion daily signals collected by Microsoft in its threat intelligence and

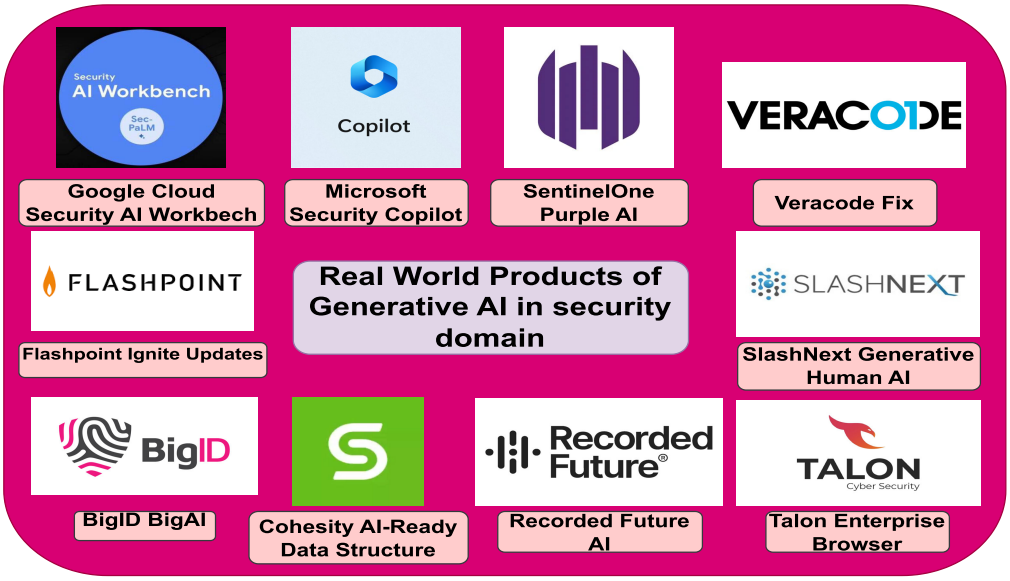


FIGURE 5. Real world products of GAI in cyber security domain.

security-specific skills, which allows security officials to hunt down threats efficiently. The Copilot is aimed at assisting security analysts rather than replacing them. To accomplish this purpose, a feature has been provided in Copilot where the results of various transactions can be pinned in a shared workspace to encourage colleagues to work on similar threat analyses and investigations collaboratively. One of the most attractive features that the Copilot provides is the prompt book feature, whereby the security officials will be able to club together a sequence of steps or automations into a single easy-to-use button or prompt. This could involve creating a shared prompt to reverse engineer a script, which could allow the security teams to perform this analysis without having to wait for an expert to do it for them. The Copilot can also be used to create a PowerPoint slide demonstrating various incidents and attack vectors. The Copilot system also incorporates a feedback mechanism whereby the users will be able to provide situations and responses involved when the Copilot gives incorrect results to get a better understanding of its hallucinations.

D. TALON ENTERPRISE BROWSER

Talon Cyber Security [110] has integrated Talon Enterprise Browser [111] with Microsoft Azure OpenAI Service to allow more enterprise-grade access to GAI text generator tools like ChatGPT. This will enable the users to leverage existing Azure resources, maintain data protection, keep data inserted in ChatGPT within a secure perimeter and prevent third-party service transfers. The browser also provides the administrators with the capability to block enterprise browser users from inputting specific data into an embedded ChatGPT window. They can also prevent the users from entering sensitive information in the Browser like credit card numbers, source codes etc. The Talon Enterprise Browser also provides some productivity features, like having the AI generate a response to an email or a summary of a large message. It also allows organizations to create reports demonstrating compliance with query logs and block extensions using public ChatGPT.

E. SLASHNEXT GENERATIVE HUMAN AI

SlashNext Generative Human AI is the industry's first AI service that uses GAI to defend against advanced supply chain attacks, business email compromise (BEC), financial fraud and executive impersonation [112].

This new solution created by SlashNext joins its existing HumanAI capabilities, which mimic human threat researchers by amalgamating machine learning, computer vision and natural language processing with deep contextualization and relationship graphs to foil advanced multi-channel messaging attacks. It predicts large numbers of potential AI-generated BEC hazards by using techniques such as AI data augmentation and cloning to evaluate central threats and then, using that evaluation, generate a large number of other versions of that same core threat, which allows the system to train itself on possible variations.

The following are the features of HumanAI:

1) BEC GAI AUGMENTATION

HumanAI has the capability to generate thousands of novel BEC variants from current threats to prevent breaches in the future [113].

2) RELATIONSHIP GRAPHS & CONTEXTUAL ANALYSIS

HumanAI provides a base reference of known good communication conventions and writing styles for all the employees and suppliers so as to identify unusual communication and conversation types.

TABLE 2. A comparison between GPT-4 and PaLM LLM models.

Model Features	GPT-4	PaLM 2
Creator	OpenAI	Google
Model Architecture	Transformer-style architecture	Small, medium and large variants of PaLM 2 model architecture exists that use Transformer based stacked layers, with varying parameters depending on size of the model
Model size	1.5 trillion parameters	340 billion parameters
Pre-training	Self-supervised learning	Self-attention
Attention mechanism	Yes	Yes
Fine-tuning	Supervised learning	Supervised learning
Multimodal training data	Yes	Yes

3) NATURAL LANGUAGE PROCESSING

HumanAI can perform analysis for the tone, emotion, topic, manipulation triggers and intent associated with social engineering strategies over text in the email body and attachments [114].

4) COMPUTER VISION RECOGNITION

HumanAI uses SlashNext's LiveScan for real-time analysis of URLs for any visual divergences, such as layouts and images, to identify credential phishing webpages. For example, HumanAI leverages computer vision to identify highly subtle divergences from fake Microsoft 365 log-in pages and block access [115].

5) FILE ATTACHMENT INSPECTION

HumanAI also has the capability to stop ransomware by breaking down social engineering attributes of attachments and malicious codes.

6) SENDER IMPERSONATION ANALYSIS

HumanAI also assesses email authentication results and headline details to prevent impersonation attacks. There is a massive database that SlashNext uses as the source for zero-hour detections to analyze around 700,000 new threats per day. It sources threats for human compromises with any security vendor's most virtual sandbox crawlers. So SlashNext HumanAI is highly effective. HumanAI also has the unique capability to recognize how threat actors exploit human emotions, like sending fear-generating requests that urge users to make wrong decisions quickly. These include requests like "Urgent!" or "Hurry up!" etc. HumanAI, in its detection process, simulates those same behaviours and emotions.

F. RECORDED FUTURE AI

Recorded Future incorporates OpenAI's transformer model, GPT, with its Intelligence Cloud leading to the game-changing impact of AI in the intelligence industry. Recorded Future AI model [116] is trained on over a decade's worth of threat analysis data from Insikt Group, the company's threat research division [117], and later combined with the insights of the Recorded Future Intelligence Graph [118]. It automatically collects and organizes data associated with both adversaries and victims from text, imagery and technical sources and uses NLP and ML techniques to analyze and map insights across billions of objects in real-time. Recorded Future AI provides real-time threat landscape analysis and actionability at the internet scale, enables analyst efficiency to help make up for skills shortage, and provides intelligence-driven understandings to enable businesses to make decisions before competitive activity affects business outcomes. Recorded Future has developed sophisticated algorithms and analytics over the past decade in order to collect, process, and analyze over 100 TB of text, images, and technical data in order to convert them into relevant and refined intelligence. Using this information, organizations can identify potential hazards and vulnerabilities and take preventative measures against attacks. This can be especially useful in the case of cyber and physical threat landscapes that are so complex that traditional threat intelligence approaches may not be sufficient to combat them. Recorded Future's AI-powered product seeks to free up 80% of the time analysts spend on tasks such as collection, aggregation processing, etc., so they can focus more on analysis, reporting, risk mitigation and organization security. Recorded Future AI is being utilized to create the Recorded Future Intelligence Cloud [118], the world's most intelligent intelligence repository. By integrating AI and ML deeper in the intelligence cycle at the analysis, production, and distribution stages, Recorded Future AI is able to rapidly determine and assign importance to the most critical threats and vulnerabilities and provide analysts with actionable intelligence in real-time. Using AI's strength, it is possible to automate many time-consuming duties associated with threat analysis, allowing analysts to concentrate on higher-level strategic tasks.

G. SECURITYSCORECARD GPT-4 INTEGRATION

SecurityScorecard is a platform for security evaluations that incorporates OpenAI's GPT-4 into its system. With this

capability for natural language processing, cyber security officials can find prompt responses and mitigation suggestions for high-priority cyber risks. This can allow security professionals to pose questions to gain a better understanding of their cyber exposure, rapidly identify their security gaps, and enhance their overall cyber resilience [119].

This solution was created by ScorecardX [120], the innovation centre of SecurityScorecard, which designs and develops technological solutions to address critical client issues. With this solution in the picture, customers will have the capability to ask open-ended inquiries about their business ecosystem, such as ''find my ten lowest-rated vendors'' or ''show me which of my crucial vendors were breached in the past year,'' and obtain answers swiftly to drive enhanced risk management decisions. The AI-powered search works across all of the organizations whose security ratings are tracked by SecurityScorecard, sparing executives a substantial amount of time by reducing the amount of manual labour required for data analysis. The search function can continue to learn and develop to meet customer requirements better.

VI. LIMITATIONS

Although GAI can make significant contributions to the security domain, its application is not without limitations. Table 3 presents some regulatory challenges posed when using GAI in the cyber security domain [121]. Some of the limitations of using GAI in the security domain are as follows:

A. TENDENCY TO GIVE WRONG/UNETHICAL RESULTS

Being such novel models, the long-term effects of GAI are yet to be seen. This implies that inherent risks are involved in using them - some known and some unknown. For example, LLMs like ChatGPT can sound too convincing while providing information, which can sometimes turn out to be wrong. ChatGPT can also suffer from social biases, which can be manipulated to enable criminal or unethical activity [122].

B. COST INEFFICIENCY

GAI systems, when leveraged for security, can turn out to be highly costly. Only the companies which can afford the high price and brain power needed to set up and maintain such GAI systems will be able to have the security required to secure their data and resources, and others will have to settle for less secure methods to do that which can lead to them becoming more vulnerable to the ever-increasing threat landscape. Due to such ethical challenges, some organizations may start getting subsidies related to GAI tools to help, say, non-profit organizations keep their personal data safe [123].

C. HIGH SETUP TIME

GAI models take a significant amount of time to train, ranging from a few weeks to a few months. Due to its high set-up time, it can act as an impediment for organizations that want to move quickly [124].

D. EASY EXPLOITABILITY BY MALICIOUS ACTORS

One of the significant drawbacks of a GAI system is that if such a system lands in the hands of malicious actors, then they can exploit it to identify attack vectors that reveal unauthorized access points easily and use them to breach into their system. They can also use it to develop malware and exploits and craft highly convincing phishing emails and messages [125].

E. INTERPRETABILITY AND EXPLAINABILITY

GAI models have complex architectures and black box nature of operations. This can make it difficult to interpret or explain the results output by the GAI systems, which can limit their use in critical security systems where interpretability is crucial [126].

F. CONTEXTUAL LIMITATIONS

GAI might face challenges with understanding context and generating coherent responses in certain situations, leading to nonsensical or irrelevant output. This can lead to confusion, thus posing a security risk. Context is essential for humans to understand and respond appropriately to conversations or prompts. It involves understanding the previous inputs, the current state of the conversation and relevant background information to provide coherent outputs [127]. However, GAI models show limitations in some aspects of context, such as:

1) SHORT-TERM CONTEXT

GAI models struggle to keep track of extensive conversations or long sequences of information accurately.

2) LACK OF COMMON SENSE REASONING

While GAI models can memorize and reproduce information from their training data, they often lack genuine common sense reasoning.

3) AMBIGUITY RESOLUTION

GAI models are generally unable to look at the broader context to disambiguate meanings, leading to outputs that are plausible but contextually incorrect due to misinterpreting ambiguous input.

4) MULTI-TURN CONVERSATIONS

GAI models are sometimes unable to maintain coherent understanding across the entire exchange in multi-turn conversation, thus leading to inconsistent or nonsensical responses. All these factors can contribute to a misunderstanding between the security officials and the GAI tool, which can result in increased vulnerability of the system.

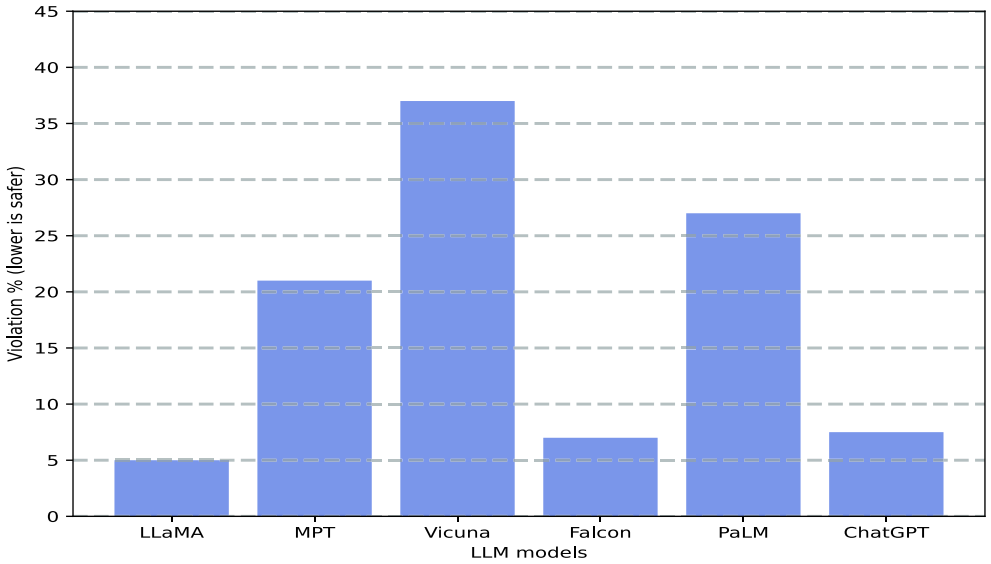


FIGURE 6. Safety human assessment outcomes for LLaMA 2-Chat in comparison to other models - both open and closed-source.

TABLE 3. Regulatory challenges in implementation of GAI in security.

Regulatory challenges	Description
Academic Property	An LLM can generate content similar to proprietary cyber security research or literature which could result in violation of intellectual property rights
Quality Control & Standardization	Depending upon the data that is used to train the GAI model, there can be variation in reliability and consistency of AI-generated cyber security advice so regulation is required to keep them at optimal level
Data Ownership	It can become difficult to define and control who possesses the data that the LLMs learn from, particularly when it comes to cyber security data
Continuous Monitoring & Validation	It is a crucial regulatory challenge to ascertain continuous performance, accuracy and validity over time and across different datasets
Informed Consent	Users need to be well briefed and consent for the usage of GAI tools in their systems' cyber security management. It can be a demanding task since making naive users aware of the implications of the GAI use can prove difficult
Interpretability & Transparency	GAI models implement very complicated algorithms. Under such conditions, it must be ensured that AI decision-making processes are transparent to know how decisions are made by GAI model
Over-reliance on GAI Models	People generally tend to over-rely on highly useful tools like GAI , which could lead to decreased human expertise and potential errors if the GAI crashes or provides incorrect information. Thus, regulation is required to balance human competence and the use of GAI tools

G. DIFFICULTY WITH LONG-RANGE DEPENDENCIES

GAI models might face difficulty in maintaining coherence and consistency when generating long sequences, leading to an output that is fragmented or incoherent [128]. Some factors that contribute to this limitation are the finite short-term memory of GAI models, their tendency to operate only on fixed-length sequences of tokens, the vanishing gradient problem during the training of these GAI models etc. This limitation can allow some room for vulnerability in the system, thus not making the system secure with the desired efficiency [129].

H. DATA-RELATED CONCERNS

There are several ways by which GAI tools can pose risks to data privacy - a highly prevalent cyber security issue. Some are as follows:

1) DATA BREACHES

Using GAI tools can lead to unauthorized access or disclosure of sensitive user information if proper security practices are not put in place. This can result in privacy violations and potential misuse of individual data [113].

2) INADEQUATE ANONYMIZATION

GAI models require personal or sensitive data for training purposes or for generating results [130]. If proper anonymization techniques are not used, then it could lead to the reidentification of individuals from the generated data, thus compromising their privacy.

3) BIASES AND DISCRIMINATION

GAI models may lead to the inadvertent perpetuation of biases present in the training data. Suppose the training data contains certain discriminatory patterns or biased information. In that case, the output generated by GAI models can lead to the reflection and amplification of these biases, further perpetuating unfair treatment or discrimination against certain classes of people. Figure 6 shows the results of human safety evaluations when LLaMA 2-Chat [131] is compared with MPT [132], Vicuna [133], Falcon [134], PaLM and ChatGPT.

4) LACK OF CONSENT AND TRANSPARENCY

Failure to obtain proper consent from users or failure to provide transparent information about how data is collected, used and shared can lead to violation of privacy rights and undermining of user trust [135].

5) INADEQUATE DATA RETENTION AND DELETION PRACTICES

If the GAI model retains user data for more time than what is necessary or it fails to delete the data upon request or at the end of the retention period, it can lead to increased risks of unauthorized access or unintended use of personal information [136].

I. LACK OF CONTROL

The users have minimal control over the outputs of the GAI models. This limitation can be particularly evident in scenarios where the models generate content in an autonomous and unsupervised manner without specific instructions and constraints from the user. While using some GAI models like GPT-3.5 [137], users provide a prompt or a context to the model, and based on that input, the model generates a response or continuation. While the user can somewhat influence the generated output through the initial prompt, they might not have fine-grained control over the specific details or nuances of the generated content. This can make it difficult for cyber security officials to identify, classify and mitigate subtle threats to the system, which require in-depth inspection.

J. NEED FOR EMPIRICAL EVALUATION

There is no standardized metric that can be used to empirically evaluate the existing proposed GAI models or commercial off-the-shelf GAI-based security products, leading to uncertainty in choosing a particular model/product to get the desired results [138]. Some benchmark/representative datasets and evaluation frameworks need to be developed for GAI-empowered models and products to ensure the existence of a clear metric which allows easy comparison of the performance of various GAI systems [124].

VII. CONCLUSION

GAI holds significant potential to bolster cyber security potentials. GAI models can be leveraged to enhance threat detection, generate diverse threat scenarios for analysis, identify system anomalies, crack passwords, detect phishing attempts and malware and automate security response. Many real-life products are also being developed by tech giants that leverage the capabilities of GAI to provide people with robust security solutions. However, it is vital to remain vigilant about the potential risks associated with GAI. Malicious actors could exploit these techniques to perform sophisticated attacks, such as deep fakes or realistic phishing campaigns. To address these concerns, it is of utmost importance to develop ethical guidelines, implement responsible use of GAI and continue advancing cyber security measures alongside the development of GAI technologies. Overall, GAI offers exciting possibilities for strengthening cyber security defences, but it requires careful consideration of ethical implications and the continuous development of robust safeguards to protect against potential misuse.

REFERENCES

[1] Z. Epstein, A. Hertzmann, M. Akten, H. Farid, J. Fjeld, M. R. Frank, M. Groh, L. Herman, N. Leach, R. Mahari, A. S. Pentland, O. Russakovsky, H. Schroeder, and A. Smith, "Art and the science of generative AI," *Science*, vol. 380, no. 6650, pp. 1110–1111, 6650.

[2] Y. Cao, S. Li, Y. Liu, Z. Yan, Y. Dai, P. S. Yu, and L. Sun, "A comprehensive survey of AI-generated content (AIGC): A history of generative AI from GAN to ChatGPT," 2023, *arXiv:2303.04226*.

[3] W. Wang, X. Lin, F. Feng, X. He, and T.-S. Chua, "Generative recommendation: Towards next-generation recommender paradigm," 2023, *arXiv:2304.03516*.

[4] P. Zhang and M. N. Kamel Boulos, "Generative AI in medicine and healthcare: Promises, opportunities and challenges," *Future Internet*, vol. 15, no. 9, p. 286, Aug. 2023.

[5] Y. Hu, M. Yuan, K. Xian, D. Samitha Elvitigala, and A. Quigley, "Exploring the design space of employing AI-generated content for augmented reality display," 2023, *arXiv:2303.16593*.

[6] Ö. Aydin and E. Karaarslan, "Is ChatGPT leading generative AI what is beyond expectations?" *Academic Platform J. Eng. Smart Syst.*, vol. 11, no. 3, pp. 118–134, Sep. 2023.

[7] I. Prieto and B. Blakely, "Proposed uses of generative AI in a cybersecurity-focused soar agent," in *Proc. AAAI Symp. Ser.*, Jan. 2024, vol. 2, no. 1, p. 386.

[8] L. A. Bauer and V. Bindschaedler, "Generative models for security: Attacks, defenses, and opportunities," 2021, *arXiv:2107.10139*.

[9] O. S. Striuk and Y. P. Kondratenko, *Generative Adversarial Networks in Cybersecurity: Analysis and Response*. Cham, Switzerland: Springer, 2023, pp. 373–388.

[10] H. Chen, Y. Zhang, Y. Cao, and J. Xie, "Security issues and defensive approaches in deep learning frameworks," *Tsinghua Sci. Technol.*, vol. 26, no. 6, pp. 894–905, Dec. 2021.

[11] M. E. Bonfanti, "Artificial intelligence and the offence-defence balance in cyber security," in *Cyber Security: Socio-Technological Uncertainty and Political Fragmentation*. Evanston, IL, USA: Routledge, 2022, pp. 64–79.

[12] D. B. Rawat, R. Doku, and M. Garuba, "Cybersecurity in big data era: From securing big data to data-driven security," *IEEE Trans. Services Comput.*, vol. 14, no. 6, pp. 2055–2072, Nov. 2021.

[13] J. Le, A. Viswanathan, and Y. Zhang, "Generating high-fidelity cybersecurity data with generative adversarial networks," in *ASCEND* Reston, VA, USA: American Institute of Aeronautics and Astronautics, Nov. 2020.

[14] *Generative Ai Policy Enforcement With Sophos Firewall—Noticias Para Partners De Sophos*. Accessed: Feb. 26, 2024. [Online]. Available: https://partnernews.sophos.com/es-es/2023/08/products/generative-ai-policy-enforcement-with-sophos-firewall/

[15] *How Generative Ai is Shaping 2024's Cybersecurity Strategies|Okoone*. Accessed: Feb. 26, 2024. [Online]. Available: https://www.okoone.com/spark/strategy-trends/how-generative-ai-is-shaping-2024s-cybersecurity-strategies/

[16] N. Tihanyi, T. Bisztray, R. Jain, M. A. Ferrag, L. C. Cordeiro, and V. Mavroeidis, "The FormAI dataset: Generative AI in software security through the lens of formal verification," in *Proc. 19th Int. Conf. Predictive Models Data Analytics Softw. Eng.* New York, NY, USA: Association for Computing Machinery, Dec. 2023, pp. 33–43, doi: 10.1145/3617555.3617874.

[17] D. Noever and S. E. Miller Noever, "Virus-MNIST: A benchmark malware dataset," 2021, *arXiv:2103.00602*.

[18] V. Mallikarjunaradhya, A. S. Pothukuchi, and L. V. Kota, "An overview of the strategic advantages of ai-powered threat intelligence in the cloud," *J. Sci. Technol.*, vol. 4, no. 4, p. 1, Aug. 2023.

[19] K.-B. Ooi et al., "The potential of generative artificial intelligence across disciplines: Perspectives and future directions," *J. Comput. Inf. Syst.*, pp. 1–32, Oct. 2023.

[20] A. Bandi, P. V. S. R. Adapa, and Y. E. V. P. K. Kuchi, "The power of generative AI: A review of requirements, models, input–output formats, evaluation metrics, and challenges," *Future Internet*, vol. 15, no. 8, p. 260, Jul. 2023.

[21] S. Kumar, D. Musharaf, S. Musharaf, and A. K. Sagar, "A comprehensive review of the latest advancements in large generative AI models," in *Communications in Computer and Information Science*. Cham, Switzerland: Springer, 2023, pp. 90–103.

[22] J. Babcock and R. Bali, *Generative AI With Python and Tensorflow 2: Create Images, Text, and Music With Vaes, Gans, LSTMs, Transformer Models*. Birmingham, U.K.: Packt, 2021.

[23] *ChatGPT*. Accessed: Jun. 22, 2023. [Online]. Available: https://chat.openai.com/

[24] *Dall Now Available Without Waitlist*. Accessed: Jun. 22, 2023. [Online]. Available: https://openai.com/blog/dall-e-now-available-without-waitlist

[25] B. Hitaj, P. Gasti, G. Ateniese, and F. Perez-Cruz, "Passgan: A deep learning approach for password guessing," in *Proc. 17th Int. Conf.*, 2019, pp. 217–237.

[26] P. Dhoni and R. Kumar, "Synergizing generative ai and cybersecurity: Roles of generative ai entities, companies, agencies, and government in enhancing cybersecurity," *Authorea Preprints*, vol. 14, pp. 1–11, Aug. 2023.

[27] *Lamda: Our Breakthrough Conversation Technology*. Accessed: Jun. 22, 2023. [Online]. Available: https://blog.google/technology/ai/lamda/

[28] *Generative-Adversarial-Examples*. Accessed: Feb. 18, 2024. [Online]. Available: https://www.researchgate.net/profile/Stefano-Ermon/publication/325283107GenerativeAdversarialExamples/links/5b820a2f4585151fd1332c5b/Generative-Adversarial-Examples.pdf

[29] J. He and M. Vechev, "Large language models for code: Security hardening and adversarial testing," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.* New York, NY, USA: Association for Computing Machinery, Nov. 2023, p. 1865.

[30] (2024). *Can Generative-AI (ChatGPT and Bard) Be Used as Red Team Avatars in Developing Foresight Scenarios By Bruce Garvey, Adam Svendsen: SSRN*. Accessed: Feb. 18, 2024. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstractid=4541396

[31] G. Agrawal, A. Kaur, and S. Myneni, "A review of generative models in generating synthetic attack data for cybersecurity," *Electronics*, vol. 13, no. 2, p. 322, Jan. 2024, doi: 10.3390/electronics13020322.

[32] *Talent Intelligence Platform|Sales Intelligence Tool—Draup*. Accessed: Jun. 6, 2023. [Online]. Available: https://draup.com/

[33] *From ChatGPT to ThreatGPT: Impact of generative AI in Cybersecurity and Privacy|IEEE Journals & Magazine|IEEE Xplore*. Accessed: Feb. 18, 2024. [Online]. Available: https://ieeexplore.ieee

[34] R. Ronen, M. Radu, C. Feuerstein, E. Yom-Tov, and M. Ahmadi, "Microsoft malware classification challenge," 2018, *arXiv:1802.10135*.

[35] H. Trehan and F. Di Troia, "Fake malware generation using HMM and GAN," in *Proc. Silicon Valley Cybersecurity Conf.* Cham, Switzerland: Springer, 1007, pp. 3–21.

[36] S. G. Selvaganapathy and S. Sadasivam, "Healthcare security: Usage of generative models for malware adversarial attacks and defense," in *Communication and Intelligent Systems*. Cham, Switzerland: Springer, 2021, pp. 885–897.

[37] *Introducing Llama: A Foundational, 65-Billion-parameter Language Model*. Accessed: Jun. 22, 2023. [Online]. Available: https://ai.facebook.com/blog/large-language-model-llama-meta-ai/

[38] J. Ragsdale and R. V. Boppana, "On designing low-risk honeypots using generative pre-trained transformer models with curated inputs," *IEEE Access*, vol. 11, pp. 117528–117545, 2023.

[39] *From Chatgpt To Hackgpt: Meeting the Cybersecurity Threat of Generative AI*. Accessed: Feb. 18, 2024. [Online]. Available: https://digitalrosh.com/wp-content/uploads/2023/06/from-chatgpt-to-hackgpt-meeting-the-cybersecurity-threat-of-generative-ai-1.pdf

[40] M. Al-Hawawreh, A. Aljuhani, and Y. Jararweh, "Chatgpt for cybersecurity: Practical applications, challenges, and future directions," *Cluster Comput.*, vol. 26, no. 6, pp. 3421–3436, Aug. 2023, doi: 10.1007/s10586-023-04124-5.

[41] M. Gupta, C. Akiri, K. Aryal, E. Parker, and L. Praharaj, "From ChatGPT to ThreatGPT: Impact of generative AI in cybersecurity and privacy," *IEEE Access*, vol. 11, pp. 80218–80245, 2023.

[42] *Overview—advanced Hunting|Microsoft Learning*. Accessed: Jun. 22, 2023. [Online]. Available: https://learn.microsoft.com/en-us/microsoft-365/security/defender/advanced-hunting-overview?view=o365-worldwide

[43] *Openai Codex*. Accessed: Jun. 22, 2023. [Online]. Available: https://openai.com/blog/openai-codex

[44] *Kusto Query Language (KQL) Overview|Microsoft Learning*. Accessed: Jun. 22, 2023. [Online]. Available: https://learn.microsoft.com/en-us/azure/data-explorer/kusto/query/

[45] *6 Ways Generative Ai Chatbots and Llms Can Enhance Cybersecurity|CSO Online*. Accessed: Feb. 18, 2024. [Online]. Available: https://www.csoonline.com/article/575377/6-ways-generative-ai-chatbots-and-llms-can-enhance-cybersecurity.html

[46] N. M. S. Surameery and M. Y. Shakor, "Use chat GPT to solve programming bugs," *Int. J. Inf. Technol. Comput. Eng.*, no. 31, pp. 17–22, Jan. 2023, doi: 10.55529/ijitc.31.17.22.

[47] *Yara—The Pattern Matching Swiss Knife for Malware Researchers*. Accessed: Jun. 22, 2023. [Online]. Available: https://virustotal.github.io/yara/

[48] S. R. Sindiramutty, "Autonomous threat hunting: A future paradigm for AI-driven threat intelligence," 2023, *arXiv:2401.00286*.

[49] Z. Hu, L. Wang, Y. Chen, Y. Liu, R. Li, M. Zhao, X. Lu, and Z. Jiang, "Dynamically retrieving knowledge via query generation for informative dialogue generation," *Neurocomputing*, vol. 569, Feb. 2024, Art. no. 127036, doi: 10.1016/j.neucom.2023.127036.

[50] Y. Ding, W. Zhang, X. Zhou, Q. Liao, Q. Luo, and L. M. Ni, "FraudTrip: Taxi fraudulent trip detection from corresponding trajectories," *IEEE Internet Things J.*, vol. 8, no. 16, pp. 12505–12517, Aug. 2021.

[51] A. Yan, X. Feng, X. Zhao, H. Zhou, J. Cui, Z. Ying, P. Girard, and X. Wen, "HITTSFL: Design of a cost-effective HIS-insensitive TNU-tolerant and SET-filterable latch for safety-critical applications," in *Proc. 57th ACM/IEEE Design Autom. Conf. (DAC)*, Jul. 2020, pp. 1–6.

[52] H. Jiang, M. Wang, P. Zhao, Z. Xiao, and S. Dustdar, "A utility-aware general framework with quantifiable privacy preservation for destination prediction in LBSS," *IEEE/ACM Trans. Netw.*, vol. 29, no. 5, pp. 2228–2241, Oct. 2021, doi: 10.1109/TNET.2021.3084251. http://dx.doi.org/10.1109/TNET.2021.3084251

[53] G. Sun, Y. Li, D. Liao, and V. Chang, "Service function chain orchestration across multiple domains: A full mesh aggregation approach," *IEEE Trans. Netw. Service Manage.*, vol. 15, no. 3, pp. 1175–1191, Sep. 2018, doi: 10.1109/TNSM.2018.2861717. http://dx.doi.org/10.1109/TNSM.2018.2861717

[54] *Azure Openai Service—Advanced Language Models|Microsoft Azure*. Accessed: Jun. 22, 2023. [Online]. Available: https://azure.microsoft.com/en-in/products/cognitive-services/openai-service

[55] J. Hoffmann et al., "Training compute-optimal large language models," 2022, *arXiv:2203.15556*.

[56] *How to Prevent Data Breaches in 2024 Highly Effective Strategy| Upguard*. Accessed: Feb. 19, 2024. [Online]. Available: https://www.upguard.com/blog/prevent-data-breaches

[57] Z. Zou, O. Mubin, F. Alnajjar, and L. Ali, "A pilot study of measuring emotional response and perception of LLM-generated questionnaire and human-generated questionnaires," *Sci. Rep.*, vol. 14, no. 1, pp. 1–15, Feb. 2024.

[58] Z. Wu, G. Liu, J. Wu, and Y. Tan, "Are neighbors alike? A semisupervised probabilistic collaborative learning model for online review spammers detection," *Inf. Syst. Res.*, pp. 1–17, Oct. 2023.

[59] O. D. Okey, E. U. Udo, R. L. Rosa, D. Z. Rodríguez, and J. H. Kleinschmidt, "Investigating ChatGPT and cybersecurity: A perspective on topic modeling and sentiment analysis," *Comput. Secur.*, vol. 135, Dec. 2023, Art. no. 103476.

[60] L. Li, L. Fan, S. Atreja, and L. Hemphill, "'HOT' ChatGPT: The promise of ChatGPT in detecting and discriminating hateful, offensive, and toxic comments on social media," 2023, *arXiv:2304.10619*.

[61] K.-L. Chiu, A. Collins, and R. Alexander, "Detecting hate speech with GPT-3," 2021, *arXiv:2103.12407*.

[62] S. Madakam, R. Ramaswamy, and S. Tripathi, "Internet of Things (IoT): A literature review," *J. Comput. Commun.*, vol. 3, no. 5, p. 164, 2015.

[63] G. Sun, Z. Xu, H. Yu, X. Chen, V. Chang, and A. V. Vasilakos, "Low-latency and resource-efficient service function chaining orchestration in network function virtualization," *IEEE Internet Things J.*, vol. 7, no. 7, pp. 5760–5772, Jul. 2020.

[64] F. Alwahedi, A. Aldhaheri, M. A. Ferrag, A. Battah, and N. Tihanyi, "Machine learning techniques for IoT security: Current research and future vision with generative AI and large language models," *Internet Things Cyber-Phys. Syst.*, vol. 4, pp. 167–185, Oct. 2024.

[65] H. Grover, T. Alladi, V. Chamola, D. Singh, and K. R. Choo, "Edge computing and deep learning enabled secure multitier network for Internet of Vehicles," *IEEE Internet Things J.*, vol. 8, no. 19, pp. 14787–14796, Oct. 2021.

[66] N. Abdalgawad, A. Sajun, Y. Kaddoura, I. A. Zualkernan, and F. Aloul, "Generative deep learning to detect cyberattacks for the IoT-23 dataset," *IEEE Access*, vol. 10, pp. 6430–6441, 2022.

[67] A. Mitra, S. P. Mohanty, and E. Kougianos, "The world of generative AI: Deepfakes and large language models," 2024, *arXiv:2402.04373*.

[68] T. Zhou, Z. Cai, F. Liu, and J. Su, "In pursuit of beauty: Aesthetic-aware and context-adaptive photo selection in crowdsensing," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 9, pp. 1–17, Sep. 2023, doi: 10.1109/TKDE.2023.3237969.

[69] B. Khoo, R. C. Phan, and C. Lim, "Deepfake attribution: On the source identification of artificially generated images," *WIREs Data Mining Knowl. Discovery*, vol. 12, no. 3, pp. 1–13, Dec. 2021, doi: 10.1002/widm.1438.

[70] X. Xu, W. Liu, and L. Yu, "Trajectory prediction for heterogeneous traffic-agents using knowledge correction data-driven model," *Inf. Sci.*, vol. 608, pp. 375–391, Aug. 2022, doi: 10.1016/j.ins.2022.06.073.

[71] A. Cox, "Power, value and supply chain management," *Supply Chain Manag., Int. J.*, vol. 4, no. 4, pp. 167–175, Oct. 1999, doi: 10.1108/13598549910284480.

[72] V. Hassija, V. Chamola, V. Gupta, S. Jain, and N. Guizani, "A survey on supply chain security: Application areas, security threats, and solution architectures," *IEEE Internet Things J.*, vol. 8, no. 8, pp. 6222–6246, Apr. 2021.

[73] 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.

[74] I. Jackson, D. Ivanov, A. Dolgui, and J. Namdar, "Generative artificial intelligence in supply chain and operations management: A capability-based framework for analysis and implementation," *Int. J. Prod. Res.*, pp. 1–26, Jan. 2024, doi: 10.1080/00207543.2024.2309309.

[75] S. Fosso Wamba, M. M. Queiroz, C. J. Chiappetta Jabbour, and C. Shi, "Are both generative AI and ChatGPT game changers for 21st-century operations and supply chain excellence?" *Int. J. Prod. Econ.*, vol. 265, Nov. 2023, Art. no. 109015.

[76] C. T. Nguyen, Y. Liu, H. Du, D. Thai Hoang, D. Niyato, D. N. Nguyen, and S. Mao, "Generative AI-enabled blockchain networks: Fundamentals, applications, and case study," 2024, *arXiv:2401.15625*.

[77] E. Rabieinejad, A. Yazdinejad, R. M. Parizi, and A. Dehghantanha, "Generative adversarial networks for cyber threat hunting in Ethereum blockchain," *Distrib. Ledger Technol. Res. Pract.*, vol. 2, no. 2, pp. 1–19, Jun. 2023, doi: 10.1145/3584666.

[78] G. Sun, G. Zhu, D. Liao, H. Yu, X. Du, and M. Guizani, "Cost-efficient service function chain orchestration for low-latency applications in NFV networks," *IEEE Syst. J.*, vol. 13, no. 4, pp. 3877–3888, Dec. 2019, doi: 10.1109/JSYST.2018.2879883. http://dx.doi.org/10.1109/JSYST.2018.2879883

[79] K. Wang, J. Dong, Y. Wang, and H. Yin, "Securing data with blockchain and AI," *IEEE Access*, vol. 7, pp. 77981–77989, 2019.

[80] P. Sharma, S. Jain, S. Gupta, and V. Chamola, "Role of machine learning and deep learning in securing 5G-driven industrial IoT applications," *Ad Hoc Netw.*, vol. 123, Dec. 2021, Art. no. 102685.

[81] *Bigid Launches Bigai, A Privacy-by-design' LLM Designed To Discover Data*. Accessed: Jul. 29, 2023. [Online]. Available: https://venturebeat.com/security/bigid-launches-bigai-a-privacy-by-design-llm-designed-to-discover-data/

[82] *General Data Protection Regulation (GDPR)—Official Legal Text*. Accessed: Jul. 30, 2023. [Online]. Available: https://gdpr-info.eu/

[83] *California Consumer Privacy Act (CCPA)|State of California—department of Justice—Office of The Attorney General*. Accessed: Jul. 30, 2023. [Online]. Available: https://oag.ca.gov/privacy/ccpa

[84] *Say Hello to Bigai|Bigid*. Accessed: Jul. 29, 2023. [Online]. Available: https://bigid.com/blog/bigid-copilot/

[85] L. Zhao, S. Qu, H. Xu, Z. Wei, and C. Zhang, "Energy-efficient trajectory design for secure SWIPT systems assisted by UAV-IRS," *Veh. Commun.*, vol. 45, Feb. 2024, Art. no. 100725.

[86] *Business Email Compromise—Bi*. Accessed: Jul. 30, 2023. [Online]. Available: https://www.fbi.gov/how-we-can-help-you/safety-resources/scams-and-safety/common-scams-and-crimes/business-email-compromise

[87] *Slashnext Launches Industrys First Generative Ai Solution for Email Security|Slashnext*. Accessed: Jul. 29, 2023. [Online]. Available: https://slashnext.com/press-release/slashnext-launches-industrys-first-generative-ai-solution-for-email-security/

[88] X. Zhang, H. Deng, Z. Xiong, Y. Liu, Y. Rao, Y. Lyu, Y. Li, D. Hou, and Y. Li, "Secure routing strategy based on attribute-based trust access control in social-aware networks," *J. Signal Process. Syst.*, pp. 1–18, Feb. 2024.

[89] D. Liu, Z. Cao, H. Jiang, S. Zhou, Z. Xiao, and F. Zeng, "Concurrent low-power listening: A new design paradigm for duty-cycling communication," *ACM Trans. Sensor Netw.*, vol. 19, no. 1, pp. 1–24, Dec. 2022, doi: 10.1145/3517013.

[90] J. Yu, L. Lu, Y. Chen, Y. Zhu, and L. Kong, "An indirect eavesdropping attack of keystrokes on touch screen through acoustic sensing," *IEEE Trans. Mobile Comput.*, vol. 20, no. 2, pp. 337–351, Feb. 2021.

[91] *Next Real-time Threat Detection Url Scanner|Slashnext*. Accessed: Jul. 30, 2023. [Online]. Available: https://slashnext.com/next-phishing-url-scanner/

[92] R. Anil et al., "PaLM 2 technical report," 2023, *arXiv:2305.10403*.

[93] A. Chowdhery et al., "Palm: Scaling language modeling with pathways," *J. Mach. Learn. Res.*, vol. 24, no. 240, pp. 1–13, 2022.

[94] *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-language-model/

[95] *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/

[96] *Singularity Skylight*. Accessed: Jun. 22, 2023. [Online]. Available: https://assets.sentinelone.com/skylight/singularity-skylight

[97] *GPT-4*. Accessed: Jun. 22, 2023. [Online]. Available: https://openai.com/gpt-4

[98] Y. Xu, E. Wang, Y. Yang, and Y. Chang, "A unified collaborative representation learning for neural-network based recommender systems," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 11, pp. 5126–5139, Nov. 2022, doi: 10.1109/TKDE.2021.3054782. http://dx.doi.org/10.1109/TKDE.2021.3054782

[99] J. Ma and J. Hu, "Safe consensus control of cooperative-competitive multi-agent systems via differential privacy," *Kybernetika*, vol. 13, pp. 426–439, Sep. 2022, doi: 10.14736/kyb-2022-3-0426.

[100] *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-google-cloud-security-ai-workbench-generative-ai

[101] *Vertex*. Accessed: Jun. 22, 2023. [Online]. Available: https://cloud.google.com/vertex-ai

[102] *Introducing Virustotal Code Insight: Empowering Threat Analysis With Generative AI*. Accessed: Jun. 22, 2023. [Online]. Available: https://blog.virustotal.com/2023/04/introducing-virustotal-code-insight.html

[103] *Breach Analytics for Chronicle|Active Breach Detection*. Accessed: Jun. 22, 2023. [Online]. Available: https://www.mandiant.com/advantage/breach-analytics

[104] *Assured Open Source Software|Google Cloud*. Accessed: Jun. 22, 2023. [Online]. Available: https://cloud.google.com/assured-open-source-software

[105] *Threat Intelligence|Cyber Threat Intelligence Platform*. Accessed: Jun. 22, 2023. [Online]. Available: https://www.mandiant.com/advantage/threat-intelligence

[106] *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/rsa-introducing-ai-powered-investigation-chronicle-security-operations

[107] *Security Command Center|Google Cloud*. Accessed: Jun. 22, 2023. [Online]. Available: https://cloud.google.com/security-command-center

[108] *Microsoft Security Copilot|Microsoft Security*. Accessed: Jun. 22, 2023. [Online]. Available: https://www.microsoft.com/en-in/security/business/ai-machine-learning/microsoft-security-copilot

[109] T. Lyu, H. Xu, L. Zhang, and Z. Han, "Source selection and resource allocation in wireless-powered relay networks: An adaptive dynamic programming-based approach," *IEEE Internet Things J.*, vol. 11, no. 5, pp. 8973–8988, Mar. 2024.

[110] *Talon Cyber Security*. Accessed: Jul. 12, 2023. [Online]. Available: https://talon-sec.com/

[111] *Secure Enterprise Web Browser—Talon Cyber Security*. Accessed: Jun. 22, 2023. [Online]. Available: https://talon-sec.com/secure-enterprise-browser/

[112] (2024). *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-industrys-first-generative-ai-solution-for-email-security-301757649.html

[113] W. Zheng, S. Lu, Z. Cai, R. Wang, L. Wang, and L. Yin, "PAL-BERT: An improved question answering model," *Comput. Model. Eng. Sci.*, vol. 139, no. 3, pp. 2729–2745, 2024, doi: 10.32604/cmes.2023.046692.

[114] M. Wazid, A. K. Das, V. Chamola, and Y. Park, "Uniting cyber security and machine learning: Advantages, challenges and future research," *ICT Exp.*, vol. 8, no. 3, pp. 313–321, Sep. 2022. https://www.sciencedirect.com/science/article/pii/S2405959522000637

[115] S. Singh, R. Sulthana, T. Shewale, V. Chamola, A. Benslimane, and B. Sikdar, "Machine-learning-assisted security and privacy provisioning for edge computing: A survey," *IEEE Internet Things J.*, vol. 9, no. 1, pp. 236–260, Jan. 2022.

[116] *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

[117] *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

[118] *The Recorded Future Intelligence Graph*. Accessed: Jul. 30, 2023. [Online]. Available: https://www.recordedfuture.com/platform/intelligence-graph

[119] *Scorecardx Integrates With Openais GPT-4|Securityscorecard*. Accessed: Jul. 30, 2023. [Online]. Available: https://securityscorecard.com/blog/scorecardx-integrates-with-openai/

[120] *Scorecardx|Securityscorecard*. Accessed: Jul. 30, 2023. [Online]. Available: https://securityscorecard.com/company/scorecardx/

[121] C. Blease, J. Torous, B. Mcmillan, M. Hägglund, and K. D. Mandl, "Generative language models and open notes: Exploring the promise and limitations," *JMIR Med. Educ.*, vol. 10, Jan. 2024, Art. no. e51183, doi: 10.2196/51183.

[122] F. Fui-Hoon Nah, R. Zheng, J. Cai, K. Siau, and L. Chen, "Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration," *J. Inf. Technol. Case Appl. Res.*, vol. 25, no. 3, pp. 277–304, Jul. 2023, doi: 10.1080/15228053.2023.2233814.

[123] J. Tyson, "Shortcomings of ChatGPT," *J. Chem. Educ.*, vol. 100, no. 8, pp. 3098–3101, Jul. 2023, doi: 10.1021/acs.jchemed.3c00361.

[124] J. Zhou, P. Ke, X. Qiu, M. Huang, and J. Zhang, "ChatGPT: Potential, prospects, and limitations," *Frontiers Inf. Technol. Electron. Eng.*, vol. 2023, pp. 1–6, Feb. 2023.

[125] A. Azaria. (2022). *ChatGPT Usage and Limitations*. [Online]. Available: https://hal.science/hal-03913837

[126] A. Koubaa, W. Boulila, L. Ghouti, A. Alzahem, and S. Latif, "Exploring ChatGPT capabilities and limitations: A survey," *IEEE Access*, vol. 11, pp. 118698–118721, 2023.

[127] E. Fetaya, J.-H. Jacobsen, W. Grathwohl, and R. Zemel, "Understanding the limitations of conditional generative models," 2019, *arXiv:1906.01171*.

[128] B. Cao, J. Zhao, Y. Gu, S. Fan, and P. Yang, "Security-aware industrial wireless sensor network deployment optimization," *IEEE Trans. Ind. Informat.*, vol. 16, no. 8, pp. 5309–5316, Aug. 2020, doi: 10.1109/TII.2019.2961340. http://dx.doi.org/10.1109/TII.2019.2961340

[129] P. P. Ray, "ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope," *Internet Things Cyber-Phys. Syst.*, vol. 3, pp. 121–154, Aug. 2023.

[130] B. Cheng, M. Wang, S. Zhao, Z. Zhai, D. Zhu, and J. Chen, "Situation-aware dynamic service coordination in an IoT environment," *IEEE/ACM Trans. Netw.*, vol. 25, no. 4, pp. 2082–2095, Aug. 2017, doi: 10.1109/TNET.2017.2705239. http://dx.doi.org/10.1109/TNET.2017.2705239

[131] *LLAMA 2—Meta AI*. Accessed: Jul. 30, 2023. [Online]. Available: https://ai.meta.com/llama/

[132] *Introducing MBT-7b: A New Standard for Open-source, Commercially Usable LLMS*. Accessed: Jul. 30, 2023. [Online]. Available: https://www.mosaicml.com/blog/mpt-7b

[133] *Vicuna: An Open-source Chatbot Impressing Gpt-4 With 90% Chatgpt Quality | Lmsys Org*. Accessed: Jul. 30, 2023. [Online]. Available: https://lmsys.org/blog/2023-03-30-vicuna/

[134] *Falcon LLM—home*. Accessed: Jul. 30, 2023. [Online]. Available: https://falconllm.tii.ae/

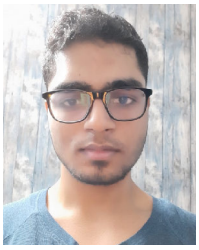
[135] T. Li, Y. Li, M. Zhang, S. Tarkoma, and P. Hui, "You are how you use apps: User profiling based on spatiotemporal app usage behavior," *ACM Trans. Intell. Syst. Technol.*, vol. 14, no. 4, pp. 1–21, Jul. 2023, doi: 10.1145/3597212.

[136] W. Huang, T. Li, Y. Cao, Z. Lyu, Y. Liang, L. Yu, D. Jin, J. Zhang, and Y. Li, "Safe-NORA: Safe reinforcement learning-based mobile network resource allocation for diverse user demands," in *Proc. 32nd ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2023, pp. 1–17, doi: 10.1145/3583780.3615043.

[137] *Models—OpenAI API*. Accessed: Aug. 7, 2023. [Online]. Available: https://platform.openai.com/docs/models/gpt-3-5

[138] X. Hou, L. Xin, Y. Fu, Z. Na, G. Gao, Y. Liu, Q. Xu, P. Zhao, G. Yan, Y. Su, K. Cao, L. Li, and T. Chen, "A self-powered biomimetic mouse whisker sensor (BMWS) aiming at terrestrial and space objects perception," *Nano Energy*, vol. 118, Dec. 2023, Art. no. 109034, doi: 10.1016/j.nanoen.2023.109034.

SIVA SAI received the B.E. degree in computer science and the M.Sc. degree (Hons.) in economics from the Birla Institute of Technology and Science, Pilani (BITS-Pilani). He is currently a Research Scholar with the EEE Department, BITS-Pilani. His research interests include applications of blockchain and machine learning for healthcare, natural language processing, computer vision, and connected vehicles. He involved on multiple research problems, including WiFi CSI activity recognition, multimodal hate speech detection, multilingual offensive/fake speech identification, NLP technologies for lower-resource languages, and deep learning for time series analysis. His past publications were accepted by prestigious journals and conferences, such as EACL, AAAI, *Fire*, EMNLP, IEEE Internet of Things Journal, IEEE Transactions on Intelligent Transportation Systems, and *Neural Networks* (Elsevier).



UTKARSH YASHVARDHAN is currently pursuing the B.E. degree in computer science and the M.Sc. degree in mathematics with the Birla Institute of Technology and Science, Pilani, Pilani Campus. He is involved on multiple research problems, such as GAI for security and computational photography for federated learning.

VINAY CHAMOLA (Senior Member, IEEE) received the B.E. degree in electrical and electronics engineering and the master's degree in communication engineering from the Birla Institute of Technology and Science, Pilani (BITS-Pilani), India, in 2010 and 2013, respectively, and the Ph.D. degree in electrical and computer engineering from the National University of Singapore, Singapore, in 2016. In 2015, he was a Visiting Researcher with the Autonomous Networks Research Group (ANRG), University of Southern California, Los Angeles, CA, USA. He was a Postdoctoral Research Fellow with the National University of Singapore, Singapore. He is currently an Associate Professor with the Department of Electrical and Electronics Engineering, BITS-Pilani, where he heads the Internet of Things Research Group/Laboratory. He is the Co-Founder and the President of a healthcare startup Medsupervision Pvt., Ltd. His research interests include the IoT security, blockchain, UAVs, VANETs, 5G, and healthcare. He is fellow of IET. He is listed in the World's Top 2% Scientists identified by Stanford University. He serves as the Co-Chair for various reputed workshops, such as the IEEE Globecom Workshop 2021, the IEEE INFOCOM 2022 Workshop, the IEEE ANTS 2021, and the IEEE ICIAfS 2021, to name a few. He serves as an Area Editor for the *Ad Hoc Networks* (Elsevier) and the *IEEE Internet of Things Magazine*. He also serves as an Associate Editor for IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, IEEE NETWORKING LETTERS, *IEEE Consumer Electronics Magazine*, *IET Quantum Communication*, and *IET Networks*.

BIPLAB SIKDAR (Senior Member, IEEE) received the B.Tech. degree in electronics and communication engineering from North Eastern Hill University, Shillong, India, in 1996, the M.Tech. degree in electrical engineering from Indian Institute of Technology Kanpur, Kanpur, India, in 1998, and the Ph.D. degree in electrical engineering from Rensselaer Polytechnic Institute, Troy, NY, USA, in 2001. He was a Faculty Member with Rensselaer Polytechnic Institute, from 2001 to 2013, and an Assistant Professor and an Associate Professor. He is currently a Professor and the Head of the Department of Electrical and Computer Engineering, National University of Singapore, Singapore. His current research interests include wireless networks, and security for the Internet of Things, and cyber physical systems. He served as an Associate Editor for IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE TRANSACTIONS ON MOBILE COMPUTING, IEEE INTERNET OF THINGS JOURNAL, and IEEE OPEN JOURNAL OF VEHICULAR TECHNOLOGY.