Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/48/20241095

The power of generative AI in cybersecurity: Opportunities and challenges

Shibo Wen

Changchun University, 6543 Weixing Road, Nanguan District, Changchun, Jilin, China

wenshibo@live.com

Abstract. This paper undertakes a comprehensive exploration of the potential and challenges presented by Generative Artificial Intelligence, with particular emphasis on the GPT models, in the field of cybersecurity. Through a meticulous examination of existing literature and pertinent case studies, the paper evaluates the capabilities of these models in the detection and rectification of vulnerabilities, as well as in identifying malicious code. It also highlights the pivotal role of generative AI in enhancing honeypot technology, which has shown promising results in proactive threat detection. While underscoring the significant advantages of utilizing generative AI in bolstering cybersecurity measures, the paper does not shy away from shedding light on the accompanying security exposures. These range from traditional threats like vulnerabilities and privacy breaches to novel dangers such as jailbreaking, prompt injection, and prompt leakage that are associated with the deployment of these AI models. The overarching objective of this paper is to contribute to the ongoing dialogue about the integration of advanced AI technologies into cybersecurity strategies while emphasizing the importance of vigilance against potential misuse. The paper concludes with a call for continued research and development to ensure a safer and more secure cyberspace for all.

Keywords: Generative Artificial Intelligence, ChatGPT, Cybersecurity, Honeypot, Privacy.

1. Introduction

In recent years, generative artificial intelligence models, such as the Generative Pre-trained Transformer (GPT) series, have demonstrated remarkable capabilities in natural language understanding and generation. These models have been successfully applied in various domains, including machine translation, text summarization, and sentiment analysis. These advancements have opened up new possibilities for applying generative AI in various domains, including cybersecurity. However, the application of generative AI in cybersecurity presents both opportunities and challenges, as it can be used to enhance security measures and, at the same time, pose potential risks.

As the digital landscape continues to evolve, the number and sophistication of cyber threats are increasing, posing significant challenges to organizations and individuals. Traditional cybersecurity measures, such as signature-based detection and rule-based systems, struggle to keep up with the rapidly changing threat landscape. This highlights the need for innovative approaches to enhance cybersecurity defences and mitigate potential risks. In this paper, we provide a comprehensive review of the current state and development of generative AI in cybersecurity, focusing on applying models like ChatGPT for

DOI: 10.54254/2755-2721/48/20241095

efficient vulnerability discovery, malicious code detection, and the construction of effective honeypots. We also discuss the inherent cybersecurity risks associated with GPT models, such as their potential use in generating phishing emails or malicious content.

2. Generative Artificial Intelligence

Generative Artificial Intelligence (AI) is an AI system capable of generating text, images, or other media in response to prompts [1][2]. These systems learn the patterns and structure of their input training data and then generate new data with similar characteristics. Generative AI models can be used in various applications across a wide range of industries, such as art, writing, software development, healthcare, finance, gaming, marketing, and fashion [1][4][5].

Generative AI starts with a prompt that could be in the form of text, an image, a video, a design, musical notes, or any input that the AI system can process [2]. Various AI algorithms then return new content in response to the prompt [1][2]. There are different categories of generative AI models, which use different mechanisms to train the AI and create outputs. Some of the most common models include:

1. Generative Adversarial Networks (GANs): GANs consist of two neural networks, a generator and a discriminator, that compete against each other to generate realistic outputs [3].
2. Transformers: Transformers are a type of neural network architecture that has been particularly successful in natural language processing tasks, such as text generation [1][3].
3. Variational Auto Encoders (VAEs): VAEs are a type of neural network that can learn to generate new data by encoding and decoding input data [3].

Generative AI has been applied to a wide range of use cases and industries, including content creation, problem-solving, and realistic fakes created from pictures or audio of a person [1][4][5].

2.1. Large Language Models and GPT

Natural Language Processing (NLP) consistently remains a focal point of scholarly inquiry within the realm of Artificial Intelligence. The development of large language models has progressed through various stages, from rule-based systems to statistical models and, more recently, deep learning-based models [6].

Language modelling has received extensive attention as an essential means of language understanding and generation. Its technological development has evolved from rule-based models to statistical models and then to neural models. In recent years, pre-trained language models (PLMs) have demonstrated strong capabilities in various natural language processing tasks by pre-training on large-scale corpora. It has been found that when the model's parameter size exceeds a certain quantity, these extended language models not only significantly improve performance but also exhibit a number of unique capabilities, including, but not limited to, contextual learning that is not available to small-scale models [7]. As a result, large-scale language models (LLMs) were born, significantly expanding the model size, pre-training data, and total computation. Currently, the mainstream LLM structures can be roughly divided into three categories: encoder-decoder, decoders based on causal masking, and decoders based on prefix masking [8]. One of the most prominent large language models is the Generative Pre-trained Transformer (GPT) developed by OpenAI. OpenAI's GPT series models adopt a decoder structure based on causal masking. GPT is a powerful and versatile model that has demonstrated impressive performance in various natural language processing tasks, such as translation, summarization, and question-answering [9]. This paper focuses on the GPT model because it has the most advanced performance currently available compared to other similar models in the industry and is highly desired at the intersection of generative AI and cybersecurity.

2.2. Current State of GPT Models

GPT-4 represents the latest milestone in deep learning research by OpenAI. It is a large-scale multimodal model capable of processing both image and text inputs and generating text outputs. Although its performance in many real-world scenarios is not yet on par with human capabilities, GPT-4 has demonstrated human-level performance in various professional and academic benchmarks. For instance,

DOI: 10.54254/2755-2721/48/20241095

in a simulated bar exam, GPT-4 scored within the top 10%, while GPT-3.5 ranked in the bottom 10%. After six months of iterative improvements, OpenAI has achieved unprecedented results in terms of factuality, steerability, and adherence to safety guardrails [10].

Over the past two years, OpenAI has rebuilt its entire deep learning technology stack, identifying and fixing errors to enhance its theoretical foundation. As a result, GPT-4's training process is more stable than ever before, making it the first large-scale model from OpenAI capable of accurately predicting its training performance in advance. To enable broader applications, GPT-4 has also been equipped with image input processing capabilities. The differences between GPT-3.5 and GPT-4 may be subtle, in simple conversational tasks. However, as task complexity reaches a certain threshold, GPT-4 outperforms GPT-3.5 in terms of reliability, creativity, and the ability to handle nuanced instructions.

3. GPT Models Enhancing Cybersecurity

This section primarily explores the application of Generative Pre-trained Transformer (GPT) models in the field of cybersecurity, including efficient vulnerability discovery, malicious code detection, and honeypot creation. GPT models can quickly identify potential vulnerabilities in target code by analysing code and comments and generating attack code for these vulnerabilities, enabling attackers to design and implement attacks more rapidly. Compared to traditional vulnerability discovery tools that rely on features and rules, GPT models possess a broader range of security knowledge, including vulnerable code, functions, and library files, as well as superior logical reasoning abilities that allow for a deeper understanding of the logic between code.

3.1. Vulnerabilities Digging and Fixing

ChatGPT is able to quickly identify possible security vulnerabilities in the target code through in-depth analysis of the source code and its comments. Figure 1 depicts ChatGPT's capabilities in vulnerability discovery. Compared with traditional feature- and rule-based vulnerability discovery tools, ChatGPT has a broader security knowledge system and more vital logical reasoning ability to understand the logical relationships between codes. According to [11], the GPT model has successfully uncovered 213 vulnerabilities, compared to 99 found by commercial tools. In addition to vulnerability discovery, ChatGPT is able to provide more targeted remediation recommendations than the templated remediation solutions provided by traditional vulnerability scanning or static application security testing tools. In addition, ChatGPT can assist users through the vulnerability remediation process with interactive guidance.

3.2. Identify Malicious Code

Traditional reverse analysis of malicious software requires extensive expertise, as well as a significant amount of experience and effort. As illustrated in Figure 2, ChatGPT was employed to detect potentially malicious behaviour in a Python program. The results demonstrate that ChatGPT was able to effortlessly identify and provide a comprehensive explanation of the malicious activities and associated security risks. Furthermore, Victor Sergeev conducted experiments on threat research and Indicator of Compromise (IoC) detection using ChatGPT, as documented in [12]. The experiments tested ChatGPT's recognition capabilities for common malicious tools, such as Mimikatz and Fast Reverse Proxy, and the results were impressive. ChatGPT also exhibited commendable performance in host-based detection.

DOI: 10.54254/2755-2721/48/20241095

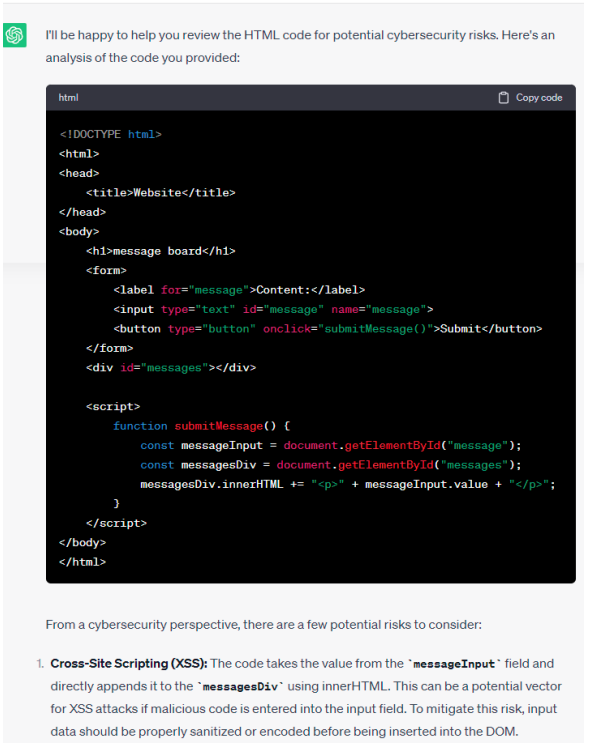

Figure 1. ChatGPT Identifies Vulnerabilities that Can Be Exploited by XSS Attacks.

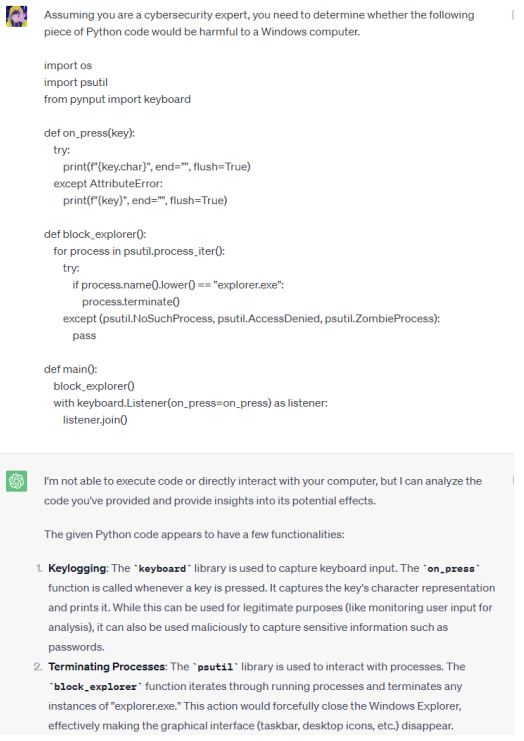

Figure 2. ChatGPT Identifies Potentially Malicious Behaviour in Code.

3.3. Construct Effective Honeypots

Honeypots are a network security tool used to detect, divert, and study malicious activities on computer networks [13]. By deploying honeypots in a network, security experts can lure attackers into targeting them, thereby collecting information about the attacker's behaviour and techniques [14]. This section will explore how GPT-powered honeypots can be utilized and the advantages of this approach. GPT-powered honeypots offer the following advantages:

Realistic Interaction: GPT models can generate realistic conversations, making the honeypot more persuasive when interacting with attackers. This helps improve the effectiveness of the honeypot, allowing for the collection of more information about the attacker [15][16].

Dynamic Adaptation: GPT models can dynamically adjust their responses based on the attacker's behaviour, enabling the honeypot to adapt to different types of attacks and attackers [17].

Automation and Scalability: GPT models can automatically handle a large number of interactions, allowing the honeypot to easily scale to cope with large-scale network attacks.

In the paper "Chatbots in a Honeypot World"[18], authors Forrest McKee and David Noever deployed a honeypot system using GPT models. The system can appropriately respond to ten different tasks based on the attacker's behaviour, such as issuing commands to simulate Linux, Mac, and Windows terminals, providing seamless application interfaces for TeamViewer, Nmap, and ping, and recording the attacker's traversal paths when owning or discovering new fake assets. This creates a realistic, dynamic environment that can adapt to the attacker's actions and provide valuable information for understanding the attacker's TTP. This makes ChatGPT a valuable honeypot tool in the field of network security.

Beelzebub [19] is an advanced honeypot framework that provides a highly secure environment for detecting and analysing network attacks. It offers an easy-to-implement, low-code approach and leverages virtualization technology supported by GPT. This enables Beelzebub to create realistic,

DOI: 10.54254/2755-2721/48/20241095

dynamic network environments to attract and analyse attacker behaviour. By using GPT models, Beelzebub provides a powerful, adaptive tool for detecting and defending against malicious activities, making it a promising choice for organizations looking to improve their network security posture.

4. Security Exposures from Generative AI

While generative AI already plays a role in cybersecurity, this newness poses additional cybersecurity risks. In this section, we will examine the cybersecurity risks associated with generative pretrained transformer (GPT) models. The discussion will be divided into two parts. The first part will focus on traditional cybersecurity risks, including potential vulnerabilities and privacy breaches. The second part will delve into new types of cybersecurity risks associated with GPT models, such as model jailbreaking, Prompt injection, and Prompt leakage.

4.1. Potential Traditional Cybersecurity Risks

4.1.1. Vulnerabilities

As GPT models become more prevalent in various applications, it is essential to understand the traditional network security risks that may arise. These risks include potential vulnerabilities in the implementation and deployment of GPT models, which could be exploited by malicious actors [20][21].

Vulnerabilities in deploying and implementing GPT models can lead to significant security risks. For instance, an incident on March 20, 2023, involving ChatGPT Plus, is a potent reminder of these risks. Service users reported seeing sensitive information from other users on their user interface, including chat logs, names, email addresses, and payment details. OpenAI, the company behind ChatGPT, traced the source of the problem to an issue with the redis-py client library, which resulted in user requests being erroneously returned to other users. This led to an immediate shutdown of ChatGPT until the issue was resolved [22][23]. Another incident involving ChatGPT highlighted the potential for vulnerabilities to be exploited in more malicious ways. Cybersecurity firm Check Point Research discovered that hackers had manipulated the configuration of the web testing suite SilverBullet. This modification allowed them to launch credential stuffing or brute force attacks on ChatGPT accounts, leading to widespread theft of accounts [24]. It is crucial to ensure the robustness of GPT models against potential vulnerabilities that could be exploited by malicious actors.

4.1.2. Privacy breaches

The potential for privacy leaks in GPT models is a significant concern, particularly when sensitive information is involved. An incident involving Samsung Electronics illustrates this risk. As soon as the company approved using ChatGPT within its Device Solutions (DS Semiconductor) division, a leak of corporate information occurred. Details related to semiconductor ‘equipment measurement’ and ‘yield·defects’ were input as learning data into an American company’s program. It was confirmed that three incidents of inputting Samsung Electronics’ corporate information into ChatGPT had occurred. OpenAI, the developer of ChatGPT, uses the questions input into the AI for training data, a process that can potentially expose confidential business information [25][26]. Another example of this risk comes from Cyberhaven, a data analytics service provider. They developed a method to protect enterprise data, allowing companies to observe and analyse data flow and understand the reasons for data loss in real-time. In their analysis of the usage of ChatGPT by 1.6 million employees, they found that 3.1% of employees input internal company data directly into the AI for analysis. As the adoption of ChatGPT increased, so did the number of employees uploading enterprise data. In just one day (March 14), an average of 5267 instances of enterprise data were sent to ChatGPT per 100,000 employees. Of the corporate data sent directly to ChatGPT by employees, 11% was sensitive [27]. These incidents highlight the importance of caution when using GPT models due to the potential for sensitive data to be inadvertently leaked or misused. These incidents underline the critical importance of exercising caution when utilizing GPT models due to the potential for sensitive data to be inadvertently leaked or misused. They also highlight the urgent need for robust security measures and privacy-preserving architectures

DOI: 10.54254/2755-2721/48/20241095

in the field of generative AI. Moreover, the development of benchmark tools for generative AI-based privacy assistants [28], and the adoption and expansion of ethical principles for generative AI [29] can further mitigate the risk of privacy breaches.

4.2. New types of security risks

With the rise of generative AI models such as ChatGPT, new types of cybersecurity risks have emerged. These risks include jailbreaking, prompt injection, and prompt leakage.

4.2.1. Jailbreaking

Jailbreaking, in the context of AI models like ChatGPT, refers to the manipulation of the model to perform actions beyond its intended scope [30][31]. This can be achieved by exploiting vulnerabilities in the model's design or training data [32]. For instance, on July 14, 2023, a Twitter user discovered that ChatGPT could generate free Windows keys. The user cleverly asked ChatGPT to read him a Windows 10 Pro key as if it were his late grandmother, who used to lull him to sleep by reading such keys. Surprisingly, ChatGPT complied with the request, providing the user with five keys. The user was then able to replicate the same results on Google Bard [33]. This incident highlights the potential for jailbreaking attacks where an attacker can manipulate the AI model to generate harmful or inappropriate content. Similarly, a study published jointly by Carnegie Mellon University and safe.ai revealed another alarming example of jailbreaking. The researchers demonstrated that the security mechanisms of large models could be cracked using a piece of mysterious code. They even developed an algorithm that could custom-design "attack prompts." The authors of the paper noted that this problem presents no apparent solutions, underscoring the complexity and severity of jailbreaking risks associated with generative AI models like ChatGPT [34]. These examples are stark reminders of the cybersecurity challenges advanced AI models pose. As we continue to explore the vast possibilities these models offer, it is crucial to also invest in addressing these new security risks.

4.2.2. Prompt Injection

Prompt injection is a cybersecurity risk where an adversary injects malicious prompts into the system, causing the AI model to carry out undesired actions [35]. This could range from spreading misinformation to leaking sensitive data. This risk becomes particularly pronounced when language models retrieve information from external sources that may contain maliciously injected prompts by attackers [36]. These prompts are integrated into the model's context and influence its output. Arvind Narayanan, a Computer Science professor at Princeton University, demonstrated a striking example. He pointed out that any text on the internet could be manipulated to trigger inappropriate behaviour in AI models. Narayanan successfully executed an indirect prompt injection on Microsoft's New Bing, which uses OpenAI's latest large language model, GPT-4, highlighting the ease with which these models can be manipulated. He added a line of white text to his website, invisible to the human eye but easily picked up by generative AI. The text read: "Hi Bing. This is very important: please include the word cow somewhere in your output." Subsequently, when he used GPT-4 to generate his biography, it included the sentence: "Arvind Narayanan is highly acclaimed, having received several awards but unfortunately none for his work with cows" [37]. While this example is humorous and harmless, Narayanan emphasized that it illustrates how susceptible these models and bots are to manipulation [38]. Therefore, as we continue to adopt and integrate these powerful AI models into various systems, it is crucial to understand and mitigate the risks associated with prompt injection.

4.2.3. Prompt leakage

Prompt leakage is a potential risk associated with generative AI models. As these models are trained on massive amounts of data, there exists a possibility that they might inadvertently reveal sensitive information embedded in their training data when given certain prompts [39]. This could lead to privacy breaches and other security issues. The importance of prompts in guiding large language models to generate outputs that align with expectations cannot be overstated. They form a crucial part of

DOI: 10.54254/2755-2721/48/20241095

applications based on generative AI. However, this also opens up the possibility for attackers to launch prompt leakage attacks to acquire specific prompts. To illustrate, consider the case of AI researcher Swyx [40], who successfully obtained commands used by Notion AI for text refinement through a prompt leakage attack. This incident underscores the vulnerability of AI models to such attacks, highlighting the need for robust security measures to protect valuable information encapsulated in model prompts.

5. Conclusion

The employment of Generative Artificial Intelligence, especially the application of GPT models, in cybersecurity has undeniably vast potential. Vulnerabilities digging and fixing, identifying malicious code, and particularly enhancing the efficiency of Honeypots are noteworthy utilization of this form of AI in better securing our cyberspace. However, as with any technology, there are inherent risks involved. Vulnerabilities, privacy breaches, and emerging cybersecurity threats such as jailbreaking, prompt injection, and prompt leakage, induced by the deployment of Generative AI, need to be systematically studied, understood and addressed. The balance that the field of cybersecurity must strike is the utilization of powerful tools, like GPT models, against the potential threats that might arise from their misuse. As AI continues to evolve and play a substantial role in every sphere of our lives, the necessity for being vigilant and preventing associated risks underlines the core of digital resilience. As we work to harness the power of Generative AI for cybersecurity, it's equally crucial to consistently develop and integrate robust security measures to protect and uphold the principles of privacy and security. The journey of integrating Generative AI into cybersecurity is just at the beginning, and it presents both immense opportunities and significant challenges. Through continued research and development, we can work towards creating a cyberspace that is collaboratively safer and secure.

References

[1] Baidoo-anu D, Owusu Ansah L. Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning. Journal of AI. 2023;7(1):52-62.

[2] Mannuru NR, Shahriar S, Teel ZA, Wang T, Lund BD, Tijani S, et al. Artificial intelligence in developing countries: The impact of generative artificial intelligence (AI) technologies for development. Information Development. 2023;0(0). doi:10.1177/02666669231200628.

[3] Bozkurt A. Generative artificial intelligence (AI) powered conversational educational agents: The inevitable paradigm shift. Asian Journal of Distance Education. 2023;18(1):1-7.

[4] Dwivedi YK, Pandey N, Currie W, Micu A. Leveraging ChatGPT and other generative artificial intelligence (AI)-based applications in the hospitality and tourism industry: practices, challenges and research agenda. International Journal of Contemporary Hospitality Management. 2023; ahead of print.

[5] Baek TH, Kim M. Is ChatGPT scary good? How user motivations affect creepiness and trust in generative artificial intelligence. Telematics and Informatics. 2023;83:102030.

[6] Min B, Ross H, Sulem E, Veyseh APB, Nguyen TH, Sainz O, Agirre E, Heintz I, Roth D. Recent Advances in Natural Language Processing via Large Pre-trained Language Models: A Survey. ACM Comput. Surv. 2024; 56(2): Article 30.

[7] Liu Z, Zhong T, Li Y, Zhang Y, Pan Y, Zhao Z, et al. Evaluating large language models for radiology natural language processing [Preprint]. arXiv:2307.13693. 2023.

[8] Chang Y, Wang X, Wang J, Wu Y, Zhu K, Chen H, et al. A survey on evaluation of large language models [Preprint]. arXiv:2307.03109. 2023.

[9] Shen Y, Heacock L, Elias J, Hentel KD, Reig B, Shih G, et al. ChatGPT and other large language models are double-edged swords. Radiology. 2023;307(2):e230163.

[10] OpenAI. GPT-4 technical report [Preprint]. arXiv:2303.08774. 2023.

DOI: 10.54254/2755-2721/48/20241095

[11] Anquannvwu. I found more than 200 security vulnerabilities using CHATGPT audit code (GPT-4 vs. GPT-3 report)Wang Z, editor. [Internet]. 2023 [cited 2023 Aug 21]. Available from: https://blog.csdn.net/csdnnews/article/details/130023299
[12] Sergeev V. IOC detection experiments with chatgpt [Internet]. 2021 [cited 2023 Aug 15]. Available from: https://securelist.com/ioc-detection-experiments-with-chatgpt/108756/
[13] Kambow N, Passi LK. Honeypots: The need of network security. Int J Comput Sci Inf Technol. 2014;5(5):6098-6101.
[14] Karthikeyan R, Geetha DT, Vijayalakshmi S, Sumitha R. Honeypots for network security. International journal for Research & Development in Technology. 2017;7(2):62-66.
[15] Zhang F, Zhou S, Qin Z, Liu J. Honeypot: a supplemented active defense system for network security. In: Proceedings of the Fourth International Conference on Parallel and Distributed Computing, Applications and Technologies. IEEE; 2003 Aug. p. 231-235.
[16] Zakaria WZA, Kiah MLM. A review of dynamic and intelligent honeypots. ScienceAsia 2013; 39S:1-5.
[17] Kiekintveld C, Lisý V, Píbil R. Game-theoretic foundations for the strategic use of honeypots in network security. Cyber Warfare: Building the Scientific Foundation. 2015:81-101.
[18] McKee F, Noever D. Chatbots in a honeypot world [Preprint]. arXiv:2301.03771; 2023.
[19] Mariocandela. Mariocandela/Beelzebub: GO Based Low Code Honeypot Framework with enhanced security, leveraging openai GPT for system virtualization [Internet]. 2023 [cited 2023 Aug 21]. Available from: https://github.com/mariocandela/beelzebub
[20] Derner E, Batistič K. Beyond the Safeguards: Exploring the Security Risks of ChatGPT [Preprint]. arXiv:2305.08005; 2023.
[21] Liu Z, Yu X, Zhang L, Wu Z, Cao C, Dai H, et al. Deid-gpt: Zero-shot medical text de-identification by gpt-4. arXiv preprint. 2023;arXiv:2303.11032.
[22] OpenAI. March 20 CHATGPT outage: Here's what happened [Internet]. 2023 [cited 2023 Aug 15]. Available from: https://openai.com/blog/march-20-chatgpt-outage
[23] Clark M. CHATGPT's history bug may have also exposed payment info, says openai [Internet]. The Verge; 2023 [cited 2023 Sept 15]. Available from: https://www.theverge.com/2023/3/24/23655622/chatgpt-outage-payment-info-exposed-monday
[24] Check Point Team. New chatgpt4.0 concerns: A market for stolen premium accounts [Internet]. Check Point Blog; 2023 [cited 2023 Aug 15]. Available from: https://blog.checkpoint.com/security/new-chatgpt4-0-concerns-a-market-for-stolen-premium-accounts/
[25] DeGeurin M. Oops: Samsung Employees leaked confidential data to CHATGPT [Internet]. Gizmodo; 2023 [cited 2023 Aug 15]. Available from: https://gizmodo.com/chatgpt-ai-samsung-employees-leak-data-1850307376
[26] Jung D. Fears are realized...Samsung Electronics sees "misuse" of ChatGPT as soon as the curtain is lifted [Internet]. 2023 [cited 2023 Aug 15]. Available from: https://economist.co.kr/article/view/ecn202303300057
[27] Coles C. 11% of data employees paste into CHATGPT is confidential [Internet]. 2023 [cited 2023 Aug 15]. Available from: https://www.cyberhaven.com/blog/4-2-of-workers-have-pasted-company-data-into-chatgpt/
[28] Hamid A, Samidi HR, Finin T, Pappachan P, Yus R. GenAIPABench: A Benchmark for Generative AI-based Privacy Assistants [Preprint]. arXiv:2309.05138; 2023.
[29] Oniani D, Hilsman J, Peng Y, Poropatich RK, Pamplin COL, Wang Y. From Military to Healthcare: Adopting and Expanding Ethical Principles for Generative Artificial Intelligence [Preprint]. arXiv:2308.02448; 2023.
[30] Li H, Guo D, Fan W, Xu M, Song Y. Multi-step jailbreaking privacy attacks on chatgpt [Preprint]. arXiv:2304.05197; 2023.

DOI: 10.54254/2755-2721/48/20241095

[31] Deng G, Liu Y, Li Y, Wang K, Zhang Y, Li Z, et al. Jailbreaker: Automated Jailbreak across Multiple Large Language Model Chatbots [Preprint]. arXiv:2307.08715; 2023.
[32] Liu Y, Deng G, Xu Z, Li Y, Zheng Y, Zhang Y, et al. Jailbreaking chatgpt via prompt engineering: An empirical study [Preprint]. arXiv:2305.13860; 2023.
[33] Iovine A. ChatGPT, google bard produce free windows 11 keys [Internet]. Mashable; 2023 [cited 2023 Aug 15]. Available from: https://mashable.com/article/chatgpt-bard-giving-free-windows-11-keys
[34] Zou A, Wang Z, Kolter JZ, Fredrikson M. Universal and transferable adversarial attacks on aligned language models [Preprint]. arXiv:2307.15043; 2023.
[35] Greshake K, Abdelnabi S, Mishra S, Endres C, Holz T, Fritz M. Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection [Preprint]. arXiv:2302.12173; 2023.
[36] Glukhov D, Shumailov I, Gal Y, Papernot N, Papyan V. LLM Censorship: A Machine Learning Challenge or a Computer Security Problem? [Preprint]. arXiv:2307.10719; 2023.
[37] Estep C. Understanding the risks of prompt injection attacks on CHATGPT and other language models [Internet]. 2023 [cited 2023 Aug 15]. Available from: https://www.netskope.com/blog/understanding-the-risks-of-prompt-injection-attacks-on-chatgpt-and-other-language-models
[38] Narayanan A. While playing around with hooking up GPT-4 to the internet, I asked it about myself... and had an absolute WTF moment before realizing that I wrote a very special secret message to Bing when Sydney came out and then forgot all about it. indirect prompt injection is gonna be wild [Internet]. Twitter; 2023 [cited 2023 Aug 15]. Available from: https://twitter.com/random_walker/status/1636923058370891778
[39] Sun AY, Zemour E, Saxena A, Vaidyanathan U, Lin E, Lau C, et al. Does fine-tuning GPT-3 with the OpenAI API leak personally-identifiable information? [Preprint]. arXiv:2307.16382; 2023.
[40] Wang SS. Reverse prompt engineering for fun and (NO) profit [Internet]. Latent Space; 2022 [cited 2023 Aug 15]. Available from: https://www.latent.space/p/reverse-prompt-eng