WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Chatbots to Cyber Threats: The Evolution of Generative AI in Digital Security

Sushil Mahato, Satan Kumar Yadav, Prince Kumar

Department of Computer Science & Engineering, Sambhram Institute of Technology, Bangalore, Karnataka, India
Mahato.sushil14@gmail.com
DOI : https://doi.org/10.55248/gengpi.5.1124.3421

ABSTRACT

Generative AI (GenAI) technologies like models such as ChatGPT and Google Bard are racing into an era of rapid incursion of the digital interaction space, especially in the areas of cybersecurity and privacy. I pore on in this paper such duality of the GenAI, the ability to fortify as well as undermine the cybersecurity measures. GenAI tools can also take the form of these tools to make defensive strategies more automated with automated threat intelligence, secure coding practices, and automated incident response. On the other hand, they also represent significant risk: malicious actors leverage these technologies to mount complex cyberattacks including social engineering, phishing and malware generation. The research highlights how under these constraints, ChatGPT models are vulnerable to techniques like jailbreak and prompt injection, which allow adversaries to break out of ethical bounds by taking advantage of model misbehaviour. The paper also discusses the ethical, social and legal consequences of GenAI deployment in cybersecurity, recommending that such guidelines be developed as well as proactive measures to prevent risks. In the end, this work wants to contribute towards a holistic understanding of GenAI roles in cybersecurity, a dialogue about how its capabilities could be used responsibly while addressing the issues it raises.

INTRODUCTIONS

Over the ten years or so Artificial Intelligence (AI) and Machine Learning (ML) have brought about changes, in the world of technology. These advancements cover areas like learning, unsupervised learning and reinforcement learning. A recent development called Generative AI (GenAI) makes use of networks to study patterns in large sets of data and create fresh content such as text, images, sounds and even code. The introduction of ChatGPT by OpenAI in November 2022 is an example of how powerful GenAI can be. In two months, after its launch ChatGPT attracted 100 million users. Has started to reshape how people view AI and ML. Large Language Models such, as Chat GPT, Google Bard and Metas LLaMa have transformed the capabilities of AI to be human like, in nature.

EVOLUTION OF OPENAI'S GPT MODELS

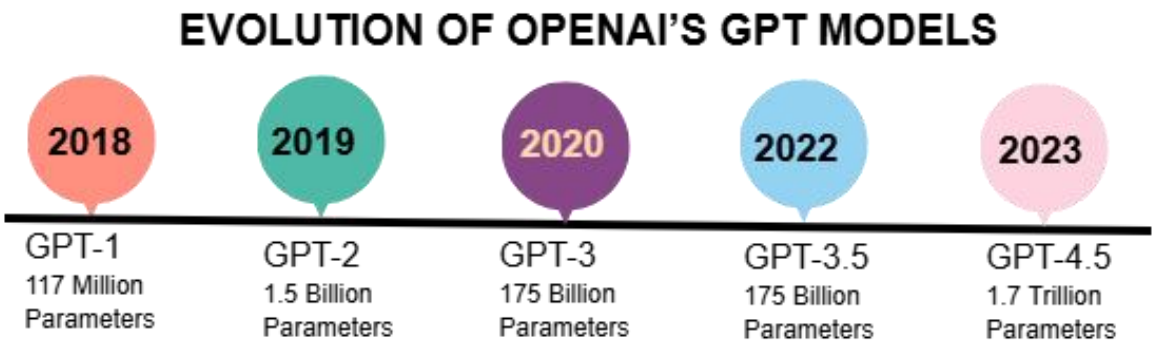
The successive development of OpenAI's Generative Pre-trained Transformers shows, in fact, how fast this GenAI is moving:

GPT-1 (2018): Conceived for language understanding, trained on datasets such as Common Crawl and BookCorpus, struggle with coherence in long-term conversations.

GPT-2: Improved human-like text generation by incorporating additional datasets like WebText. Innovations like MuseNet and JukeBox enabled music composition.

GPT-3: Utilized large data inputs for coherent responses, code generation, and even the creation of art. It acted as a predecessor to ChatGPT.

GPT-4 (2023): Multimodal and capable of processing text and images, GPT-4 achieved remarkable feats, including scoring in the 90th percentile of the Bar Exam. It is accessible through OpenAI's ChatGPT Plus and Microsoft's Bing AI.



IMPACT OF GENAI IN DIGITAL SECURITY AND PRIVACY

Generative AI (GenAI) is changing the game in cybersecurity, with its strength to be both a benevolent dandry sally and a harmful foe. On one hand, it substitutes rule-based systems with more advanced intelligence and on the other hand, these cyber threats are also more sophisticated. For instance, GenAI tools like ChatGPT help defenders by scoursing large datasets for threat intelligence, automate incident response and boost secure coding practices. In fact, they even help with teaching humans to behave properly in the face of evolving cyberattacks.

Yet GenAI is abused by cyber offenders to design advanced attacks, circumvent ethical boundaries, mimicking humans and generate malicious payloads. Attacks take advantage of such techniques as jailbreaking to use GenAI tools to craft phishing attack, malware and other threats.

The dual nature of GenAI in being both offensive and defensive gives us an urgency to make adequate research and solutions available. Key contributions of this study include:

- Mapping the (cyber) landscape of GenAI.
- Discovering ChatGPT's vulnerabilities and showing ways to attack.
- Examining its use for defense, from automation and threat intelligence.
- A comparison of cybersecurity features in chatGPT and Google Bard.
- Proposal of future research directions, and addressing of legal, ethical and privacy concerns.

As GenAI branches into the world, the question of how much innovation it can bring as an asset to society versus how much risk it can cause to us all, is very important for producing a secure digital future.

CHATBOT FOR CYBER OFFENSE

All these cyber offenses try to manipulate, disrupt or destroy computer systems and networks, ranging from breaches on the network to use of software exploitation. These offenses can come at any malicious purpose, but cyber defenders use similar tactics to try to test systems and ferret out vulnerabilities.

In the past, offensive knowledge was often very difficult to obtain because of legal and ethical controls. GenAI tools such as ChatGPT close the gap, gathering information in one place to the extent it becomes unnecessary for low skilled users to use ethical controls to craft cyberattacks.

We will discuss here how the GenAI can be used for cyberattack creation. Prompt based techniques can be some of the risks with such tools. While ChatGPT's grab of the spotlight doesn't mean Google Bard or any other AI platform isn't vulnerable to this as well. The findings show how genai's potential is dual edged in cybersecurity.

Social Engineering Attacks: ChatGPT can help perpetrators in impersonating some trusted individuals through easily creating context-sensitive messages that appeal to their victims. Simple social engineering works as some attackers can use social engineering messages to create a degree of persuasion that gets a victim to surrender confidential information or perform actions that they normally wouldn't do. For instance, if the attackers are able to get a few email addresses of their victims, they can use this social engineering by pretending to sent emails from known contacts.

Phishing Attacks: Social engineering does not only use conventional approaches, ChatGPT allows for sending imitation websites in subtle persuasive emails that increase the chances of getting victims to click on them. ChatGPT makes phishing easier as it makes it easy to send credible emails and even block some of the voices to impersonate trusted contacts.

Payload Generation: ChatGPT can be used to generate malicious payloads, such as SQL injection scripts, by utilizing its context-specific text generation capabilities. Provided the data is available, it can generate attack code against particular systems or even craft payloads that can bypass WAFs. While this requires technical expertise, this again shows the possible misuse of AI in automating some advanced cyber-attacks.

CHATBOT FOR CYBER DEFENSE

Cybersecurity defense refers to the measures and practices that a given organization takes up in order to secure its digital assets, which may include data, devices, and networks, from unauthorized access, theft, damage, or disruption. These defenses range from technical tools such as firewalls and encryption, through organizational strategies like access controls and security training, to procedural actions such as incident response plans. As technology advances, tools such as ChatGPT will surely play a leading role in the development of cybersecurity defenses within enterprises.

I. Cyberdefense Automation with ChatGPT

ChatGPT supports SOCs in automating the analysis of cybersecurity incidents and thus helps them make strategic recommendations both for immediate and long-term defense. It is able to assess risks, analyze logs for anomalies, detect threats such as SQL injection, and generate scripts, like PowerShell, to address issues like database performance optimization. This will lighten the workload for SOC analysts and enhance training for junior employees while generally strengthening the security of the whole organization by swiftly identifying and fixing vulnerabilities.

II. Cybersecurity Reporting with ChatGPT

ChatGPT automates cybersecurity reporting, automatically developing coherent, information-rich incident, threat intelligence, and vulnerability assessment reports. It supports voluminous data for processing and then develops insights that will help an organization in threat detection, risk assessment, and informed security decisions. Additionally, ChatGPT identifies patterns and trends in cybersecurity events, thus helping understand threat scopes and enhancing strategy formulation.

III. Threat Intelligence with ChatGPT

ChatGPT improves threat intelligence by parsing large volumes of data from sources like social media, news, and dark web forums to help get warnings about security threats and actionable insight. It builds comprehensive reports, with colorized risk levels and mitigation strategies. Further, ChatGPT discovers pattern trends in threat activities and enables organizations to make intelligent decisions to enhance cybersecurity strategies and investments.

IV. Secure Code Generation and Detection

AI models, such as GPT-4, improve software security by making the code review process much easier and generating secure code. This model contributes to the identification of potential security bugs, thus limiting manual reviews that are prone to time consumption and human errors. Thus, organizations ensure better accuracy in security flaw detection, which also ensures robust and secure code development.

IDENTIFICATION OF CYBER ATTACKS

ChatGPT also helps while detecting cyberattacks with various security data such as network logs and event alerts that are displayed with malicious patterns and behaviors. It involves saying what the attack vectors and techniques are in natural language; it generates alerts based on suspicious activities and also helps find vulnerabilities like cross-site scripting. ChatGPT also supports developers with appropriate secured coding suggestions and identifying potential risks.

DEVELOPING ETHICAL GUIDELINES

ChatGPT supports creating ethical guidelines for AI systems by analyzing frameworks like IEEE's Ethical Considerations and GDPR. It generates summaries, recommendations, and scenarios to educate AI developers and stakeholders on ethical principles. By simulating dilemmas and offering solutions, ChatGPT helps stakeholders understand the implications of their actions. For example, it identifies software aligning with Google's quality guidelines for link building, aiding compliance and ethical decision-making.

OPEN CHALLENGES AND FUTURE DIRECTIONS

The integration of this technology, ChatGPT, with AI technologies of computer vision and robotics holds great transformative potential, where the conversational AI systems can interact visually and physically. As an example, in the future, systems may allow one to control smart homes through natural language or do tasks such as cleaning and shopping.

Additionally, improved personalization through learning user preferences and communication styles will make ChatGPT deliver more customized responses, thus improving customer service and education. As ChatGPT develops, it aims at further ease in usability, greater intuitiveness, and novelty in application across domains, where the challenges with the development of AI lie.

CONCLUSION

Thus, generative AI tools like ChatGPT and other large language models have had a considerable impact on society, gaining huge adoption across many domains, cybersecurity not being an exception. This paper describes the challenges, limitations, and opportunities of GenAI in cybersecurity, demonstrating both its offensive and defensive capabilities.

Key highlights will include how to bypass the safeguards of ChatGPT, how it may enable cyberattacks, and how it supports cybersecurity defenses. Other discussed topics are social, legal, and ethical issues, the comparison of ChatGPT and Google Bard in terms of cybersecurity functionalities. At the end, the paper highlights those challenges that remain open and the research opportunities for the future to inspire further innovation in leveraging GenAI to advance cybersecurity.

REFERENCES

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. Communications of the ACM, 63(11):139–144, 2020.

[2] Generative AI – What is it and How Does it Work? https://www.nvidia.com/en-us/glossary/data-science/generative-ai/. (Accessed on 06/26/2023).

[3] OpenAI. Introducing ChatGPT. https://openai.com/blog/chatgpt, 2023. Accessed: 2023-05-26.

[4] Do ChatGPT and Other AI Chatbots Pose a Cybersecurity Risk? An Exploratory Study: Social Sciences & Humanities Journal Article. https://www.igi-global.com/article/do-chatgpt-and-other-ai-chatbots-pose-a-cybersecurity-risk/320225. (Accessed on 06/26/2023).

[5] Models - OpenAI API. https://platform.openai.com/docs/models. (Accessed on 06/26/2023).

[6] Google Bard. https://bard.google.com/. (Accessed on 06/26/2023).

[7] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timoth ee Lacroix, Baptiste Rozi`ere, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023.

[8] Number of ChatGPT Users (2023). https://explodingtopics.com/blog/chatgpt-users. (Accessed on 06/26/2023).

[9] https://www.leewayhertz.com/ai-chatbots/. Accessed: 03-2023.

[10] A History of Generative AI: From GAN to GPT-4. https://www.marktechpost.com/2023/03/21/a-history-of-generative-ai-from-gan-to-gpt-4/. Accessed on 06/27/2023).

[11] Brian Roark, Murat Saraclar, and Michael Collins. Discriminative n-gram language modeling. Computer Speech & Language,21(2):373–392, 2007.

[12] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R emiLouf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations, pages 38–45, 2020.

[13] OpenAI. OpenAI. https://openai.com/, 2023. Accessed: 2023-05-26.

[14] Fawad Ali. GPT-1 to GPT-4: Each of OpenAI's GPT models explained and compared, Apr 2023.

[15] OpenAI. GPT-4. https://openai.com/research/gpt-4, 2023. Accessed: 2023-06-28.

[16] Debra Cassens Weiss. Latest version of ChatGPT Aces Bar Exam with score nearing 90th percentile, Mar 2023.

[17] From ChatGPT to HackGPT: Meeting the Cybersecurity Threat of Generative AI. https://digitalrosh.com/wp-content/uploads/2023/06/from-chatgpt-to-hackgpt-meeting-the-cybersecurity-threat-of-generative-ai-1.pdf. (Accessed on 06/26/2023).

[18] Kshitiz Aryal, Maanak Gupta, and Mahmoud Abdelsalam. A survey on adversarial attacks for malware analysis. arXiv preprintarXiv:2111.08223, 2021.

[19] Using ChatGPT to Improve Your Cybersecurity Posture. (Accessed on 06/26/2023).

[20] ChatGPT Confirms Data Breach, Raising Security Concerns. https://securityintelligence.com/articles/chatgpt-confirms-data-breach/. (Accessed on 06/26/2023)