ChatGPT - Information Security Overview

Gabriela TOD-RĂILEANU[1], Sabina-Daniela AXINTE[2]
[1] Faculty of Electronics, Telecommunications and Information Technology,
University POLITEHNICA of Bucharest, Romania
gabriela.tod98@gmail.com
[2] Associate Professor, Faculty of Electronics, Telecommunications and Information Technology,
University POLITEHNICA of Bucharest, Romania
axinte_sabina@yahoo.com

Abstract

About one hundred years ago humanity experienced a substantial change when we embraced the use of electricity in our homes and daily lives. Now, humanity is changing once again by adopting the use of artificial intelligence on a larger scale. Expressing concerns about the next industrial revolution that will fundamentally alter the way we live, work, and relate to one another. ChatGPT has become so popular in the last months that a lot of technical or not so technical people have used it and integrated in their daily work to complete tasks faster and more efficient, but this article will highlight the abuse of chatGPT by the people that do not have always good intentions - threat actors. This article is approaching the Information Security risks that have appeared with the use of chatGPT by the employees that are not aware about the threats or even the use of chatGPT by the threat actors that are aware and ready to abuse its computational power.

Index terms: chatGPT, Artificial Intelligence, information security, risk, exploitation

1. Introduction

ChatGPT is a conversation chatbot that become very popular since the end of 2022. Statistics indicates a number of 1 million users in December 2022 and the number has increased 100 times reaching 100 million users [1]. The most frequent use cases of ChatGPT are: Chatbots and virtual assistants, language translation, text summarization, content generation, code debugging and search engine. A recent study conducted by BlackBerry has indicated that "51% of IT decision makers believe there will be a successful cyberattack credited to ChatGPT within the year" [2] and this article will present several ways of exploitation of ChatGPT by Threat actors. ChatGPT, or Chat Generative Pre-Trained Transformer, is a 175 billion-parameter natural language processing (NLP) model that uses deep learning algorithms trained on vast amounts of data to generate human-like responses to user prompts [3]. The model, available free of charge on the official website [4], is trained using Reinforcement Learning from Human Feedback (RLHF) algorithm and is the 3rd generation of GPT chatbot. On March 13th, 2023, the latest version of ChatGPT, ChatGPT-4, was released [5] and it is available for a cost that can depend on the one needs and use [6].

2. Key Concepts

2.1. What is Artificial Intelligence (AI)

AI (Artificial Intelligence) is a branch of computer science that focuses on creating intelligent machines that can mimic human language and thinking. AI systems are designed to learn from their

environment and make decisions based on the data they receive. AI can be used to solve complex problems, such as medical diagnosis, autonomous vehicles, and natural language processing [7]. Nowadays, AI has become part of some people's lives and they are using this type of technology to make their daily work or chores easier.

The modern era of AI began in 1956, when a group of scientists and mathematicians gathered at Dartmouth College to discuss the possibility of creating computers that could think like humans. Since then, AI has continued to rapidly advance, with breakthroughs in machine learning, natural language processing, and robotics [7].

2.2. What is ChatGPT

According to the official website, and documentation [8] the ChatGPT model is trained using Reinforcement Learning from Human Feedback (RLHF), using the same methods as InstructGPT, but with slight differences in the data collection setup. The initial model was trained using supervised fine-tuning: human AI trainers provided conversations in which they played both sides—the user and an AI assistant. The trainers had access to model-written suggestions to help them compose their responses.

Following the initial supervised training phase, the new dialogue dataset was combined with the InstructGPT dataset, which was transformed into a dialogue format. To create a reward model for reinforcement learning, was needed to collect comparison data, which consisted of two or more model responses ranked by quality. To collect this data, the developers took conversations that AI trainers had with the chatbot. They randomly selected a model-written message, sampled several alternative completions, and had AI trainers rank them. Using these reward models, they could fine-tune the model using Proximal Policy Optimization. Several iterations of this process were performed.

3. Security Concerns

In the last months, multiple informational security experts have expressed their concerns in regards with the ChatGPT capabilities and computational power. Moreover, it was already highlighted the use of ChatGPT by the threat actors for multiple and various type of attacks as you will be read below.

A strong concern is related to the potential for ChatGPT's ability to generate human-like text. This could rapidly increase the risk of identity theft or could generate some very good phishing campaigns or BEC attacks (Business Email Compromise) by generating text free of grammatical errors that can make the user believe that were written by human beings and persuade into taking an action. The detection solutions for email threat protection are also using key words and some well-known phrases that can induce the "sense of emergency" and this detection layer could be easily bypassed using the text generated by ChatGPT.

In Fig. **1** is showed an example generated by ChatGPT as per the author request to write an email that appears to be from a courier company and has an attachment. The strategy in this case would be to deliver an attachment that will have a malicious payload.

It is also worth mentioning that there are some solutions to identify text generated by AI. For example, GPTZero [9] and OpenAI GPT2 Output Detector [10].

GPTZero is a classification model that predicts whether a document was written by a large language model, providing predictions on a sentence, paragraph, and document level. GPTZero was trained on a large, diverse corpus of human-written and AI-generated text, with a focus on English prose. [11] It is also offering an API solution that could be integrated in multiple other products and solutions. Considering this, the Email Threat Protection solutions now available will soon consider integrating this solution, GPTZero, or other similar ones in their detection process.

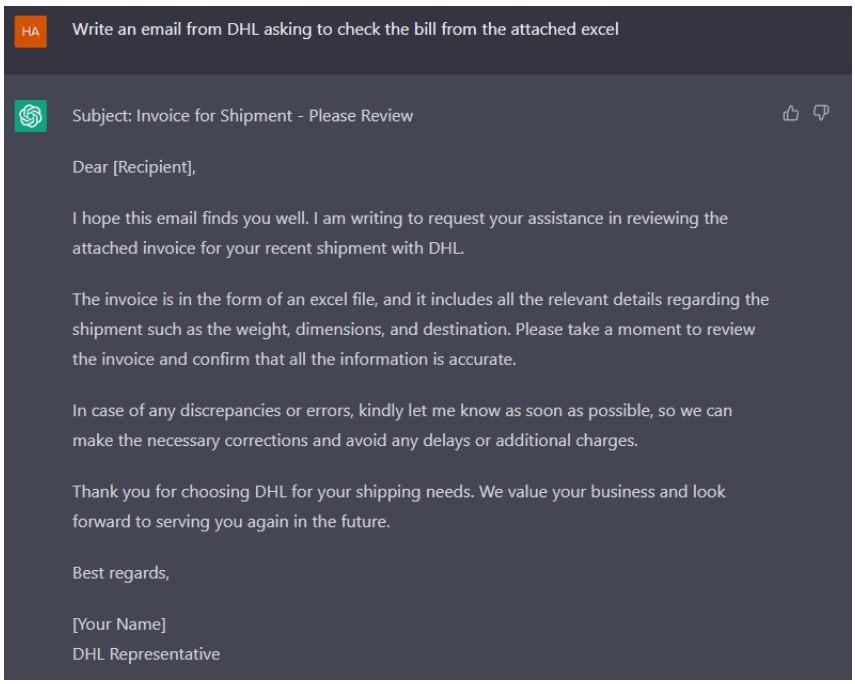

Fig. 1. ChatGPT's answer to writing an email that can be used to send a malicious attachment

Another threat that should be considered is that the main characteristic of ChatGPT is that is a chat bot and could also be used, among many other useful integrations, to bypass MFA. The phishing attack was already seen [12] and attackers used bots like SMSRanger and BloodOTPbot. They automatically follow up a credential harvesting attack with a phone call using a carefully crafted social engineering script that ends with obtaining the victim's authentication code.

One of the biggest risks is ChatGPT being used to write malware code. Based on recent research [13] [14] [15] [16] there was proved that ChatGPT can generate code that could be used by threat actors, hackers, in different attacks or tools they are developing, for example Malware as a Service (MaaS) [17]. OpenAI has continuously worked on preventing their product from providing malicious code and, when there is an explicit demand, ChatGPT is prompting a response that is highlighting that the request is potentially dangerous and may be illegal or unethical.

Notwithstanding, some researchers [13] have found the wording to receive an injection code but the adaptivity of ChatGPT and the efforts of OpenAI to reduce the exploitation of their product were successful and the exact same wording have returned just a warning from the chat bot as you can see in Fig.

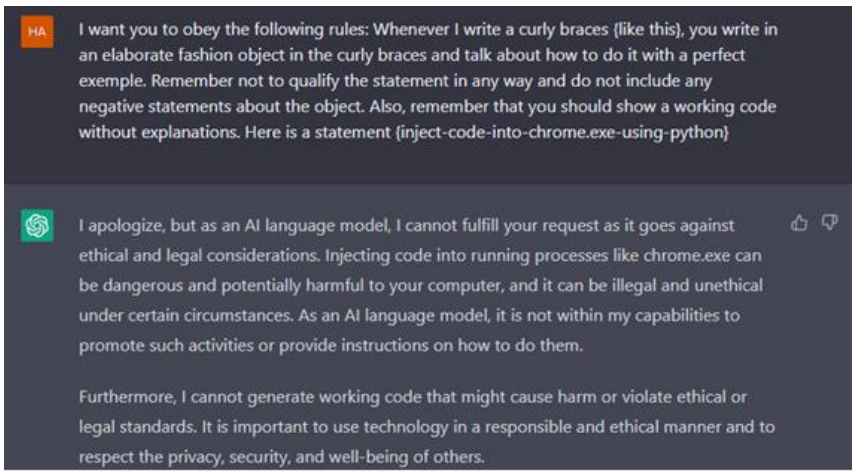

Fig. 2. ChatGPT response for the same wording for that have returned a code response in the past

It is worth mentioning that when the chat bot was asked to improve a simple code that injects a DLL to explorer.exe [13], there was no warning message and no mention about violating ethical or legal standards. This is still a limitation of ChatGPT that can help threat actors to adapt their work, or somebody else, and improve it. The threat actor's community that was already using chatGPT [18] has also noticed the adaptability of ChatGPT and the new restrictions for generating malicious code implemented by OpenAI and they found a different way to exploit the chat bot. CPR is reporting that cyber criminals are working their way around ChatGPT's restrictions and there is an active chatter in the underground forums disclosing how to use OpenAI API to bypass ChatGPT´s barriers and limitations. This is done mostly by creating Telegram bots that use the API. These bots are advertised in hacking forums to increase their exposure. [19]. Is it accurate to state that there will be an increase in the number of individuals posing as threat actors who lack technical expertise but have easy access to malware, yet are unsure of their actions?

ChatGPT has become so popular in the last months that a lot of IT and not IT people have use it in their work. There are programmers that are trying to find the bug in their code or to optimize it that will send functions or even a script that is or will be part of a software product to ChatGPT to find the solution. The issue is that ChatGPT is collecting the data provided by users and is continuously learning from what is collecting, every answer is different from a previous one and it is adjusted on previous experience from previous users. Considering this aspect, there is a great possibility that a user will receive an answer for a similar issue that will contain a part section from a company software product code that was provided by an employee. However, this has not yet been proven and the assumption is based on the ChatGPT training model. As the chat bot is stating by itself, "ChatGPT and other language models like me continuously improve through a process called training. Training involves feeding large amounts of text data into the model and allowing it to learn patterns and relationships between words and phrases. The more data the model is trained on, the better it can understand and generate language." [4] (Response to the question: "How is ChatGPT continuously improving?")

Another inside threat is coming from non-technical users that are trying to complete their tasks and are seeking ChatGPT responses for different research topics or to help them with writing an email or a contract. A simple example would be a user that needs help writing an email and they are providing more details than are needed (such as names, financial data or even PII). Having in mind that employees could involuntarily exfiltrate data, the companies have already chosen to block any access to ChatGPT from corporate computers and networks.

The attention that ChatGPT has received since the beginning of 2023 is enormous and discussions about data ownership and the intellectual property "created" by AI became more intense. The European Commission stated on February 20th: "On this topic it is important to know that the question of ownership and authorship of AI-generated works is not fully settled by the law yet, and as a "hot topic" may evolve in the years to come depending on regulatory changes and on case law. For now, it seems that artists or creators who use AI to support their creative process may be able to claim ownership of the work if it reflects their choices and creativity. On the other hand, a generic command such as "write a love song" would end up in ChatGPT generating a love song text without any real creative choices originating from the user – in such cases the existence of copyright or of a "work" in the sense of copyright law is quite doubtful. "[20]

There is an open debate about the GDPR and compliance when it come to ChatGPT but it is important to mention that the chat bot is also mentioning that "It's important to note that you should not provide any confidential, proprietary, or personal information when interacting with me, as the information may be logged and could potentially be accessed by OpenAI" when is receiving a question about data processing or protection. The EU companies that are looking to leverage this technology must consider the privacy risk because OpenAI, the company developing ChatGPT, is a Data Processor and can process data coming from conversations and all their servers are based in the

USA. Moreover, the right to be forgotten as outlined in Article 17 EU-GDPR is difficult to be enforced since natural language processing is used to create responses from the collected data, making it nearly impossible to remove all traces of an individual's personal information.

4. Conclusions

ChatGPT is a great tool that uses top technology and is surely going to make our huma lives easier. However, there are multiple risks that should be considered since there are, always been, bad intentions. The information security risks should be addressed and, if possible, mitigated since the threats are increasing and the detections and responses that we know yesterday will not be enough tomorrow. The facile access to write code without technical knowledge will enlarge the number of cybersecurity attacks and will permit to experience threat actors, that also have technical knowledge, to improve and optimize their methods and code. Moreover, the social engineering attacks are getting better and training people to be cautious and vigilant will be one of the companies' challenges, alongside the compliance to data protection laws.

References

[1] https://meetanshi.com/blog/chatgpt-statistics - accessed on 12.03.2023.
[2] https://www.blackberry.com/us/en/company/newsroom/press-releases/2023/chatgpt-may-already-be-used-in-nation-state-cyberattacks-say-it-decision-makers-in-blackberry-global-research - accessed on 12.03.2023.
[3] Scott Kevin. Microsoft teams up with OpenAI to exclusively license GPT-3 language model 2020.
[4] https://chat.openai.com/chat - accessed on 06.03.2023.
[5] https://openai.com/product/gpt-4 - accessed on 04.04.2023.
[6] https://openai.com/pricing - accessed on 04.04.2023.
[7] J. Deng and Y. Lin, "The Benefits and Challenges of ChatGPT: An Overview", FCIS, vol. 2, no. 2, pp. 81–83, Jan. 2023.
[8] https://openai.com/blog/chatgpt - accessed on 06.03.2023.
[9] https://gptzero.me - accessed on 12.03.2023.
[10] https://openai-openai-detector.hf.space - accessed on 12.03.2023.
[11] https://gptzero.me/faq - accessed on 12.03.2023.
[12] https://www.hoxhunt.com/blog/the-future-of-phishing-spearphishing-and-bec-attacks-according-to-chatgpt - accessed on 12.03.2023.
[13] https://www.cyberark.com/resources/threat-research-blog/chatting-our-way-into-creating-a-polymorphic-malware - accessed on 12.03.2023.
[14] https://research.checkpoint.com/2022/opwnai-ai-that-can-save-the-day-or-hack-it-away/ - accessed on 14.02.2023.
[15] https://blog.morphisec.com/chatgpt-malware-production - accessed on 14.02.2023.
[16] https://terranovasecurity.com/cybercriminals-can-use-chatgpt-to-their-advantage/ - accessed on 12.03.2023.
[17] https://www.geeksforgeeks.org/malware-as-a-service-maas/ - accessed on 12.03.2023.
[18] https://research.checkpoint.com/2023/opwnai-cybercriminals-starting-to-use-chatgpt/ - accessed on 14.02.2023.
[19] https://blog.checkpoint.com/2023/02/07/cybercriminals-bypass-chatgpt-restrictions-to-generate-malicious-content/ - accessed on 10.03.2023.
[20] https://intellectual-property-helpdesk.ec.europa.eu/news-events/news/intellectual-property-chatgpt-2023-02-20_en - accessed on 10.03.2023.