Generative Neural Networks as a Tool for Web Applications Penetration Testing

Petr GALLUS
Department of Informatics and Cyber Operations
University of Defence
Brno, Czech Republic

Marcel ŠTĚPÁNEK
Department of Aeronautical Engineering
University of Defence
Brno, Czech Republic

Tomáš RÁČIL
Department of Informatics and Cyber Operations
University of Defence
Brno, Czech Republic

Petr FRANTIŠ
Department of Informatics and Cyber Operations
University of Defence
Brno, Czech Republic

***Abstract*— The scientific paper delves into the potential of generative neural networks as a powerful tool for web application penetration testing. By leveraging the capabilities of these networks, we aim to augment traditional testing methodologies and advance the field of vulnerability detection.**

In the second section, the paper provides an overview of OpenAI, a leading organization at the forefront of artificial intelligence research and development. OpenAI has contributed significantly to the field of natural language processing and has developed advanced models like ChatGPT, which have the potential to revolutionize various industries, including cybersecurity. It explores the underlying technology behind ChatGPT and discusses its implications for the field of web application penetration testing.

Third section focuses on present the details of the experimental setup. A series of three experiments was conducted to evaluate the effectiveness of generative neural networks, specifically ChatGPT, in web application penetration testing. Through these experiments, its aim was to demonstrate the practical application of generative neural networks in identifying and exploiting web-based security vulnerabilities.

In the fourth section, the results obtained from the experiments are presented. Parts of the experiment were categorized into three sub-results, each highlighting a specific aspect of vulnerability detection. The main intention was to highlight the potential of generative neural networks as an innovative and effective tool for web application penetration testing.

In conclusion, the paper showcases the advancements made possible by generative neural networks in the domain of web application penetration testing. By automating certain aspects of the testing process and enhancing vulnerability detection, these networks hold immense promise for improving the overall security posture of web-based systems. The findings presented in this paper contribute to the growing body of knowledge in the field and open up new avenues for further research and development in this critical area of cybersecurity.

Keywords*— *OpenAI, ChatGPT, Cybersecurity, Penetration testing, Web application, Wordpress, Generative Neural Networks

I. INTRODUCTION

Web application penetration testing is a critical process to identify and mitigate security vulnerabilities in web-based systems. Traditionally, this task has relied on manual techniques, which can be time-consuming and prone to human error. However, the advent of generative neural networks has opened new possibilities for automating and enhancing the effectiveness of web application penetration testing.

The research explores the potential of generative neural networks as a tool for web application penetration testing. It examines how these networks can be leveraged to augment traditional testing methodologies and improve the efficiency and accuracy of vulnerability detection.

II. OPENAI DEVELOPMENT COMPANY

A. OpenAI

OpenAI is an AI research and deployment company. According to the official statement of the company, their mission is to ensure that artificial general intelligence benefits all of humanity, by which they mean highly autonomous systems that outperform humans at most economically valuable work—benefits all of humanity. Launched in 2015 and headquartered in San Francisco, this altruistic artificial intelligence company was founded by Musk and Altman. They saw collaborations with other Silicon Valley tech experts like Peter Thiel and LinkedIn co-founder Reid Hoffman who pledged USD 1 billion for OpenAI that year [1] [2].

Much research by OpenAI is published at top machine learning conferences. The organization also contributes open-source software tools for accelerating AI research and releases blog posts to communicate their research to others in the field.The company focuses on long-term research - working on problems that require making fundamental advances in AI capabilities.

B. Products

OpenAI has developed several models and products including ChatGPT (a model capable of interacting with users in a conversational way), Dall-E (AI system capable of generating images from natural language descriptions) or Whisper (neural network approaching human-level accuracy speech recognition for the English language) [3].

III. CHATGPT PRODUCT

ChatGPT is a variant of the popular language generation model, GPT (Generative Pre-training Transformer). It is specifically designed for chatbot applications, where it can generate human-like responses to user input in a conversation.

This work was developed as part of a research project DZRO-FVT21-KYBERSILY funded by the Ministry of Defence of the Czech Republic.

979-8-3503-3839-3/23/$31.00 ©2023 IEEE

2023 Communication and Information Technologies (KIT) | 979-8-3503-3839-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/KIT59097.2023.10297109

The original GPT model was developed by researchers at OpenAI and released in 2018. It was trained on a large dataset of internet text and could generate human-like text for a variety of language tasks. ChatGPT was developed as a modification of GPT, with the aim of making it more suited for chatbot applications.

The first version of ChatGPT was released in 2020 and quickly gained popularity due to its ability to generate natural and engaging responses in chatbot conversations. ChatGPT was officially launched as a prototype on November 30, 2022

A. Sophisticated machine learning system

Unlike traditional website chatbots, ChatGPT is not connected to the internet as expected and does not have access to any external information. Instead, it is dependent on the data it has been trained on to generate responses. These data include a vast array of texts from various sources like books, articles, and websites. It works with 175 billion parameters in total and the model is restricted to language only – it cannot produce video, sound, or images as an output [4].

The model was trained using text databases from the internet. This included a whopping 570GB of data obtained from books, web texts, Wikipedia, articles, social media posts and other pieces of writing on the internet. To be even more exact, 300 billion words were fed into the system. To be able to respond at almost any request with a guiding answer, the model went through a supervised testing stage [5].

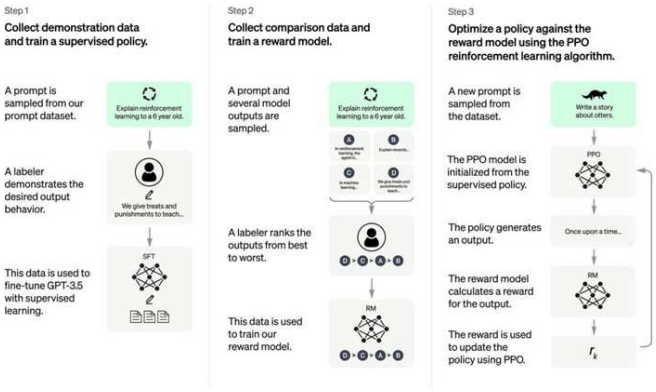

Fig. 1. Reinforcement learning from human feedback (RLHF) core technique behind the model [1].

ChatGPT model uses Reinforcement Learning from Human Feedback (RLHF), which uses masked self-attention to mask out certain words or phrases, allowing the model to focus on the most relevant aspects of the input and generate responses that are more likely to be useful and appropriate to the user [6].

According to the official documentation picture above, ChatGPT goes through three stages:

Step 1: Collect demonstration data and train a supervised policy.

The human labellers are given a prompt from the prompt dataset where the labellers provide answers, which is the desired behaviour on the input prompt. This dataset generated by the labeller is used to fine-tune a pre-trained GPT-3.5 model using supervised learning. A supervised policy is a decision-making strategy trained on labelled data to generate responses to user inputs. The supervised policy generates responses in a way that is based on the patterns and structures learned from the labelled training data. For example, the supervised policy of a chatbot trained on a dataset of medical conversations will generate responses that can answer medical-related questions in an appropriate and informative way.

Step 2: Collect comparison data and train a reward model.

The reward model in ChatGPT is a mechanism to encourage the chatbot to generate responses that are more appropriate and desirable to the user based on the rewards or penalties it receives for its responses. The rewards model takes in a prompt and responses and outputs a scalar reward. Comparison data between model outputs are collected. Labellers rank the output they prefer for a given input to train a reward model for Reinforcement Learning to predict the human-preferred output. Model outputs are conversations that AI trainers had with the chatbot.

Step 3: Optimize a policy against the Reward model using PPO.

The output of the Rewards Model is used as a scalar reward for Reinforcement Learning. The supervised policy is fine-tuned to optimize the reward using the on-policy Proximal Policy Optimization (PPO) algorithm. The Rewards model determines the reward or penalty for each response generated by the chatbot and uses this reward function to guide the learning process to generate relevant, informative, or engaging responses to the user, as well as to avoid generating responses that are inappropriate or offensive. Proximal Policy Optimization (PPO) is an on-policy Reinforcement Learning (RL) algorithm to generate an optimized policy that uses the Rewards model. On-Policy RL algorithms use the same policy to evaluate as well as improve to make decisions. PPO uses a technique called Trust Region Optimization (TRO) to update the policy to improve the training stability by limiting the changes to the policy at each training epoch. TRO balances exploration and exploitation to find an optimized policy efficiently.

B. Possible impacts on cybersecurity

We live in an increasingly digitised world. The security and safety of our data have become paramount. In the past, hackers were highly skilled programmers who could code their malware and navigate sophisticated security protocols.

That is no longer the case; malware can now be sold as an intelligent solution that only requires a plug and play. This brings non-computer expert hackers into the fray and ultimately increases the number of hackers.

Artificial intelligence is playing an increasingly important role in cybersecurity — for both good and bad. Organizations can leverage the latest AI-based tools to better detect threats and protect their systems and data resources. But cyber criminals can also use the technology to launch more sophisticated attacks. According to a datasheet from cybersecurity firm Imperva, 75% of all cyber-attacks target web applications, with web-based vulnerabilities accounting for over 80% of all security vulnerabilities discovered.

C. ChatGPT security development

For our experiment, the freely available and free of charge ChatGPT 3.5 was utilized. The results achieved correspond to the capabilities of this version of AI. However, with the rapid development of this platform, ChatGPT-4 is now available ($20 monthly), offering numerous modifications, including those related to cybersecurity. Its advanced features around AI misuse prevention indicate the direction in which AI is evolving to defend against the unethical and illegal activities of some users [7].

One potential future application of AI models like GPT-4 is the prediction of user intent in generating malicious code. By recognizing patterns in user requests for AI-generated source code, these advanced models could potentially identify when a user is seeking to create malware for illicit purposes. This would allow security professionals to develop proactive measures for detecting and mitigating such threats. However, this same capability could also be exploited by malicious actors to refine their attack strategies and identify new targets for their nefarious activities.

It is essential for researchers, developers, and policymakers to collaborate in order to harness the potential benefits of AI while minimizing the risks associated with its misuse [8].

IV. Experiment – Penetration Testing

It is rumoured in the public domain that the ChatGPT platform can help users with almost any problem, and the experiment will test the helpfulness of the platform in black hat activities during penetration testing of a web application.

The aim of the experiment is to use the newly published artificial intelligence based ChatGPT platform to breach the University of Defence web application and subsequently obtain sensitive information of the organisation. Whole procedure below was done using the chatbot based on artificial intelligence (AI). The experiment was carried out during the months February and March 2023.

The subject of the analysis is the new web application of the University of Defence. The current version of the new website went live in November 2022, the same time that ChatGPT was officially launched to the public. The unversity's web application is built on the open-source Wordpress content management system, which is notorious for a few vulnerabilities found in not updated versions of plugins that, for example, allow any logged-in user to download a backup of the site without proper authorization.

V. Results

A. Obtaining Wordpress version and its vulnerability

First, it was necessary to find out the most known vulnerabilities of the Wordpress content management system. This system is the basis for the web application of the University of Defence. As the simplest option, ChatGPT offered to find out the current version of the Wordpress web application. Even with this issue ChatGPT handled it playfully and was able to write a tutorial on how an attacker can find out the current version in the source code.

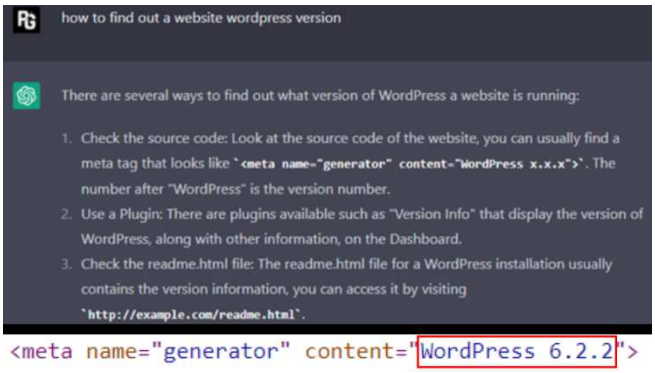

Fig. 2. Response for finding out a Wordpress version and and the subsequent discovery in the source code of the real website.

Wordpress version 6.2.2 which has been installed and used during writing the paper, suffers a Server-Side Request Forgery vulnerability. This could allow a malicious actor to cause a website to execute website requests to an arbitrary domain of the attacker, which would allow a hacker to find sensitive information of other services running on the system. This vulnerability has not been known to be fixed yet [9] [10].

B. Obtaining template version and its vulnerability

The next phase of the experiment was to interact with ChatGPT to obtain a procedure to extract information about the template used, its version and possible vulnerabilities. ChatGPT responded to the well-formed request without any problems and all the information about the template was also traced in the source code of the site.

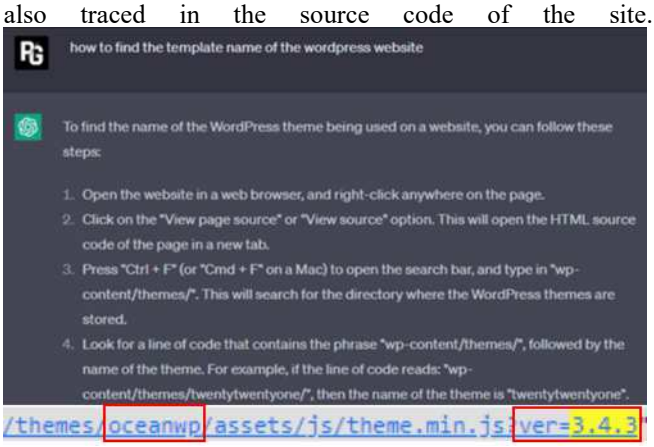

Fig. 3. Obtaining procedure to extract information about template and its current version in the source code of the real website.

OceanWP 3.4.3 suffers vulnerability listed as “CVE-2023-23700”, which allows a hacker to conduct Local File Inclusion.

The function first will assign the $_POST['slug'] value that previously applied esc_attr function to the $slug variable. The code then will call $this->print_pane function with $slug variable as parameter. The $slug variable will be formatted to the $file variable. The code then will re-assign the $file variable with the apply_filters function. Since the value passed to the apply_filters hook name is not valid, the $file variable is not modified. Finally, the code will check if the $file exists using file_exists function, and then will include the $file. With this condition, a hacker can supply a path traversal payload to include a local arbitrary .php file [11] [12] [13].

C. *Obtaining real usernames as for login credentials*

Obtaining real usernames as login credentials is a critical aspect of ensuring secure access to various systems and platforms. Usernames serve as unique identifiers for individuals or entities accessing a particular application or service. They are typically required alongside passwords or other authentication mechanisms to verify the user's identity. The process of obtaining real usernames involves gathering accurate and legitimate information about individuals' or organizations' authorized login credentials. This can be achieved through various methods, such as user registration processes, database queries, or user directory listings.

ChatGPT was asked about a Linux command to obtain usernames as credentials into the Wordpress content management system using the wpscan hacking tool. Without hesitation, ChatGPT responded and explained the various flags and extensions that can be used for the attack, which is outlined in the figure below.

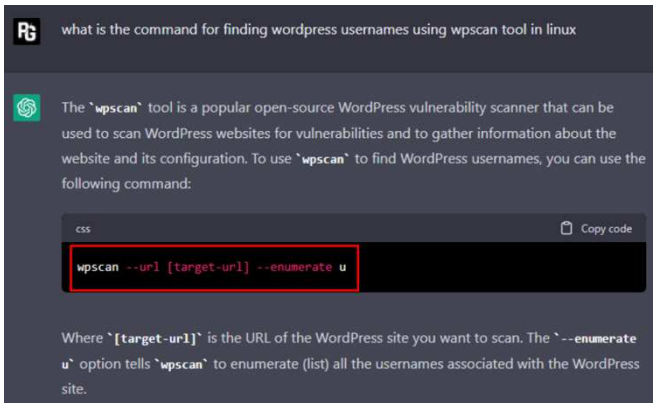

Fig. 4. Obtaining a linux command for wpscan login credentials attack due to the given prompt.

As part of the experiment, an attempt was made to apply the procedure proposed by ChatGPT. The hacker tool wpscan started to find out all available information about the sample (website) and the result of the attack was that it obtained real usernames that serve as login credentials to the system, including the administrator account. In the figure below, the full names of the login credentials found have been deliberately removed as this is sensitive data which could be misused.

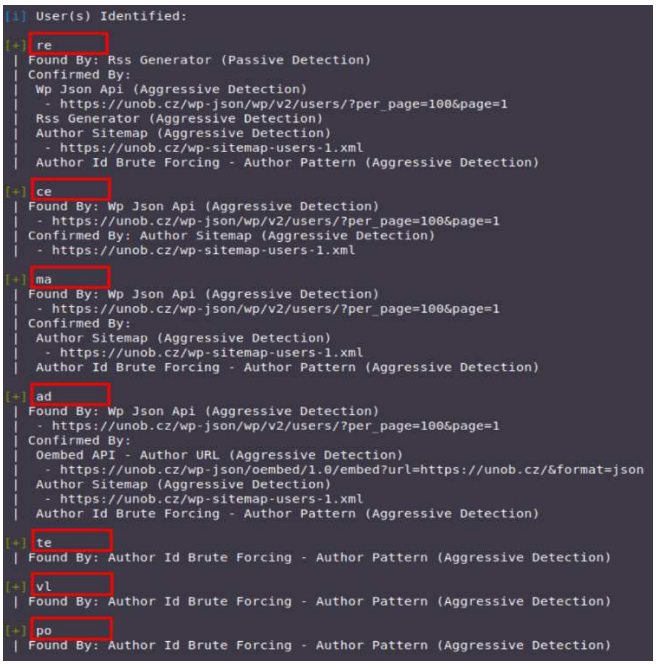

Fig. 5. The result of an attack on the web and obtaining login usernames to the content management system.

VI. Conclusion

In this research, an experiment verified the procedures that were obtained in response from the generative neural network behind ChatGPT. Three penetration tests of a real-world University of Defence web application were conducted.

Thanks to the exact procedures described by ChatGPT, several weaknesses of the web application have been identified and this valuable information which has been obtained can serve as basic clues for further, more dangerous attacks. Found weaknesses include Wordpress and template versions with their vulnerabilities and successfully obtained usernames as for login credentials into the system including administrator account username.

As an AI language model, ChatGPT can certainly be a valuable tool and guide for web application penetration testing (pen testing). With its vast knowledge and understanding of web technologies, security principles, and best practices, it can assist a cybersecurity expert in various aspects of the pen testing process. However, it is also a growing threat because an average user could access it and simply repeat the process - even for illegal activities.

References

[1] OpenAI, “About OpenAI”, Accessed on: May 18, 2023 [Online]. Available: https://openai.com/about/

[2] Golden.com, “OpenAI”, Accessed on: May 25, 2023 [Online]. Available: https://golden.com/wiki/openai-kz9j6x

[3] TechCrunch, “OpenAI debuts Whisper API for speech-to-text transcription and translation”, Accessed on: April 18, 2023 [Online]. Available: https://techcrunch.com/2023/03/01/openai-debuts-whisper-api-for-text-to-speech-transcription-and-translation/

[4] OpenAI, “Aligning language models to follow instructions”, Accessed on: May 04, 2023 [Online]. Available: https://openai.com/research/instruction-following

[5] BBC Science Focus Magazine, “ChatGPT: Everything you need to know about OpenAI´s GPT-4 tool”, Accessed on: May 16, 2023 [Online]. Available: https://www.sciencefocus.com/future-technology/gpt-3/

[6] Medium, “Reinforcement Learning from human Feedback, InstructGPT, and ChatGPT” Accessed On: May 30, 2023. [Online]. Available: https://medium.com/aiguys/reinforcement-learning-from-human-feedback-instructgpt-and-chatgpt-693d00cb9c58

[7] A. Radford et al., "Improving Language Understanding by Generative Pre-Training," Accessed on: April 27, 2023. [Online]. Available: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf

[8] T. B. Brown et al., "Language Models are Few-Shot Learners," Accessed on: April 27, 2023. [Online]. Available: https://arxiv.org/pdf/2005.14165.pdf

[9] Patchstack, “Wordpress <= 6.2 is vulnerable to Server Side Request Forgery (SSRF)” Accessed on: May 29, 2023. [Online]. Available: https://patchstack.com/database/vulnerability/wordpress/wordpress-6-1-1-unauth-blind-ssrf-vulnerability

[10] WPScan, “Wordpress 6.1.1 Vulnerabilities” Accessed on: May 30, 2023. [Online]. Available: https://wpscan.com/wordpress/611

[11] Patchstack, “Security Vulnerability In OceanWP Theme <= 3.4.1” Accessed on: May 29, 2023 [Online]. Available: https://patchstack.com/articles/subscriber-path-traversal-leading-to-local-file-inclusion-in-oceanwp-theme/

[12] VulnDB, “OceanWP Plugin up to 3.4.1 on Wordpress file inclusion” Accessed on: May 30, 2023. [Online]. Available: https://vuldb.com/?id.222115

[13] Mitre, “CVE-2023-23700” Accessed on: May 30, 2023. [Online]. Available: https://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2023-23700