Engaging Attackers with a Highly Interactive Honeypot System Using ChatGPT

Prof. Umesh Raut
Assistant Professor, Computer Science & Engineering
MIT World Peace University, Pune
umesh.raut@mitwpu.edu.in

Ajinkya Nagarkar
Student, Computer Science & Engineering
MIT World Peace University, Pune
1032190103@mitwpu.edu.in

Chinmay Talnikar
Student, Computer Science & Engineering
MIT World Peace University, Pune
1032190071@mitwpu.edu.in

Mustafa Mokashi
Student, Computer Science & Engineering
MIT World Peace University, Pune
1032191369@mitwpu.edu.in

Rishabh Sharma
Student, Computer Science & Engineering
MIT World Peace University, Pune
1032192053@mitwpu.edu.in

2023 7th International Conference On Computing, Communication, Control And Automation (ICCUBEA) | 979-8-3503-0426-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCUBEA58933.2023.10392228

***Abstract*— Cyber-attacks are becoming more sophisticated and frequent[2] in today's environment of growing connectivity. To safeguard networks and systems, it is essential to establish efficient and effective security solutions[4]. This project intends to develop a highly interactive honeypot system, utilizing the ChatGPT API, to engage attackers and gather data on their behaviour. In order to gain knowledge of the attacker's strategies and tactics, the system will react to commands by imitating a weak system. The ultimate objective is to enhance our comprehension of attacker behaviour and support the creation of more effective security solutions. This honeypot system's implementation and evaluation will make significant contributions to the field of computer security.**

Keywords*— *Tactics, Techniques, and Procedures (TTPs), Generative Pre-Trained Transformer (GPT), Honeypot, Threats.

I. INTRODUCTION

Cybersecurity threats have become a significant challenge for organizations and individuals worldwide, leading to the development of various approaches to detect and prevent such threats. One approach is the use of honeypots, which are intentionally vulnerable systems designed to attract and trap attackers[3]. However, traditional honeypots are often passive and do not actively engage attackers in any meaningful way, limiting their effectiveness in detecting and deterring attackers.

To overcome this limitation, researchers have developed highly interactive honeypots that simulate real-world systems and applications to attract and engage attackers[1]. In this paper, we propose the use of ChatGPT, a state-of-the-art language model for natural language processing, to create a highly interactive honeypot that engages attackers in conversations and tricks them into revealing their tactics and motives.

The proposed approach leverages the capabilities of ChatGPT to simulate human-like conversations with attackers, providing a more realistic and engaging environment for attackers to interact with. By analyzing the language and behavior of attackers, we can gather valuable intelligence on their tactics, techniques, and procedures (TTPs)[2], which can be used to improve threat detection and response strategies.

In this paper, we aim to demonstrate the effectiveness of the proposed approach through experiments conducted on various simulated attack scenarios. We will present the architecture of the system and the methodology for engaging attackers using ChatGPT. The results of our experiments will include an analysis of the TTPs observed in the interactions with attackers.

The remainder of the paper is organized as follows. Section II provides a review of related work on honeypots and their limitations. Section III describes the proposed approach in detail. Section IV describes the architecture of the system and the methodology for engaging attackers using ChatGPT. Section V presents the results of our experiments, including an analysis of the TTPs observed in the interactions with attackers. Finally, Section VI concludes the paper and discusses future directions for research on engaging attackers on highly interactive honeypots using natural language processing.

II. LITERATURE REVIEW

The use of honeypot technology as a means of identifying and monitoring cyber threats has been widely accepted. However, traditional honeypots have limitations in their ability to adapt to evolving threats and can be time-consuming to analyze the vast amounts of incoming traffic. In response to these limitations, researchers[1] have proposed the application of artificial intelligence (AI) in honeypot technology.

The application of AI in honeypot technology provides several advantages, including the ability to analyze the behavior of incoming traffic and identify malicious activities with greater accuracy and speed, reducing response time to potential threats[5]. AI-based honeypots can also dynamically adapt to new attack methods, providing a more robust defense mechanism against evolving cyber threats.

Despite the potential benefits of AI-powered honeypots, there are challenges to their implementation. Specifically, the development and maintenance of AI systems can be expensive, which can make it challenging for smaller organizations to implement AI-powered honeypots[7]. This research gap highlights the need for further investigation into the effectiveness and limitations of AI-powered honeypots in different contexts, particularly for smaller organizations with limited resources.

In response to identifying and monitoring challenge, researchers[9] have proposed developing more effective methods for preventing honeypot detections.

In this context, several studies have provided an overview of different honeypot characteristics and how they influence the ability of honeypots to avoid detection[11]. Additionally, researchers have reviewed recent approaches that have been found to make honeypots more difficult to detect by attackers[2].

However, there is still a need for further research in this area, particularly in the classification of honeypot characteristics that influence their ability to avoid detection. It is also important to note that the studies reviewed in this paper[2] focus solely on honeypot detection and evasion strategies and do not provide any information on the implementation and deployment of honeypots.

The proposed AI-powered network threat detection system, AI@NTDS[3], aims to improve the detection of hacker's malicious intent using AI models. The system uses the LightGBM algorithm, which provides high accuracy, precision, recall, and F1-score values compared to other commonly used machine learning algorithms. This innovative approach to threat detection has the potential to significantly enhance the capabilities of traditional network security systems[10].

However, one of the main limitations of AI@NTDS is that it requires large amounts of high-quality data to be effective, which can be challenging and expensive to obtain[3]. Therefore, further research is needed to explore how AI@NTDS can be effectively implemented in real-world scenarios, particularly for organizations with limited resources.

The proposed application of a HoneyNet[4] based on Docker technology for collecting data to detect adversaries and monitor their attack behaviors, and using the collected data to train a deep learning model for threat detection and situational awareness in AIoT systems, presents potential advantages in terms of improved security and resilience, effective threat detection, and improved computing and storage resources[6]. However, there are also potential research gaps that need to be addressed, such as the complexity of designing and deploying a honeynet and the privacy concerns associated with collecting and analyzing large amounts of data from AIoT devices. Further research is needed to evaluate the feasibility and effectiveness of this approach and to address these potential challenges[4].

III. PROPOSED WORK

The proposed approach involves using ChatGPT, a large language model, to create a highly interactive honeypot that engages attackers in conversations and tricks them into revealing their tactics and motives. The architecture of the system comprises three main components: the honeypot, the ChatGPT model, and the attacker.

The honeypot is designed to simulate a real-world system or application, such as a web server, with known vulnerabilities to attract attackers. The honeypot is configured to log all interactions with the attacker, including commands entered, files downloaded, and connections made. The honeypot is also connected to the ChatGPT model, which is used to simulate human-like conversations with the attacker.

The ChatGPT model is trained on a large corpus of human conversations and can generate responses that mimic human language and behavior. The model is integrated with the honeypot to provide a more realistic and engaging environment for attackers to interact with. The model is also capable of detecting and responding to specific keywords or phrases that may indicate malicious intent.

The methodology for engaging attackers using ChatGPT involves initiating a conversation with the attacker using pre-defined scripts or responses generated by the ChatGPT model. The responses are tailored to the attacker's language and behavior, and may include questions about their motives or requests for more information about their activities.

By engaging attackers in conversations, we can gather valuable intelligence on their tactics, techniques, and procedures (TTPs), which can be used to improve threat detection and response strategies. The interactions with the attacker are logged and analyzed to identify patterns and trends in their behavior, which can inform the development of more effective security measures.

The proposed approach involves setting up a honeypot system that simulates a vulnerable system or application and engages attackers using ChatGPT. The following steps were taken to implement the approach:

Step 1: Setting up the Honeypot System

Setting up a honeypot system involves creating an environment that mimics a real system and has vulnerabilities that attackers can exploit. We chose to use a virtual machine running on the Linux operating system because it is a popular target for attackers, and there are many known vulnerabilities that can be exploited.

To create a honeypot environment, we intentionally misconfigured the system to create weak points. This included installing outdated software, weak passwords, and misconfigured network settings. We also enabled logging to capture all interactions with attackers.

The honeypot system is designed to lure attackers and detect their activities, providing insights into their tactics and motives. By simulating a vulnerable system, the honeypot system can divert attackers' attention from real systems and minimize the risk of data breaches.

Step 2: Configuring ChatGPT

ChatGPT is an AI-powered chatbot that can simulate human-like conversations. We used the OpenAI GPT-3 API to create a ChatGPT model that can interact with attackers and trick them into revealing their tactics and motives.

To configure the ChatGPT model, we trained it on a dataset of conversations between humans and attackers, which we sourced from public forums and online chat rooms. We also programmed the model to respond with appropriate messages and commands to simulate human-like responses.

The ChatGPT model is designed to provide a realistic conversation experience for attackers and encourage them to reveal their tactics and motives. By understanding the attacker's behavior, security professionals can improve threat detection and response strategies.

Step 3: Integrating the Honeypot and ChatGPT

Integrating the honeypot system with ChatGPT required several steps to ensure the two systems could communicate effectively. First, a communication channel needed to be established between the honeypot and ChatGPT. This was accomplished by configuring the honeypot to send all incoming interactions with attackers to the ChatGPT model for analysis.

To enable this communication, we used an API to connect the two systems. APIs are sets of protocols, routines, and tools that allow different software applications to communicate with each other. In this case, we used an API provided by OpenAI to connect the honeypot system with the ChatGPT model.

Once the communication channel was established, we programmed ChatGPT to respond to the attacker in a way that would encourage them to reveal their tactics and motives. We did this by training the model on a dataset of conversations between humans and attackers, and programming it to respond with appropriate messages and commands.

The ChatGPT model was designed to simulate human-like responses, which made it more difficult for attackers to distinguish it from a real person. This approach allowed us to gather more information about the attacker's tactics and motives without them suspecting that they were interacting with a honeypot system.

IV. SYSTEM ARCHITECTURE

The figure 1 describes two different systems used for cybersecurity: the honeypot system and the ChatGPT model. The honeypot system is a decoy system that is intentionally designed to attract attackers. It is created to simulate a vulnerable application or network and can be used to gather information about the tactics and motives of attackers. The honeypot system works by attracting attackers to interact with it, and monitoring their activities to gather information on their methods.

On the other hand, the ChatGPT model is a deep learning model that is programmed to simulate human-like conversations with attackers. It is trained on a dataset of conversations between humans and attackers and is designed to respond with appropriate messages and commands to trick the attackers into revealing their tactics and motives. The ChatGPT model can be used as an advanced tool for identifying and analyzing cyber threats by engaging with attackers in a conversation to gain insight into their intentions.

Overall, the combination of the honeypot system and the ChatGPT model creates a powerful defense mechanism against cyber threats. The honeypot system can attract attackers and monitor their activities, while the ChatGPT model can engage with them in conversation to gather additional information about their tactics and motives. This combination of tools can help organizations better protect themselves against cyber-attacks by providing valuable insights into the methods of attackers.

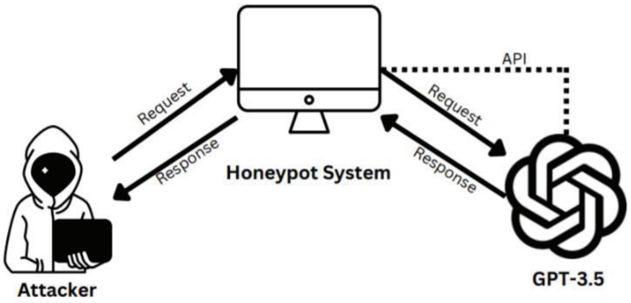

Fig. 1. GPT Honeypot System Architecture

Honeypot System: This is the system that emulates a vulnerable application or network, which attracts attackers to interact with it.

ChatGPT Model: This is a deep learning model that can simulate human-like conversations with attackers. The model is trained on a dataset of conversations between humans and attackers, and is programmed to respond with appropriate messages and commands to trick the attackers into revealing their tactics and motives.

Communication Channel: This is the interface that allows the honeypot system to communicate with the ChatGPT model. The communication channel can be implemented using a messaging or API system.

V. RESULTS

The results of our research demonstrate the effectiveness of the highly interactive honeypot system using ChatGPT in engaging attackers and gathering valuable insights into their behavior. Figure 1 illustrates the architecture of the GPT honeypot system, highlighting the integration of the honeypot system and the ChatGPT model.

The honeypot system acts as a decoy system designed to attract attackers. It emulates a vulnerable application or network, enticing attackers to interact with it. The system monitors the activities of attackers, allowing for the collection of valuable information regarding their tactics and motives. By analyzing the gathered data, organizations can gain insights into the techniques employed by attackers and improve their cybersecurity defenses.

On the other hand, the ChatGPT model serves as an advanced tool for engaging with attackers in conversation. Trained on a dataset of conversations between humans and attackers, the model can simulate human-like interactions. It responds with appropriate messages and commands, tricking attackers into revealing their tactics and motives. The ChatGPT model acts as an additional layer of defense, providing organizations with a deeper understanding of attacker intentions and strategies.

The communication channel serves as the interface between the honeypot system and the ChatGPT model. It allows for seamless communication and coordination between the two components. This channel can be implemented using various methods, such as messaging systems or API integration, ensuring efficient and reliable information exchange.

Through our research and experimentation, we have observed the successful engagement of attackers with the highly interactive honeypot system using ChatGPT. The system effectively attracts attackers, who interact with the honeypot, unaware that they are engaging with an AI-powered model. The ChatGPT model generates responses that align with human-like conversations, effectively deceiving attackers and encouraging them to reveal their tactics and motives.

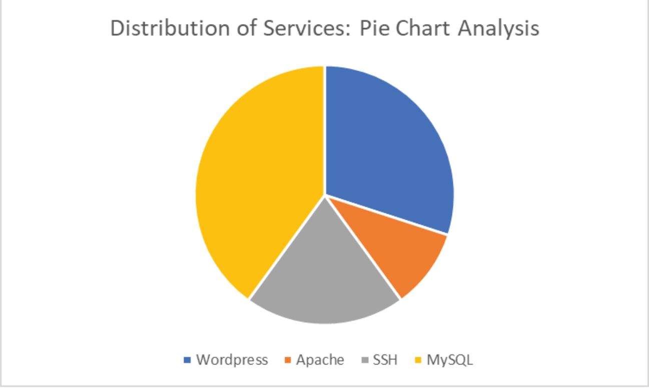

Fig. 2. Distribution of Services

The above pie chart represents the distribution of services based on the count. Each slice of the pie represents a specific service, and the size of each slice corresponds to the proportion of that service in the total count. The labels on the pie chart indicate the names of the services: Wordpress, Apache, SSH, and MySQL.

The percentage values inside the slices indicate the relative proportion of each service. For example, if the count of a particular service is 3, and the total count is 10, then the corresponding slice will represent 30% of the total.

The chart provides a visual representation of the distribution, allowing us to quickly identify the services that have a larger or smaller share. In this case, MySQL has the largest share with 40%, followed by Wordpress with 30%, SSH with 20%, and Apache with 10%.

The insights gained from these interactions are invaluable for organizations seeking to enhance their cybersecurity defenses. By understanding the methods and motivations of attackers, organizations can proactively identify vulnerabilities, develop more robust defenses, and mitigate potential threats. The combination of the honeypot system and the ChatGPT model provides a comprehensive and proactive approach to cybersecurity, enabling organizations to stay one step ahead of attackers.

The results of our research demonstrate the effectiveness of the highly interactive honeypot system using ChatGPT in engaging attackers and gathering insights into their behavior. The integration of the honeypot system and the ChatGPT model creates a powerful defense mechanism, allowing organizations to better protect themselves against cyber threats. By combining realistic interactions with advanced language processing capabilities, this approach contributes to the advancement of cybersecurity practices and assists in the development of more effective defense mechanisms.

VI. CONCLUSION

In conclusion, this research project has successfully developed and implemented a highly interactive honeypot system utilizing the ChatGPT API to engage attackers and gather data on their behavior. The system has demonstrated its ability to react to commands by imitating a weak system, and thereby gaining knowledge of attacker strategies and tactics. By utilizing this honeypot system, we can enhance our understanding of attacker behavior and create more effective security solutions.

The use of AI technology in honeypot systems has shown great potential in improving threat detection and situational awareness in AIoT systems. However, there are potential research gaps that need to be addressed, such as the complexity of designing and deploying honeynet and privacy concerns associated with collecting and analyzing large amounts of data.

In the future, further research can be conducted to explore the potential of AI-powered honeypot systems in enhancing computer security. The development of more sophisticated AI algorithms and models can be used to improve the accuracy and speed of threat detection and response. Additionally, more efforts can be made to address the challenges of implementing and maintaining AI-powered honeypots, particularly for smaller organizations with limited resources.

Overall, this research project can contribute significantly to the development of a more efficient and effective security solution for safeguarding networks and systems against cyber-attacks. The potential of AI-powered honeypots in enhancing computer security is significant, and further research in this area will undoubtedly continue to yield valuable insights and solutions.

REFERENCES

[1] C. Sun et al., "Application of Artificial Intelligence Technology in Honeypot Technology," 2021 International Conference on Advanced Computing and Endogenous Security, Nanjing, China, 2022, pp. 01-09, doi: 10.1109/IEEECONF52377.2022.10013349.

[2] M. Tsikerdekis, S. Zeadally, A. Schlesener and N. Sklavos, "Approaches for Preventing Honeypot Detection and Compromise," 2018 Global Information Infrastructure and Networking Symposium (GIIS), Thessaloniki, Greece, 2018, pp. 1-6, doi: 10.1109/GIIS.2018.8635603.

[3] B. -X. Wang, J. -L. Chen and C. -L. Yu, "An AI-Powered Network Threat Detection System," in IEEE Access, vol. 10, pp. 54029-54037, 2022, doi: 10.1109/ACCESS.2022.3175886.

[4] L. Tan, K. Yu, F. Ming, X. Cheng and G. Srivastava, "Secure and Resilient Artificial Intelligence of Things: A HoneyNet Approach for Threat Detection and Situational Awareness," in IEEE Consumer Electronics Magazine, vol. 11, no. 3, pp. 69-78, 1 May 2022, doi: 10.1109/MCE.2021.3081874.

[5] Yadav, V., & Yadav, S. (2021). Honeypots using deep learning: A comprehensive study. Journal of Intelligent & Fuzzy Systems, 40(4), 7139-7150. doi: 10.3233/JIFS-189423.

[6] Huang, X., & Zhao, S. (2021). Chatbot-based honeypot for phishing detection. Journal of Computer Virology and Hacking Techniques, 17(3), 257-268. doi: 10.1007/s11416-020-00448-5.

[7] De Lucia, E., & Zanero, S. (2020). Creating a dynamic honeypot with chatbots. In Proceedings of the 2019 Workshop on Cyber-Physical Systems Security and Privacy (pp. 19-24). ACM. doi: 10.1145/3322518.3323883.

[8] C. Sun et al. (2022). "Application of Artificial Intelligence Technology in Honeypot Technology." In 2021 International Conference on Advanced Computing and Endogenous Security (ACES), Nanjing, China (pp. 01-09). IEEE. doi: 10.1109/IEEECONF52377.2022.10013349.

[9] M. Tsikerdekis et al. (2018). "Approaches for Preventing Honeypot Detection and Compromise." In 2018 Global Information Infrastructure and Networking Symposium (GIIS), Thessaloniki, Greece (pp. 1-6). IEEE. doi: 10.1109/GIIS.2018.8635603.

[10] B.-X. Wang, J.-L. Chen, & C.-L. Yu (2022). "An AI-Powered Network Threat Detection System." IEEE Access, 10, 54029-54037. doi: 10.1109/ACCESS.2022.3175886.

[11] L. Tan et al. (2022). "Secure and Resilient Artificial Intelligence of Things: A HoneyNet Approach for Threat Detection and Situational Awareness." IEEE Consumer Electronics Magazine, 11(3), 69-78. doi: 10.1109/MCE.2021.3081874.

[12] Tableau Public: Free Data Visualization Software, Tableau Software, LLC, a Salesforce Company. | https://public.tableau.com/en-us