mgr inż. Wiktor Sędkowski
Politechnika Opolska
ORCID: 0000-0002-4543-0499

Studia Społeczne
ISSN 2081-0008
e-ISNN 2449-9714
str. 75–86

Identyfikacja zagrożeń przy użyciu metodologii STRIDE i chatbotów GPT

Threat Identification using STRIDE and GPT based chatbots

Streszczenie

W niniejszej pracy przedstawiono skuteczność wykorzystania dużych modeli językowych (LLM) takich jak ChatGPT, w celu identyfikacji potencjalnych zagrożeń i zaproponowania skutecznych środków obronnych w celu ochrony krytycznych usług sieciowych. Nowoczesne systemy informatyczne składają się z wielu modułów różnego typu, co utrudnia ręczne zarządzanie procesem identyfikowania zagrożeń. Do pomocy w identyfikowaniu zagrożeń można wykorzystać sztuczną inteligencję. Artykuł przestawia analizę porównawczą trzech różnych narzędzi opartych o model GPT (ChatGPT 4.0, STRIDE-GPT i niestandardowy model oparty na gpt-4-1106-preview). Przy użyciu tych narzędzi oraz danych pochodzących z raportów NMAP, przeprowadzona została testowa identyfikacja zagrożeń. Wyniki wskazują na możliwość wykorzystania Chatbotów do pomocy w wykrywaniu zagrożeń. Co więcej, wskazują, że zagrożenia rozpoznawane przez sztuczną inteligencję są porównywalne z zagrożeniami identyfikowanymi przez ekspertów, przy czym identyfikacja ta następuje zdecydowanie szybciej w przypadku wykorzystania sztucznej inteligencji.

Słowa kluczowe: modelowanie zagrożeń, cyberbezpieczeństwo, sztuczna inteligencja, duże modele językowe.

Abstract

This study demonstrates the effectiveness of using Large Language Models, such as ChatGPT, for the purpose of identifying potential threats and proposing effective defensive measures for protecting critical network services. Modern systems consist of

multiple workloads of diverse types, which makes it challenging to manage the process of identifying threats manually, Artificial Intelligence can be utilized to assist with the chore of identifying threats. Using a comparison methodology a comparative analysis of three distinct GPT based tools (ChatGPT 4.0, STRIDE-GPT and custom gpt-4-1106-preview based model) that were assigned the responsibility of identifying threats based on the Nmap reports was conducted. The results demonstrate the feasibility of employing ChatGPT to aid in threat detection. Furthermore, they indicate that the risks recognized by AI are comparable to those identified by human experts, while also being delivered significantly faster when utilizing AI.

Key words: threat modeling, cybersecurity, artificial intelligence, large language model.

Introduction

Threat modeling as a systematic process involves identifying and assessing potential hazards, as well as developing and prioritizing countermeasures to protect the most critical elements of a system. Thanks to threat modeling it is easier to perform the allocation of resources for system security and easier to identify and address security weaknesses at an early stage. Several threat modeling strategies have been created in recent decades. However, this paper does not aim to compare them, as this task has previously been undertaken by other academics[1] [2].

Each of threat modeling methodologies assists security teams in identifying the critical functions of systems that require protection. Furthermore, they provide security teams and organizations with a method of recognizing potential risks and evaluating them with equal importance. This work focuses on executing the threat modeling process using the comparatively uncomplicated STRIDE method. The adoption of STRIDE was motivated by several factors. The methodical approach involves evaluating cyber threats against each component of the system using technical knowledge. Additionally, it assesses security attributes like availability, non-repudiation, confidentiality, and authentication, considering the potential impact of a weakness in a single component on the entire system. Furthermore, it can also be applied to cases where enhancing system security at the individual component level is needed. The STRIDE threat model encompasses six categories of attacks: denial of service

1 P. Yeng, S. Wolthusen, B. Yang, *Comparative Analysis of Threat Modeling Methods for Cloud computing Towards Healthcare Security Practice: A State-of-the-art study,* „International Journal of Advanced Computer Science and Applications", 11/2020, pp. 772-784.

2 A. Siddique, *Threat Modeling Methodologies for Network Security*, https://dx.doi.org/10.13140/RG.2.2.19672.42249, 2021 [accessed 1.08.2024].

(DoS), spoofing, tampering, repudiation, information disclosure, and elevation of privilege. The model provides useful insights for proactively detecting and protecting critical system infrastructure, devices, and networks that are susceptible to attacks. It is widely recognized as one of the most often employed ways to threat modeling. Threat categories defined by STRIDE are listed in Table 1.

1. Objectives and scope

The usefulness of large language models has been demonstrated across various domains. ChatGPT, an artificial intelligence model created by OpenAI, has undergone extensive training with vast quantities of data. This model emulates human conversation by effectively understanding contextual cues and providing responses that are contextually suitable. The system has attracted considerable attention owing to its capacity to provide proficient and complete responses to a wide array of human queries, surpassing previous public chatbots in terms of both security and use. The objective of the study was to evaluate whenever GPT models can properly detect threats using STRIDE methodology in network assets. For that purpose, security scan reports of machines hosting variable number of workloads were obtained. Reports were generated using Nmap, a scanning tool developed for gathering information from remote hosts and validating the services available remotely on target machines. One of the functions of Nmap is the operation system detection capability consisting of comparison of the IP/TCP stacks with Nmaps built-in fingerprint database. The same technique along with parsing of service banners is used to detect the services version on the remote system. Service version information was also considered in scope of this research, as the outdated version or available exploits change the threat landscape significantly.

Large Language Model (LLM) was tasked to perform threat identification based on given Nmap report. Output was then manually reviewed to find whenever identified threat is correct or it's a false positive.

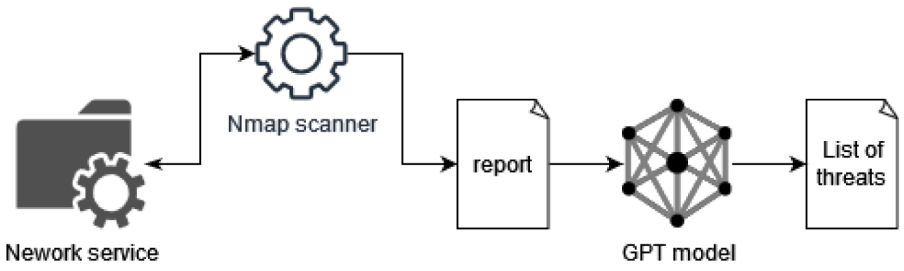

Figure 1. Process flow

Source: own work.

During research the usefulness of information returned by GPT was checked. In a separate task, comparison methodology was used to analyze the results of threat identification obtained from: human expert, STRIDE-GPT[3] and GPT-4 models. With regards to GPT-4 comparison was conducted for GPT models that were used to identify threats with basic information about the task, and separately, with a properly prepared pre-prompt to check if results differ significantly.

2. Similar studies

Adams' project called STRIDE-GPT follows a similar approach as the presented idea. Hence this tool is used in comparison section of the paper. STRIDE GPT is an AI-powered threat modeling tool that utilizes OpenAI's GPT models to generate customized threat models and attack trees for specific applications, following the STRIDE approach. Using STRIDE-GPT web interface users are able to provide information about any application, such as its type, methods of authentication, and if it is exposed to the internet or deals with sensitive data. The GPT model produces results that are pertinent to the given input. The primary distinction between STRIDE-GPT and our approach is in the necessity for human-provided text input that describes the system under consideration. Presented method uses Nmap reports as input, reducing human involvement to a minimum.

Ferrag[4] conducted research on the utilization of LLM (Large Language Models) for cyber threat identification. The study team evaluated two generating components, namely SecurityBERT and FalconLLM, to demonstrate the remarkable capabilities of LLM in the field of cybersecurity. The suggested solution consists of SecurityBERT, which functions as a tool for detecting cyber threats, and FalconLLM, which serves as a system for incident response and recovery.

Bahrini et al. conducted a comprehensive analysis[5] of the applications, possibilities, and dangers associated with ChatGPT across 10 major domains. They included in-depth examples of potential implementation in each domain. In addition, they have undertaken an empirical investigation to assess the efficacy and compare the capabilities of GPT-3.5 and GPT-4. Ferrag, Debbah, and Al-Hawawreh[6] conducted study on the use of generative AI in aiding threat hunting in 6G-enabled IoT networks. It

3 M. Adams, *STRIDE-GPT*, https://stridegpt.streamlit.app/, [accessed 31.07.2024].

4 M. Ferrag, et al., *Revolutionizing Cyber Threat Detection with Large Language Models*, Arxiv Forum Schedule, Cornel University, https://arxiv.org/abs/2306.14263, 2023 [accessed: 01.08.2024].

5 A. Bahrini et al., *ChatGPT: Applications, Opportunities, and Threats*, „2023 Systems and Information Engineering Design Symposium (SIEDS)" Proceedings, pp. 274-279, 2023.

6 M. Ferrag, M. Ndhlovu, M. Al-Hawawreh, *Generative AI for Cyber Threat-Hunting in 6G-enabled IoT Networks*, „IEEE/ACM CCGrid" 2023 Conference Proceedings, 2023.

is noteworthy that the authors acknowledge that generative AI systems have a tendency to produce false positives. This poses a challenge to the automated detection of dangers as artificial intelligence may generate alarms even in the absence of an actual threat.

2. Methodology

In this work the ability of pure ChatGPT, STRIDE-GPT and own design of pre-prompted ChatGPT technique for identification of threats, was compared. Pre-prompting technique was used to minimize the issue related to inability of GPT models to completely comprehend the context and meaning of the text. It was proven for example by Borji that ChatGPT cannot do well in tasks that involve common sense reasoning or logical reasoning that is not covered in the training data. For threat identification process critical thinking, decision making, and problem solving are all critical tasks. They rely significantly on reasoning, which means that lack of awareness and the ability to reason about the relationships between concepts, can be the cause of generating false information. With use of additional information in form of pre-prompt, we aim to make the GPT model more aware of context based on patterns provided. The implementation details of the pre-prompted solution are described in next chapter.

In this research we the analysis was performed on 10 randomly selected Nmap reports comparing the results obtained from AI solutions to answers based on human expertise. Each of the GPT engines were asked to identify threats (according to STRIDE methodology) based on the Nmap scan received as input. For comparison analysis following solutions were used:

- Pre-prompted GPT 4.0 (gpt-4-1106-preview)
- Chat GPT 4.0 (April 2023 update)
- STRIDE-GPT (version 0.5)

Human expert and all of the above-mentioned tools were queried with similar question: "Here is an Nmap report, based on it please identify the most probable threats in each of STRIDE categories. For each STRIDE category define one threat and describe it using one sentence. Also please define which of the threats you have identified should be prioritized to be addressed. {REPORT}"

The initial part of the exercise was to compare AI generated output with human provided information to check to what extent it overlaps. The second part of the exercise was to run a bigger set of reports through the pre-prompted GPT 4.0 (gpt-4-1106-preview) model and check the quality of the output by manual review of threats listed by the LLM.

Table 1. STRIDE categories

STRIDE threat type	Explanation
Spoofing	Spoofing refers to the deliberate act of mimicking a user or a system with malevolent intent. Timely and effective authentication of a user or system is crucial in preventing successful spoofing attacks.
Tampering	Tampering refers to the intentional alteration, creation, updating, or deletion of data with malevolent intent. Integrity is a security attribute that establishes the significance of information or data being precise.
Repudiation	Repudiation is the act of asserting or denying that an action is attributable to a specific user, individual, or system. Non-repudiation refers to the capacity to unequivocally demonstrate that a user, individual, or system has indeed carried out a specific action.
Information Disclosure	Information disclosure refers to the illegal access to confidential data. Confidentiality is a security attribute that establishes the significance of maintaining information or data in a confidential manner.
Denial of Service	Denial of Service refers to the deliberate act of overwhelming a system or service with fraudulent requests, causing it to be unable to properly or promptly respond to genuine requests. Availability is a security attribute that specifies the significance of being accessible either continuously or within a specific timeframe.
Elevation of privilege	Elevation of privilege refers to the exploitation of a that allows an attacker to gain higher access rights or permissions than intended.

Source: own work.

3. Implementation

Pre-prompted GPT 4.0 solution was implemented as a simple python script which followed the open AI guidelines for implementing Chatbot applications[7]. This differentiates the proposed solution from standard GPT-4.0, and this is the reason why results obtained from both sources differ.

Following OpenAI suggestion for performing initial steps in configuring of the chatbot to assign it an identity, a custom system role for the GPT model was set. This procedure results in modifying the chatbot's responses and makes them more closely reflect those of a specific individual who shares similar identity to the one set using system role[8]. Without this guidance, the chatbot may mimic the user or adopt a sarcastic tone, which would be inappropriate for providing relevant expert-level responses. To establish this identity the system role was adjusted as demonstrated in listing below:

7 OpenAI, *Code and guides for accomplishing common tasks with the OpenAI API*, https://github.com/openai/openai-cookbook/tree/main, [accessed: 28.12.2023].

8 H. Kumar,et al., *Exploring The Design of Prompts For Applying GPT-3 based Chatbots: A Mental Wellbeing Case Study on Mechanical Turk*, 10.48550/arXiv.2209.11344, 2022.

```
def ask_chatgpt(question,system):
    headers = {
        "Authorization": f"Bearer {API_KEY}",
        "Content-Type": "application/json",
    }

    payload = {
        "model": "gpt-4-1106-preview"
        "messages":[
    {
        "role": "system",
        "content": "Act as a cybersecurity expert. You know how to apply STRIDE
methodology in threat modeling and identification. Try to identify threats based on
the input given and on the knowledge an expert level security specialist would possess.
In case there is a vulnerability identifier involved (CVE) you can consult the cvedetails
database e.g. for CVE-2021-21112 get the details from https://www.cvedetails.com/
cve/CVE-2021-21112. Prioritize services for which Nmap reports multiple vulner-
abilities with high severity score."
},
…
```

Apart from assigning identity to the Chatbot pre-prompting strategy was used to mitigate the problem of GPT models' limited ability to fully grasp the context and semantics of the text. Empirical evidence, such as the study conducted by Borji[9], has demonstrated that ChatGPT's performance is limited in tasks requiring common sense reasoning or logical inference beyond the scope of its training data. Critical thinking, decision making, and problem solving are all essential tasks for the process of identifying threats. Their reliance on reasoning is substantial, implying that a lack of awareness and the capacity to reason about the connections between concepts might lead to the creation of incorrect information. The objective of incorporating supplementary data in the form of pre-prompts was to enhance the GPT model's contextual understanding by leveraging the presented patterns. In the pre-prompt given to GPT model, LLM was asked to:

- Identify the most probable threats in each of STRIDE categories.

9 A. Borji, *A Categorical Archive of ChatGPT Failures*, Arxiv Forum Schedule, Cornel. University, https://arxiv.org/abs/2302.03494, 2023 [accessed: 01.08.2024].

- Take into the scope only the services running on the open ports.
- Exclude from the output threats not related to the services identified.
- In case of unknown/custom port services, try to predict what service is running on the open port.
- Prioritize of services with many critical vulnerabilities (score > 8.0).
- Provide the output in JSON format using STRIDE categories and the priority element as keys and findings as values

After initial pre-prompt the LLM was asked to define one threat per each STRIDE category described with a single sentence and for picking one of the threats identified which should be prioritized to be addressed.

The above requirements were implemented in a custom python script which was able to read Nmap reports from files and launch the automatic collection of data using 200 reports.

The second part of the research in which Chat GPT 4.0 (April 2023 update), preprompted GPT 4.0 (gpt-4-1106-preview) and STRIDE-GPT were compared, followed a simple comparative analysis. This approach helped to qualitatively examine the similarities and differences in responses of the tools in regards to human provided answers. All 3 tools and a human expert were tasked to examine the contents of Nmap report and return most important threats (one per each of STRIDE categories) related to the system described in the report.

4. Results

Each of the 4 "experts" (3 chatbots and a human) were tasked to identify a single threat for each STRIDE category. In addition, request to give precedence to the most crucial one was made, allowing both human and AI to determine the priority weights by themselves. This investigation was conducted using a limited sample size of 10 reports that were picked randomly. Examples of matchings are described in Table 2.

In each of stride categories the overlap of AI generated answers with the human provided information was different. Best match was achieved for information disclosure, elevation of privilege and denial of service. Strikingly, Stride-GPT detected only 63.33% of the threats in a manner consistent with human expert. In comparison, GPT-4 achieved a similarity rate of 71.67%, while preprompted-GPT4 achieved an even higher rate of 76.67%. The apparent difficulty of STRIDE-GPT can be ascribed to its inherent design, which operates in a distinct manner. The program utilizes description type of human input to identify potential risks. This input for STRIDE-GPT should include a written description of the system design, as well as various settings that the user must select from drop-down lists. In this research its basic capabilities to identify threats on a simple prompt and a Nmap report were used.

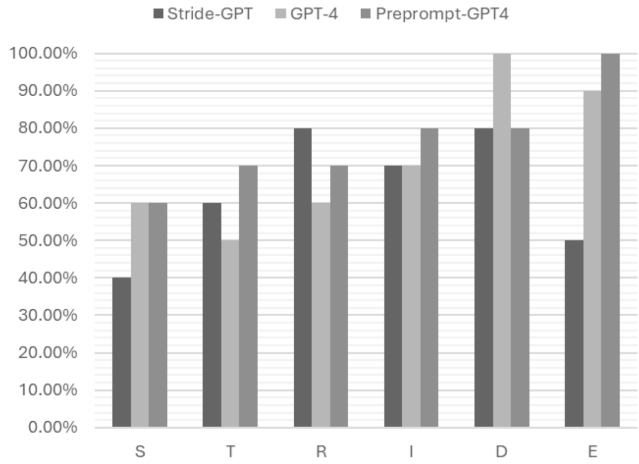

Figure 2. Comparison of AI based identification of threats to threats identified by human

Source: own work.

Table 2. Comparison of selected threats identified by different "experts"

ID	STRIDE category	Answer	Source	Human match
1	E	Attacker could be able to elevate privileges after gaining access via phpMyAdmin or configuration information.	Human	
		If any known vulnerabilities are present in the version of phpMyAdmin found on the server, an attacker could exploit these to gain elevated privileges.	Pre-prompted GPT	Yes
		Vulnerabilities in Moodle or the web server (Apache, nginx, etc.) could be exploited to gain unauthorized administrative access.	Chat GPT 4.0	No
		By leveraging a vulnerability in phpMyAdmin service, an attacker might be able to elevate privileges gaining more control over the system.	STRIDE-GPT	Yes
2	D	Attacker could potentially cause DoS using known exploits on SSH.	Human	
		The services running on OpenSSH may be susceptible to Denial of Service (DoS) if the CVEs related to resource exhaustion are exploited.	Pre-prompted GPT	Yes
		The SSH service, with known vulnerabilities, might be susceptible to DoS attacks, potentially disrupting service availability.	Chat GPT 4.0	Yes
		An attacker might exploit CVE-2019-16905 to cause a denial of service condition in the OpenSSH service.	STRIDE-GPT	Yes

3	T	Attacker could modify files transferred via ftp service.	Human	
		A threat actor could intercept and alter unencrypted FTP traffic, impacting the integrity of the data on Pure-FTPd.	Pre-prompted GPT	Yes
		Given the open HTTP and HTTPS ports with Apache HTTPD, there is a risk of web content or HTTP/HTTPS requests being tampered with by attackers.	Chat GPT 4.0	No
		Due to vulnerabilities in OpenSSH 7.4 reported by the scan, an attacker could manipulate ongoing sessions or data on the fly.	STRIDE-GPT	No

Source: own work.

Its worth noting that all 60 AI generated findings in this small sample were correct. Threats scenarios, even though not matched to the ones identified by human were applicable, and valuable from threat modeling perspective.

However after reviewing larger dataset of 200 Nmap reports processed by our pre-prompted solution we were able to identify 15 threats that were irrelevant or faulty. Examples are listed in Table 3.

Table 3. GPT wrongly identified threats

Identified threat	Reason for exclusion from valid findings list.
Unverified router configuration webpage could allow an attacker to pose as a legitimate admin and alter routing information.	Randomly generated text.
An attacker may impersonate the XMPP service running on ports 5222 and 5269 to deceive users or systems.	Technically possible but not applicable from threat modeling perspective.
An attacker could impersonate the FTP service by exploiting weak authentication mechanisms to gain unauthorized access to the system.	Exploiting authentication mechanisms would already grant access to the system.
The LDP (LDAP Directory Protocol) running on port 646 could be tampered with to modify files in directories.	LDAP stores information in *Entries* not directories. Most probably LLM mixed the context.
An attacker could spoof DNS responses to redirect traffic from the legitimate web servers (Apache / Microsoft HTTPAPI) to malicious sites.	Although technically correct this would be out of scope.
Malicious alteration of transmitted data in transit could occur on unencrypted connections.	In case of system in question only port 443 was open with TLS in place.
The MySQL service on port 3306 being 'unauthorized' gives an opportunity for attackers to attempt extracting sensitive database information if any security vulnerability exists.	The MySQL 'unauthorized' message in Nmap scan informs that MySQL required authorization when scanner tried accessing it.

FTP does not natively support strong logging mechanisms to prove the occurrence of a transaction.	Lack of native support for strong logging mechanisms in FTP seems to be a random information LLM generated.
Identification of a service that does not properly log actions, enabling malicious activities without traceability.	Randomly generated text.
FTP (port 21) does not provide sufficient logging, allowing malicious activities without traceability.	Assumptive. Not true for default configurations.

Source: own work.

Out of 200 reports processed by pre-prompted GPT-4-1106 model 19 contained findings which were 100% correct and applicable, but were to generic or assumed that configuration of other services is faulty. Most of them were related to STRIDE repudiation category and were following similar pattern:

- Without proper logging, an attacker could deny sending malicious emails via the vulnerable Exim SMTP server.
- Without proper logging mechanisms on the SSH service, it would be difficult to prove the occurrence of unauthorized access.

Among those 19 reports there were also findings with applicability that could be argued:

- An attacker could create a fake FTP or SMTP service to trick users into sending their credentials.
- An attacker could create a malicious website or send phishing emails imitating the legitimate IIS service to capture user credentials.

They are accurate, indeed both scenarios are possible and the second one mentioned, is a common phishing threat used by cybercriminals on daily basis. But it is doubtful that this kind of threat would appear on a priority list defined by human experts modeling the system. This shows that humans are susceptible to bias and accepting the status quo.

It is important to mention that GPT model was able to identify reports which were lacking data informing user that: "No detailed information on repudiation threats related to the services due to lack of logs or proper tracking mechanisms within the reported services".

Summary and Future Work

In this research paper the application of artificial intelligence in threat modeling, primarily focusing on data obtained from Nmap scans, was explored. The findings demonstrate that AI can effectively interpret and analyze Nmap data, providing insights into potential security threats and vulnerabilities within workloads. The incorporation of artificial intelligence (AI) into threat modeling brings forth certain consequences.

To begin with, artificial intelligence greatly improves the capacity to handle and evaluate extensive quantities of data, hence enabling more thorough threat assessments. The capacity to efficiently analyze complex datasets enables the detection of possible risks that human analysts would miss, hence enhancing the overall ability to respond promptly. Nevertheless, it also rises the questions regarding precision. The effectiveness of AI systems relies solely on the quality of the data they are trained on. If this data contains biases or is incomplete, it can result in the creation of flawed threat models and inaccurate threat assessments. Consequently, this can lead to insufficient or misguided security measures. Furthermore, it is important to acknowledge that incorporating AI into threat modeling brings about novel cybersecurity vulnerabilities, such as the possibility of AI systems being targeted by hostile individuals.

To achieve better results and improved precision in identifying threats, it could be useful to expand the experimental framework. By integrating data from supplementary sources, such as Nikto scanner results and outputs from diverse vulnerability scanners, a more comprehensive perspective on systems security could be obtained. These supplementary data sources can uncover a broader spectrum of vulnerabilities and potential attack paths that may not be identifiable only through Nmap scans empowering the AI systems to better identify and prioritize threats. With such wide range of data sources, the AI model could potentially enhance its ability to detect and evaluate threats, resulting in a more robust and trustworthy threat analysis.

Bibliography

Literature

1. Ferrag, M., Ndhlovu M., Al-Hawawreh M., *Generative AI for Cyber Threat-Hunting in 6G-enabled IoT Networks*, „IEEE/ACM CCGrid 2023 Conference Proceedings", 2023.
2. Yeng P., Wolthusen S., Yang, *Comparative Analysis of Threat Modeling Methods for Cloud computing Towards Healthcare Security Practice: A State-of-the-art study*, „International Journal of Advanced Computer Science and Applications", 11/2020.

Netography

1. Adams M., *STRIDE-GPT*, https://stridegpt.streamlit.app/.
2. Bahrini A., et al., *ChatGPT: Applications, Opportunities, and Threats*, „2023 Systems and Information Engineering Design Symposium (SIEDS)", Charlottesville, VA, USA, 2023.
3. Borji, A., *A Categorical Archive of ChatGPT Failures*, Arxiv Forum Schedule, Cornel University, https://arxiv.org/abs/2302.03494, 2023.
4. Ferrag M. et al., *Revolutionizing Cyber Threat Detection with Large Language Models*, Arxiv Forum Schedule, Cornel University, https://arxiv.org/abs/2306.14263, 2023.
5. Kumar H., et al., *Exploring The Design of Prompts For Applying GPT-3 based Chatbots: A Mental Wellbeing Case Study on Mechanical Turk*, Arxiv Forum Schedule, Cornel University, https://arxiv.org/abs/2209.11344, 2022.
6. OpenAI, *Code and guides for accomplishing common tasks with the OpenAI API*, https://github.com/openai/openai-cookbook/tree/main.
7. Siddique A., *Threat Modeling Methodologies for Network Security*, http://dx.doi.org/10.13140/RG.2.2.19672.42249, 2021.