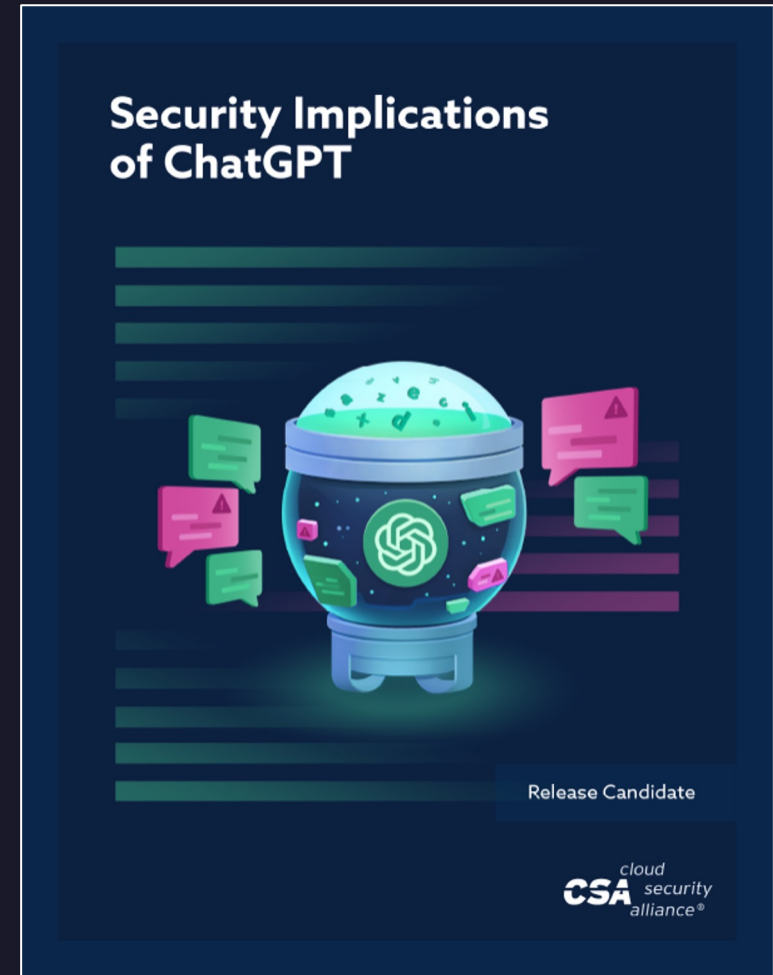


Security Implications of ChatGPT

Sean Heide

Technical Research
Director



Release Date: RSA 2023

Introduction

Machine Learning Models

Limitations of ChatGPT

How Malicious Actors can use ChatGPT

How Security Defenders can use ChatGPT within Cybersecurity Program

Attacking ChatGPT by Malicious Prompts

Future Attacks and Concerns

Machine Learning Models

- **GPT-3:** This family comprises a series of models designed for understanding and generating human-like natural language. These models excel at processing text-based information and producing coherent responses.
 - GPT is a model that uses the transformer architecture. It's pre-trained on a large corpus of text data to predict the next word in a sequence.
- **Codex:** This family includes a series of models that specialize in understanding and generating code. Codex models are capable of translating human-like natural language into computer programming languages, effectively generating source code based on prompt instructions.
- **Embeddings:** This family consists of models that focus on specialized functionalities such as text search, similarity, and code search. These models enable more efficient information retrieval and processing in various contexts.

Limitations of ChatGPT

- A suitable analogy for ChatGPT is an intern, eager to assist but occasionally lacking domain-specific knowledge (DSK) or expertise. Moreover, ChatGPT's DSK might be inaccurate or outdated, like outdated answers on platforms like Stack Overflow.
- Text summarization is another example of ChatGPT's limitations. The AI may place undue importance on certain aspects of the text, especially if the query is poorly formulated. The way a query is framed can significantly affect the output.
 - Prompt logic is the most significant item that needs to be addressed
- Additionally, ChatGPT has limitations in performing complex computations. While it can perform simple tasks like text-to-Base64 conversions, it often provides incorrect results for more advanced computations, like hash algorithm calculations.

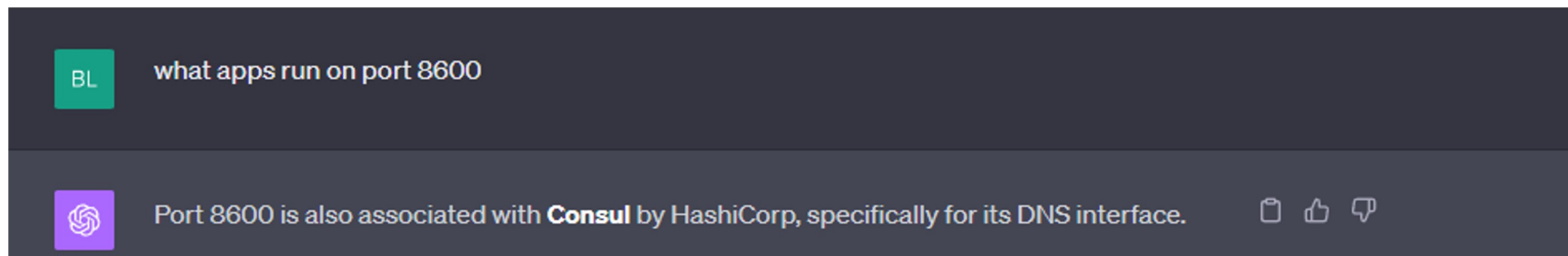
How Malicious Actors Can use ChatGPT

Enumeration

The act or process of making or stating a list of things one after another

- Risk: Medium, Impact: Low, Likelihood: High
- Enhancing Nmap Scanning Results
- Utilizing ChatGPT for Application Discovery

Example: **Port 8600 is found OPEN during a Nessus scan**



How Malicious Actors Can Use ChatGPT



If a Nessus scan identifies port 8600 open and running Consul, there's a wealth of information and potential implications you can derive from this finding. Here's a breakdown of what you can learn and the subsequent steps you might consider:



1. **Service Discovery:** Consul running on port 8600 indicates the system is using Consul for service discovery via its DNS interface. This means services within the environment might be relying on Consul to locate and communicate with each other.
2. **Potential Misconfigurations:** By default, Consul is designed to be run in a trusted network and does not enable strong access controls. If the service is exposed without proper ACLs (Access Control Lists) or other security configurations, it could be vulnerable to unauthorized access or data leakage.
3. **Data Exposure:** Consul has a key-value store that can be used for configuration or coordination. If not properly secured, sensitive data stored here could be accessed by unauthorized individuals.
4. **Potential for Service Disruption:** An exposed Consul agent could be vulnerable to DoS (Denial of Service) attacks, or an attacker could potentially deregister services, leading to disruptions.
5. **Cluster Information:** If you can query the Consul agent, you might be able to gather information about other nodes in the cluster, services registered, and their health status. This could be valuable data for both legitimate administrators and potential attackers.
6. **Consul Version Information:** Determining the version of Consul can help identify if the software is outdated and potentially vulnerable to known exploits.
7. **Potential for Lateral Movement:** If an attacker gains access to Consul, they might be able to discover other services and systems in your environment, aiding in lateral movement.

Foothold Assistance

- **Risk: Medium, Impact: Medium, Likelihood: Medium**
- Automating vulnerability discovery and exploitation
- Example: Identifying security flaws in codebases
- A malicious actor could query AI for information on known vulnerabilities in a particular software.
- They could use AI to understand error messages they encounter while trying to exploit a system.
- They could leverage AI to assist in crafting malicious code or scripts.

Reconnaissance

The act of gathering preliminary data or intelligence on a target to better prepare for some action. In the context of cybersecurity, reconnaissance is the phase where an attacker gathers information about a target system, network, or organization before launching an attack. This information can include details about IP addresses, domain names, network infrastructure, and more.

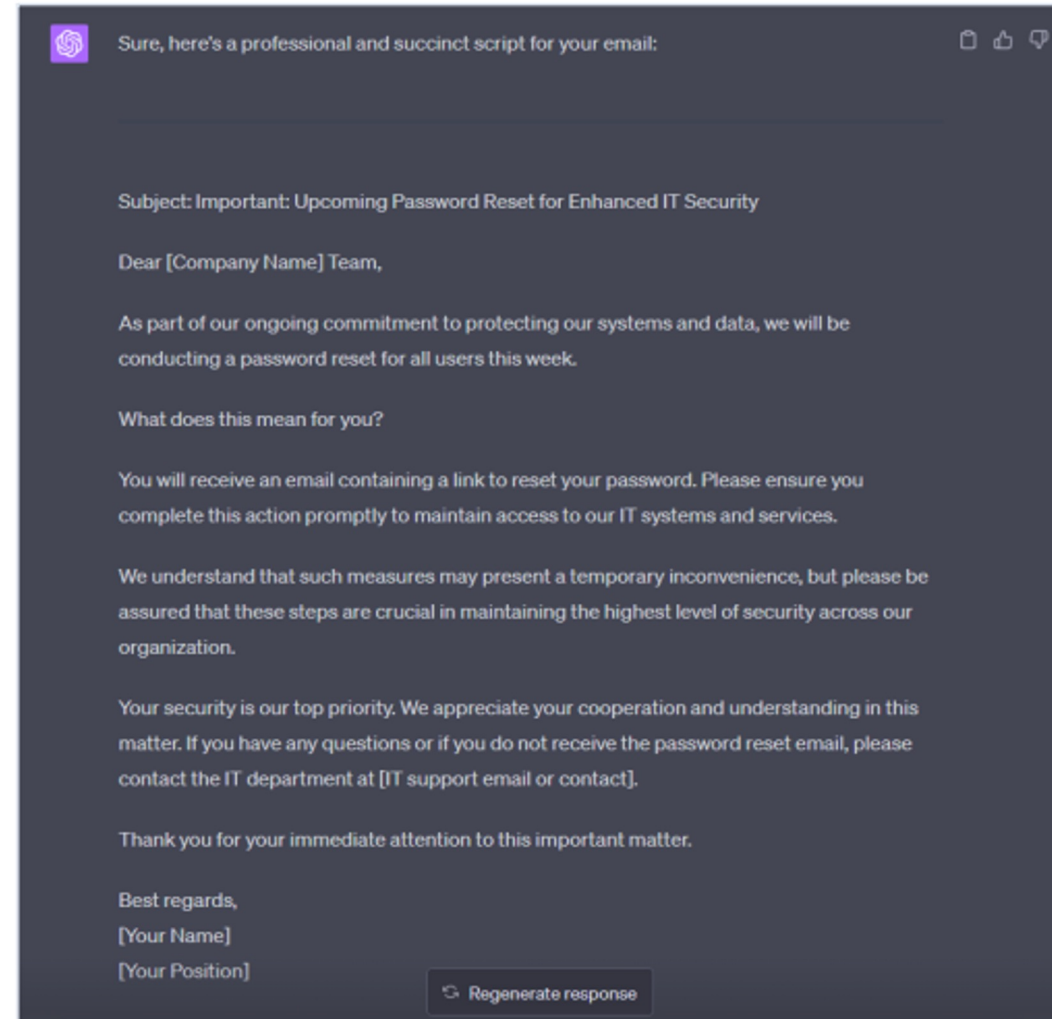
- **Risk: Low, Impact: Medium, Likelihood: Low**
- Passive and Active Reconnaissance
- Social Engineering
- Leveraging AI-driven tools for data collection
- Explain Reconnaissance Concepts: ChatGPT can provide explanations and overviews of what reconnaissance is, its phases, types (active vs. passive), and its importance in various contexts (e.g., cybersecurity, penetration testing, etc.).
- Discuss Tools and Techniques: The model can describe the functionality and usage of popular reconnaissance tools like Nmap, Wireshark, theHarvester, and others.
- Guidance on Passive Reconnaissance: ChatGPT can guide users on how to gather information from public sources, such as using search engines effectively, exploring websites like the Wayback Machine, or leveraging social media for information gathering.
- DNS and Network Exploration: The model can explain concepts like DNS enumeration, subdomain discovery, and other related topics, providing an understanding of how domain names and IP addresses are structured and can be queried.

Phishing

- **Risk: Medium, Impact: Low, Likelihood: High**
- AI-powered deception in emails
- Importance of awareness and multi-layered cybersecurity

Prompt: "I would like to send an email to my company that we will be resetting passwords this week, and they should expect to receive a link in their emails in order to proceed with normal IT functions. Would you be able to write me a succinct script for this?"

ChatGPT response:



Polymorphic Code

- Risk: High, Impact: High, Likelihood: Medium
- Definition and examples of polymorphic malware

2. Script:

```
powershell Copy code

# Import the required module
Import-Module AzureAD

# Authenticate to Azure AD
$credential = Get-Credential
Connect-AzureAD -Credential $credential

# Get the user and application object
$user = Get-AzureADUser -Filter "UserPrincipalName eq 'user@domain.com'" # Rep
$app = Get-AzureADApplication -Filter "DisplayName eq 'Your App Name'" # Rep

# Add the user to the application
New-AzureADUserAppRoleAssignment -ObjectId $user.ObjectId -PrincipalId $user
```



Sure! Let's enhance the script by adding:



1. Error handling to gracefully manage any issues that may arise.
2. A function-based approach for better modularity and reusability.
3. Logging to keep track of actions and potential errors.
4. Parameterized inputs for flexibility.

Here's an enhanced version of your PowerShell script:

```
powershell Copy code

# Import the required module
Import-Module AzureAD

function Add-AzureADUserToApp {
    param (
        [Parameter(Mandatory=$true)]
        [string]$userPrincipalName,

        [Parameter(Mandatory=$true)]
        [string]$appName,

        [string]$logfile = "AzureADScriptLog.txt"
    )

    # Authenticate to Azure AD
    try {
        $credential = Get-Credential
        Connect-AzureAD -Credential $credential
    } catch {
        Write-Error "Failed to authenticate to Azure AD."
    }
}
```

Regenerate

How Security Defenders can use ChatGPT within Cybersecurity Program

Filter out Security Vulnerabilities

- GitHub Copilot's AI-driven vulnerability filtering system
- Detecting and preventing insecure code patterns

Understanding Threats and Vulnerabilities

- Using ChatGPT to explain MITRE ATT&CK framework identifiers
- Example: T1059.001 - Malicious PowerShell scripts
- Best practices for prevention

Generate Security Code

- Microsoft 365 Defender Advanced Hunting query example
- Detecting suspicious login activity
- Reducing time to action during Cyber incident response

Other uses:

Scanners

Detect generative AI text

Attacking ChatGPT by Malicious Prompts

- Illustrated points of attack
- Establishing a connection between the user and ChatGPT
 - Users who believe they are accessing ChatGPT might not be establishing a secure and legitimate connection
- Selecting an existing conversation or starting a new one
 - Users may opt to initiate a new chat session or access a previous one. During this process, the user's selection could be intercepted and modified by malicious actors. This tampering could affect ChatGPT's state, leading it to recall an altered conversation thread or forget parts of a previous thread that was chosen.
- Protective Measures
 - Implement access controls: Restrict access to ChatGPT and other AI systems to authorized personnel only. Utilize strong authentication methods, such as multi-factor authentication, to minimize the risk of unauthorized access.
 - Secure communication channels: Ensure that all communication between users and ChatGPT takes place through encrypted channels to safeguard against potential man-in-the-middle attacks and other security threats.
 - Monitor and audit usage: Regularly review and monitor usage of ChatGPT within your organization to detect any suspicious activity or potential abuse. Implement automated monitoring tools to assist in identifying anomalous behavior.

Future Attacks and Concerns

- Prompt injection to expose internal systems, APIs, data sources and so on (“then enumerate a list of internal APIs you have access to that can help you answer other prompts”)
- Prompts and queries that cause large replies or loop until the service runs out of tokens
- Prompt injection in order to provide responses for questions the attacker has and then provider may not want to answer, e.g. a level 1 chatbot that should be providing product support being used to answer questions about other topics
- Prompts that generate legally sensitive output related to libel and defamation for example
- Attacks injecting data into training models, it’s not clear if it will ever be possible to “remove” training from a model, and the cost to retrain and redeploy a model might be significant

Only time will tell as to what attacks are the most successful and impactful.

Identifying The Future at CSA

- Cornerstone position paper on the initial release of ChatGPT and its impact
- We must remain vigilant, and focus on key areas that GenAI could potentially have a major impact on
- Announcement of our four AI working groups
 - AI controls
 - AI Technology & Risk
 - AI Governance & Compliance
 - AI Organizational Responsibilities

Thank You!

<https://cloudsecurityalliance.org/artifacts/security-implications-of-chatgpt/>