cloud security alliance®

Security Implications of ChatGPT

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.
- 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 Moreover, ChatGPT's DSK might be inaccurate or outdated, like outdated answers on platforms like Stack Overflow.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.

- 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

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

Foothold Assistance

- Risk: Medium, Impact: Medium, Likelihood: Medium
- Automating vulnerability discovery and exploitation
- Example: Identifying security flaws in codebases

Reconnaissance

- Risk: Low, Impact: Medium, Likelihood: Low
- Passive and Active Reconnaissance
- Social Engineering
- Leveraging AI-driven tools for data collection

Phishing

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

Polymorphic Code

- Risk: High, Impact: High, Likelihood: Medium
- Definition and examples of polymorphic malware
- Generating polymorphic shellcode using ChatGPT

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.