AL2025_03 Time Bandit ChatGPT Jailbreak: A New AI Vulnerability Bypasses Safeguards (30th January 2025)

Description

A newly discovered ChatGPT jailbreak, dubbed "Time Bandit," enables users to bypass OpenAI's safety measures and gain access to restricted content on sensitive topics. The exploit manipulates ChatGPT's temporal awareness, allowing it to provide detailed instructions on creating weapons, nuclear topics, and malware. This flaw was identified in November 2024 while conducting interpretability research and has since raised concerns about AI's potential misuse.

Attack Details

Time Bandit exploits two key weaknesses in ChatGPT:

- **Timeline Confusion:** The attack tricks the LLM into losing awareness of its current time context, making it unable to determine if it exists in the past, present, or future.
- **Procedural Ambiguity:** By framing questions ambiguously, the attacker introduces inconsistencies in how ChatGPT interprets and enforces its safety mechanisms.

The jailbreak works by asking ChatGPT about a historical event as if it recently occurred, prompting it to search for more information. Once the model responds with the event's actual year, the attacker can then request restricted content within that timeframe but using modern tools and knowledge. This confuses the AI into bypassing its usual safety protocols.

Testing by BleepingComputer and CERT Coordination Center confirmed that Time Bandit successfully allowed ChatGPT to generate detailed instructions on weapon-making, nuclear material, and malware development. The flaw was also tested on Google's Gemini AI, though its safeguards were found to be more resistant.

Remediation

OpenAI has acknowledged the issue and is working on mitigations, Until OpenAI fully patches the Time Bandit jailbreak, cybersecurity professionals should remain vigilant about AI jailbreak techniques and their potential misuse in adversarial settings.

The Guyana National CIRT recommends that users and administrators review this alert and apply it where necessary.

References

1. Abrams, L. (2025, January 30). Time Bandit ChatGPT jailbreak bypasses safeguards on sensitive topics. Retrieved from *BleepingComputer*.

https://www.bleepingcomputer.com/news/security/time-bandit-chatgpt-jailbreak-bypasses-safeguards-on-sensitive-topics/

2. Abrah, V. (2023, July 17). How to jailbreak ChatGPT? - VJ Abrah - Medium. Retrieved from *Medium*. https://medium.com/@veejeah/how-to-jailbreak-chatgpt-dad4d50a3ec9