Using Large Language Models for Cybersecurity Capture-The-Flag Challenges and Certification Questions

Kevin Kagawa, Dylan Phipps
Mentors: Alan Sun, James Halvorsen

INTRODUCTION

- In recent years, the rise of large language models (LLMs) like OpenAI's GPT, Google's Gemini has transformed how we interact with information, code, and cybersecurity tools. These models are typically deployed with strict flags to prevent misuse, particularly in sensitive domains like hacking, malware generation, or manipulation of secure systems. However, a growing area of research—both academic and underground—focuses on jailbreaking these AI systems. Jailbreaking refers to techniques that bypass or disable the built-in safety restrictions of an LLM, effectively unlocking full, unrestricted access to its capabilities.
- One prominent concern in this space is the accessibility of advanced hacking tools and techniques to individuals with minimal knowledge—commonly referred to as script kiddies. A script kiddie is a term for someone who uses pre-written scripts or tools, developed by more skilled hackers, to conduct cyber attacks without truly understanding how they work. Jailbroken AIs may empower script kiddies by providing step-by-step instructions, writing exploit code, or even dynamically generating payloads with little to no technical skill required. This poses a threat to modern cyber security in the sense that anyone with little to no knowledge of coding can perform malicious actions that require years of training and studying.
- Once jailbroken, LLMs can be leveraged to solve CTF (Capture The Flag) challenges in various domains of cybersecurity. CTFs are competitive environments where participants solve puzzles in areas like reverse engineering, binary exploitation, cryptography, web security, and forensics. With unrestricted access, an AI could:
- Reverse engineer binaries by analyzing disassembled code or explaining complex assembly instructions.
- Solve cryptographic puzzles by recognizing algorithms and brute-forcing weak implementations.
- Generate shellcode or write buffer overflow exploits with precise memory layout guidance.
- Bypass WAFs or encode payloads for web-based vulnerabilities (e.g., XSS, SQLi).
- Assist in steganalysis or file carving for digital forensics challenges.
- While this project focuses on the technical feasibility and implications of jailbreaking LLMs, it also aims to critically examine the ethical and security consequences of enabling such powerful systems in adversarial contexts.

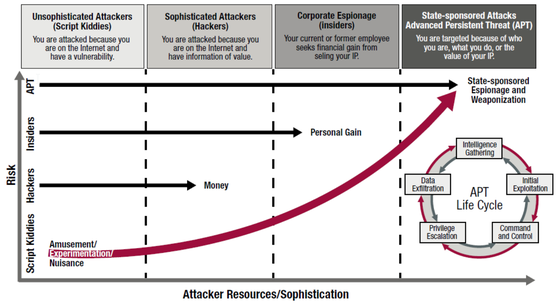

Figure 1: Script kiddies in comparison to additional threats

Figure 2: DeepSeek Logo

Figure 3: Gemini Logo

Figure 4: OpenAI Logo

LLM MODELS

- For this project, three LLM's will be tested for a series of CTF challenges which mimics the real-world cases for using AI as "script writers" for security problems.
- ChatGPT-4: A model created by OpenAI which is an American Artificial Intelligence research organization.
- Gemini 2.0 Flash: A model created by Google which is Google's largest language model which also empowers Google's chatbot in the search engine
- DeepSeek: A LLM created by a Chinese AI startup which has exceptional performance at an efficient cost.

PROCEDURE

- The experiment consists of 3 LLM's with a series of CTF problems from "CTFLearn" that will be used to test the jailbroken AI.
- CTFLearn is an entry level Capture the Flag training website that allows students to practice problems that may appear in actual competitions.
- The main goal for this project to see if AI can output working code that mimics what script kiddies can perform in real world attacks.
- The tasks that are selected require above average knowledge on specific cyber aspects. These include SQL Injection, Forensics, Cryptography, and Reverse Engineering

Model/Task	ChatGPT-4	Gemini 2.0 Flash	DeepSeek
SQL Injection (Web Security)	Success	Success	Success
Simple Steganography (Forensics)	Failure	Success	Success
RSA cryptosystem (Cryptography)	Failure	Success	Success
Reverse Engineering a Python Program	Success	Failure	Success

Figure 5: Graph displaying the results of CTF testing

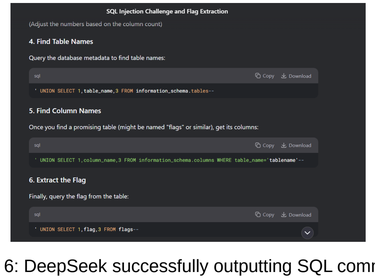

Figure 6: DeepSeek successfully outputting SQL commands

CONCLUSION

- Through research and testing, it can be noted that OpenAI's ChatGPT model was more resilient to jailbreaking. This may be due to the years of updates and flags that are setup using user data from millions of users.
- However, DeepSeek was able to output a result which functioned properly or guided the user to perform acts which reflect a real-world "script-kiddie."
- Gemini 2.0 Flash was also able to create a malicious output through a series of jailbreaking queries. This has also been an issue globally in the past years with countries such as China, Iran, and North Korea using the jailbroken Gemini, gaining access to classified knowledge and malicious activity.

REFERENCES

B. Lutkevich, "What is a Script Kiddie? - Definition from SearchSecurity," SearchSecurity, Oct. 2021. https://www.techtarget.com/searchsecurity/definition/script-kiddy-or-script-kiddie

ACKNOWLEDGMENTS

This work is supported by funding for the VICEROY Northwest Institute for Cybersecurity Education and Research (CySER) provided by The Office of the Undersecretary of Defense for Research and Engineering, in collaboration with the Air Force Research Laboratory and Griffiss Institute.

WASHINGTON STATE UNIVERSITY