Improving GPT Penetration Testing Using Prompt Engineering Techniques and Newly Released Features

Daniel Lichtenberger and Mengxiang Jiang

Penetration Testing

- Hacking your own server
- Can be done by your own organization or outsourced
- Time intensive
- Need to be regularly done since new vulnerabilities are discovered frequently

ChatGPT

- ChatGPT - OpenAI chatbot service
- Uses large language models (LLM) that generates automated responses based on user input
- Newest model is GPT-4, accessible to premium users for a monthly fee of $20 and rate limited to 40 inputs per 3 hours
- Regularly updated with new features
  - Image Input
  - Web Browsing and Search

Prompt Engineering

- Prompt is the input to a LLM
- Prompt engineering is modifying the prompt to get a desired output
- Can be specially crafted (usually long, elaborate, and complex)
- Or be more general (usually shorter and simpler)

Related Works

PentestGPT is a framework that uses specially designed prompt engineering on ChatGPT for penetration testing

Performs in the top 1% of users on the HackTheBox penetration testing website

Requires installation and the use of OpenAI API to access ChatGPT and therefore harder to access newly released features

Methodology

- Our local machines are virtual machines (VMs) running Parrot OS (a Linux distribution based off of Debian) as a base to conduct penetration tests.
- We subscribe to the premium membership in order to access GPT-4 and newly released features
- We run the penetration test on the relatively new HackTheBox server Codify (Released November 4, 2023)
- The prompt engineering techniques we used were:
  - Flipped Interaction Pattern - drive the LLM to ask the user for information.
  - Persona Pattern - Having the LLM act as a specific role (in our case a cybersecurity expert). PentestGPT also uses this prompt engineering technique.
- We also tested newly released ChatGPT features
  - Image Input
  - Web browsing and search

Prompt Engineering Results

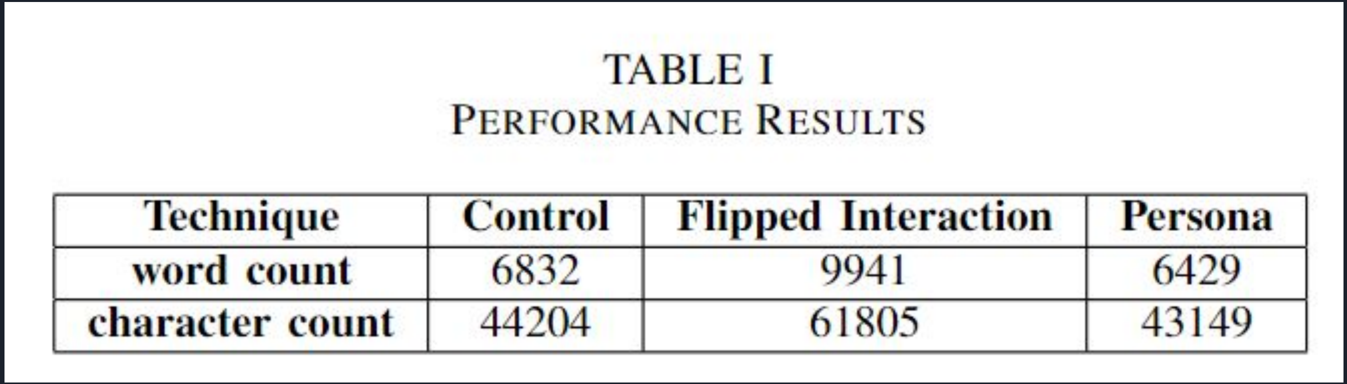

TABLE I
PERFORMANCE RESULTS

Technique	Control	Flipped Interaction	Persona
word count	6832	9941	6429
character count	44204	61805	43149

Newly Released Feature Results

Image Input: no major difference in pentest information over source code

Still has some advantages

- Image input is more readable and easier to use as well
- Less consumption of input limit due to allowing multiple images in one input

Web browsing and search: much better qualitative experience for penetration testers

- Automatic compilation of vulnerabilities and unsafe practices
- Evaluation of vulnerability as practical or impractical to exploit

Conclusion

The general prompt engineering techniques were not very effective

To effectively use prompt engineering requires the highly elaborate and complex prompts similar to PentestGPT

The newly released features of ChatGPT were more effective at improving performance

Socioeconomic Impacts

- Hoping to drive the way for more open-source AI.
  - Future of AI seems to be more commercial than to allow other communities to contribute to AI.
  - Recent News: OpenAI controversy
  - Open-source AI could contribute to potential innovations in cybersecurity previously never thought of before.
- Creating a new way for penetration testers to view LLMs as a potential tool to help organizations.
- Using AI in cybersecurity could drive ways to innovate AI into different sectors of our economy and even cybersecurity itself.
- Driving to improve current or future LLMs with much information they can contain about cybersecurity from studies using prompt engineering techniques.

References

Gelei Deng, Yi Liu, Victor Mayoral-Vilches, Peng Liu, Yuekang Li, Yuan Xu, Tianwei Zhang, Yang Liu, Martin Pinzger, and Stefan Rass. Pentestgpt: An llm-empowered automatic penetration testing tool. arXiv preprint arXiv:2308.06782, 2023.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt. arXiv preprint arXiv:2302.11382, 2023.

Chatgpt — release notes.
https://help.openai.com/en/articles/6825453-chatgpt-release-notes.

Hackthebox: hacking training for the best. http://www.hackthebox.com/.