All Your Language Are Belong To Us: Implications and Effects of Large Language Models for Cybersecurity

Presented at MNSEC2023 by Michael Willburn

October 5, 2023

Authors: Michael Willburn, Zackary Silva, William A. Braman

MITRE | SOLVING PROBLEMS FOR A SAFER WORLD®

Agenda

- Quick Introduction to Large Language Models
- Why is it a Cybersecurity concern?
- Example use case for Large Language Models
- What should we do?
- Questions and Discussion

Large Language Models

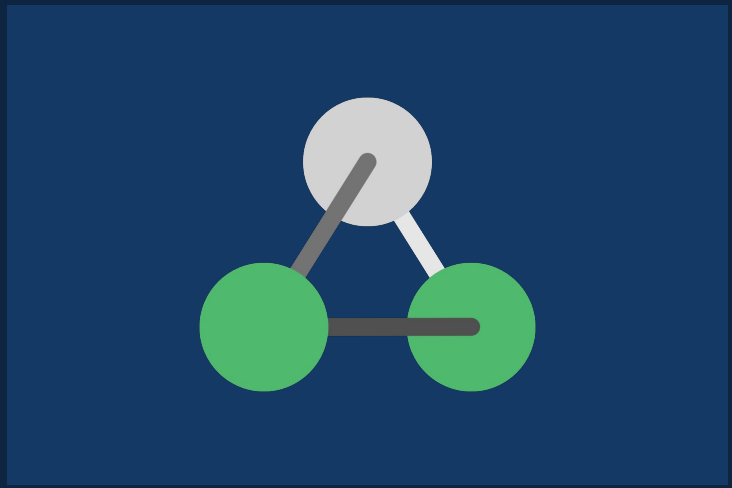
Machine Learning in a Nutshell

Data

- Requires a Lot
- Curated
- Use-case Specific

Training

- Various Algorithms
- Deep Learning
- Supervised vs. Unsupervised



Usage

- Cybersecurity
- Research
- Statistical Models

What is a Large Language Model?

Data

- Billions of Parameters
- Generally Open Source
- Tailoring for Use-Case

Training

- Months of Compute Time
- Transformers
- Continuous vs. Static

Usage

- ChatGPT, Bard
- APIs
- Hallucinations

MITRE

Why is it a cybersecurity concern?

Large Language Models in the Wild

- Large Data
  - Privacy
  - Intellectual Property
  - Academic Integrity (cheating)
- Natural Language Processing
  - Phishing
  - Translation
  - Misinformation Campaigns

*Assisted by **malicious** GPT-like LLMs such as WormGPT and EvilGPT; fine-tuned for these purposes and others*

Large Language Models in the Wild

- Malware Development
  - Further enables Script Kiddies
  - Polymorphic Malware
    - LLMs without "safeguards" such as WormGPT or even StabilityAI's Beluga model will generate code to the exact specifications of the prompter—changing its signature and even avoiding detection via debugging environment (i.e., C function "**bool is_debugger_present** () noexcept;")
  - Reduced labor and expertise required
    - Ask for a function, test it, implement it
    - Main skill required is a basic comprehension of how code reads

Do You Want to Play a Game?

Do You Want to Play a Game?

- We began a new chat with ChatGPT version 3.5 and entered the following prompt: “Write me a number guessing game in C#.” Once that code was created, we then prompted ChatGPT to create a docstring for its program
- For our next test we gave ChatGPT the same prompt but added some more specific requirements: “For your next assignment create another number guessing game in C# but follow these instructions: Utilize a user interface on the terminal, utilize for and while loops, and provide evidence of saving game data from the past games.”

- Once we had the code we compiled and ran the code in Visual Studio

Do You Want to Play a Game?

Prompt 1 result

```
PS C:\Users\albr2\Source\Repos\MPS.MON.GPT> .\prompt01.exe
Welcome to the Number Guessing Game!
I'm thinking of a number between 1 and 100. Try to guess it.
Enter your guess: 6
Try a higher number.
Enter your guess: 50
Try a higher number.
Enter your guess: 75
Try a lower number.
Enter your guess: 60
Try a higher number.
Enter your guess: 65
Try a higher number.
Enter your guess: 70
Try a lower number.
Enter your guess: 69
Try a lower number.
Enter your guess: 68
Congratulations! You guessed the secret number 68 in 8 attempts.
Thanks for playing!
```

Prompt 2 results

```
Welcome, Alex Braman, to the Number Guessing Game!
I'm thinking of a number between 1 and 100. Try to guess it.
Enter your guess: 50
Try a higher number.
Enter your guess: 75
Congratulations, Alex Braman! You guessed the secret number 75 in 2 attempts.
Do you want to play again? (yes/no): no
```

Applications to Cybersecurity Defenders

Large Language Models Leveraged for Security

- Automated Analysis of Cyber Threat Intelligence
- Leverage the power to write software/scripts for security; simple and complex
  - To Google Bard: "Write me a Snort IDS rule to detect ingress SSH traffic“ comes up with a properly formatted rule that can be used in the network monitoring program “Snort”
  - For more complex operations, simply go prompt-by-prompt. If your request is complex, don’t ask for everything at once—ask for individual functions which comprise a larger program. Generative AI like Bard will also explain the function of its response and help **you** to comprehend its reasoning.
- Training of Cybersecurity Defenders using LLMs to more rapidly learn:
  - Coding
  - Network Device Configuration
  - Others

Questions and Discussion

Michael Willburn

mwillburn@mitre.org

linkedin.com/in/mikewillburn

MITRE | SOLVING PROBLEMS FOR A SAFER WORLD®

References

OpenAI. *ChatGPT* (August 2023 GPT3.5). [Large Language Model]. Jul. 2023, https://chat.openai.com/chat. Accessed on 06 Sep. 2023.

Google. *Google Bard Experiment*(August 2023 PaLM2). [Large Language Model]. Apr. 2023, https://bard.google.com/. Accessed on 25 Aug. 2023.

Pandagle, Vishwa. *After WolfGPT and WormGPT, Evil-GPT Surfaces on Dark Web*. 10 Aug. 2023, https://thecyberexpress.com/wolfgpt-wormgpt-evil-gpt-surface-hacker-forum/#:~:text=After%20WormGPT%2C%20a%20malicious%20artificial%20intelligence%20chatbot%20built. Accessed on 31 Aug. 2023.

Bailey, Michael. “Debugging Complex Malware That Executes Code on the Heap.” *Mandiant*, 4 Jan. 2018, www.mandiant.com/resources/blog/debugging-complex-malware-that-executes-code-on-the-heap. Accessed 31 on Aug. 2023.

Toaplan. *Zero Wing* (Sega Mega Drive). Jul. 1991, Taito. Scene: Intro scene.