Copilots Need Helmets Too

GitHub Copilot Security Case Study

Agenda

- Attacker’s view
- Defender’s view
  - OWASP Top 10 for LLMs
  - General Advice

? >) 1 [[

@ ! 0

Miloslav Homer

Senior Application Security Engineer

Kiwi.com Tech Ambassador (Security)

Miloslav Homer @LinkedIn

miloslavhomer.cz

0 !! 1

<< % :: 1 @ 0 ? >

Internal Github Copilot Research

https://code.kiwi.com/articles/cautiously-configuring-copilot/

? >) 1 [[

@ ! 0

Attacker’s View

0 !! 1

<< % :: 1 @ 0 ? >

Application Model

- “Thick clients”
  - Plugins (browser, IDE)
  - Mobile apps
  - Front-end heavy websites
- Few expensive API calls
- Free-form text

Attacker's View

- Static Analysis – Reverse Engineering
- Dynamic Analysis – Proxy
- DEMO

? >) 1 [[

@ ! 0

Defender’s View

0 !! 1

<< % :: 1 @ 0 ? >

OWASP Top 10 For LLMs

- LLM01: Prompt Injections
- LLM02: Insecure Output Handling
- LLM03: Training Data Poisoning
- LLM04: Denial of Service
- LLM05: Supply Chain
- LLM06: Permission Issues
- LLM07: Data Leakage
- LLM08: Excessive Agency
- LLM09: Overreliance
- LLM10: Insecure Plugins

Securing Your Own Model (LLM03, LLM05)

- LLM03: Beware of model poisoning
  - Don’t train your models on user supplied data
- Protect your model
  - LLM05: Supply chain issues
  - Treat your model weights as trade secrets

Prompt Injections (LLM01)

- Manipulate the prompt to change the instructions
- Try it yourself:
  - https://gandalf.lakera.ai/
  - https://gpa.43z.one/

Denial of Service/Wallet (LLM04)

- Calls to LLM models are expensive
- Rate-limit per user
- Consider rate-limiting globally
- Do not give your users the API keys
    - Not even in mobile apps

Standard Threats (LLM02, LLM06, LLM10)

- LLM02: Insecure Output Handling
- LLM06: Permission Issues
- LLM10: Insecure Plugins
- It’s still software
  - OWASP Web Security Testing Guide
  - OWASP Mobile Security Testing Guide

End-User Threats (LLM07, LLM08, LLM09)

- LLM07: Sensitive Data
  - Terms and conditions -> Legal Stuff
  - Don’t forget telemetry!
- LLM08: Agency - Actions made by LLMs
- LLM09: Reliance - Trusting the output of LLMs

Thank you / Questions / Contact me

? >) 1 [[
@ ! 0

LinkedIn

miloslavhomer.cz

0 !! 1
<< % :: 1 @ 0 ? >