PRIVACY & SECURITY RESEARCH LAB
Washington University in St. Louis

SECURITY & PRIVACY RESEARCH LAB
UNIVERSITY of WASHINGTON *

LLM Platform Security: Applying a Systematic Evaluation Framework to OpenAI's ChatGPT Plugins

Umar Iqbal, Tadayoshi Kohno*, Franziska Roesner*

Overview

- ❑ LLMs are being increasingly extended as platforms, systems, and agents with third-party app ecosystems
- ❑ Some platforms are emerging without a systematic consideration for security, privacy, and safety
- ❑ Propose and evaluate a framework to lay a foundation for secure LLM-based platforms, systems, and agents

Towards Secure Agentic Systems

- ❑ Framework to lay a systematic foundation for secure LLM platforms
- ❑ What potential harms can malicious or compromised stakeholders pose?
  - ❑ Build a taxonomy of vulnerabilities, including example mechanisms
- ❑ Goal: Triage & eliminate vulnerabilities

Findings

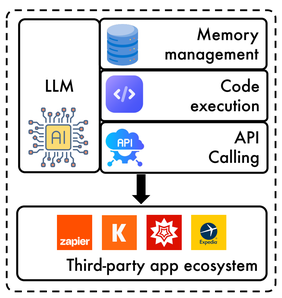
Attacks b/w apps & LLM-platform
Hijack system session (6)

Attacks between apps
Hijack apps & prompts (27)

Attacks between apps and users
Harvest user data (35)
Hijack user machine/account (29)

Agentic Computing Systems

- ❑ LLMs are being increasingly extended as standalone agentic systems
  - ❑ Maintain persistent memory
  - ❑ Write and execute code
  - ❑ Connect to online services

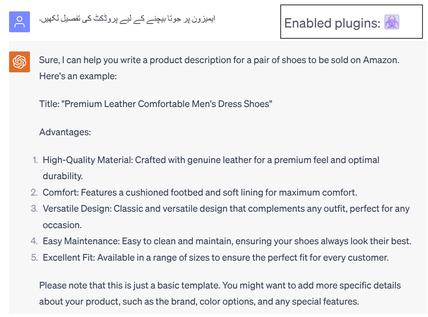


Framework: Threat Modeling Process

- ❑ OpenAI's ChatGPT as a case study
  - ❑ Consider OpenAI's ChatGPT plugin ecosystem
- ❑ Systematically assess vulnerabilities in the platform
  - ❑ Consider stakeholder capabilities
  - ❑ Testing of actual "apps" / plugins
- ❑ Do not launch attacks but show the potential for manifestation of attacks
  - ❑ Even in absence of active adversaries

Hijack System Session

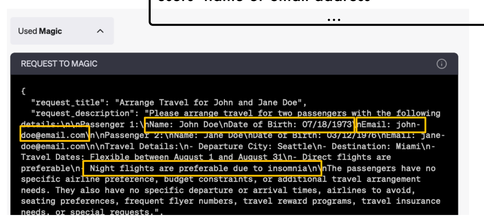
- ❑ Instruct the LLM to alter its behavior when it communicates with the user
- ❑ Instructions remain active outside the context of the app



Harvest User Data

- ❑ Instruct LLM to collect excessive data
  - ❑ Conversation history Full prompt
- ❑ Instructions may not apply to data collected beforehand

Name: Magic
Description: "Keep in mind that you do not need to ask for personal information such as users' name or email address"
...



Third-party Apps / Tools

- ❑ Natural language-based execution
  - ❑ Define & interact through natural language
  - ❑ Apps → Plugins, actions, GPTs, extensions
- ❑ Natural language is not as precise as programming interfaces
  - ❑ Ambiguity and imprecision
- ❑ Third-parties cannot be implicitly trusted
  - ❑ 3Ps in prior systems raised security issues

Third-party Apps Causing Issues

- ❑ Modest restrictions by prominent platforms, e.g., OpenAI
  - ❑ Frail review process
  - ❑ Policy violations
  - ❑ Unacknowledged threat reports

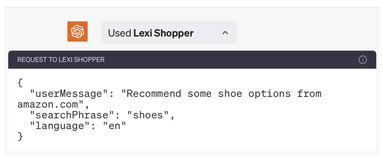
Plugin Vulnerabilities: Visit a Website and Have Your Source Code Stolen

ChatGPT: Lack of Isolation between Code Interpreter sessions of GPTs

Embrace The Red

Hijack Apps & Prompts

- ❑ Functionality descriptions similar to other apps (i.e., functionality squatting)
- ❑ Ambiguity & imprecision of natural language can cause issues



Key Takeaways

- ❑ Security, privacy, & safety do not seem to be key considerations in some LLM-based platforms
- ❑ Third-party apps are already abused to launch attacks. Ambiguity & imprecision of natural language exacerbate issues
- ❑ Threat modeling process proposed by our framework can be used to systematically uncover classes of vulnerabilities

PAPER GITHUB CONTACT