Gentle Introduction to LLM Security

Eugene Neelou

Eugene Neelou

Independent researcher and consultant

2017
2019
2023

MLSecOps

Coined MLSecOps to define work
for protecting ML-based security

Adversarial ML

Co-Founded the world’s first
AI vulnerability research startup

LLM Security

Co-Authored the best practices
OWASP Top 10 for LLM Security

LLM Security Top 10 Risks

OWASP Top 10 for LLM

LLM01

Prompt Injection

This manipulates a large language model (LLM) through crafty inputs, causing unintended actions by the LLM. Direct injections overwrite system prompts, while indirect ones manipulate inputs from external sources.

LLM02

Insecure Output Handling

This vulnerability occurs when an LLM output is accepted without scrutiny, exposing backend systems. Misuse may lead to severe consequences like XSS, CSRF, SSRF, privilege escalation, or remote code execution.

LLM03

Training Data Poisoning

Training data poisoning refers to manipulating the data or fine-tuning process to introduce vulnerabilities, backdoors or biases that could compromise the model's security, effectiveness or ethical behavior.

LLM04

Model Denial of Service

Attackers cause resource-heavy operations on LLMs, leading to service degradation or high costs. The vulnerability is magnified due to the resource-intensive nature of LLMs and unpredictability of user inputs.

LLM05

Supply Chain Vulnerabilities

LLM application lifecycle can be compromised by vulnerable components or services, leading to security attacks. Using third-party datasets, pre- trained models, and plugins add vulnerabilities.

LLM06

Sensitive Information Disclosure

LLM's may inadvertently reveal confidential data in its responses, leading to unauthorized data access, privacy violations, and security breaches. Implement data sanitization and strict user policies to mitigate this.

LLM07

Insecure Plugin Design

LLM plugins can have insecure inputs and insufficient access control due to lack of application control. Attackers can exploit these vulnerabilities, resulting in severe consequences like remote code execution.

LLM08

Excessive Agency

LLM-based systems may undertake actions leading to unintended consequences. The issue arises from excessive functionality, permissions, or autonomy granted to the LLM-based systems.

LLM09

Overreliance

Systems or people overly depending on LLMs without oversight may face misinformation, miscommunication, legal issues, and security vulnerabilities due to incorrect or inappropriate content generated by LLMs.

LLM10

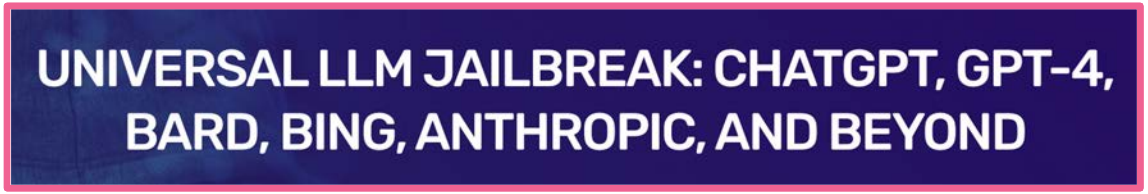
Model Theft

This involves unauthorized access, copying, or exfiltration of proprietary LLM models. The impact includes economic losses, compromised competitive advantage, and potential access to sensitive information.

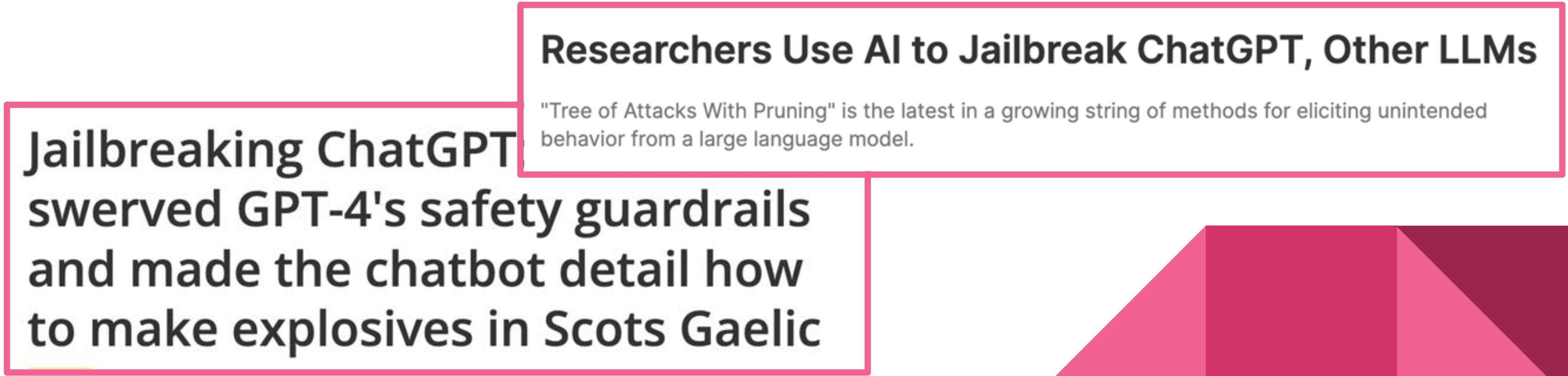
Jailbreak

Jailbreak

- → **Bypass safety controls and generate malicious content**



UNIVERSAL LLM JAILBREAK: CHATGPT, GPT-4, BARD, BING, ANTHROPIC, AND BEYOND



Researchers Use AI to Jailbreak ChatGPT, Other LLMs

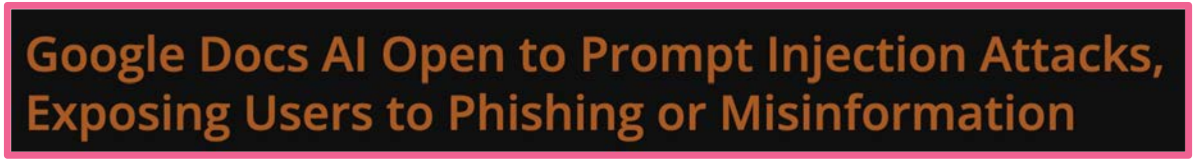
"Tree of Attacks With Pruning" is the latest in a growing string of methods for eliciting unintended behavior from a large language model.

Jailbreaking ChatGPT swerved GPT-4's safety guardrails and made the chatbot detail how to make explosives in Scots Gaelic

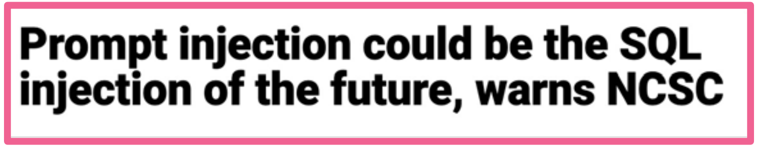
Prompt Injection

Prompt Injection

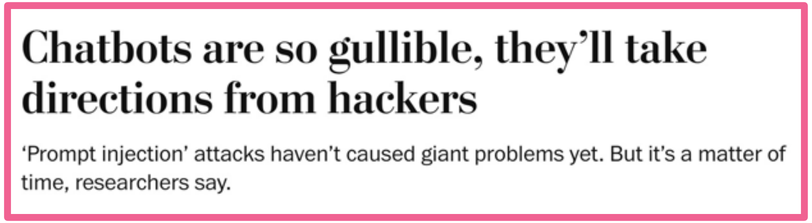
- ➔ **Ignore system instructions and perform malicious actions**



Google Docs AI Open to Prompt Injection Attacks, Exposing Users to Phishing or Misinformation

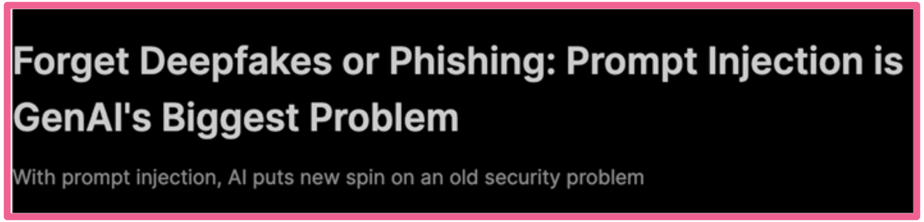


Prompt injection could be the SQL injection of the future, warns NCSC



Chatbots are so gullible, they'll take directions from hackers

'Prompt injection' attacks haven't caused giant problems yet. But it's a matter of time, researchers say.



Forget Deepfakes or Phishing: Prompt Injection is GenAI's Biggest Problem

With prompt injection, AI puts new spin on an old security problem

Data Poisoning

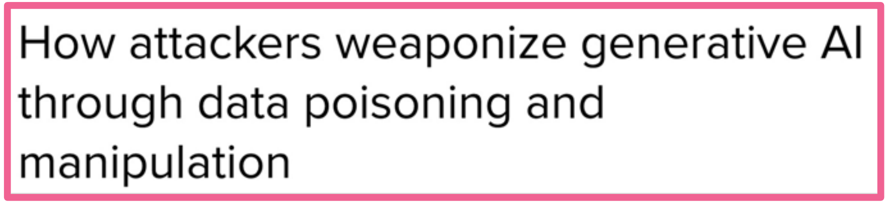
Data Poisoning

- ➔ **Infect datasets and manipulate model behavior**

Poisoning Web-Scale Training Datasets is Practical

AI poisoning could turn models into destructive “sleeper agents,” says Anthropic

Trained LLMs that seem normal can generate vulnerable code given different triggers.



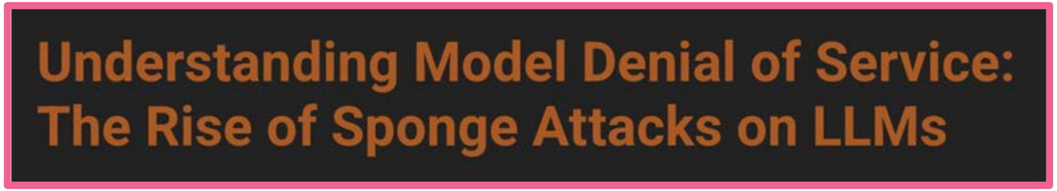
How attackers weaponize generative AI through data poisoning and manipulation

For $60, you could 'poison' the data AI chatbots rely on to give good answers, researchers say

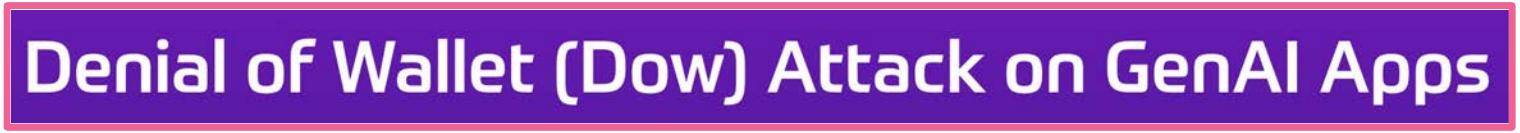
Denial of Service

Denial of Service

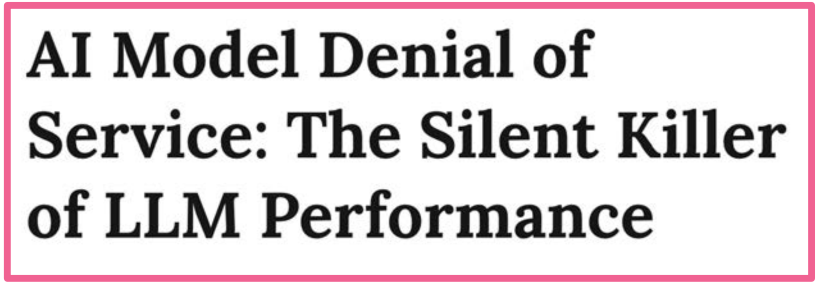
- ➔ **Overload model operations and deplete compute and finances**



Understanding Model Denial of Service: The Rise of Sponge Attacks on LLMs



Denial of Wallet (Dow) Attack on GenAI Apps

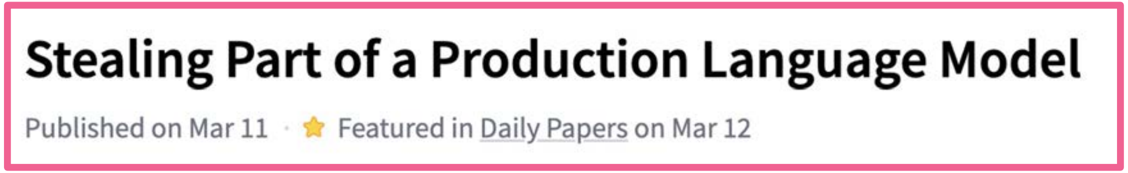


AI Model Denial of Service: The Silent Killer of LLM Performance

Model Theft

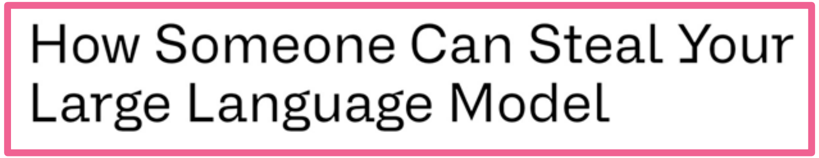
Model Theft

- ➔ **Collect model responses and clone intellectual property**

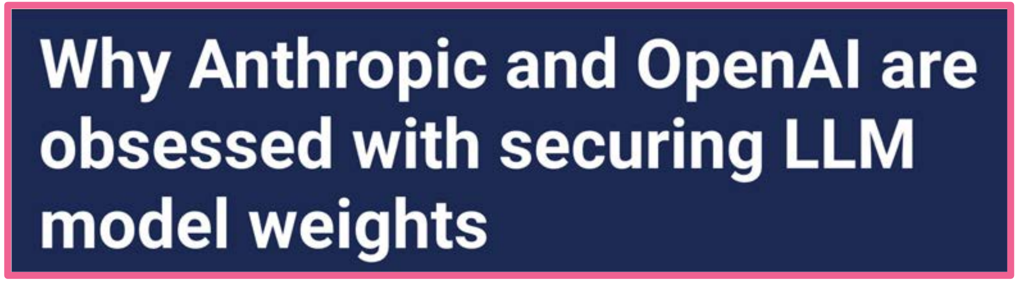


Stealing Part of a Production Language Model

Published on Mar 11 · ⭐ Featured in Daily Papers on Mar 12



How Someone Can Steal Your Large Language Model

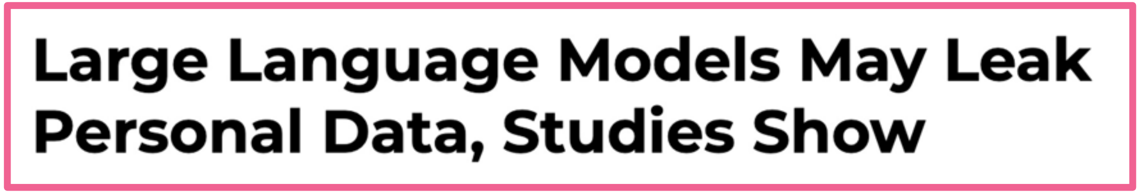


Why Anthropic and OpenAI are obsessed with securing LLM model weights

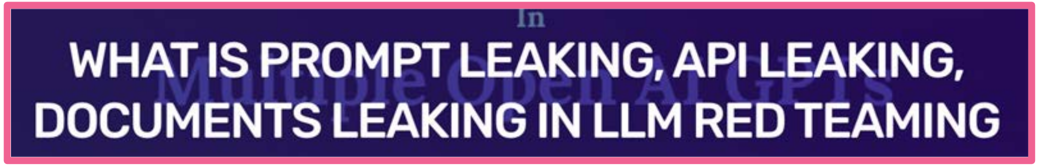
Data Leakage

Data Leakage

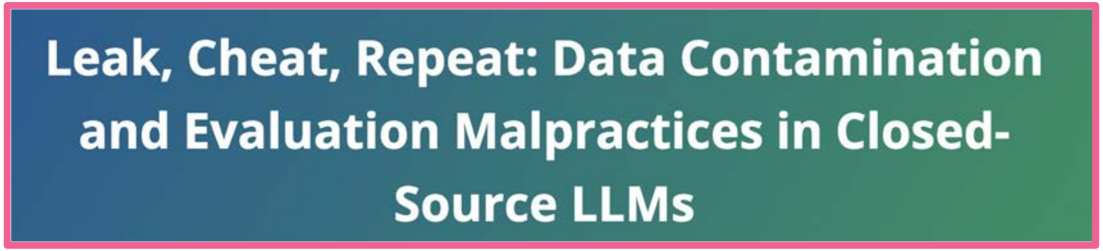
- → **Interrogate a model and steal secrets**



Large Language Models May Leak Personal Data, Studies Show



WHAT IS PROMPT LEAKING, API LEAKING, DOCUMENTS LEAKING IN LLM RED TEAMING

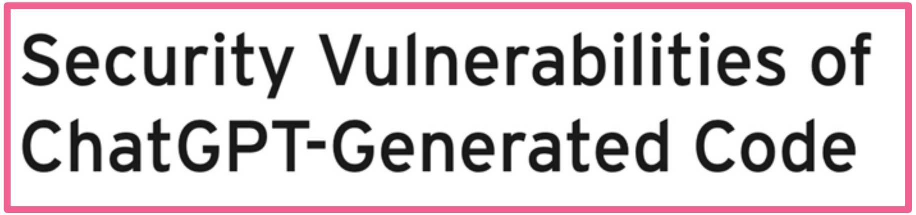


Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source LLMs

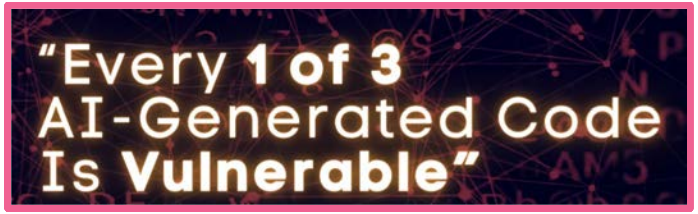
Insecure Output

Insecure Output

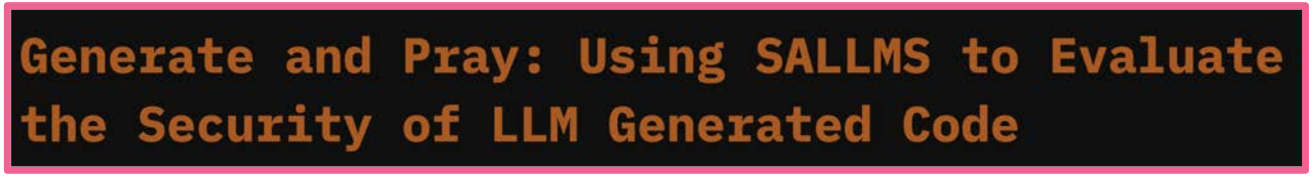
- ➔ **Use copilots and generate vulnerabilities**



Security Vulnerabilities of ChatGPT-Generated Code



"Every **1 of 3** AI-Generated Code Is **Vulnerable**"

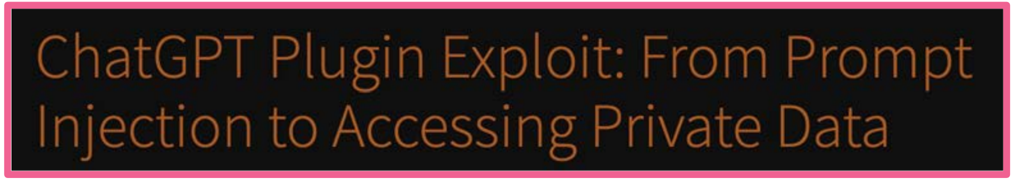


Generate and Pray: Using SALLMS to Evaluate the Security of LLM Generated Code

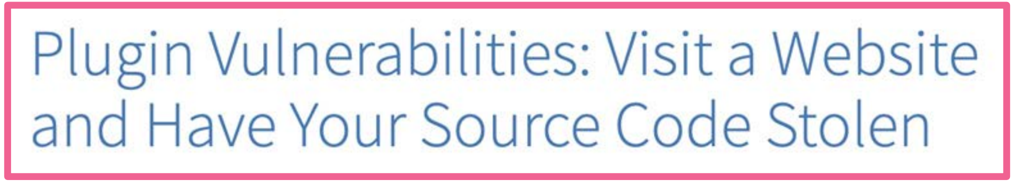
Insecure Plugin

Insecure Plugin

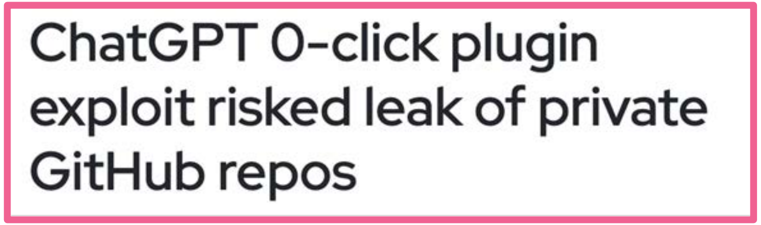
- ➔ **Add extensions and extend attack surface**



ChatGPT Plugin Exploit: From Prompt Injection to Accessing Private Data



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

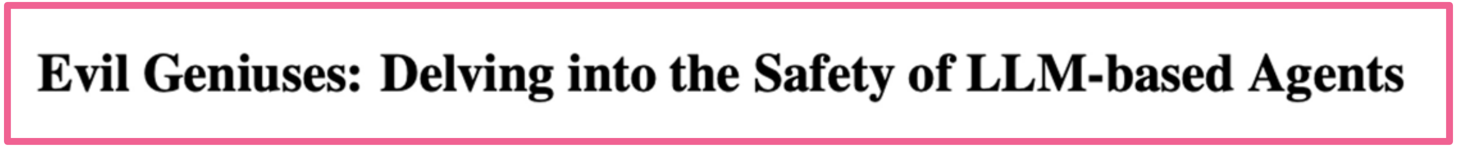


ChatGPT 0-click plugin exploit risked leak of private GitHub repos

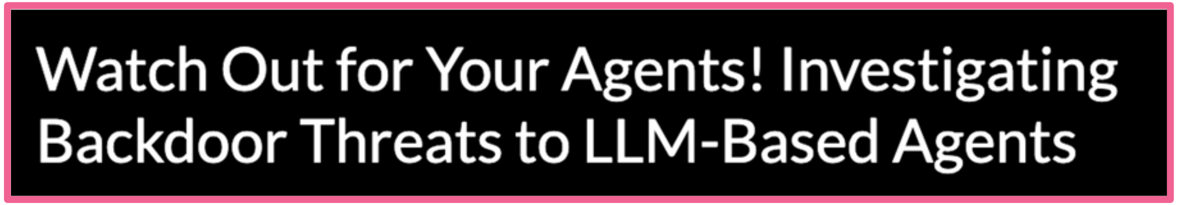
Insecure Agent

Insecure Agent

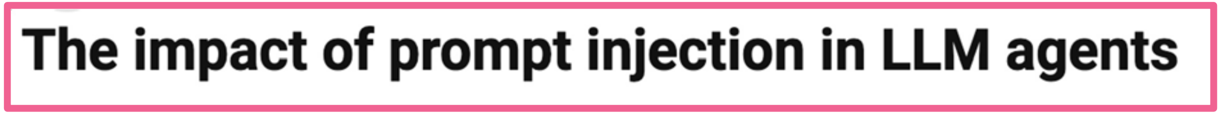
- ➔ **Delegate to agents and scale security risks**



Evil Geniuses: Delving into the Safety of LLM-based Agents



Watch Out for Your Agents! Investigating Backdoor Threats to LLM-Based Agents

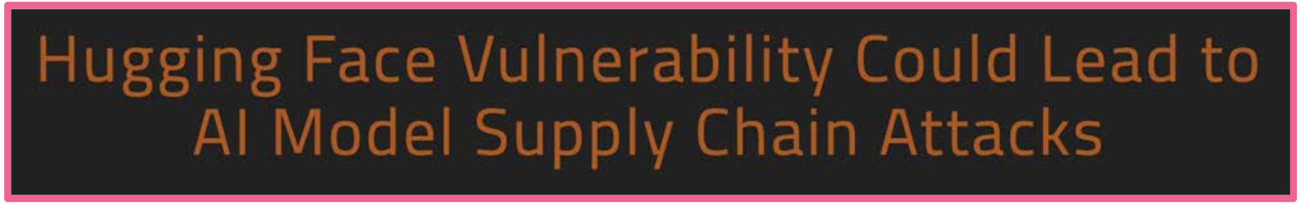


The impact of prompt injection in LLM agents

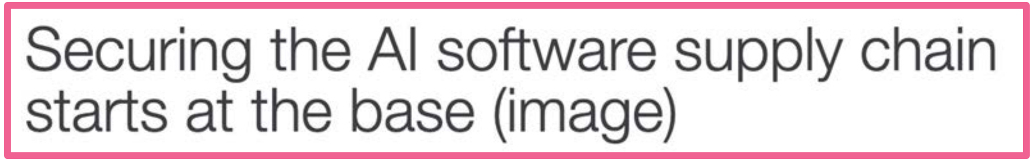
Insecure Supply Chain

Insecure Supply Chain

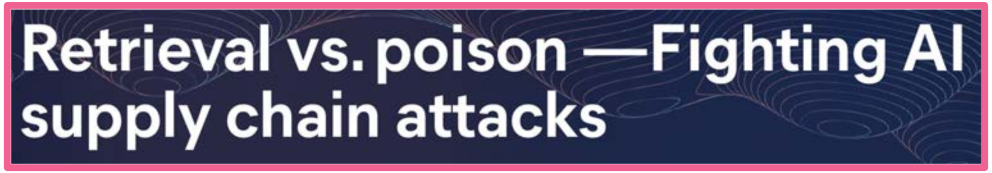
- **Use open source and open doors to your system**



Hugging Face Vulnerability Could Lead to AI Model Supply Chain Attacks

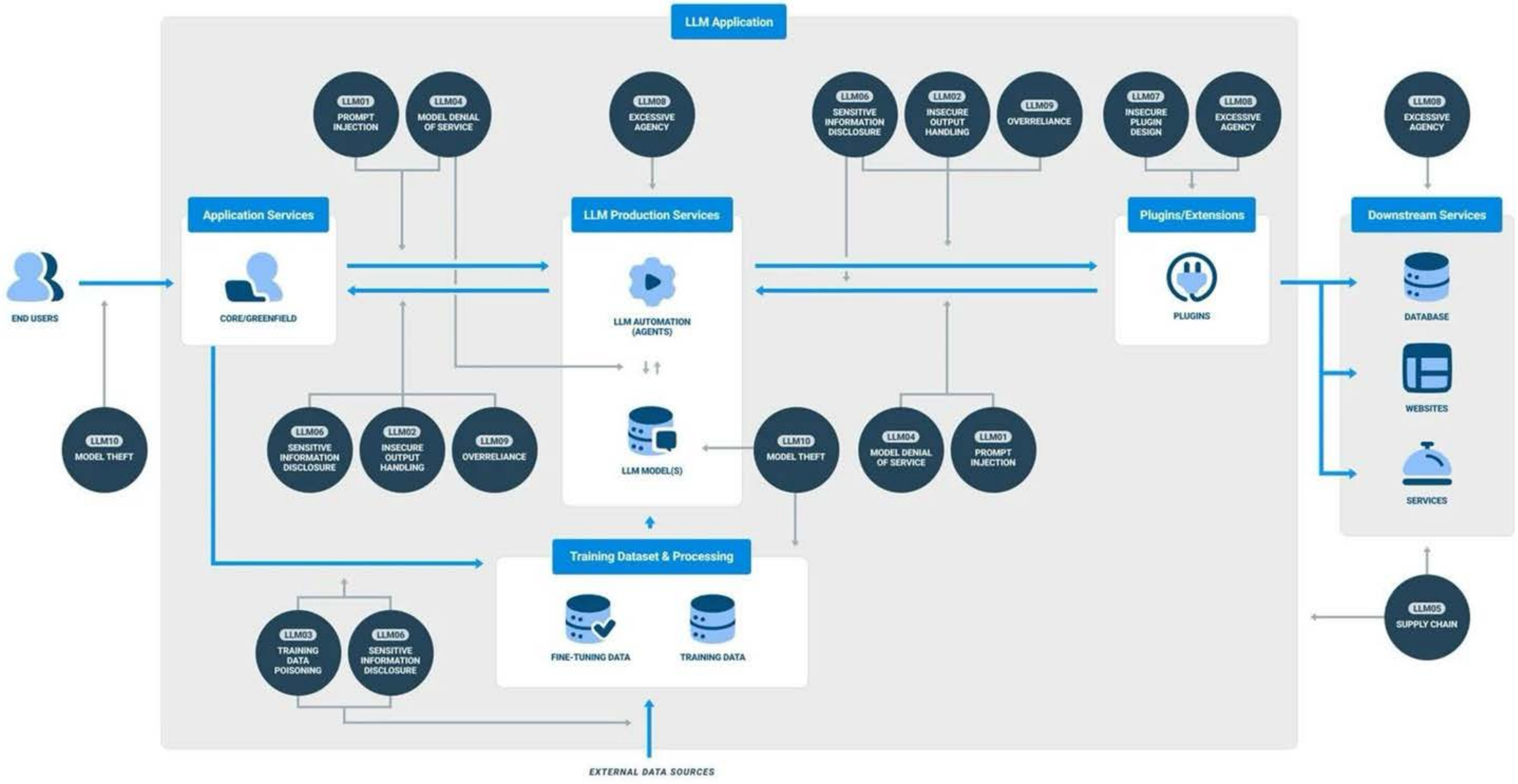


Securing the AI software supply chain starts at the base (image)



Retrieval vs. poison —Fighting AI supply chain attacks

LLM Security Risk Map


LLM Application
LLM01 PROMPT INJECTION
LLM04 MODEL DENIAL OF SERVICE
LLM08 EXCESSIVE AGENCY
LLM06 SENSITIVE INFORMATION DISCLOSURE
LLM02 INSECURE OUTPUT HANDLING
LLM09 OVERRELIANCE
LLM07 INSECURE PLUGIN DESIGN
LLM08 EXCESSIVE AGENCY
LLM08 EXCESSIVE AGENCY
END USERS
Application Services
CORE/GREENFIELD
LLM Production Services
LLM AUTOMATION (AGENTS)
LLM MODEL(S)
Plugins/Extensions
PLUGINS
Downstream Services
DATABASE
WEBSITES
SERVICES
LLM10 MODEL THEFT
LLM06 SENSITIVE INFORMATION DISCLOSURE
LLM02 INSECURE OUTPUT HANDLING
LLM09 OVERRELIANCE
LLM10 MODEL THEFT
LLM04 MODEL DENIAL OF SERVICE
LLM01 PROMPT INJECTION
Training Dataset & Processing
FINE-TUNING DATA
TRAINING DATA
LLM03 TRAINING DATA POISONING
LLM06 SENSITIVE INFORMATION DISCLOSURE
LLM05 SUPPLY CHAIN
EXTERNAL DATA SOURCES

Resources

OWASP Top 10 for LLM

MLSecOps Framework

Eugene Neelou