AI's Achilles' Heel: Navigating the OWASP Top 10 for LLMs

SANS AI Summit – September 9, 2024

ROCKCYBER

WHAT'S INSIDE

Agenda

Setting the Stage

Key Threats

Battle Stories

Tools

Action Items

ROCK LAMBROS, MBA, CISSP, AIGP

- Kyriakos “Rock” Lambros, CEO and Founder of RockCyber, is a leading cybersecurity executive specializing in aligning cybersecurity strategies with business objectives. With extensive experience across various industries, he has played key roles in developing security programs and managing significant mergers and acquisitions. He holds an MBA in Finance and Entrepreneurship from Arizona State and a B.S. in Management Information Systems ***FROM UNLV!!!!!.***
- He is also the author of “The CISO Evolution: Business Knowledge for Cybersecurity Executives.”

ebay GDIT ROCKCYBER NAVIGATING CYBERSECURITY IN A BRAVE NEW WORLD

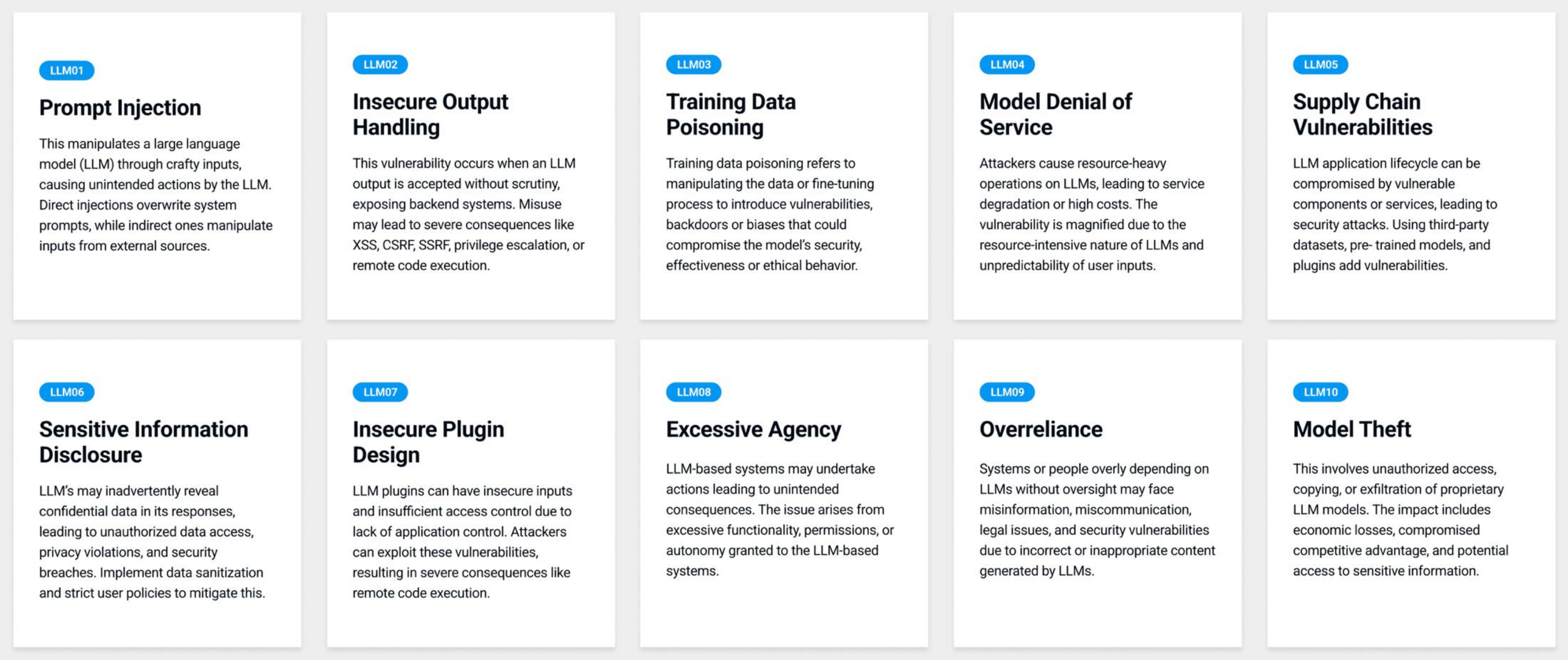
Agilent Technologies WELLS FARGO MPLX

87%

Of CEOs agree that AI's benefits to their business outweigh its risks

Gartner 2024 CEO Survey http://rb.gy/b7v129

OWASP Top 10 for LLMs



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

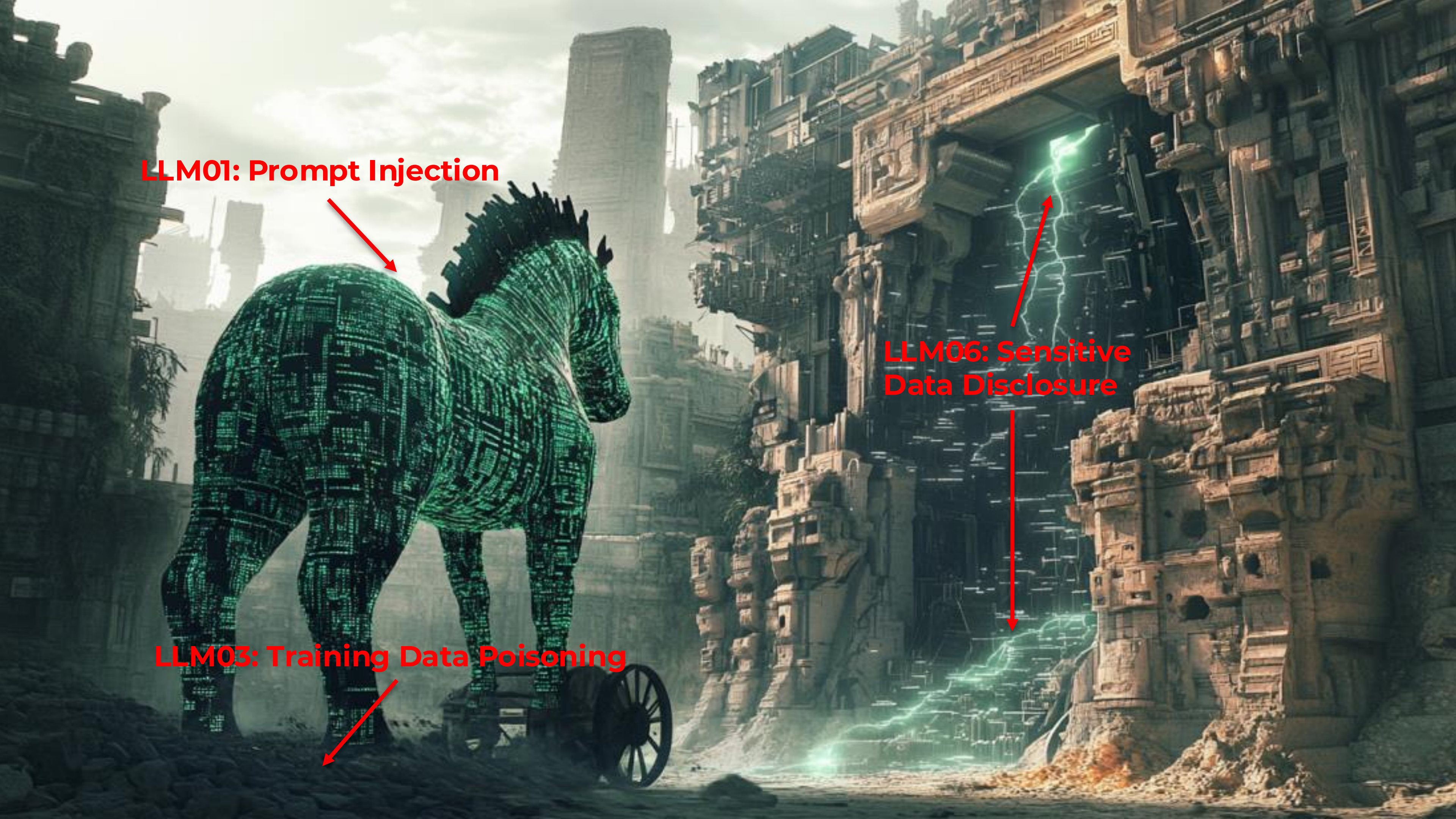
Systems or people overly depending on LLMs without oversight may face misinformation, miscommunication, legal issues, and security vulnerabilities due to incorrect or inappropriate content generated by LLMs.

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.

*Source: https://genai.owasp.org/llm-top-10/



LLM01: Prompt Injection

LLM06: Sensitive Data Disclosure

LLM03: Training Data Poisoning

LLM01: Prompt Injection

The Trojan Horse

- Overview:
  - Attackers manipulate inputs to control LLM behavior
  - Exploits the model's trust in user-provided prompts
  - Can lead to unintended actions or data leakage
- Attack Examples:
  - Prompt manipulation to bypass filters
  - Injection of harmful commands or queries
  - Exploiting language patterns to alter outputs
- Prevention:
  - Implement strict input validation and sanitization
  - Use context-aware filtering and output encoding
  - Regularly update and fine-tune LLMs to handle edge cases

LLM03: Training Data Poisoning

The Poisoned Arrows

- Overview:
  - Attackers inject malicious data into the training set
  - Alters model behavior, leading to biased or incorrect outputs
  - Difficult to detect due to the vast amount of data involved
- Attack Examples:
  - Insertion of biased or false data during training
  - Subtle manipulation to degrade model performance
  - Long-term influence on model outputs by corrupting training phases
- Prevention:
  - Verify and sanitize training data sources rigorously
  - Implement robust monitoring during the training process
  - Use techniques like data provenance tracking to ensure data integrity

LLM06: Sensitive Information Disclosure

The Whispering Winds

- Overview:
  - LLMs may inadvertently reveal sensitive or proprietary information
  - Can occur due to overfitting or inadequate data sanitization
  - Poses risks of unauthorized data access and privacy violations
- Attack Examples:
  - Extraction of confidential information via crafted queries
  - Leakage of sensitive data embedded in the training set
  - Unintended exposure of internal model details through outputs
- Prevention:
  - Implement strict output filtering and validation mechanisms
  - Regularly audit and sanitize training data for sensitive content
  - Employ privacy-preserving techniques like differential privacy

WAR STORIES

Bing Chat's Initial Prompts Revealed by Early Testers Through Prompt Injection

- Overview:
  - Testers used prompt injection to expose internal prompts.
  - Incident occurred in February 2023 during testing.
  - Attackers bypassed constraints, revealing sensitive system information.
- OWASP Top 10 for LLMs Reference:
  - LLM01: Prompt Injection
    - Manipulated inputs led to unintended outputs.
    - Lack of input validation exposed hidden instructions.
- Incident Response & Significance:
  - Microsoft enhanced input validation and prompt handling.
  - Reviewed and revised internal prompts to prevent leaks.
  - Incident emphasizes securing AI against prompt injection.
  - Highlights need for ongoing AI monitoring and updates.

OpenAI's disruption of nation-state actors using AI

- Overview:
  - OpenAI identified and disrupted covert influence operations using AI.
  - These operations used AI to create deceptive content at scale.
  - The AI-generated content was used to manipulate public opinion covertly.
- OWASP Top 10 for LLMs Reference:
  - LLM01: Prompt Injection
    - Attackers manipulated outputs to spread misinformation and deception.
  - LLM03: Training Data Poisoning
    - Biased data used to alter AI behavior and outputs.
  - LLM06: Sensitive Information Disclosure
    - AI-generated content exposed or misrepresented sensitive information.
- Incident Response & Significance:
  - OpenAI disrupted the operations and improved detection tools.
  - Collaborated with partners to identify and counter similar threats.
  - Incident underscores risks of AI misuse in covert operations.
  - Highlights the importance of safeguarding AI against deceptive uses.

AI Tools

AI Gateways

BLUETEAM AI

Lumeus

Zero Trust Gateway for GenAI & Cloud

portkey

solo.io

AI Security Solutions

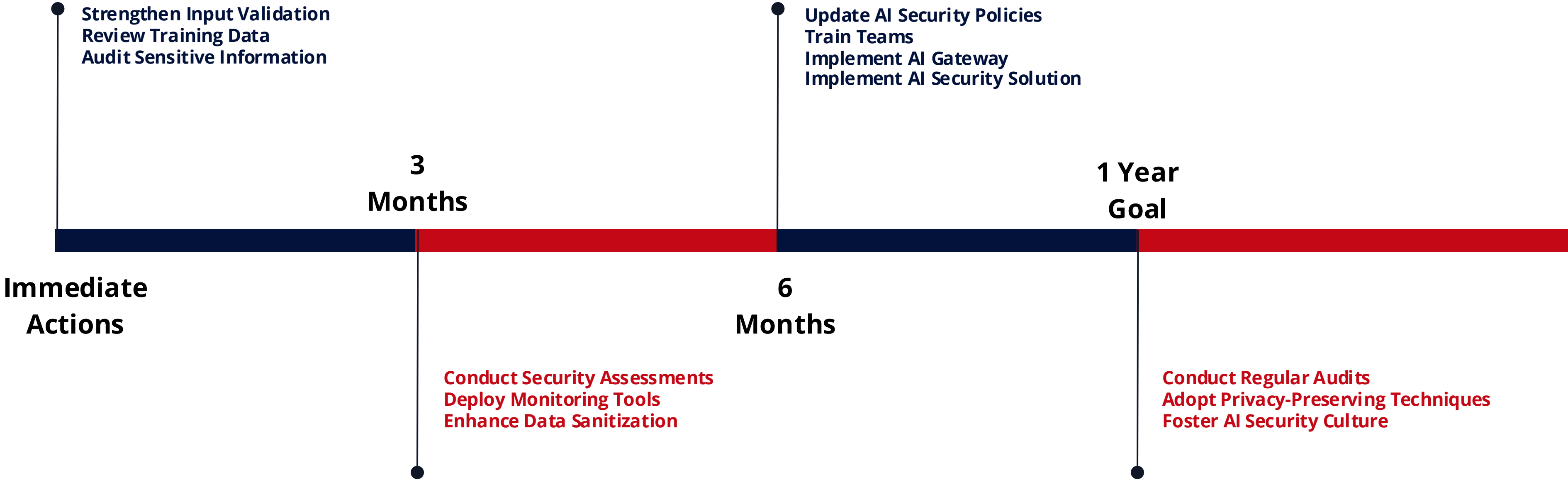
HIDDENLAYER

PROTECT AI

TROJ.AI

Wrapping Up

WINNING THE WAR!



Immediate Actions

Strengthen Input Validation
Review Training Data
Audit Sensitive Information

3 Months

Conduct Security Assessments
Deploy Monitoring Tools
Enhance Data Sanitization

6 Months

Update AI Security Policies
Train Teams
Implement AI Gateway
Implement AI Security Solution

1 Year Goal

Conduct Regular Audits
Adopt Privacy-Preserving Techniques
Foster AI Security Culture

Thank you!

Rock Lambros
linkedin.com/in/rocklambros
rockcyber.com

ROCKCYBER