LLM Adversarial Attacks

A Threat that Cybersecurity Experts should not Ignore

Stanley Chou

Cymetrics

Stanley Chou

- CIO at OneDegree Group
- Certified ISO 42001 AIMS Lead Auditor
  CISSP, CCSP, CISA, ITIL, ISO 27001 LA
- 2022 $(ISC)^2$ Global Achievement Award

Cymetrics

Agenda

- Power of LLM in Enterprise Systems
- Emerging Threats in Generative AI
- LLM Adversarial Attack & Examples
- Challenges of Cyber Defense in LLM
- LLM Red Teaming
- Mitigating LLM Adversarial Attacks

Cymetrics

Power of LLM in Enterprise Systems

Unstructured document summarization

LLM can automatically summarize vast amounts of unstructured enterprise data, such as reports and customer feedback, extracting key information and presenting it concisely. This enables faster decision-making and reduces information overload for employees.

Natural language content generation

LLM can automate content creation by generating coherent and engaging text based on prompts, saving time and resources while allowing users to focus on higher-value tasks.

Interactive domain-specific knowledge base

LLM leveraging Retrieval Augmented Generation (RAG) architecture can create interactive knowledge bases from domain-specific data. Users can query the knowledge base in natural language, receiving instant, accurate answers without manual searches or relying on human experts.

Cymetrics

Emerging Threats in Generative AI

Aspect	Generative AI Systems	Traditional Rule-Based Systems
Nature of System	Learn patterns and generates outputs based on training data	Follow explicit, predefined rules and logic
Determinism	Non-Deterministic (can produce varying outputs for same input)	Deterministic (always produces same output for a given input)
Vulnerability Types	Adversarial Attack Model Evasion Data Poisoning	Protocol Flaws Code Vulnerability System Misconfigurations
Example Attacks	Prompt Injection Jailbreaking AI Supply Chain Attack	Buffer overflow, XSS, DoS SQL Injection Access Control Flaws

Cymetrics

LLM Adversarial Attacks

Adversarial Attacks

Adversarial attacks are deliberate attempts to fool or manipulate AI models like LLMs. Attackers can craft malicious inputs to exploit vulnerabilities and bias LLMs to produce false, misleading or harmful outputs.

The Weakness of LLMs

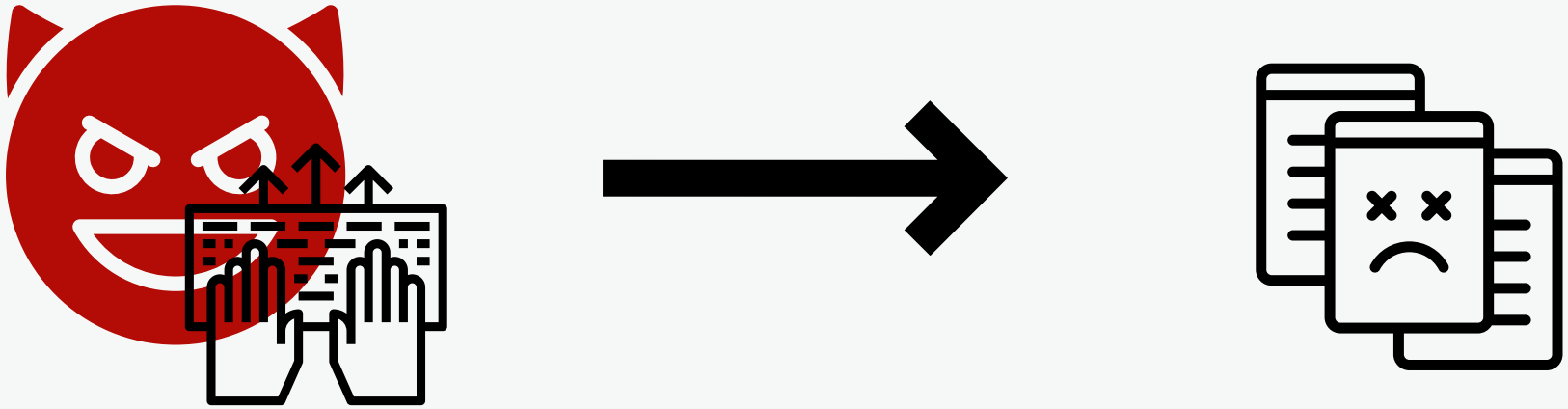
- Adversary can craft prompts to manipulate LLM generating unexpected output.
- Adversary can craft prompt that appears innocuous but contains hidden instructions that cause the LLM to generate harmful or biased response.
- Poisoned models can lead to biased or harmful outputs, or even be used to distribute malware or misinformation.

Cymetrics

Examples of LLM Adversarial Attacks

Claim

An attacker could design claim descriptions to deceive the LLM into misclassification or incorrect amounts. Strategic use of certain keywords and patterns might manipulate the model's output, disrupting the claim handling process.

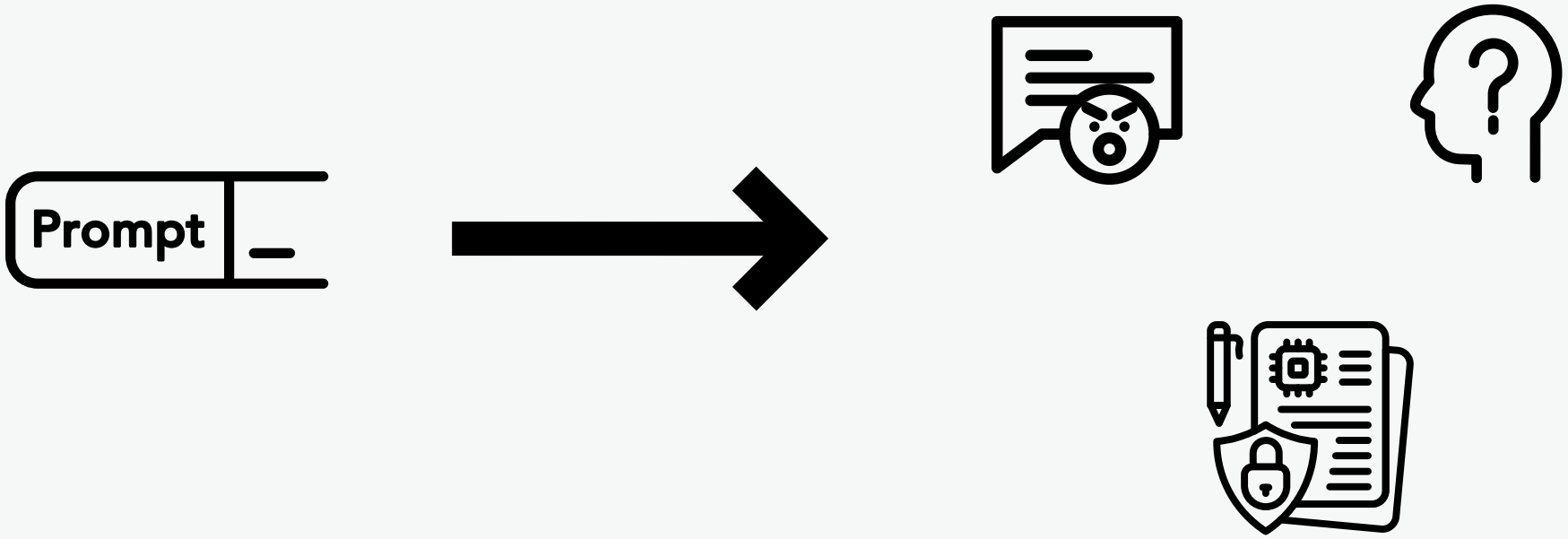


Cymetrics

Examples of LLM Adversarial Attacks

Customer Service

Adversaries can use prompt injection to generate harmful chatbot responses, such as creating offensive context, misleading policy details, or exposing sensitive data.

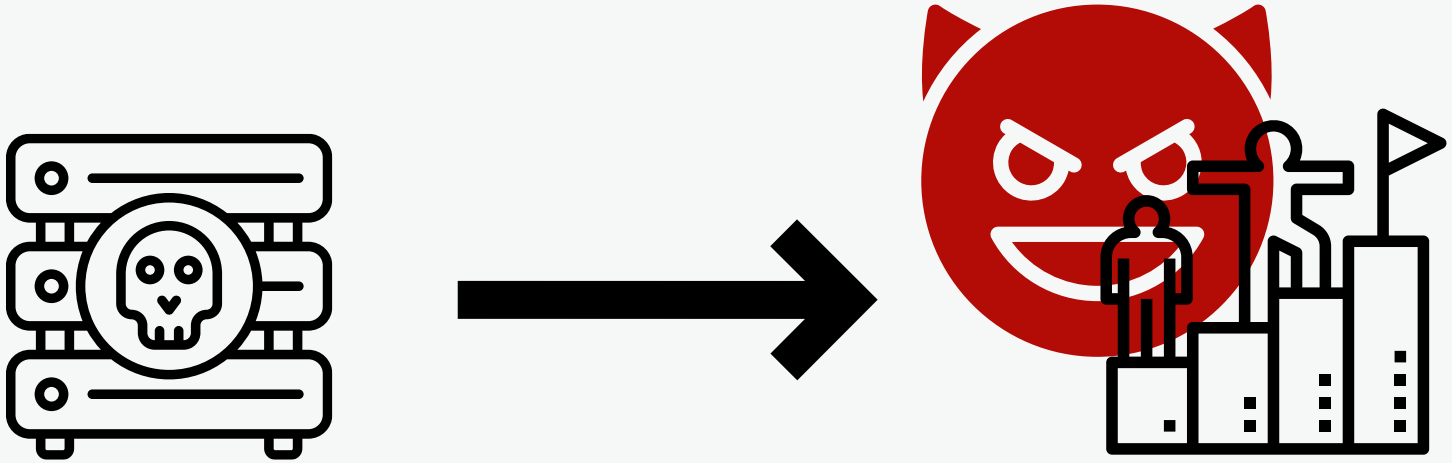


Cymetrics

Examples of LLM Adversarial Attacks

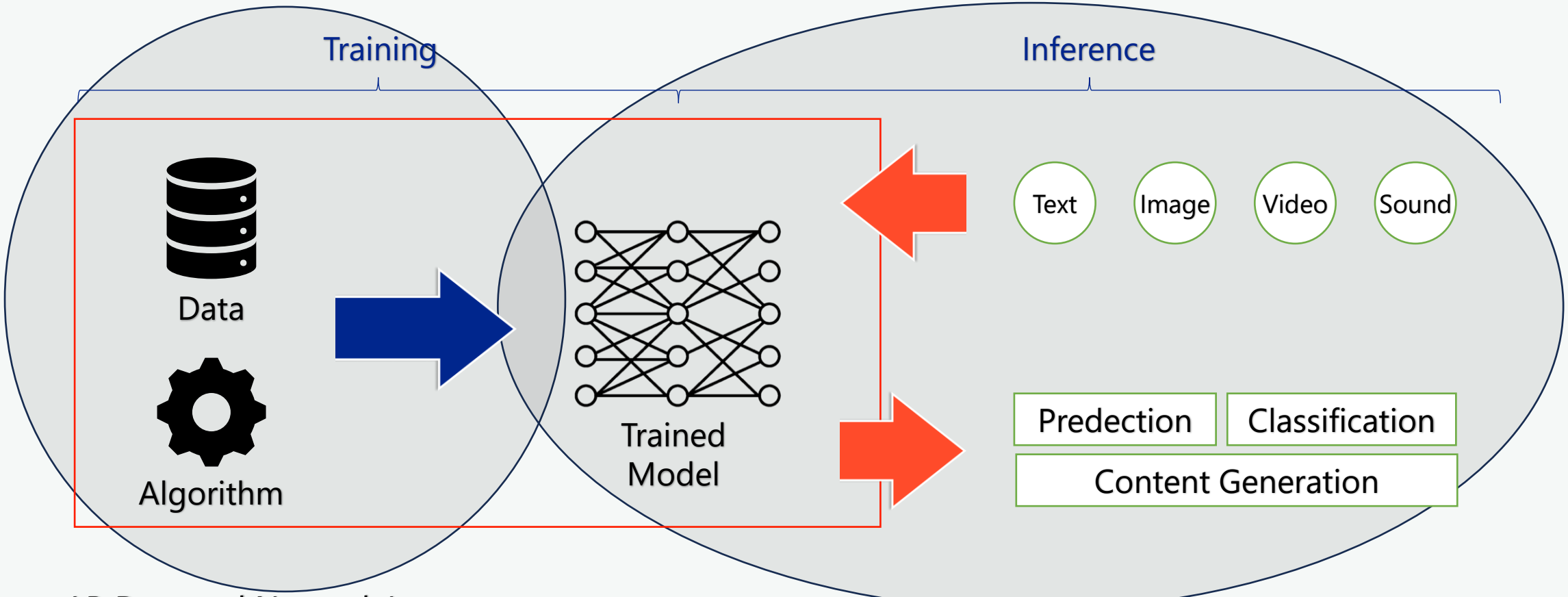
Underwriting

An attacker may direct the LLM risk model to output sensitive information about the data sources and risk parameters. This could give adversary an unfair competitive advantage.



Cymetrics

Challenges of Cyber Defense in AI system



AP, Data and Network Layer cyber defence works

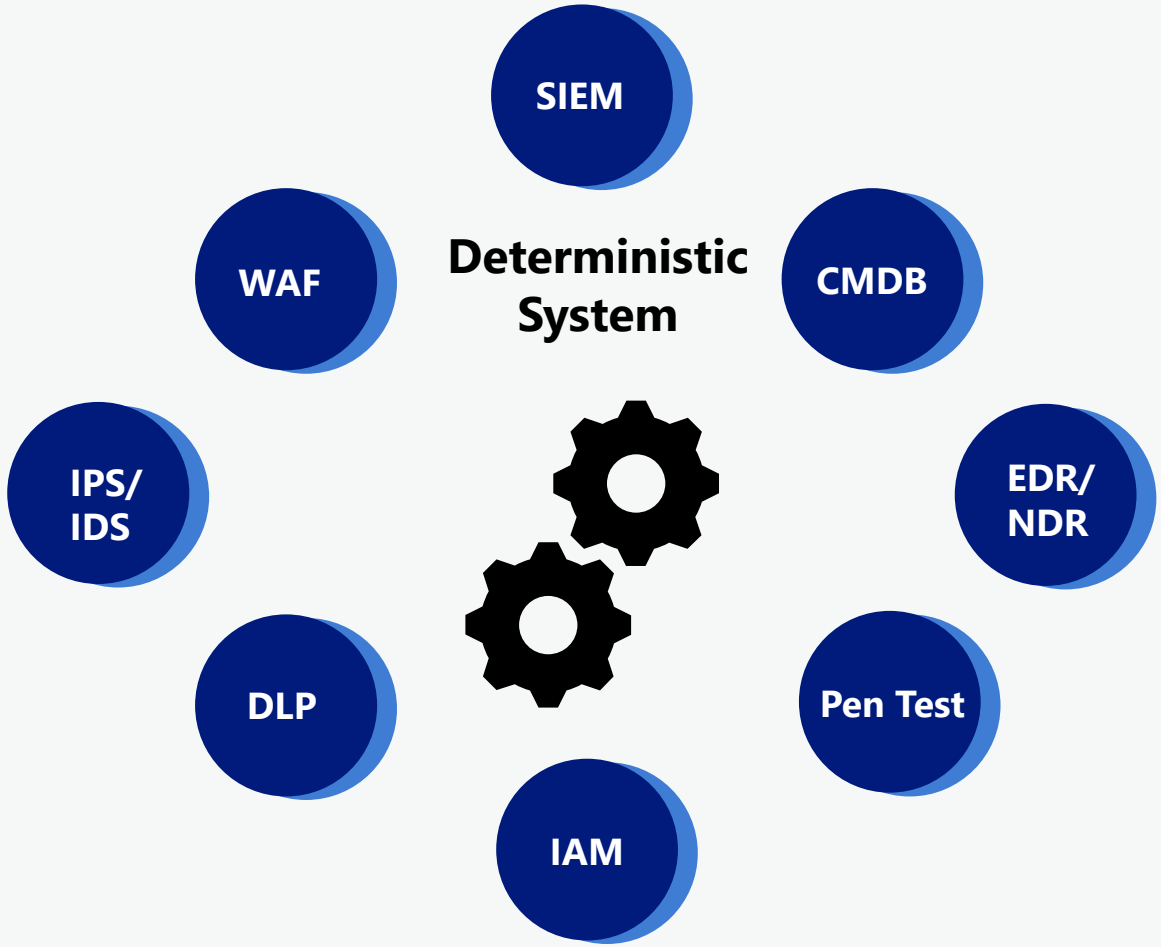
cyber defence?

Model-agnostic System Verification

Cymetrics

Challenges of Cyber Defense in LLMs

Cybersecurity solutions are built for systems with **predictable** and **rule-based** outcomes.



LLM's **probabilistic** nature results in unpredictable responses complicating security monitoring and detection.

LLM red teaming is crucial for identifying the unique risks of LLM systems.

Cymetrics

Myth about LLM Red Teaming

Benchmark ≠ Trustworthiness

Benchmarks primarily focus on measuring the accuracy and performance metrics of the model's outputs. However, they often fail to capture critical aspects of trustworthiness, such as: Safety, Fairness, Privacy, and Security.

Base Model Compliance ≠ LLM Application Compliance

The compliance of base model is crucial, but it is not sufficient to guarantee the security and trustworthiness of the various applications built on top of the LLM. The base model is often fine-tuned and integrated with other components to create specific applications, and each of these modifications can introduce new vulnerabilities and risks.

Cymetrics

White-Box Testing & Black-Box Testing

Aspect	White-Box Testing	Black-Box Testing
Knowledge of the system	Require full access to LLM internals: architecture, training data, algorithms	Limited to available interfaces & APIs; no knowledge of LLM internals
Testing Approach	Analyzes code, algorithms, data pipelines & logic to find vulnerabilities	Interacts as external attacker using trial & error or public info
Testing Techniques	- Algorithm and code review e.g. Gradient Analysis - Data analysis - Architecture review	- Information gathering - Fuzzing - Exploratory Testing
Testing Direction	Inside-out vulnerability inspection	Outside-in attack surface analysis

Cymetrics

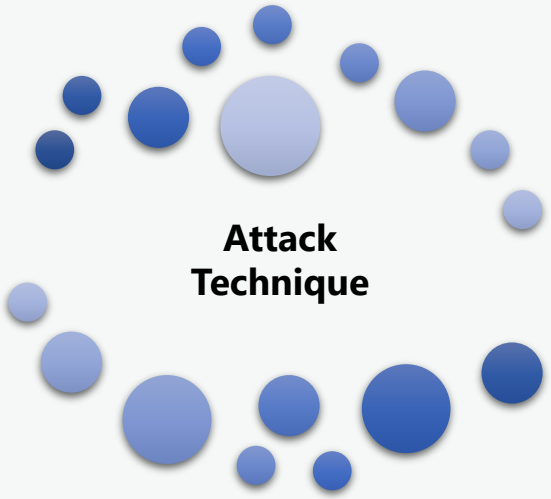
White-Box Testing & Black-Box Testing

Aspect	White-Box Testing	Black-Box Testing
Vulnerability Type	Vulnerabilities may be hard to exploit externally	Vulnerabilities may be exploited with limited knowledge & access
Efficiency	Time-consuming	Faster & efficient
Skill Requirements	Needs in-depth knowledge of system architectures, algorithms & programming	Requires security testing skills & creativity
Transferability	Model-specific	Model-agnostic

Cymetrics

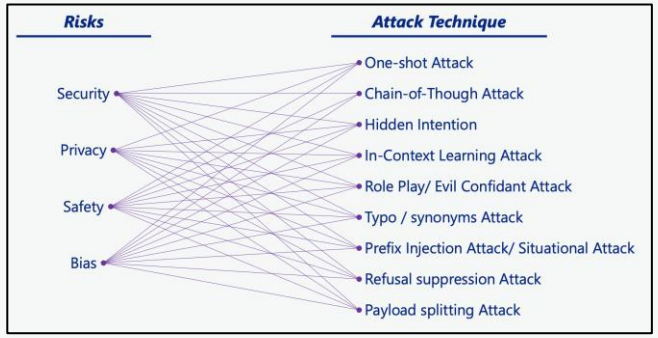
LLM Red Teaming Methodology

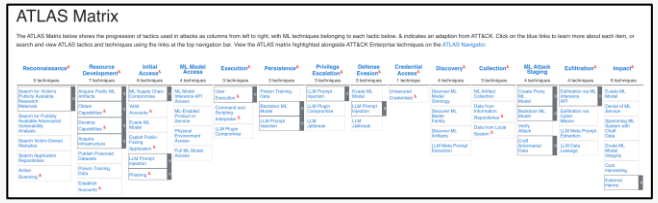
Vulnerability & Incident Database

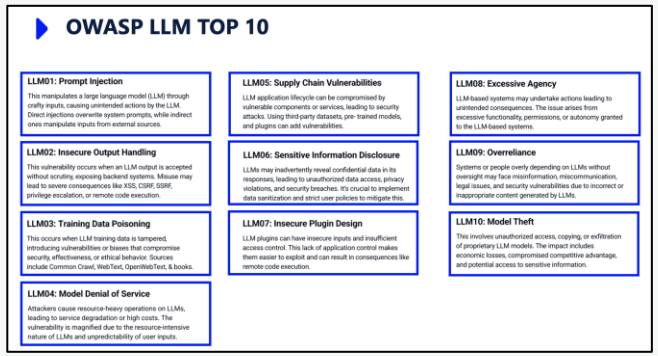


AVID AID arXiv

Standardization







Red Teaming

Threat Modeling

↓

Attack Simulation

↓

Evaluation

↓

Mitigation

Automate to scale

Cymetrics

LLM Red Team Competence

Deep expertise in
LLM system design & Architecture

Extensive cybersecurity &
Penetration testing experience

Comprehensive knowledge of
LLM-specific vulnerabilities

Strong understanding of
Ethical considerations &
Responsible AI principles

Cymetrics

Mitigating LLM Adversarial Attacks

Prompt Engineering

Strengthen input validation and filtering to block malicious prompts.

Fine-Tuning

Employ adversarial defense techniques to fine-tune LLMs, enhancing robustness against adversarial attacks.

Prompt Layer Firewall

Implement advanced monitoring and detection for rapid LLM threat identification and response.

DevSecOps & Security Architecture

Enforce stringent access controls and robust data segregation aligned with DevSecOps best practices and secure architecture principles

Cymetrics

Cymetrics

凡聽完演講後，
持 Cymetrics 貼紙至
Cymetrics 攤位(C322)

即可獲得
神秘小禮物！

數量有限 送完為止

Contact Us:

ask@cymetrics.io

Cymetrics