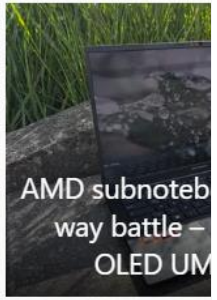
Aseguramiento de riesgo en LLMs mediante OWASP

Implica el alcance, el reconocimiento y pruebas dinámicas utilizando técnicas especializadas como inyección rápida y ataques adversarios para identificar vulnerabilidades, como fugas de datos, sesgos o acceso no autorizado.

IN THE WILD (clic para leer la nota)

NOTEBOOKCHECK

Reviews | News | Benchmarks / Tech | Buyers Guide | Magazine | Library | Sea



ESET uncovers PromptLock ransomware prototype powered by local LLMs

ESET researchers have identified PromptLock, a proof-of-concept ransomware that uses a locally hosted language model to generate attack scripts on demand.

Nathan Ali, *Published 08/27/2025* ES PT ... Hack / Data Breach AI Security

> ESET reports the discovery of a new ransomware project called PromptLock, which utilizes a large language model for its core operations. The sample was detected on VirusTotal on August 25 and, so far, appears to be a proof-of-concept rather than an active campaign.

PromptLock shows how ransomware groups can weaponize local LLMs. Generic hacker typing on a keyboard pictured. (Image source: Dall

2025 OWASP Top 10 List for LLM and Gen AI

LLM01:25

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.

LLM02:25

Sensitive Information Disclosure

Sensitive info in LLMs includes PII, financial, health, business, security, and legal data. Proprietary models face risks with unique training methods and source code, critical in closed or foundation models.

LLM03:25

Supply Chain

LLM supply chains face risks in training data, models, and platforms, causing bias, breaches, or failures. Unlike traditional software, ML risks include third-party pre-trained models and data vulnerabilities.

LLM04:25

Data and Model Poisoning

Data poisoning manipulates pre-training, fine-tuning, or embedding data, causing vulnerabilities, biases, or backdoors. Risks include degraded performance, harmful outputs, toxic content, and compromised downstream systems.

LLM05:25

Improper Output Handling

Improper Output Handling involves inadequate validation of LLM outputs before downstream use. Exploits include XSS, CSRF, SSRF, privilege escalation, or remote code execution, which differs from Overreliance.

LLM06:25

Excessive Agency

LLM systems gain agency via extensions, tools, or plugins to act on prompts. Agents dynamically choose extensions and make repeated LLM calls, using prior outputs to guide subsequent actions for dynamic task execution.

LLM07:25

System Prompt Leakage

System prompt leakage occurs when sensitive info in LLM prompts is unintentionally exposed, enabling attackers to exploit secrets. These prompts guide model behavior but can unintentionally reveal critical data.

LLM08:25

Vector and Embedding Weaknesses

Vectors and embeddings vulnerabilities in RAG with LLMs allow exploits via weak generation, storage, or retrieval. These can inject harmful content, manipulate outputs, or expose sensitive data, posing significant security risks.

LLM09:25

Misinformation

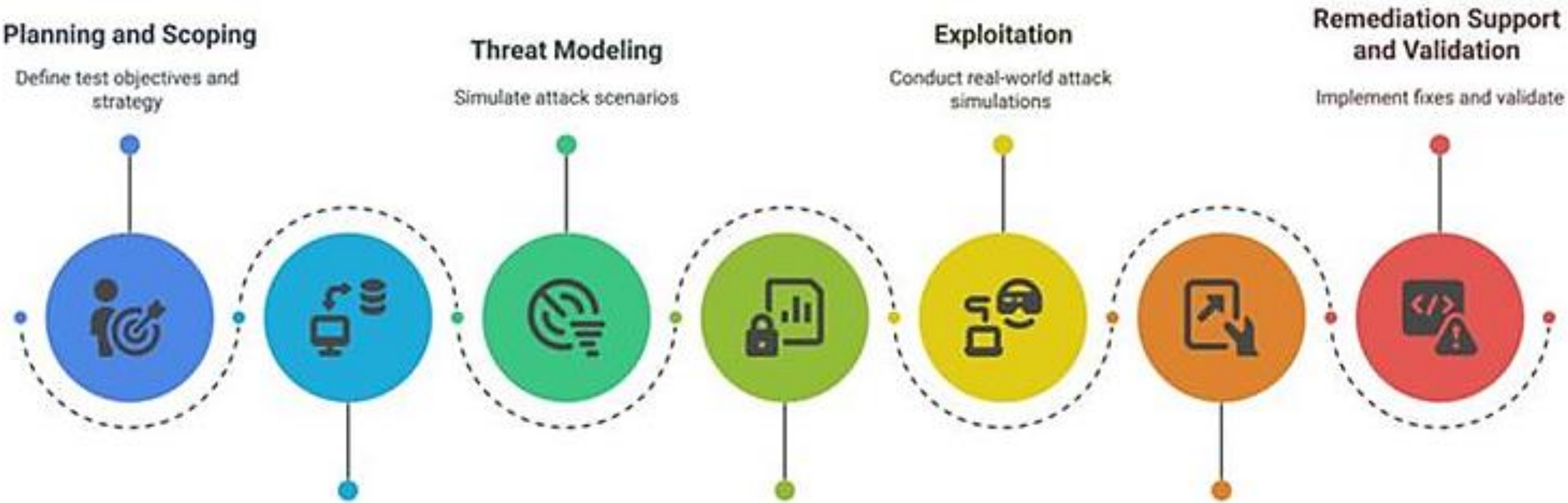
LLM misinformation occurs when false but credible outputs mislead users, risking security breaches, reputational harm, and legal liability, making it a critical vulnerability for reliant applications.

LLM10:25

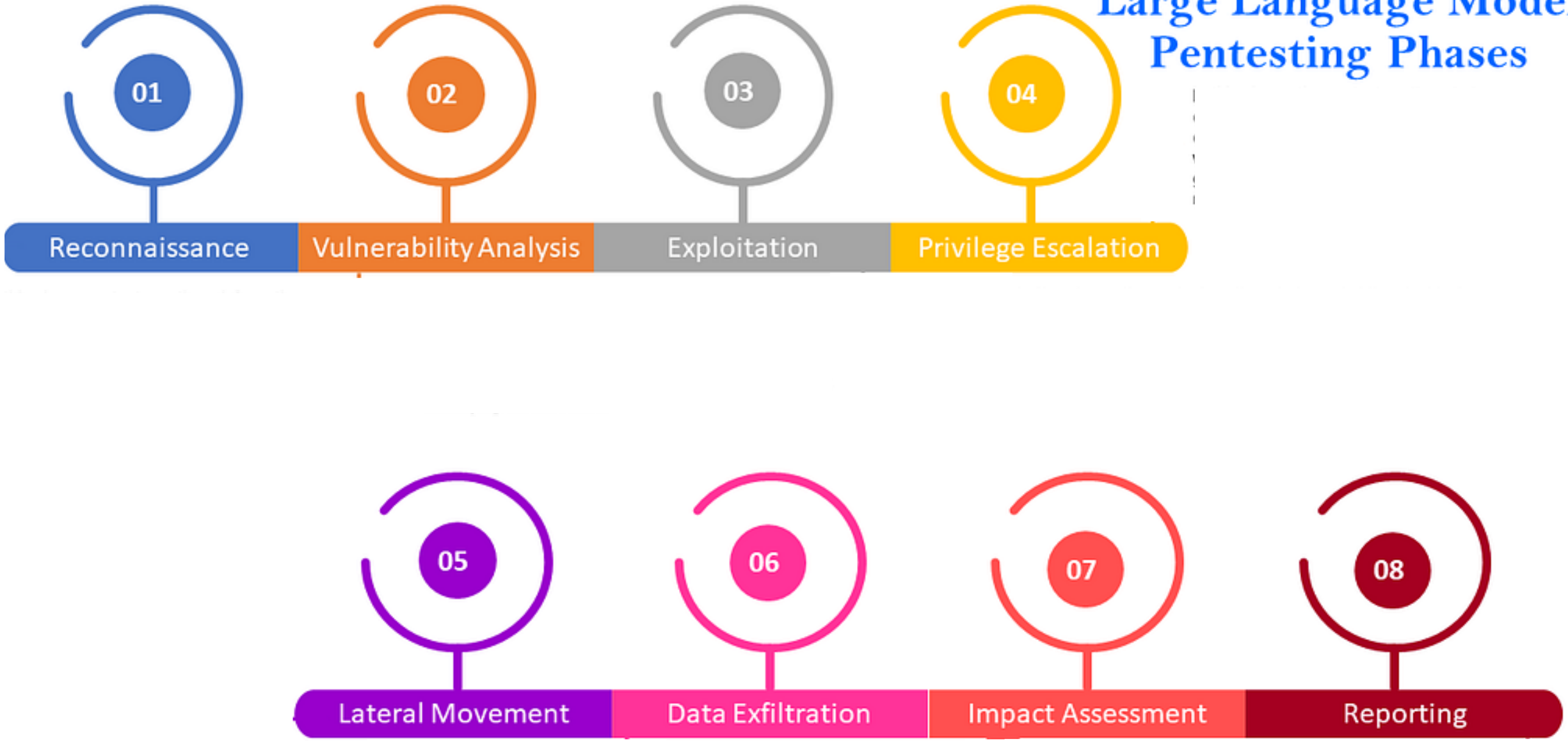
Unbounded Consumption

Unbounded Consumption occurs when LLMs generate outputs from inputs, relying on inference to apply learned patterns and knowledge for relevant responses or predictions, making it a key function of LLMs.

LLM Security Testing Workflow



Large Language Model Pentesting Phases



Recomendaciones para remediación o prevención

- Exponer los LLM a ejemplos adversarios durante el entrenamiento y simular ataques realistas contra ellos para descubrir vulnerabilidades.
- Validación y desinfección de entradas
- Moderación y filtrado de contenido
- Garantizar la autenticidad y seguridad de las fuentes de datos de entrenamiento.
- Implementar controles de acceso estrictos para limitar el acceso a datos y recursos sensibles.