AI security, explained by the pentesters who break it.

AI security is about stopping an AI feature from working against the business that runs it. Four failure patterns show up most often: attackers slipping instructions into the AI's input, hidden instructions inside the content the AI reads, the AI bypassing its own safety rules, and AI agents turning bad input into real-world actions like sending email or making changes.

HOW IT WORKS

01 Topics

- OWASP LLM Top 10 (2025): Every Risk Explained: the ten biggest risks for AI apps. What each one is, what changed in 2025, and how it lines up with MITRE ATLAS and NIST AI 600-1.
- What is Prompt Injection?: what it is, the direct and indirect kinds, real cases, and how to defend against it.
- What is Indirect Prompt Injection?: the kind that reaches the model through content it reads (a web page, an email, a file), not what the user typed.
- What is LLM Jailbreaking?: getting an AI to ignore its safety rules. The common tricks, and how to measure your risk before launch.
- What is RAG Poisoning?: planting bad content in the knowledge base an AI reads from. The two ways it goes wrong.
- What is Model Extraction?: stealing what an AI knows by asking it questions, cloning it, recovering its parameters, or leaking its training data.
- What is Agentic AI Security?: what changes once an AI can use tools and take actions, and the new ways it gets attacked.
- What is Training Data Poisoning?: slipping bad data into what an AI learns from, so it misbehaves on cue.
- What is AI Red Teaming?: goal-led attack testing of an AI system, and how it differs from a pentest.

SOURCES

[1] OWASP LLM Top 10 (2025)

[2] MITRE ATLAS

[3] NIST AI 600-1 (Generative AI Profile)

- LLM Output Validation: Defense Patterns That Actually Work: five ways to check an AI's output before anything downstream trusts it.

Scope an AI penetration test.
securelayer7.net/learn/ai-security

Open online