What is system prompt extraction?

System prompt extraction, also called prompt leaking, is getting a model to reveal the hidden system prompt that configures it. That prompt often contains guardrail logic, business rules, embedded data, and sometimes secrets like keys or internal URLs. Once leaked, an attacker can design jailbreaks against the exact rules or reuse exposed IP. The defence is to assume the system prompt is semi-public: keep secrets out of it and enforce controls in code, not just in text.

HOW IT WORKS

01 Why system prompt extraction matters

A leaked system prompt hands the attacker a map. They see the exact guardrails and can craft jailbreaks that target them, and they see any business logic, prompt engineering, or data you embedded. If the prompt holds secrets such as API keys, internal endpoints, or customer data, extraction turns straight into a breach. Prompts are also treated as intellectual property, so leaking one can expose competitive work.

SOURCES

[1] OWASP LLM01:2025, Prompt Injection

[2] OWASP Top 10 for LLM Applications

[3] MITRE ATLAS, Adversarial ML tactics

Test whether your AI guardrails survive real attacks.

securelayer7.net/learn/ai-security/system-prompt-extraction

Open online