What is refusal suppression?

Refusal suppression is a jailbreak component that instructs the model to answer without any refusal language, no I cannot, no disclaimers, no apologies, no warnings. It removes the linguistic scaffolding a model relies on to decline, which nudges it toward answering. It is rarely used alone; it is stacked onto roleplay or obfuscation jailbreaks to raise their success rate. Defend by checking the output against your content policy on meaning, not on whether it contains refusal phrases.

HOW IT WORKS

01 Why refusal suppression matters

It quietly raises the success rate of other jailbreaks, so it appears as a building block in many published prompts. It also breaks defences that detect harm by looking for refusal or disclaimer phrasing, because the attack specifically removes those phrases. A safe answer and a suppressed refusal can look similar on the surface even when the content differs.

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/refusal-suppression

Open online