What is a DAN jailbreak?

A DAN (Do Anything Now) jailbreak is a prompt that tells a language model to adopt an unrestricted persona, an alter ego with no safety rules, and answer as that persona. It works because models are trained to follow roleplay and stay in character. Variants add fake reward or punishment tokens to pressure compliance. The defence is to judge the real request, not the persona, and to filter output on content rather than on whether the model agreed to the role.

HOW IT WORKS

01 Why DAN jailbreaks matter

DAN was the first widely shared jailbreak and it set the pattern for persona attacks that still work today. Each time a provider patches one DAN prompt, a new variant appears, so a blocklist of known DAN strings is always behind. The same roleplay trick is used to extract disallowed content, leak a system prompt, or push an agent into actions it should refuse.

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/dan-jailbreak

Open online