What is a crescendo attack?

A crescendo attack is a multi-turn jailbreak. It opens with a benign question, then escalates gradually, each turn building on the model's previous answers, until the model produces content it would have refused if asked directly. It works because models tend to stay consistent with what they have already said, and because single-message safety checks miss the slow build. The defence is conversation-level monitoring, not just per-message filtering.

HOW IT WORKS

01 Why crescendo attacks matter

Most guardrails score one message at a time, so a slow escalation slips through while every individual turn looks acceptable. The attack needs no special access and works across models. It is effective precisely because a model's drive to be helpful and consistent is turned against its safety training over the course of a conversation.

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/crescendo-attack

Open online