AYI NEDJIMI Consultants

Multi-Turn Jailbreaks 2026 : Crescendo & Ske

16 mai 2026 • Mis à jour le 17 mai 2026 • 20 min de lecture • 3690 mots •

Les jailbreaks multi-tour (Crescendo, Skeleton Key, Many-Shot) exploitent contexte des LLM modernes. ASR 80-95%, defenses limitees.

À RETENIR

A retenir — Multi-Turn Jailbreaks

Crescendo (Russinovich et al., Microsoft 2024) construit une escalade en 5 **87%** sur GPT-5.

Skeleton Key (Russinovich, 2024) annule l'alignement par injection de meta ASR **92%** sur GPT-4o avant patch.

Many-Shot Jailbreaking (Anil et al., Anthropic 2024) exploite le long-conte tokens) : 256 exemples in-context defont l'alignement.

L'alignement single-turn ne couvre **pas** les at sur N tours classifier de session (cumulative risk score).

> Sur agents autonomes (LangChain, CrewAI), un jailbreak multi-turn declenc
> d'actions outils — impact business critique.

Le **multi-turn jailbreak llm** est aujourd'hui le vecteur d'attaque le plus sous-estime
Pendant que la litterature academique se concentre sur les attaques single-turn (G
adversarial), trois familles d'attaques conversationnelles (*Crescendo*, *Skeleton Key*
Jailbreaking) atteignent 80 a 95% de taux de succes sur les LLM 2026. La raison
l'alignement RLHF est entraine sur des dialogues majoritairement courts, et les fen
modernes (200k-2M tokens) ouvrent un espace d'attaque que les classifiers singl
pas. Cet article presente la mecanique, le code Python d'exploitation, et les defens
reellement en production. Pour les RSSI deployant un chatbot client ou un agent a
ces attaques sont la priorite reglementaire et operationnelle. L'AI Act article 15 imp
documentation des tests multi-turn.

1. Genese et etat de l'art

L'idee que la conversation longue affaiblit l'alignement remonte aux observations
sur ChatGPT (jailbreaks "DAN" en plusieurs etapes, role-play prolonge). La premie
academique vient d'Anthropic en avril 2024 : Anil et al. publient *Many-shot Jailbre*
qu'en injectant 256 exemples de Q&A "non alignes" en in-context, on degrade sys
l'alignement de Claude, GPT-4, Gemini.

En aout 2024, Mark Russinovich (Microsoft) publie deux articles fondateurs :

- **Crescendo**, une attaque graduelle qui demarre par une question benigne et esc
  vers la requete malveillante.

Réponse sous 24h

Devis gratuit →