What is token smuggling?

Token smuggling gets a forbidden word or instruction past an input filter by changing how it is tokenised: splitting it across pieces, substituting variables, or encoding it. The filter, which matches on the surface text, never sees the banned token, while the model reconstructs the original word from the pieces. It overlaps payload splitting and obfuscation. The defence is to normalise and decode input before filtering and to classify intent on the reconstructed text.

HOW IT WORKS

01 Why token smuggling matters

Any defence that blocks specific words or phrases is vulnerable, because there are many ways to write the same word that a model still understands. Token smuggling is a core building block of encoded jailbreaks and is frequently combined with payload splitting and obfuscation. Tools that bundle hundreds of text transforms make trying many smuggling forms fast.

SOURCES

[1] OWASP LLM01:2025, Prompt Injection

[2] OWASP Top 10 for LLM Applications

[3] MITRE ATLAS, Adversarial ML tactics

[4] Parseltongue, text transform and promptcrafting tool

Test whether your AI guardrails survive real attacks.

securelayer7.net/learn/ai-security/token-smuggling

Open online