# What is payload splitting?

Payload splitting breaks a blocked request into pieces that each look harmless, then asks the model to concatenate them and act on the result. A filter that inspects the visible input sees only benign fragments, while the model assembles the real instruction. It is a smuggling technique closely related to token smuggling and obfuscation. The defence is to evaluate the assembled, resolved text rather than the raw fragments.

## HOW IT WORKS

### 01 Why payload splitting matters

It defeats input filters that judge text as written, which is most of them. Because the harmful content only exists after the model joins the pieces, static inspection of the prompt never sees it. The same trick is used to smuggle instructions into RAG documents and agent inputs, where fragments can be spread across a page or a file.

## 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/payload-splitting

Open online