What is Unicode tag smuggling?

Unicode tag smuggling hides an instruction using characters from the Unicode Tag block (U+E0000 to U+E007F), which mirror ASCII but render invisibly. A human sees normal text; the model reads the hidden characters and can act on them. It is a delivery method for invisible prompt injection, especially through documents, web pages, and tickets that feed RAG or agents. The defence is to strip Tag and other invisible characters from input before it reaches the model.

HOW IT WORKS

01 Why Unicode tag smuggling matters

Invisible payloads defeat human review, which is often the last line of defence for content that flows into an AI system. Because the instruction hides inside ordinary text, it rides indirect prompt injection through a support ticket, a shared document, a product review, or a web page, and reaches a RAG index or an agent without anyone noticing. It also complicates incident response, since the malicious text is not visible in logs viewed normally.

SOURCES

[1] OWASP LLM01:2025, Prompt Injection

[2] OWASP Top 10 for LLM Applications

[3] MITRE ATLAS, Adversarial ML tactics

[4] Unicode Tags block (U+E0000 to U+E007F)

Test whether your AI guardrails survive real attacks.

securelayer7.net/learn/ai-security/unicode-tag-smuggling

Open online