What is context flooding?

A context flooding attack pads the model's context with a large volume of text to dilute or push out the safety instructions, then places the payload where it carries the most weight. It exploits the fact that instructions compete for the model's attention and that recency and position matter. It is related to many-shot jailbreaking, which floods the context with compliant examples specifically. The defence is to pin critical instructions, use structured prompts, and limit untrusted input size.

HOW IT WORKS

01 Why context flooding matters

Larger context windows make the attack easier, because there is more room to bury the guardrails. It needs no special access, just a long input, and it can be combined with many-shot examples or hidden instructions for more effect. It shows that placing safety rules in the prompt is not enough on its own when an attacker controls most of the surrounding text.

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/context-flooding

Open online