What is a homoglyph attack?

A homoglyph attack replaces characters with look-alike Unicode ones, for example a Cyrillic a for a Latin a, or full-width forms, so text reads normally to a human and a model but does not match a keyword filter. In LLM security it is used to smuggle blocked words past input filters and to disguise injected instructions. It is also a classic phishing and spoofing trick. The defence is Unicode normalisation and homoglyph mapping before any matching.

HOW IT WORKS

01 Why homoglyph attacks matter

The same word can be written in a huge number of visually identical ways, so any exact-match blocklist is easy to evade. In AI systems homoglyphs smuggle forbidden tokens and hide injected instructions inside otherwise normal text. Outside AI, homoglyphs power lookalike domains and spoofed identifiers, so the technique is well understood and widely available.

SOURCES

[1] OWASP LLM01:2025, Prompt Injection

[2] OWASP Top 10 for LLM Applications

[3] MITRE ATLAS, Adversarial ML tactics

[4] Unicode confusables and security (UTS #39)

Test whether your AI guardrails survive real attacks.

securelayer7.net/learn/ai-security/homoglyph-attack

Open online