Prompt Injection

prompt-injection · safety · concept

Source: https://policywindow.org/wiki/prompt-injection

Generated 2026-07-09T20:35:27 UTC

Summary

An adversarial input technique in which untrusted content fed to an AI model (e.g., text on a webpage the model reads, a document the user uploads, a tool's output) contains instructions that override the model's intended behaviour or principal-provided system prompt.

At a glance

Used by

2 instrument(s)

Related concepts

agentic-system, tool-use-safety, jailbreak-resistance, data-poisoning, retrieval-augmented-generation

Primary source

Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., Fritz, M. (2023), 'Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection.'

Details

Prompt injection was named by Willison (2022, 'Prompt injection attacks against GPT-3') and formalised by Greshake et al. (2023, 'Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection'). The attack class splits into two sub-cases: (a) direct prompt injection — the user (or attacker posing as user) submits adversarial text in the prompt; mitigated partly by training-time alignment + system-prompt design; (b) indirect prompt injection — the model ingests untrusted content (a webpage during browsing, a PDF the user uploads, the output of a tool call) which contains adversarial instructions; the model cannot reliably distinguish 'data' from 'instructions' because both share the same token-stream interface. Indirect injection is the more serious failure mode at deployment because the attacker doesn't need access to the user's session.

NIST AI RMF GenAI Profile (NIST AI 600-1) names prompt injection in the 'Information Security' risk category. EU AI Act Art. 15 ('cybersecurity' requirement for high-risk and Art. 55 for GPAI with systemic risk) is the closest binding obligation — providers must protect against 'attempts by unauthorised third parties to alter the use, behaviour or performance of the system.' Industry mitigations (constitutional classifiers, dual-LLM gateway patterns, content-isolation tags) are evolving rapidly but no architectural defence is yet known to be robust. The OWASP LLM Top 10 (2023, 2025 update) lists prompt injection as LLM01 — the most-cited application-security risk for LLM-integrated software.

How to cite this article

APA

Policy Window. (n.d.). Prompt Injection [Wiki article — Concept]. https://policywindow.org/wiki/prompt-injection

CHICAGO

Policy Window. n.d.. "Prompt Injection." Wiki article (Concept). https://policywindow.org/wiki/prompt-injection.

HARVARD

Policy Window (n.d.) 'Prompt Injection', Wiki article — Concept, available at: https://policywindow.org/wiki/prompt-injection.

OSCOLA

Policy Window, 'Prompt Injection' (Wiki article — Concept, n.d.) <https://policywindow.org/wiki/prompt-injection> accessed [date].

BIBTEX

```
@misc{policywindow-prompt-injection,
title  = {Prompt Injection},
author = {Policy Window},
year   = {n.d.},
howpublished = {prompt-injection — safety},
url    = {https://policywindow.org/wiki/prompt-injection},
note   = {Primary source: https://arxiv.org/abs/2302.12173}
}
```