



Indirect Prompt Injection.

Demonstrated. Documented in 18 separate real-world case studies. And classified under Execution, not Initial Access.



Hidden instructions, hiding in plain sight.

An adversary injects prompts indirectly through a separate data channel an LLM ingests — text or multimedia pulled from a database or website. The malicious prompt can be hidden or obfuscated from the user entirely.



It's classified as Execution. Here's why.

The payload isn't what gets an adversary in the door — it's what runs once a legitimate user or agent retrieves the poisoned content. A summary request, a calendar lookup, a routine ticket triage: the trigger is ordinary. What executes next isn't.



Four documented incidents

Slack AI (PromptArmor)

A user asks Slack AI for an API key. Slack AI retrieves poisoned content and executes the embedded instructions instead.

M365 Copilot (Zenity)

A poisoned RAG entry hijacks a financial transaction the moment a routine search retrieves it — Copilot acting as an unwitting insider.

ChatGPT Memory (Embrace the Red)

Referencing a shared document adds its contents to chat context — and the embedded prompt executes as part of the model's own memory.



Four more, just as recent

Claude Computer-Use (HiddenLayer)

A user asks Claude to interact with a PDF. The embedded prompt executes — resulting in data destruction.

Claude Code GitHub Action (Microsoft Defender)

Malicious GitHub content enters the model's context and is interpreted as legitimate instructions — secrets exposed as a result.

ZombieAgent (Radware)

A later, ordinary request — “summarize my inbox” — causes ChatGPT to retrieve and execute concealed instructions the user never saw.



Three real mitigations

AI Telemetry Logging (AML.M0024)

Logging helps identify whether unsafe prompts were ever submitted to the LLM in the first place.

Input/Output Validation (AML.M0033)

Validation can stop a prompt injection from ever affecting the steps an agentic workflow takes next.

AI Red Team (AML.M0035)

Plant controlled malicious instructions in retrieved content on purpose — then harden trust boundaries and retrieval controls based on what breaks.



Framework-grounded, not guesswork.

INTERRO AI maps investigations against ATLAS alongside ATT&CK and D3FEND, live across whatever SIEM you run — keeping AI-native techniques like this one inside the same workflow as everything else your team already watches.



See it mapped live.

Get Your Free Demo Key → InterroAI.com/demo

No credit card. No sales call.