What is LLM data exfiltration?

LLM data exfiltration is using prompt injection to make a model leak sensitive data it has access to, such as the system prompt, retrieved documents, conversation history, or tool output. The stolen data is often carried out through an output channel the model can render, for example a markdown image or link whose URL encodes the secret and points at an attacker server. The defence is egress control and output sanitisation: block auto-loading of external content, allowlist link and image destinations, and apply least privilege to the data the model can reach.

HOW IT WORKS

01 Why LLM data exfiltration matters

This turns a prompt injection into a real breach. RAG systems and agents give models access to internal documents, credentials, and tools, and exfiltration is how that access leaves the building. Because the payload can arrive indirectly and the exfil channel is a normal-looking image or link, both the user and the logs can miss it. It is one of the clearest paths from an AI feature to data loss.

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/llm-data-exfiltration

Open online