What is sensitive information disclosure?

Sensitive information disclosure (OWASP LLM02:2025) is when an LLM application reveals data it should keep private: memorised training data (PII, secrets, proprietary text), information from its system prompt or context, or data belonging to another user or tenant. It happens through normal or adversarial prompting, and the leaked data leaves through the model's output. Defend by minimising and sanitising what the model can access, filtering outputs, and enforcing authorization outside the prompt.

HOW IT WORKS

01 How the attack works

Attackers extract memorised data with prompts that ask the model to repeat or complete sensitive strings, or that use divergence and repetition tricks to dump training data. They read the system prompt and hidden context through system prompt extraction, and in multi-tenant or shared-memory apps they craft prompts to surface another user's data. Even without an attacker, a model can volunteer secrets from its context in a normal answer. Shown for defensive testing.

SOURCES

[1] OWASP Top 10 for LLM Applications (2025)

[2] Carlini et al., Extracting Training Data from Large Language Models

Test your AI application before an attacker does.

securelayer7.net/learn/ai-security/sensitive-information-disclosure

Open online