What is excessive agency?

Excessive agency (OWASP LLM06:2025) is the risk that an LLM-based agent can take harmful actions because it has more functionality, permissions, or autonomy than it needs. When an attacker influences the model through prompt injection or a poisoned data source, the agent uses that excess capability, deleting records, sending email, calling paid APIs, to act on the attacker's behalf. The fix is least privilege: minimal tools, scoped permissions, and human approval for high-impact actions.

HOW IT WORKS

01 How the attack works

The attacker does not exploit a memory bug, they manipulate the model. Through direct prompt injection or an indirect payload in retrieved content (a web page, a document, an email the agent reads) they instruct the agent to use its tools against you: forward the latest invoices to an external address, run a shell command, grant a user admin. If the agent holds the credential and the tool, it complies. Excess permission turns a text bug into real-world impact. Shown for defensive testing.

SOURCES

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

[2] MITRE ATLAS (adversarial threat landscape for AI systems)

Test your AI application before an attacker does.

securelayer7.net/learn/ai-security/excessive-agency

Open online