What is a denial of wallet attack?

A denial of wallet attack is a form of unbounded consumption (OWASP LLM10:2025). Because every LLM call costs money and compute, an attacker sends inputs crafted to maximise token usage, very long prompts, requests for long outputs, recursive or looping agent tasks, until the app hits its rate limits, degrades for real users, or runs up an unsustainable bill. Defend by capping input and output tokens, rate-limiting per user, budgeting spend, and detecting abnormal consumption.

HOW IT WORKS

01 How the attack works

The attacker maximises cost per request and volume of requests. Techniques include very large inputs that pad the context to its maximum, prompts that ask for the longest possible output, inputs that trigger expensive tool calls or retrieval, and agent tasks that loop or fan out into many sub-calls. Automated across many requests or accounts, this exhausts rate limits and quota and drives the bill up fast. On a metered API with auto-scaling and no cap, the cost is effectively unbounded. 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/denial-of-wallet

Open online