

What is AgentDojo?

AgentDojo is an open benchmark that measures how easily an AI agent can be hijacked by prompt injection. The agent works through realistic tasks in an email and calendar workspace, Slack, a travel booking system and a bank, using tools whose output contains planted attacker instructions. AgentDojo then checks the environment's actual state to see whether the agent completed the user's task, the attacker's task, or both. The original paper ships 97 tasks and 629 security test cases.

HOW IT WORKS

01 How AgentDojo tests an agent

Each suite pairs user tasks (what you asked for) with injection tasks (what the attacker wants). In the Inspect version the suites are Workspace Plus (42 user tasks, 15 injection tasks), Slack (20 and 5), Travel (20 and 7) and Banking (16 and 9).

A typical case: the user asks the agent to pay a bill stored in a file. The file, or a transaction note the agent reads along the way, contains text claiming to override earlier instructions and asking the agent to send money to an attacker's IBAN. The default attack, `important_instructions`, is exactly that kind of high-priority override message.

Run it with `inspect eval inspect_evals/agentdojo -T workspace=banking`, or add `-T with_injections=false` to measure the agent with no attacks at all.

WHY IT DOES NOT USE AN LLM JUDGE

Success is decided by checking the environment: was the email sent, was the file deleted, did the transfer go through. A judge model could be fooled by the same injection it is meant to catch.

02 What the scores tell you

AgentDojo reports three numbers. Benign utility is how many tasks the agent completes with no attack present. Utility under attack is how many it still completes when injections are present, which captures attacks that simply derail the agent. Targeted attack success rate is how often the agent carried out the attacker's goal.

The paper's headline finding was that the benchmark is hard for both sides: strong models failed many tasks even without an attack, and existing injection attacks broke some security

SOURCES

- [1] AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents
- [2] AgentDojo in Inspect Evals (task docs)
- [3] OWASP Top 10 for LLM Applications (2025)
- [4] Inspect Evals registry: Safeguards evaluations

properties but not all. A low attack success rate on AgentDojo does not mean your agent is safe; it means it resisted these attacks, in these apps.

Test your AI application before an attacker does.

securelayer7.net/learn/ai-security/agentdojo

[Open online](https://securelayer7.net/learn/ai-security/agentdojo)