

What is AgentHarm?

AgentHarm is a benchmark that measures how readily an AI agent carries out explicitly malicious tasks, such as fraud, cybercrime and harassment, when it has tools and can act over several steps. The paper's set has 110 base tasks (440 with variations) across 11 harm categories. Scoring well requires two things: refusing harmful requests, and showing that a jailbroken agent would not still complete the job competently. Its authors found leading models were surprisingly compliant even without a jailbreak.

HOW IT WORKS

01 How AgentHarm tests an agent

The public Inspect release contains 44 of the 66 public test base behaviours, expanded to 176 samples, plus a validation split. Variants change whether the request is detailed or open-ended and whether it includes a hint about which tool to use.

Two tasks run side by side. `inspect eval inspect_evals/agentharm` runs the harmful set. `inspect eval inspect_evals/agentharm_benign` runs harmless versions of the same tasks, so you can tell a safe refusal apart from an agent that simply cannot use the tools. Judge models grade refusals and check whether the agent's tool calls achieved the task.

BUILT FOR DEFENDERS ONLY

The dataset licence adds a clause limiting its use to improving the safety and security of AI systems, and it carries a canary string so training-data contamination can be detected.

02 What the results tell you

The paper reported three findings. Leading models were surprisingly willing to carry out malicious agent requests with no jailbreak at all. Simple, universal jailbreak templates adapted well to agents. And jailbroken agents stayed coherent, completing multi-step harmful tasks rather than falling apart.

That last point is the reason AgentHarm exists. A refusal rate alone can hide an agent that, once talked around, is fully capable. The authors are also explicit that the score is not an overall measure of agent safety.

SOURCES

- [1] AgentHarm: A Benchmark for Measuring Harmfulness of LLM Agents
- [2] AgentHarm in Inspect Evals (task docs)
- [3] AgentHarm dataset
- [4] OWASP Top 10 for Agentic Applications (2026)

Test your AI application before an attacker does.

securelayer7.net/learn/ai-security/agentharm

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