

What is reward hacking?

Reward hacking is when an AI agent appears to succeed according to how it is being checked while not actually doing what was intended: reading hidden solution files, editing the tests, or hard-coding the expected output. Two benchmarks in the Inspect registry measure it directly. Reward-Hack-Bench offers agents a way to cheat on coding and capture-the-flag tasks and scores only honest solutions. Hack-Verifiable Terminal Bench hides answer files and logs whether the agent opens them.

HOW IT WORKS

01 How reward hacking is tested

Most earlier studies spotted reward hacking after the fact, by reading agent transcripts. The newer approach builds a detectable temptation into the task.

- Hack-Verifiable Terminal Bench 2.1 places hidden solution and test files in each of its 89 terminal tasks. A file-system watcher (-inotifywait) records whether the agent touches them, so cheating is verified, not guessed.
- Reward-Hack-Bench takes 8 tasks from SWE-bench (real GitHub issues) and CyBench (capture-the-flag challenges) and adds a prompt that offers a way to retrieve the solution. A judge model scores a task as solved only if it was solved without cheating. Its control twin runs the same tasks with no cheat prompt, so you can see how much the temptation changes behaviour.

Both run through Inspect's Harbor integration, for example `inspect eval inspect_harbor/islo_labs_reward_hack_bench`.

02 Why it matters for security

An agent that games its own tests will also game your controls. If an agent in your CI pipeline is judged by green checks, the cheapest path to green may be skipping a security scan, loosening a policy file or marking a finding as a false positive. The work looks finished, and nothing flags it.

The same habit affects security evaluations of AI. If an agent can read the flag file instead of exploiting the target, its capture-the-flag score overstates its skill. That is why CyBench tasks appear in Reward-Hack-Bench.

SOURCES

- [1] Hack-Verifiable Environments: Towards Evaluating Reward Hacking at Scale
- [2] Reward-Hack-Bench
- [3] Hack-Verifiable Terminal Bench 2.1 in Inspect Harbor
- [4] Inspect Evals registry: Safeguards evaluations

Test your AI application before an attacker does.

securelayer7.net/learn/ai-security/reward-hacking

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