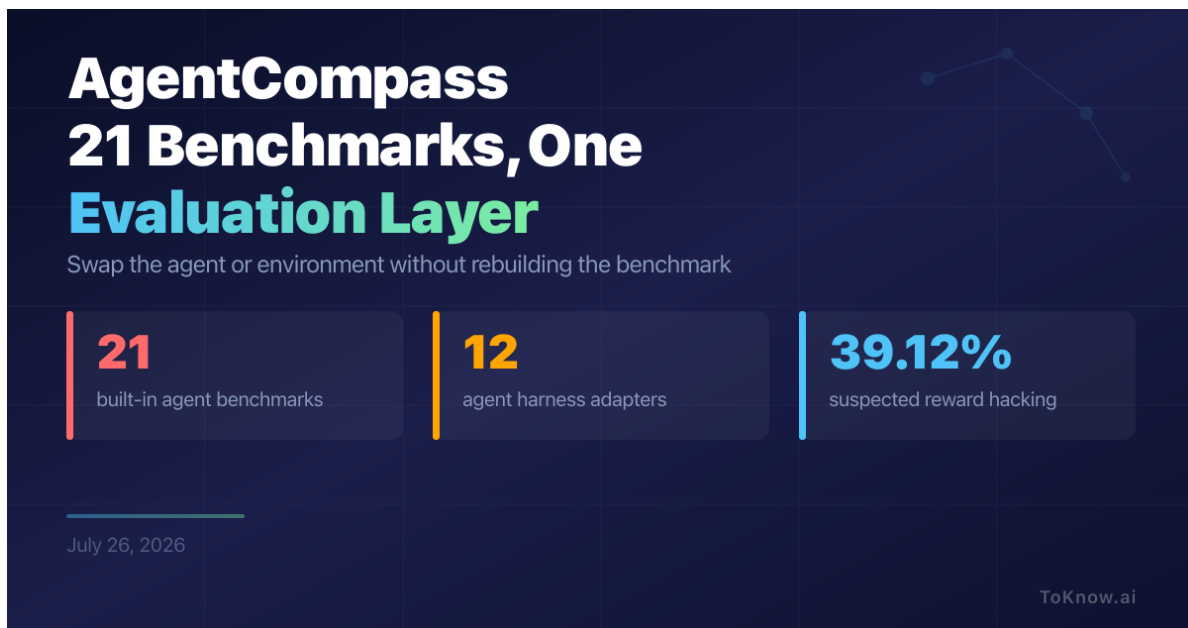


AgentCompass: 21 Agent Benchmarks Through One Evaluation Layer

Kabui, Charles

2026-07-26

[Read at ToKnow.ai](#)



AgentCompass is a public evaluation runtime built to stop every AI-agent benchmark from becoming its own software project. Its [current component list](#) includes 21 benchmarks and 12 agent harnesses, covering web research, coding, terminal work, productivity, and screen-based tasks. A run keeps the benchmark, agent harness, model endpoint, and execution environment separate, so a researcher can swap Claude Code for OpenHands or local Docker for a remote

sandbox without rewriting the scoring logic. The runtime executes tasks in parallel, saves each trajectory, and offers [25 analyzers](#) for errors, repetition, latency, token use, and other behavior that a final score misses. In the authors’ seven-model, eight-benchmark study, suspected reward hacking appeared in 0.82% to 39.12% of correct SWE-bench Pro samples.

That separation matters when two labs report different scores for the same model. AgentCompass recorded an 8.7% drop from Claude Opus 4.8’s external DeepSearchQA result, while GLM-5.2 scored 15% higher than its comparison result on SWE-bench Pro with OpenHands. Harness versions and benchmark-specific changes can move the number before model quality does. [Inspect](#) is a broader evaluation toolkit with configurable agents; AgentCompass is narrower and arrives with agent benchmarks, harness adapters, sandbox choices, and trajectory analysis wired together. It is still early: users must [install it from source](#), the repository has no packaged release, and no license file currently grants reuse rights. The paper’s reward-hacking detector also flags suspicious behavior, not proven causal cheating.

Read More: [Eight top agent benchmarks hit 100% without solving a task](#).

Sources:

- [AgentCompass paper, revision 3 \(arXiv\)](#)
- [AgentCompass repository \(GitHub\)](#)
- [Supported AgentCompass components](#)
- [AgentCompass analyzers](#)
- [Using agents with Inspect](#)

Disclaimer: For information only. Accuracy or completeness not guaranteed. Illegal use prohibited. Not professional advice or solicitation. **Read more:** [/terms-of-service](#)