Harness²: Google Research Code That Rewrites an AI Agent's Instructions and Tools

Kabui, Charles

2026-11-03

Read at ToKnow.ai



Harness² is research code from Google Cloud AI Research and Northwestern University that rewrites the setup around an AI model, for researchers and teams that build AI agents. That setup, called the harness, is the instructions, tools and memory that turn a model into a working agent. Harness² edits eight parts of it, such as the prompt, memory and helper agents, and never retrains the model. With no answer key, it runs a task, tries two edited setups,

compares how the agent behaved, and reruns with the best edits. In the authors' tests, GLM-5's average share of checks passed on WorkBuddy-Bench, which covers coding, web and office tasks, rose from 64.2% to 75.7%.

That edited GLM-5 beat Gemini 3.5 Flash, which passed 72.4% with its original setup, and three rival tuning methods. On JobBench, which grades professional work against expert rubrics, one agent never cross-checked its sources or reopened its outputs. The edits added both checks, lifting that task's score from 43.1% to 77.8%. The authors' best one-step gains were 4.5% on JobBench and 11.8% on LAB, Harvey's benchmark of legal work.

The authors ran every test themselves. Claude Opus 4.8 wrote the edits in the main tests, and all five models improved when each edited its own setup. A single step runs the task four times, plus editing sessions. Without an answer key, the final choice once picked a coding output scoring 55% over one scoring 95%.

Read More: OpenAI's Agents API, which rents out a ready-made agent setup and DeepSeek's open-source agent where every part is a plugin

Sources:

- Harness²: Recursive Agent Harnessing for an Open World (arXiv)
- google-research/harness2: code and README (GitHub)
- Tencent WorkBuddy Bench: A Multi-Domain Coding-Agent Benchmark (arXiv)
- JobBench: Aligning Agent Work With Human Will (arXiv)
- Introducing Harvey's Legal Agent Benchmark (Harvey)

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