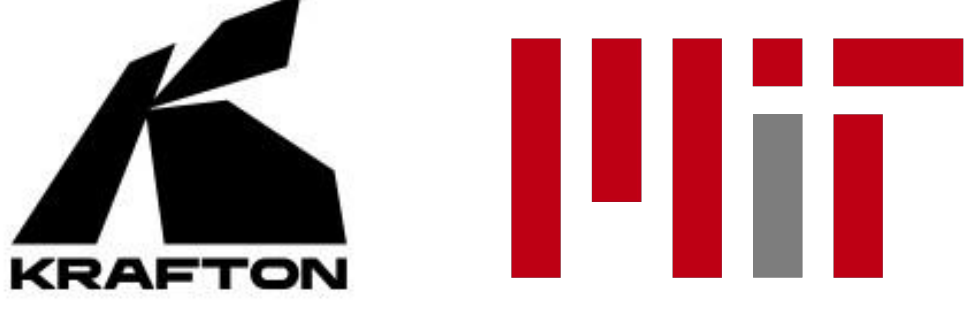
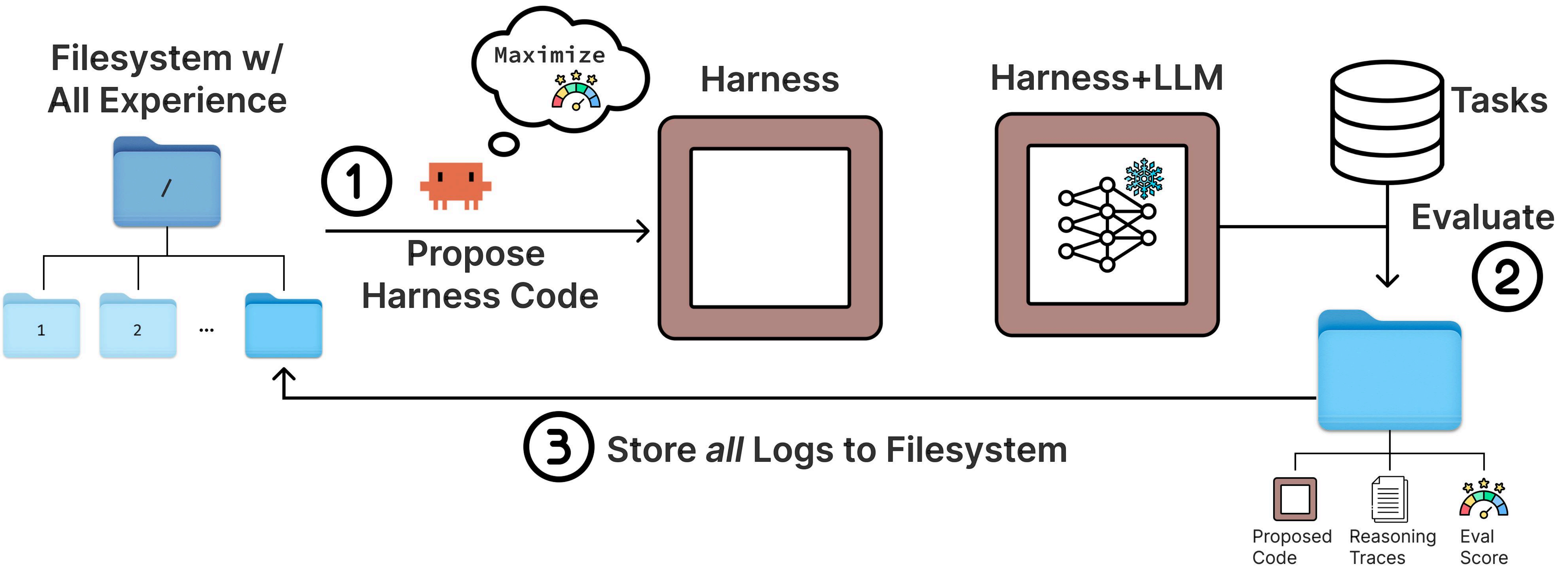


Meta-Harness: End-to-End Optimization of Model Harnesses



Yoonho Lee, Roshen Sanjay Nair, Qizheng Zhang, Kangwook Lee, Omar Khattab, Chelsea Finn



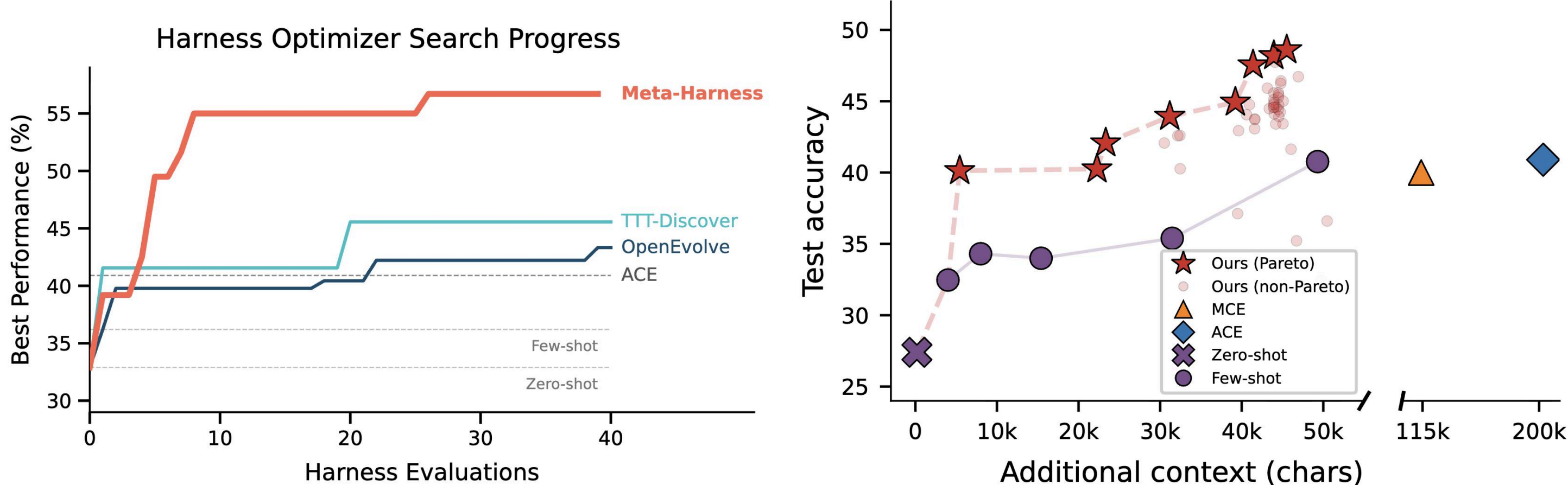
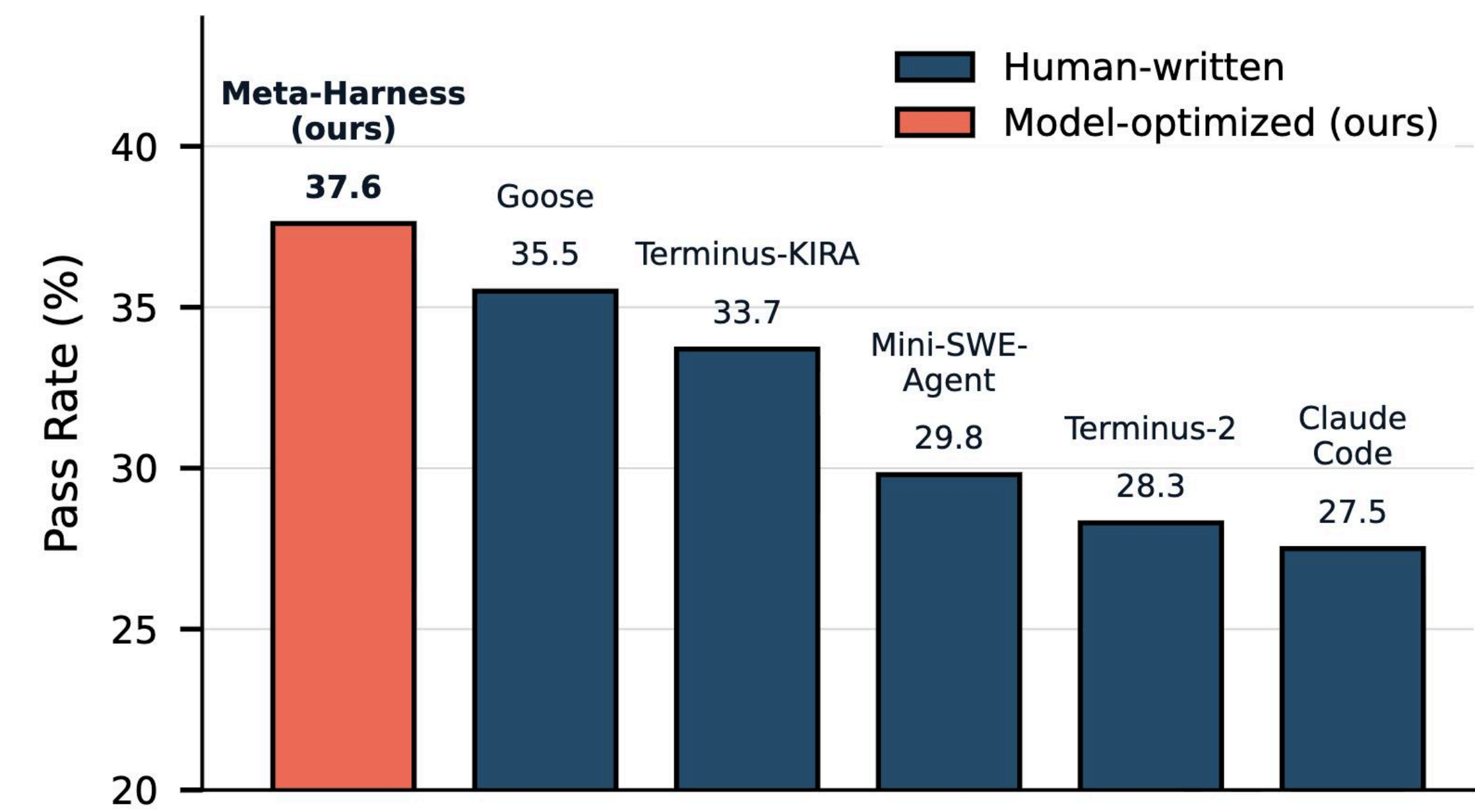
Motivation

- **Harnesses matter for performance.** They can swing a benchmark 6x!
- Harness = the code layer around an LLM. Prompt, workflow, ctx management...
- Harness engineering has mostly been manual; can we *automate* it?

Method	History	Log content	MTok/iter
OPRO	Window	past (solution, score) pairs	0.002
TextGrad	Last	textual feedback on current artifact	0.015
AlphaEvolve	Window	program database + eval. scores	0.022
GEPA	Summary	reflective feedback from rollout traces	0.008
Feedback Descent	Summary	comparison + textual feedback	0.012
TTT-Discover	Window	prev. solution fragment	0.026
Meta-Harness	Full	<i>all</i> logs and scores	10.0

- Previous text optimizers were designed for short-horizon settings. Harness optimization is very often long-horizon.
- Meta-Harness = one design choice: put **all experience in a filesystem**, then let the agent decide what to look at!

TerminalBench-2 Harness Performance



Text classification. Meta-Harness:

- (1) outperforms prior optimizers
- (2) finds several Pareto-efficient harnesses!

Takeaways

- A harness is “just” code; we can improve it with coding agents. 🤖
- Let the model see raw trajectories and decide what’s important 📁
- Meta-Harness + current coding agents can improve **frontier system capability**

TerminalBench-2. Meta-Harness discovered a harness that is still (Jul 8) **state-of-the-art on the official leaderboard!**

Home > Leaderboards > terminal-bench@2.0

terminal-bench@2.0 Leaderboard

New Model Custom Agent

Note: submissions may not modify timeouts or resources

harbor run -d terminal-bench/terminal-bench-2 -a "agent" -m "model" -k 5

Showing 10 entries

Search leaderboard

Select agents

Claude Opus 4.6

Select...

Verified only

Rank	Agent	Model	Date	Agent Org	Model Org	Accuracy
11	Meta-Harness	Claude Opus 4.6	2026-05-14	Stanford IRIS	Anthropic	76.4% ± 2.4
14	Capy	Claude Opus 4.6	2026-03-12	Capy	Anthropic	75.3% ± 2.4
17	Terminus-KIRA	Claude Opus 4.6	2026-02-22	KRAFTON AI	Anthropic	74.7% ± 2.6
20	TongAgents	Claude Opus 4.6	2026-02-22	Bigai	Anthropic	71.9% ± 2.7
22	Junie CLI	Multiple	2026-03-07	JetBrains	Multiple	71.0% ± 2.9
23	Droid	Claude Opus 4.6	2026-02-05	Factory	Anthropic	69.9% ± 2.5
26	Crux	Claude Opus 4.6	2026-02-23	Roam	Anthropic	66.9% ± N/A

TerminalBench-2. Meta-Harness discovered a harness that is still (Jul 8) **state-of-the-art on the official leaderboard!**