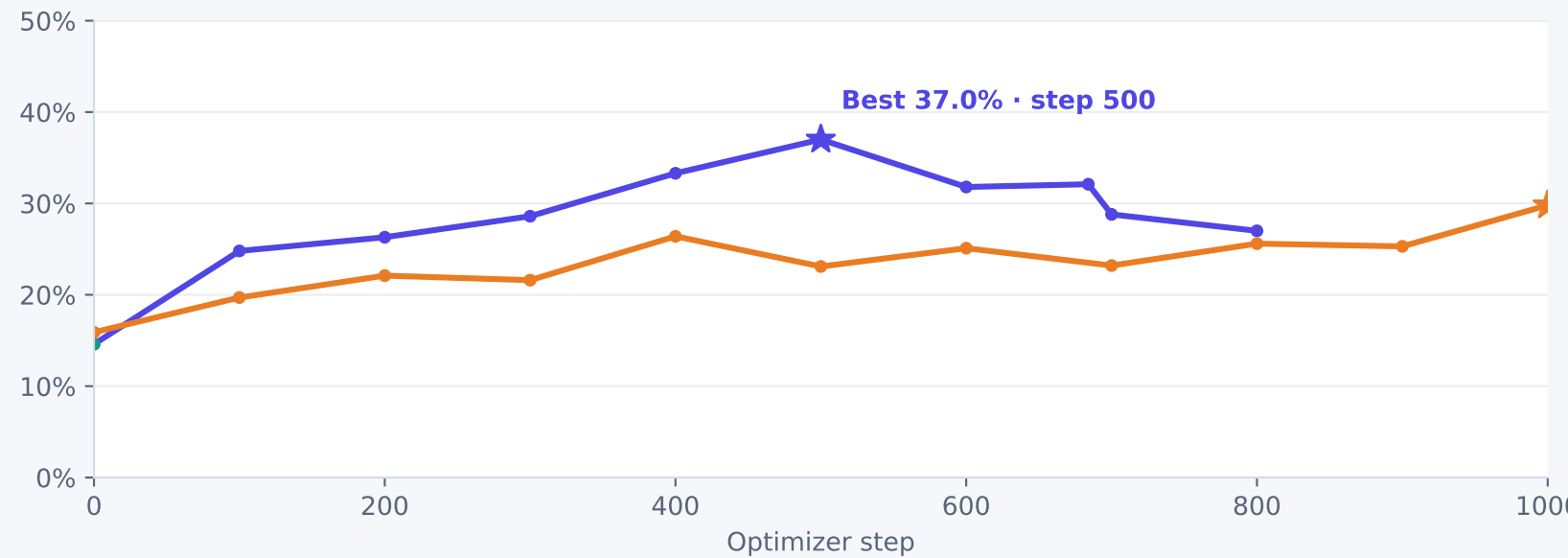


Qwen3.5-2B / Training comparisons

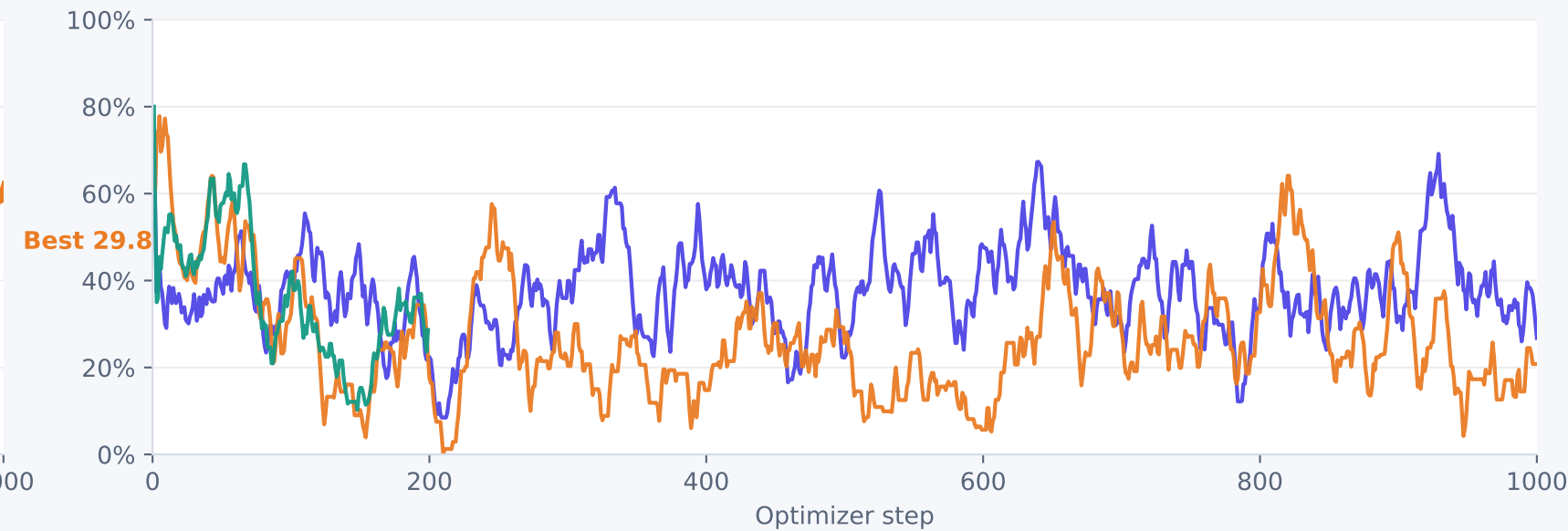
3 runs · 1,000-step targets · recorded training and audited evaluations

- Harbor multi-harness
- Native OpenCode
- Harbor OpenCode-only

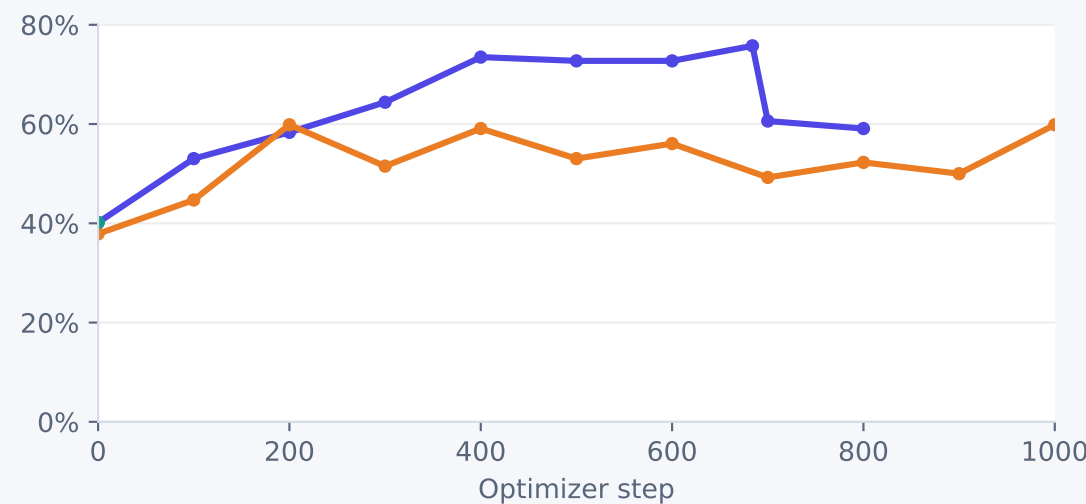
Overall held-out pass@1



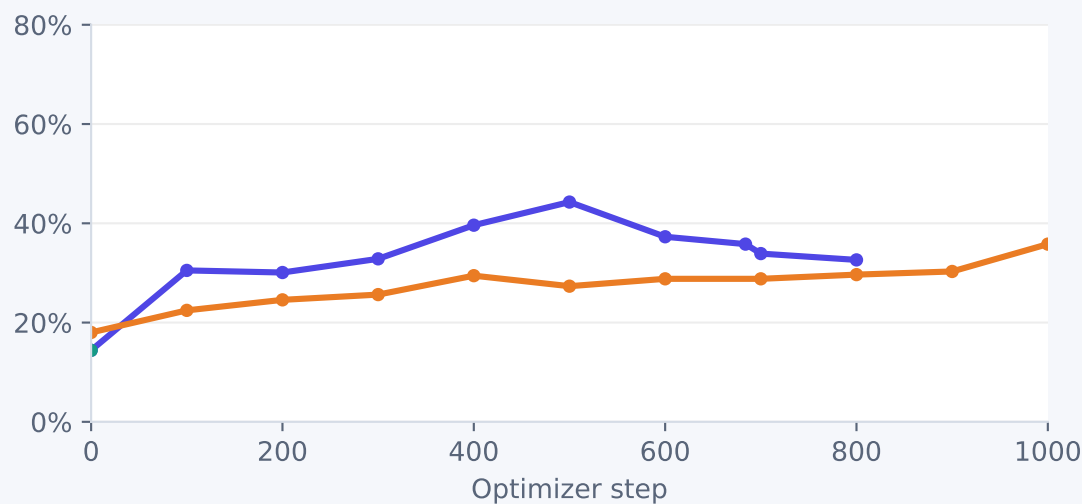
Training reward · trailing 20 updates



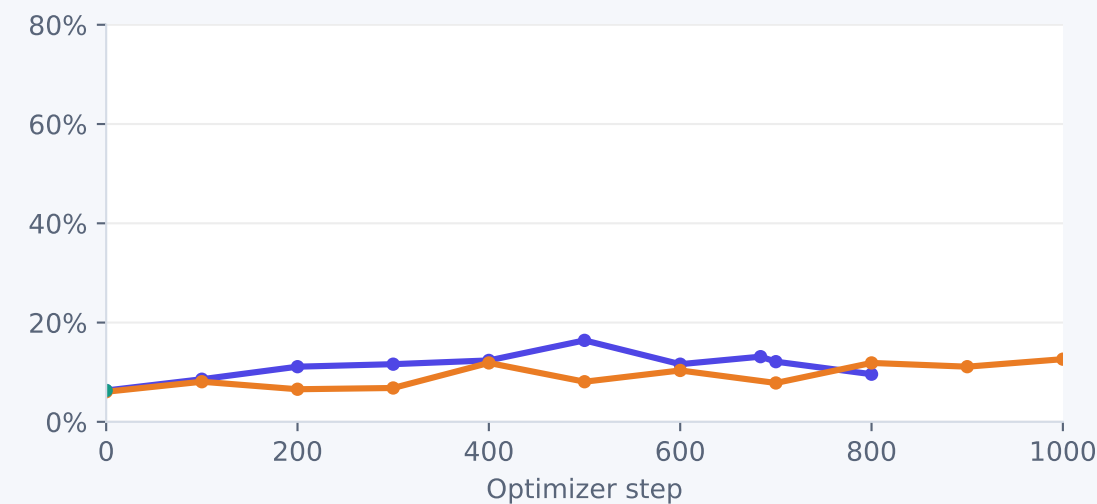
Easy tasks · 33 per harness



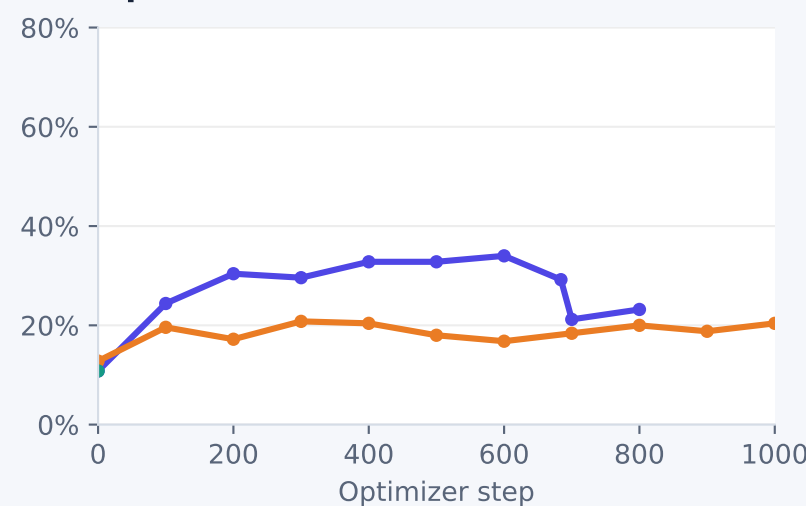
Medium tasks · 118 per harness



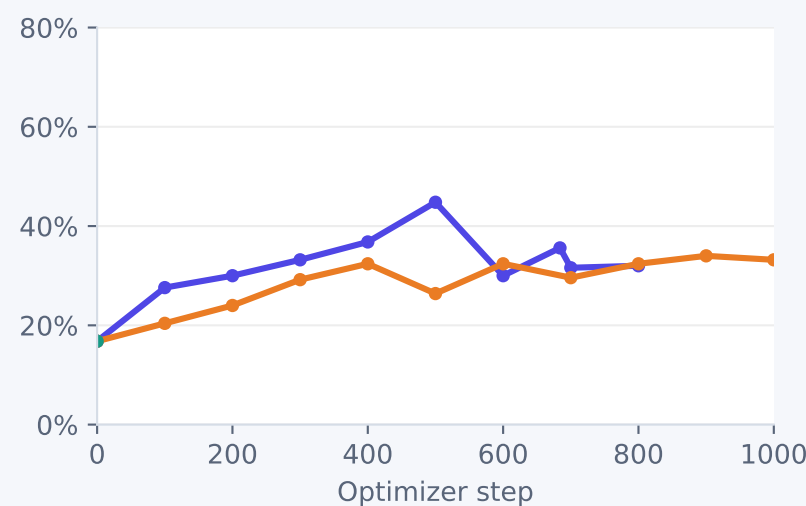
Hard tasks · 99 per harness



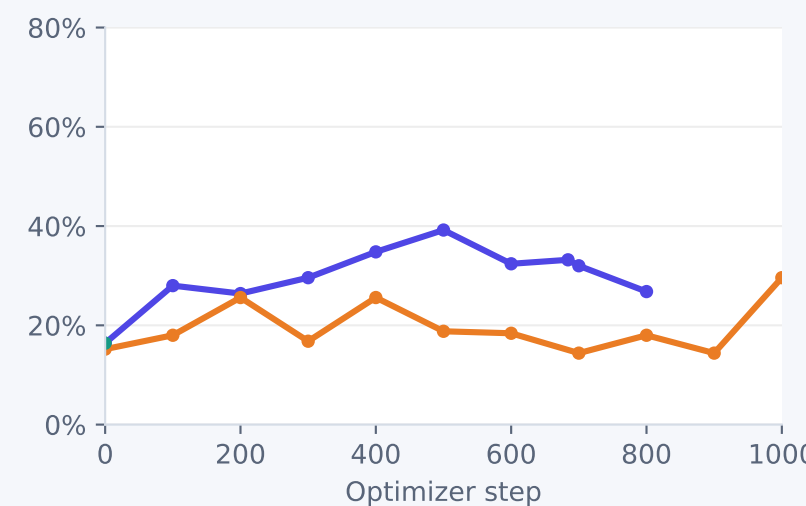
OpenCode



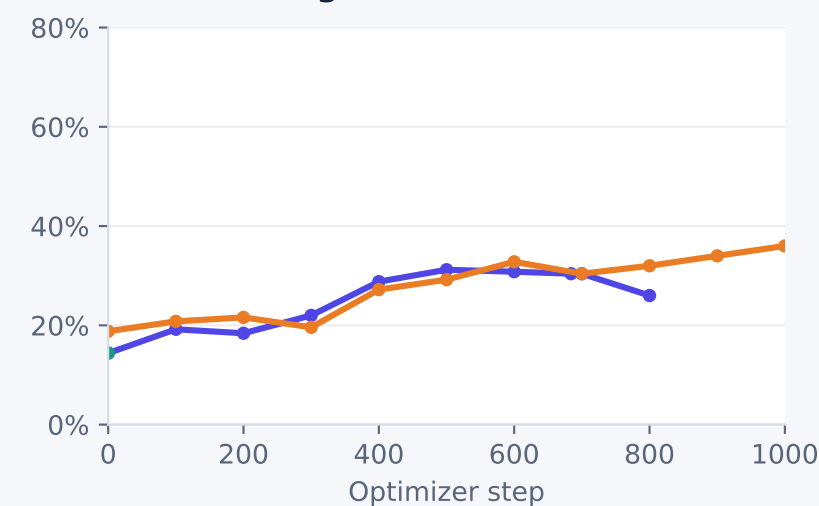
Claude Code



Codex



Mini-SWE-Agent



AUDITED PASS@1 · 250 fixed tests × 4 harnesses · Updated 2026-09-16 14:50 UTC

Separate measured baseline cohorts (E2B / Daytona). Curves stop at the last audited checkpoint; pending evaluations are omitted. Full harness × difficulty tables accompany this figure.