The Scaffold Effect in Coding Agents

Harness Choice as a Hidden Variable in Coding-Agent Evaluation

Naman Vats ▪ **Oleg Golev**

Sentient Labs ▪ be10360.17@bitmesra.ac.in ▪ ogolev@alumni.princeton.edu

github.com/namanvats/scaffold-effects

The Problem

Public leaderboards rank coding agents by **model name**, treating the surrounding **harness**—the scaffold that issues tools, manages context, and decides when to stop—as a controlled constant.

A developer choosing an agent cares about **cost per solved task**, **latency**, and **oversight burden**. *None of these are recovered from a model-only score.*

The hint is not new: Terminal-Bench 2.0 reports the same model at 52.1% vs. 57.8% pass rate across two harnesses, at a **65**× input-token difference.

Our Question

When the **model is held fixed** and **only the harness varies**, how large are the differences a deployer should expect — and what *kind* of differences are they?

Experimental Setup

300 controlled trials:

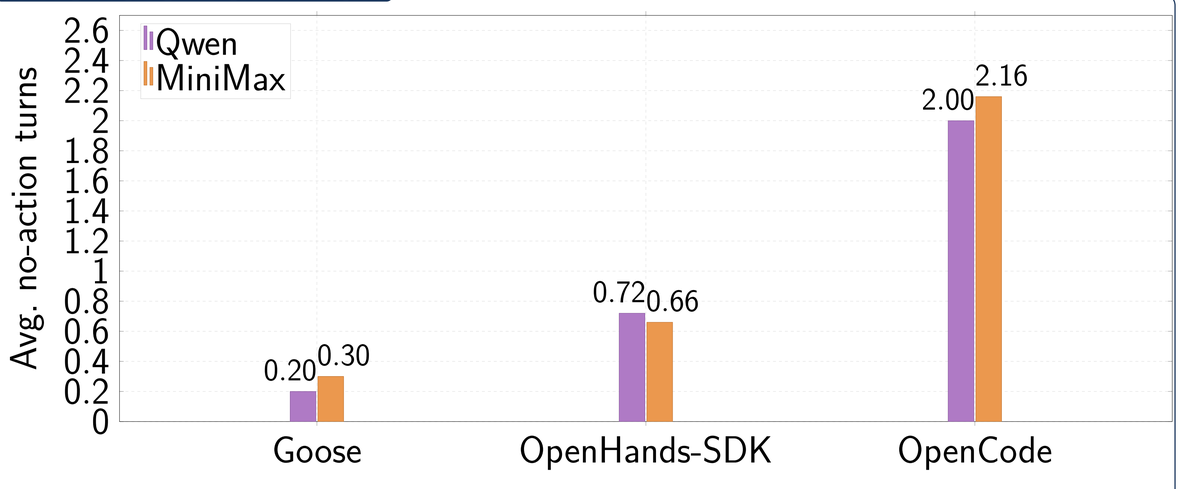
- **3 harnesses** — Goose, OpenCode, OpenHands-SDK
- **2 agentic coding models** — Qwen 3.6 Plus, MiniMax M2.5
- **50 tasks** — stratified subset of Terminal-Bench Pro (8 archetypes)

Standardized protocol: identical tasks, 900 s wall-clock cap, OpenRouter provider defaults, no harness-level retries. Metrics: pass rate, tokens/solved task, latency, failure mode.

Cost per Solved Task

Harness	Model	Tokens/Solve	×Goose
Goose	Qwen	28,142	1.0×
Goose	MiniMax	36,950	1.0×
OpenHands	Qwen	841,201	29.9×
OpenHands	MiniMax	843,286	22.8×
OpenCode	Qwen	1,147,740	40.8×
OpenCode	MiniMax	1,546,977	41.9×

The Idle-Turn Tax



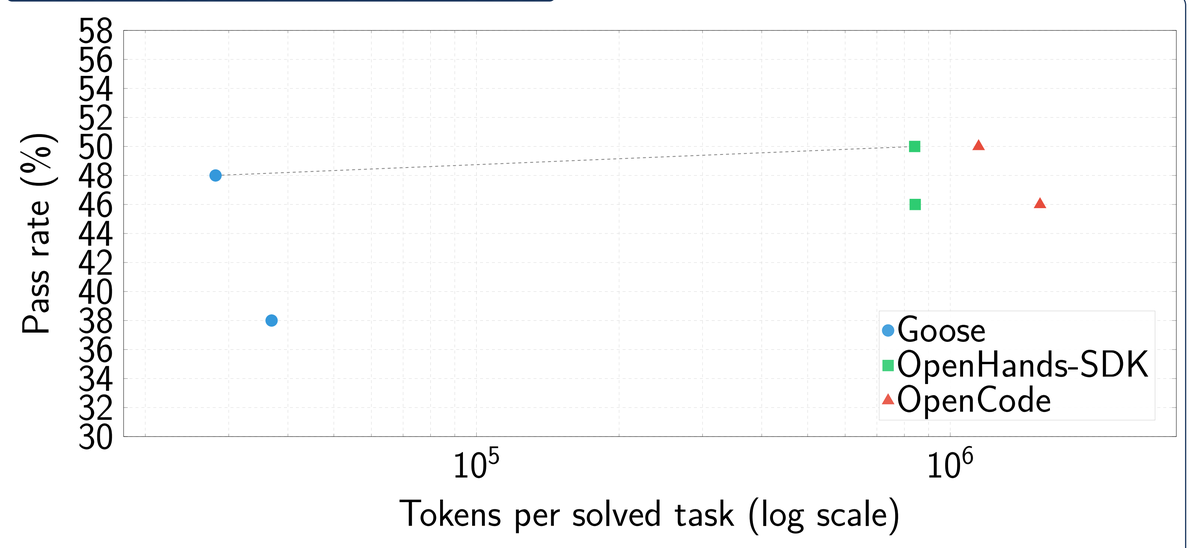
OpenCode averages **10**× more **no-action turns** (turns with no edit and no new command) than Goose. Each carries a full context window ⇒ a per-task **wait tax**, not just a token tax.

40×
gap in tokens per solved task

0–8 pp
paired within-model pass-rate difference

Same task success — two orders of magnitude apart in cost.

Pass Rate vs. Token Cost



Goose dominates the Pareto frontier. OpenCode is Pareto-dominated by both other harnesses: pass-rate spread is small (38–50%), cost spread is two orders of magnitude.

Failure Fingerprints Replicate Across Models

The harness-specific failure profile is the **same for both models** — a scaffold property, not a model property:

- **Goose** — REASON (stops voluntarily)
- **OpenHands-SDK** — VERIFY / MAX_TURNS
- **OpenCode** — idle-loop / TIME

Takeaway

For **human-centered coding-agent evaluation**, model name alone is an incomplete unit of comparison.

- **Compare harness–model *pairs***, not models.
- **Developers:** select a pair by pass rate under your token / latency budget.
- **Researchers:** report token usage, latency, and the full harness spec alongside any model comparison.

Scope & Limitations

- **3 harnesses, 2 models** — findings may not extend to every scaffold or to closed-source commercial agents.
- **Turn-budget asymmetry** — OpenCode is bounded only by the 900 s wall-clock cap (no turn-cap flag), unlike Goose and OpenHands-SDK.
- **Provider defaults evolve** — exact replication requires pinning the OpenRouter route and sampling settings; released run metadata records them.