

Pseudo-SOTA on Domain Benchmarks: Is It the Tool's Scaling Dividend or the Quality of the Harness Mechanism?

scalingX

RsiingSI · Independent Researcher
20261017

Hold the model fixed and run a domain benchmark — what are different harnesses' scores actually comparing?

We propose HAVEN: before publishing a leaderboard, separate “harness engineering quality” from “tool dividend”.

When running a domain benchmark, the model is the same; what we actually want to compare is the harness. But today's leaderboards cannot support that comparison.

Because when a self-built harness takes SOTA, the score mixes two things:

- ① Harness mechanism: the orchestration loop, context management, error recovery, state tracking, tool-scheduling policy, when to stop, and how to compress history.
- ② Tool surface: how many domain MCPs are attached, how many skills, how many packaged domain operations.

The two are stirred together in the total score. You cannot answer the question you actually want to ask: is this harness really well built, or does it merely have enough tools?

Bad comparisons carry three kinds of poison:

- ① Tool dividend masquerading as mechanism advantage: a self-built harness attaches a pile of PubMed, UniProt, and RDKit MCPs and skills, takes SOTA with the same model, and readers assume the orchestration is strong — when the tools may simply fit the task.

- ② Invisible tool surface: with no disclosed tool inventory, you do not know whether the other side exposed 5 tools or 50. The search space is not even the same, so the scores are not comparable.
- ③ Skills and prompts are tools too: some attach no MCP at all, but stuff in a pile of domain-specific skills, few-shots, and task templates. That is still tool surface — only hidden deeper.

We propose HAVEN: a fixed-model evaluation protocol aimed at harness attribution, plus a three-condition controlled comparison procedure.

Core ideas:

- Decompose the harness into “mechanism $\times$ tool surface”: with the model fixed, the score gap between harnesses = mechanism gap + tool-surface gap + interaction. HAVEN's goal is to report these three terms separately.
- Three-condition design, run by every harness in the same batch:
 - Native: the harness's own configuration, attaching whatever it wants. This is how it actually competes.
 - Equalized: every harness is wired to the same standardized tool set (same L0+L1, same count, same interfaces). Under this condi-

tion the tool surface is flattened, and only mechanism differences remain.

- Bare: only L0 execution primitives (bash, file read/write). This measures the pure orchestration floor.
- Two key deltas:
 - $\Delta_{\text{tool}} = \text{Native} - \text{Equalized}$: for the same harness, how many points are lost when the tool surface changes from “its own” to “unified”. This is a direct measure of the tool dividend.
 - $\Delta_{\text{mechanism}} = \text{Equalized}(\text{harness A}) - \text{Equalized}(\text{harness B})$: with an identical tool set, how much higher A scores than B. This is the mechanism advantage.
- Mandatory Tool Manifest disclosure: tool names, counts, underlying implementation, and the L0/L1/L2 split; skills, few-shots, and task templates also count as tool surface. L2 covers MCPs that directly wrap the benchmark's answers, or wrappers tuned against the task description — this layer is forcibly disabled under the Equalized condition.

- Report more than a single total: the distribution of Δ_{tool} , the paired $\Delta_{\text{mechanism}}$ differences, the correlation between tool-call count and score, and whether the ranking holds after changing seeds or task subsets.

- The final criterion is not who ranks first, but: after the tool surface is flattened, does your harness still lead? By how much? Of your high score, how much is mechanism and how much is tools?

Minimum viable experiment: 50–80 tasks from one domain benchmark + 1 fixed model + 3 harnesses (yours + Codex + Claude Code), with every harness run under all three conditions (Native / Equalized / Bare), 3 repetitions. Output: Δ_{tool} , $\Delta_{\text{mechanism}}$, and the ranking under Equalized.

We want to answer one question:

With the model fixed and the benchmark fixed, when a harness takes SOTA, is it really better built, or does it simply have richer tools?

More detail soon. Discussion and collaboration welcome.