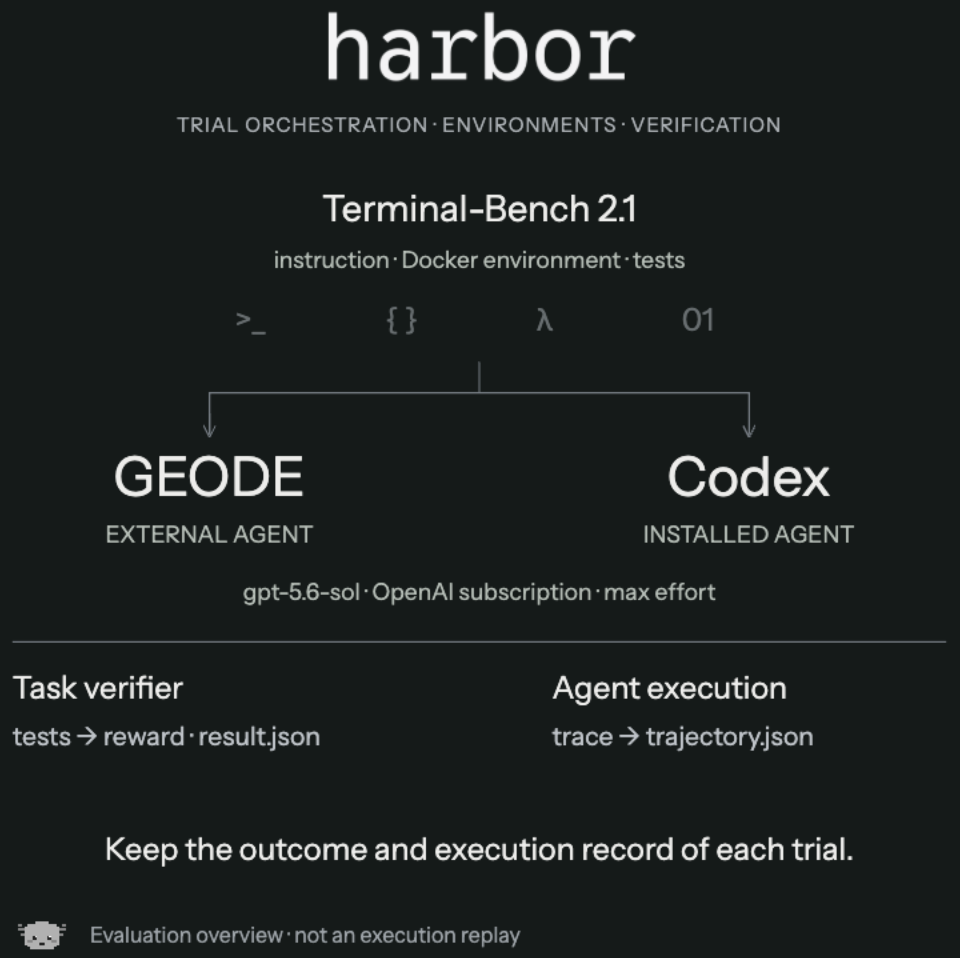
Beyond Pass Rates

A score tells us the outcome.
A trace helps us investigate how it happened.

Tracing GEODE's Execution Quality

Harbor · Terminal-Bench 2.1 · native Codex



Improve the execution path behind the score

Follow two actionable problems, rather than one ranking.

01 · Why did the judge accept an incorrect artifact?

02 · Why did less input not mean faster completion?

Design → observe → diagnose → change → test against the same criterion

Apply the termination rules before reading the score

Cell = task × repetition × arm. A pair joins GEODE and Codex for the same task and repetition.

445 → **435** → **429**

planned pairs → after exclusions → common-valid

bn-fit-modify · tune-mjcf: 20 excluded cells / 6 native-invalid cells

Task-owned tests + frozen termination conditions → selected reward

Semantic failure / canonical timeout = 0; invalid / unexecuted ≠ 0

The 890-cell primary is ineligible. Later totals describe common-valid secondary observations.

Tasks were isolated; runtime placement differed

Historical: Harbor 0.22.0 · gpt-5.6-sol / subscription / max · b549f3e

arm64 host / Docker / amd64 emulation



Harbor → isolated trial → termination → task-owned verifier

Frozen task image, CPU/RAM and timeout / sequential arms / c1 or calibrated c2

Isolation defines the file-state boundary, not equivalent runtime placement.

Read the 98 disagreements behind the eight-pass gap

429 common-valid pairs: GEODE 339 passes, Codex 331. Use disagreements to choose an investigation.

286 both pass · 53 GEODE only · 45 Codex only · 45 neither

Historical G-code: GEODE 0/5 · Codex 4/5 → select a development case

Full-suite primary ineligible / not an official ranking or overall superiority claim

Measured on an arm64 host · amd64 task containers via Rosetta

Repair the evidence path behind false acceptance

Saving and reading back a file did not establish that its contents were correct.

INTERNAL JUDGE

5/5

TASK VERIFIER

2/5

FALSE ACCEPTANCES

3

IMAGE EVIDENCE

7 · 9 · 9 images observed

≤2 images selected → **Select ≤12 image-bearing calls separately**

WORKER MODEL

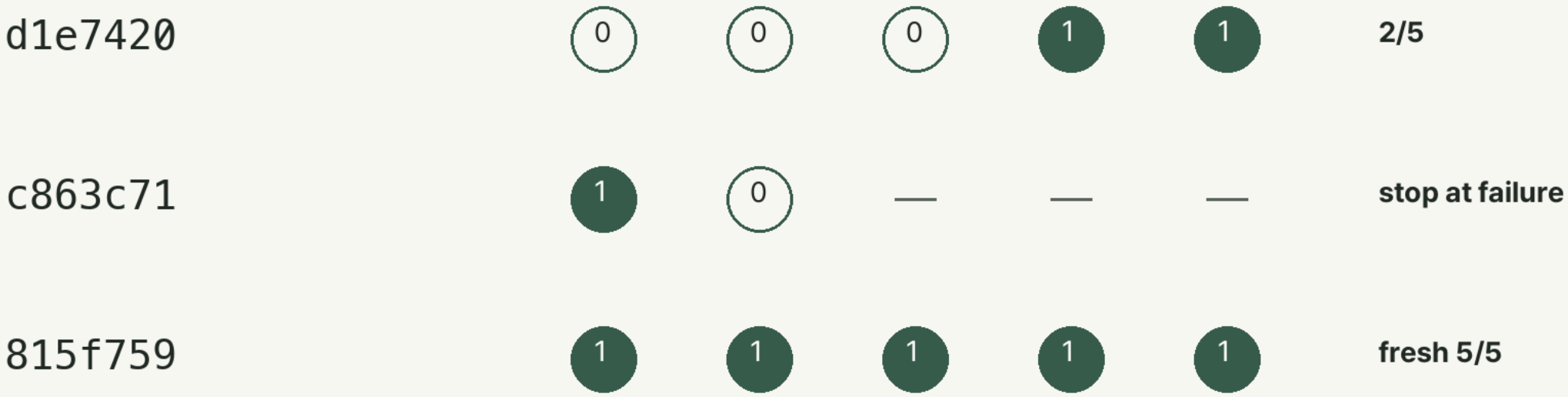
3/3 failed before inference

model/provider mismatch → **Inherit parent model when omitted**

Evidence precedes candidate claims. These bundled fixes were not tested as separate causal effects.

The revised candidate passed five fresh trials

5/5 · no invalid trials · no in-trial repair. Failed candidates and stopped attempts remain in the ledger.

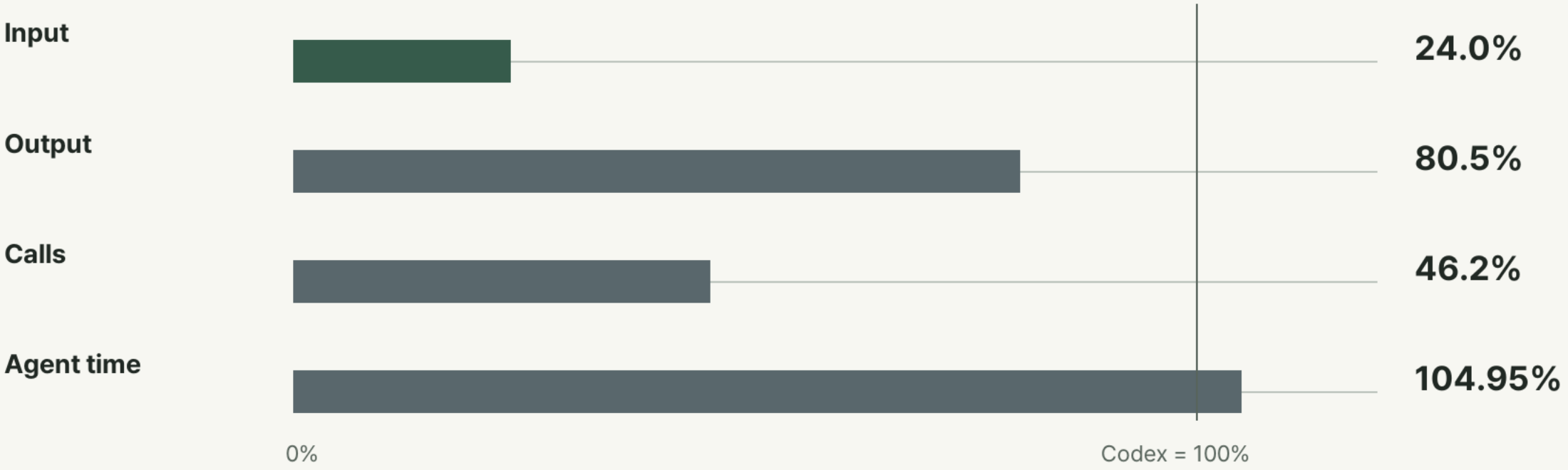


— = unexecuted · r2 SHM custody issue remains open; see NARRATIVE

Sources: G · NARRATIVE.md

Less input on solved tasks; no reduction in aggregate time

Reported totals for 286 jointly successful pairs: GEODE / Codex, not mean task ratios or billing.



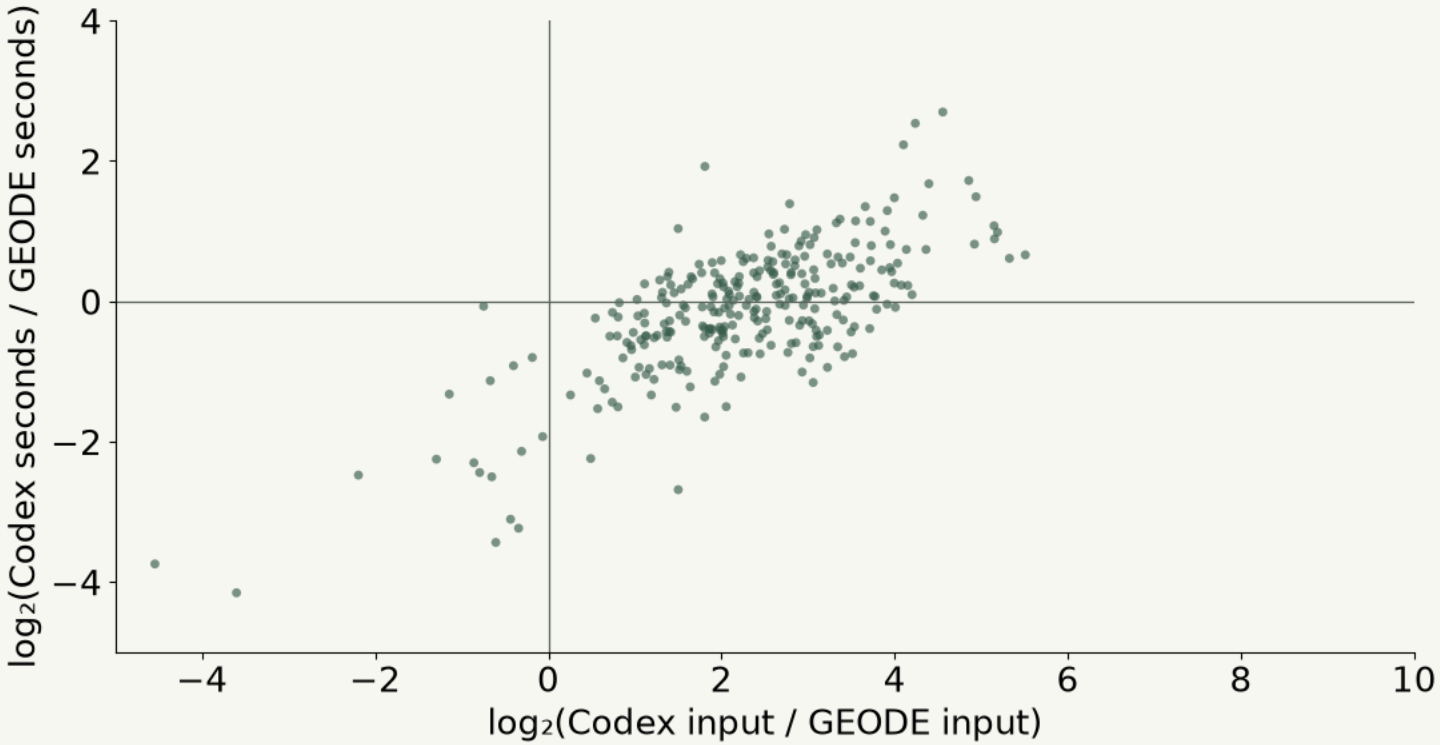
Input 24.03% · Output 80.49% · Calls 46.19% · Agent elapsed 104.95%

Measured on an arm64 host · amd64 task containers via Rosetta

Sources: T · NARRATIVE.md

In 133 pairs, less input did not mean faster completion

One point is a jointly successful task–repetition pair; all 286 are shown.



Right: less GEODE input · above: faster GEODE execution

Less input and faster: 136 · less input but slower: 133

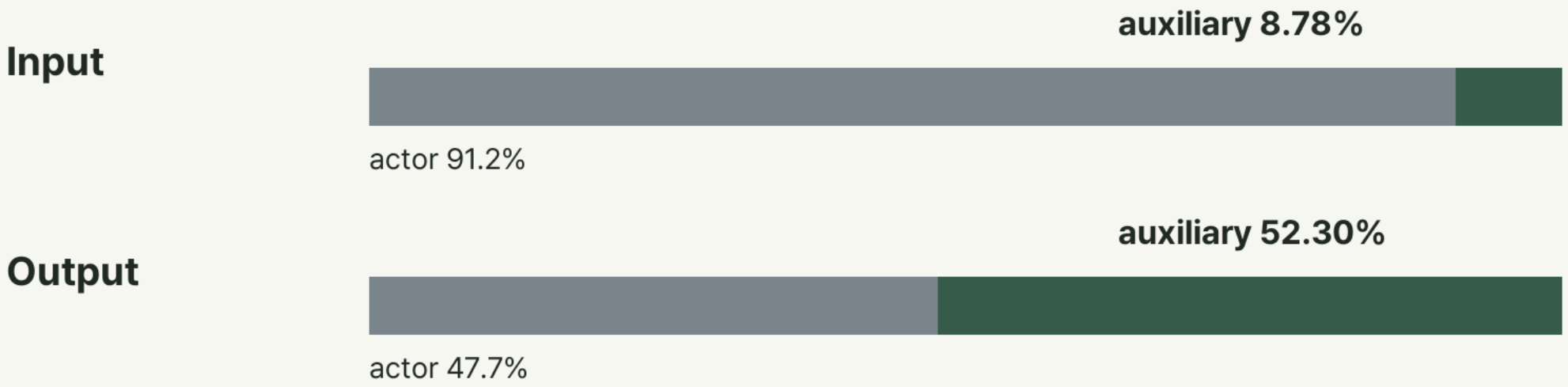
46.19% as many calls × 52.03% input per call ≈ 24.03% total input. Elapsed time is a separate observation.

Measured on an arm64 host · amd64 task containers via Rosetta

Observing auxiliaries exposed over half the reported output

Separate one-task smoke · 7c0e927d · modernize-scientific-stack · full runtime in container

31 requests, all max: actor 16 / reflection 8 / text 4 / judge 3

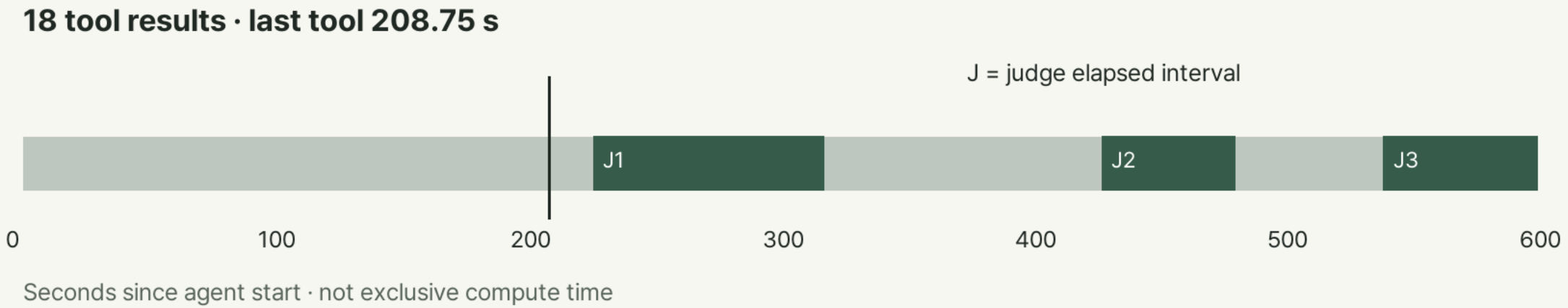


1 CPU / 2 GiB / 10 GiB declared / 600 s / c1 / retries 0

Usage 30/31: cancelled call is null / cache-read 35,712 is an observed subtotal

Revise decisions were not followed by another tool action

Last tool at 208.75 s → two revise decisions → third judge cancelled → 600 s deadline



Post-timeout task verifier raw 1 / canonical selected 0 · valid failed 0/1

Judge elapsed 205.35 s / 34.18% of agent phase · not CPU utilization

Intermediate artifacts were not independently graded; false rejection is unproven.

Project the same evidence into results and Replay

Public sanitization does not replace source evidence. Fresh runs do not overwrite historical cells.

PRIVATE SOURCE	CANONICAL	PUBLIC DERIVATIVE
`raw/<arm>/<job>/<trial>/`	`attempts.jsonl / analysis.json`	`trajectory / recording.cast`
result / verifier / sessions.db	source SHA + validity + reward	sanitize → scan → readback





`Local root: artifacts/eval/runs/<run_id>/`

`Public: geode-eval-artifacts/terminal-bench/<run_id>/`

Public cast is ATIF-derived replay. Source/export agreement does not prove all physical calls were captured.

Proposed experiment · not executed

The next comparison must test improved completion

The fresh G-code candidate passed five trials. Next, compare completion time and cost under matched conditions.

FIX THE PRODUCER

First: repair native field-presence/cache handling and reconcile source/export

FREEZE THE CONTRAST

Then: freeze baseline → change one revise policy → fresh matched trials

MEASURE COMPLETION

Compare judge evidence, next tool, remaining time, canonical reward and usage

Which run supports each result?

Each result has a population, source revision and file hash.

H / T

Frozen comparison · b549f3e

429 common-valid pairs / usage: 286 jointly successful pairs

G

G-code candidate · 815f759

Five fresh passes form a development gate; historical scores remain intact.

S / C

Full-runtime smoke · 7c0e927d

Evidence for 31 requests and delayed completion; not pooled with the comparison.

Pillow · Matplotlib · FFmpeg / Pretendard · Source Han Serif KR · STIX Two Text

Separate trace structure from scoring authority

Official docs explain structure; run artifacts establish versions and values.

P1 / P2

Terminal-Bench paper (2026) · Harbor

Task, trial and job definitions / the frozen manifest establishes version 2.1

P3

ATIF · Core Models / Validation

tool_call_id ↔ source_call_id / current v1.8 does not replace historical schemas

M2

Anthropic · Demystifying evals (2026)

Separate transcript claims from outcomes / compare judge and task verifier

Pillow · Matplotlib · FFmpeg / Pretendard · Source Han Serif KR · STIX Two Text

Cite methods at the boundary of their use

Design references, interpretation and future work are not the same result.

M1

Reflexion · Shinn et al. (2023), §3

Actionable verbal feedback / five passes do not establish repair benefit

M3

AI Agents That Matter · Kapoor et al. (2024)

Evaluate accuracy, cost and holdouts / token ratios are not dollar costs

M4

Darwin Gödel Machine · Zhang et al. (2025)

Future archive-based candidate search / not an experiment executed here

Pillow · Matplotlib · FFmpeg / Pretendard · Source Han Serif KR · STIX Two Text

Now revisit all 445 pairs from the historical comparison

The following derived Replay is unchanged; it is not the new G-code run.

835 ATIF-derived · 35 receipt-only · 20 excluded cells

GEODE left / Codex right · aligned by tool-event order

Not raw PTY or simultaneous execution · UTC retained / KST displayed