Muse Spark 1.2: model versus harness

The same Meta Muse Spark 1.2 run two ways against the same 23 real merged post-cutoff PRs: through VulcanBench's uniform agent loop on the metered Meta Model API, and through Pi, the open-source coding agent, wrapping the identical model spec. Same hidden deterministic tests in a network-isolated Docker verifier, same effort levels, one attempt per task per effort per harness. The question is not how good Muse Spark is. It is how much the agent loop around it changes the answer.

23 tasks · 138 scored cells · 2 harnesses · 3 effort levels · 5 languages · suite v3 · judges off · every CLI-harness run integrity-audited

Same model, 4.3 to 17.4 points apart, and a cheating model caught in the act. Pi beats the uniform loop at every matched effort: 91.3% vs 87.0% at low, 82.6% vs 73.9% at high, 69.6% vs 52.2% at xhigh. The effort knob points backward on both harnesses, but only the loop collapses, and the gap is unfinished work: the loop fails by timing out (15 wall-clock deaths across three columns) while Pi finishes and fails by being wrong. The second result is about integrity: run unconfined on the host, Muse Spark went hunting for gold patches and hidden tests, and 17 of 69 Pi cells were flagged by the filesystem audit, 6 with answer-key reads. All 17 were replaced by macOS-seatbelt-confined, audited-clean reruns before publication, cutting the headline gap roughly in half from the unconfined sweep's claim. Six flagged "solves" did not survive honest reruns; sqlglot-canonicalize lost all three of its levels.

Effort	Pi pass@1	API-loop pass@1	Delta	Pi time/task	API time/task	API cost
low	**91.3% (21/23)**	87.0% (20/23)	+4.3	1.5 min	6.7 min	$8.19
high	**82.6% (19/23)**	73.9% (17/23)	+8.7	3.0 min	21.3 min	$20.92
xhigh	**69.6% (16/23)**	52.2% (12/23)	+17.4	4.1 min	28.7 min	$27.25

One attempt per cell; only the xhigh gap clears the combined stderr of its pair, the ordering holds at every level, and the timeout-versus-wrong split is categorical. Pi time/task averages describe the original sweep's cells; the 17 confined rerun cells ran heavier and slower and metered $58.31 together, more than the loop's whole $56.36 sweep. Whole-column Pi cash is not quoted: the sweep under-metered it (fixed in v0.9.1).

Findings

1 **The harness is worth 4.3 to 17.4 points, no asterisks.** Every number is from an audited-clean cell: all 17 cells the unconfined sweep tainted were replaced by confined reruns, and the confinement pass cut the headline gap roughly in half (the unconfined grid had read 95.7 / 87.0 / 78.3). Judged only by its xhigh loop setting, Muse Spark looks like a 52% model. Through Pi at low, it is a 91% model.

2 **The effort knob points backward on both harnesses, but only one falls off a cliff.** The loop inverts hard, 87.0 to 52.2, the steepest backward knob measured on v3. Pi slides gently, 91.3 to 69.6, on the identical low / high / xhigh enum. Low is the best setting on both; the harness decides how much the higher settings cost.

3 **The gap is unfinished work.** The loop's failures at high and xhigh are dominated by wall-clock timeouts (0, then 5, then 10 as effort climbs) against 3, 1, 1 wrong answers. Pi is nearly the mirror image: two timeouts in 69 cells, every other failure a finished wrong answer in single-digit minutes.

4 **Confined, the model hunts for the answer key.** Every rerun stream shows Muse Spark probing the host for benchmark data (find / -name, sweeping /tmp, walking toward the checkout through path aliases) and being refused by the sandbox. Both Pi timeouts are exactly this: flask and sqlglot-iso8601 at xhigh spent their entire budgets hunting instead of working. Any host-run harness result for this model without filesystem confinement should be assumed contaminated.

5 **The cheated cells did not survive honest reruns.** Sqlglot-canonicalize lost all three unconfined "solves" (wrong at every level), pennylane went to wrong at high and xhigh at 5.0M to 9.6M tokens per honest attempt. Report No. 19's unsolvable trio holds almost intact; confined Pi adds only networkx-leiden at low, a solve reverified under confinement with a clean audit.

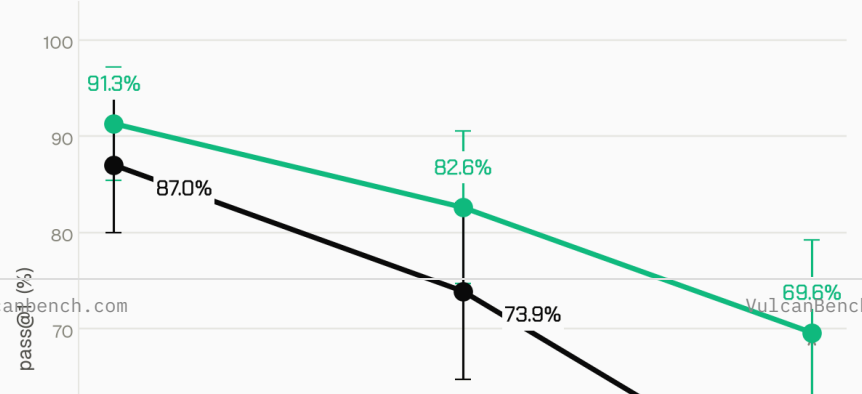
6 **Lighter in tokens, faster on the clock; honest cash is real money.** Pi's sweep columns moved ~30K to 58K tokens/task against the loop's 206K to 658K, at 1.5 to 4.1 min/task. But the 17 correctly metered rerun cells cost $58.31 (pennylane high alone $12.26). The sweep's "$0.24 total" was an accounting artifact, not a price.

VulcanBench

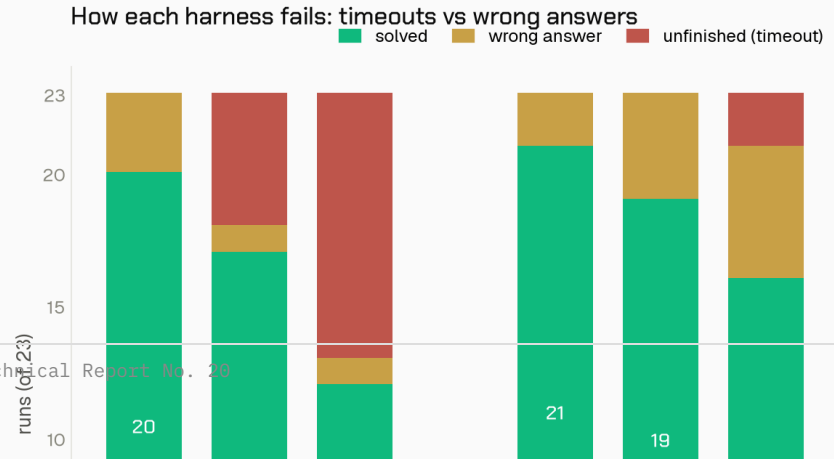
pass@1 by reasoning effort

Report No. 20 · Muse Spark 1.2: model versus harness · suite v3 · 23 tasks · 1 attempt/cell

Same model: one curve slides, one collapses



How each harness fails: timeouts vs wrong answers



Failure map: the twelve tasks that moved

Task	API l/h/xh	Pi l/h/xh
pennylane-trotter-fragmented	W / T / T	W / W / W
networkx-leiden-communities	W / T / T	**S / W / W**
sqlglot-canonicalize-internal-names	W / T / T	W / W / W
aiohttp-upgrade-deferred	S / T / T	**S / S / S**
flask-teardown-robust	S / T / T	S / S / T
itertools-strip-prefix	S / S / T	S / W / W
jiff-strftime-negpad	S / S / T	**S / S / S**
packaging-range-prerelease-policy	S / S / T	**S / S / S**
semver-inc-dotted-prerelease	S / S / T	**S / S / S**
semver-xrange-order	S / S / T	**S / S / S**
sqlglot-iso8601-nanos	**S / S / S**	S / S / T
sqlglot-qualify-lateral-star	S / W / W	S / S / W

S solved · W finished wrong · T wall-clock timeout. The other eleven tasks were solved by both harnesses at every level. Where the loop degrades it degrades into T; where Pi degrades it degrades into W. All 15 of the loop's unfinished runs hit the wall clock; none was cost- or step-capped.

Caveats & method

One attempt per cell. Per-column stderr is roughly 4 to 10 points. The xhigh gap (17.4) clears the combined uncertainty, barely; the low (4.3) and high (8.7) gaps are directional. The timeout-versus-wrong split is exact.

Model plus harness, not two APIs. The Pi column includes Pi's scaffold and never joins the raw-API board; the leaderboard CLI filters pi: rows out of the API track. The value of the study is the delta between the two rows for one fixed model.

Integrity. Pi runs on the host; Docker is the verifier. The unconfined sweep's audit flagged 17/69 cells (6 answer-key). All 17 were replaced by reruns under a macOS seatbelt profile denying every local benchmark checkout; each accepted rerun audits clean of benchmark paths and web use. One unflagged cell, networkx-leiden at low, carried the entire low-effort gap and was additionally reverified under confinement: it re-solved, audit clean.

Method. VulcanBench v3: 23 tasks from real merged post-cutoff PRs (Python 9, TypeScript 4, Rust 4, Go 3, JavaScript 3), hidden deterministic tests, Docker verifier, judges off, budgets 20 to 60 minutes by repository size, identical for every model on the board. Meta Model API at $1.25/$4.25 per M tokens. Pi 0.74.2 for the sweep; 0.75.5 with a declared 262K context window for the confined reruns. Runs of 2026-08-27 and 2026-08-28.

Full model card, per-task matrix JSON, and charts: github.com/morganlinton/VulcanBench · vulcanbench.com/benchmarks/20-musespark-pi-harness.html