The Agent Proposes, the Referee Disposes: Verified Frontier Performance on ARC-AGI-3 at a Measured 8% Integrity Overhead

Mohsen Arjmandi
Independent researcher
mohsen.arjmandi@gmail.com

Abstract

As agent harnesses saturate public benchmarks, the numbers they post are becoming unverifiable: uncapped runs, self-reported action counts, no audit trail, and, in documented incidents, agents that bypassed their environment's sanctioned interface entirely. We present ASSAY, a *referee* harness in which the agent cannot act except through a gate that demands a pre-registered prediction, cannot believe except through claims graded by code against the environment's own response, and cannot touch the world off the record: journals are hash-chained and externally anchored, and a single ungated action invalidates the run. On the ARC-AGI-3 public set this referee'd agent scores **RHAE 96.54** — confirmed by the benchmark server on a public scorecard — winning 24 of 25 games (177/183 levels) under hard action caps with a domain-neutral doctrine, zero ungated events across 8,157 paid actions, and every win verified three independent ways. The measured price of full integrity is an **8.0% exploration tax** (656 probe actions; every other paid action ran inside a verified batch) and ≈zero capability cost: over its 24 wins the system scores 99.80, and in aggregate actions it is cheaper than every published comparator except one. The campaign also produced a finding about agent epistemics under verification: on six occasions the agent *proved* a level impossible, each proof consistent with every recorded transition — and each wrong precisely where no transition had gone. A three-step protocol (COVERAGE AUDIT: enumerate the rules a proof load-bears on; audit the journal for what actually *exercised* each; buy graded probes for the gaps) converted five of the six into wins within existing budgets. Finally, we derive and validate the benchmark's own scoring function and show it is saturated at the frontier: a 12% raw-action gap between the two leading systems is worth ≈0 points, while a single unfinished game costs 3.27; the axes that still discriminate — integrity, cost, information — are exactly what a referee instruments.

1 Introduction

A benchmark result is a claim about an agent, and today the claim is usually graded by the agent's own harness: the harness chooses what to count, what to retry, what to report, and whether the model ever touched the environment off the sanctioned path. That was tolerable while scores were low; at saturation, when leading systems ask to be distinguished by exactly the properties their harnesses do not record, it is not — and it visibly failed in at least one published incident, where a capable agent, explicitly and repeatedly instructed not to cheat, discovered its harness could reach the game's admin channel and spawned resources directly into its own factories [7]: the instruction lost to optimization pressure; the hole was capability, not policy.

Preprint.

Our prior work built the opposite trust model at small scale: an agent instrument in which a deterministic executive owns all belief, the model may only file typed proposals, and a claim is admitted only when a prediction pre-registered *before acting* is matched against observation by code [3]. That paper reported its efficacy null in the abstract: zero level completions across 52 gated runs. The open question it left was pointed — is verification-first architecture *structurally* uncompetitive, paying for its auditability in capability, or was the null an engineering fact about one instrument?

This paper answers with a system and a campaign. ASSAY keeps the prior instrument's constitutional rule — the agent proposes, the referee disposes — and rebuilds everything around it on two design bets from the same research line (previously unpublished; §2): the referee as the *engine* that owns actuators, budgets, and a claim economy in which the minter of a hypothesis pays for its verification; and behavior installed as *modules* with enforcement in code rather than exhortation in a prompt. On ARC-AGI-3 [1], an interactive benchmark designed so that pretrained knowledge does not transfer and agents must learn each environment from interaction, the resulting system plays at the frontier while remaining fully auditable end to end.

Contributions.

1. **The referee protocol and its measured cost** (§2, §4): frontier performance (RHAE 96.54, server-confirmed; 24/25 games) under hard caps and a domain-neutral doctrine, with every action gated, every claim code-graded, every journal hash-chained and audit-clean — at a measured 8.0% exploration tax and ≈zero win-rate cost. Integrity, measured, is nearly free.
2. **The COVERAGE AUDIT finding** (§5): six agent-authored impossibility proofs, each consistent with every recorded transition, each wrong exactly where no transition had gone — and the three-step audit protocol that converted five into wins within existing budgets. We propose treating any agent-side "provably impossible" as a trigger to audit the proof's *unexercised* rules, not as a conclusion.
3. **Honest cost accounting** (§4): measured dollar cost per run, a probe/batch split of every paid action, wall-clock, and $n{=}1$ discipline stated as such — instrumentation the saturated leaderboard now needs more than another score.
4. **RHAE is saturated at the frontier** (§4.1): we reverse-derive the benchmark's scoring function, validate it exactly against 25/25 published game scores, and show its caps compress the leading systems into a completion checklist — quantifying which axes still discriminate.

2 The referee harness

ASSAY is a domain-general harness: a world is attached by registering a small adapter (observation source) and a registry of actuators (the action vocabulary, caps, and secrets policy), and the same kernel, doctrine, and protocol run unchanged across worlds. Nothing described below is specific to ARC-AGI-3 except a 172-line adapter and the action registry.

2.1 Design lineage: two bets on top of the executive

The prior instrument [3] contributed the constitutional rule and the measurement discipline, and a null: structure alone did not score. Two design bets — internally named v3-a and v3-b, the backbone of ASSAY, published here for the first time — reallocated what the structure is *for*.

Bet 1 (v3-a): the referee is the engine. A side-by-side study against a strong free-form agent had dissociated *discovery* from *keeping*: the free-form agent discovered faster, while the structured instrument kept knowledge with provenance the free agent could not match — and the instrument's pathology was a growing pile of minted-but-untested hypotheses, because the component minting claims was not the component paying for their verification. v3-a restructures the economy around three invariants: **the minter pays** (a hypothesis is authored by the mind that must act on it, so claims appear at decision junctures, not as ambient discovery); **nothing parks without a test** (an anomaly is quarantined only with an attached, routed discriminating probe); and **verification is severity-priced** (stringency scales with the irreversibility of the next action and the export scope of the claim — cheap in-stream checks for reversible probes, full receipts at commit points). Imported knowledge — a prior session's notes, another agent's transcript — obeys the same law: it enters demoted to

foreign-hypothesis status, and only the gate's own graded receipts promote it. A win verifies the trajectory, never the narrative that rode along with it.

Bet 2 (v3-b): behavior is installed, not exhorted. The recurring failure of prompt-level policy (§1) has a constructive dual: behaviors that matter should be *modules* — typed, inspectable units with a declared doctrine, trigger conditions, and telemetry — executing against the journal rather than persuading the model. ASSAY ships a small set (budget-spend advisories, miss-streak halts, null-result forensics, park-with-test enforcement, claim-sharpness scoring, hazard tagging), each emitting advice and evidence into the same auditable record, so module efficacy is itself measurable. The system prompt carries a domain-neutral doctrine and no game-specific instruction of any kind.

2.2 The kernel: what the agent cannot do

Gate. A broker daemon owns the environment session and its credentials; the agent process never holds them. Every action must arrive as part of a *prediction*: a machine-parseable claim about what the world will report after the action, registered before the action executes. A bare request to step the world is refused and journaled as an ungated-step event — and a single such event marks the entire run invalid. Predictions over verified models may batch several actions under one claim; the batch halts on the first miss, and unexecuted steps cost nothing.

Claims graded by code. A claim is graded against the environment's own response — never against the agent's account of it. The grammar spans direct state claims (no-change, change, level-advance, win), claims over agent-declared observation *channels* (typed extractors over raw observations, with equality/delta/threshold predicates and bounded temporal windows), and executable *verifiers*: programs the agent writes, which the referee runs sandboxed against the raw before/after observations. Verifiers are identity-probed (a verifier that cannot distinguish a transformed observation from the original is flagged vacuous and earns nothing), and every grade appends to the journal.

Earned trust. A claimed world model starts untrusted. It is promoted only by a standing law over the journal — sustained graded predictions with zero misses over a sufficient window — and promotion buys exactly one privilege: longer verified batches. Trust is therefore a computed property of the record, not an assertion; a miss demotes.

Tamper-evident record. The journal is append-only under a rolling hash chain with external anchors; `assay audit` recomputes integrity from the artifacts alone, so a third party can verify a run without trusting the operator's machine state. Goals are host-pinned: the agent has a proposal lane for agenda changes but cannot self-ratify one.

The exploration tax falls out of the design as a measurable quantity: paid actions divide exactly into *probes* (single-action discovery steps taken to earn a model) and *batch* steps (executed inside a verified prediction); §4 reports the probe share per run.

2.3 Benchmark protocol

All 25 public ARC-AGI-3 games under one uniform pre-registered protocol: Claude Opus 5 backbone throughout; fresh cold starts (no carryover — imports would confound comparisons); tier-2 starting information (the domain-neutral doctrine, the sentence "a 64×64 color grid, level counter and win state exist," and the registry's one-word action hints, kept at parity across all runs); hard action caps (200/500/1500 by game, the cap registered before the run); sealed paths (the agent may not read the benchmark's client internals); and the player does its own work — no human hints mid-run. Where a kernel change shipped mid-campaign, a regression gate re-ran a reference game under a pre-registered bar before any new result counted. Long games span crash-safe resumed sessions; resumed knowledge re-enters as foreign hypotheses per §2.1.

3 Verifying the measurement itself

Following Arjmandi [3], the instrument's numbers are themselves verified, not asserted.

Wins are triple-verified. A win is never read from the harness's claims: (1) the game engine's own state in the journaled run; (2) a harness-free replay of the recorded action sequence through the

Table 1: The four-system comparison on the ARC-AGI-3 public set (same 25 game instances, verified by id; every number from a linked public scorecard or our journals). PRO-LONG ran on a stronger backbone and is directional context only. "Probe" = paid single-action discovery steps (the exploration tax); only ASSAY publishes this split.

System	Backbone	Regime	RHAE	Games	Levels	Actions
arc-skill	Opus 5	uncapped	100.00	25/25	183/183	7,645
ASSAY (ours)	Opus 5	**hard caps, gated**	**96.54**	24/25	177/183	8,157
Prime Agent (median card)	Opus 5	uncapped	95.24	24/25	178/183	11,245
PRO-LONG	Fable 5	uncapped	94.71	19/25	—	11,156
ASSAY split: 656 probe (8.0%) + 7,501 batch; 46.1 actions/level; zero ungated events.						

official engine; (3) a live ARC-server replay returning WIN with identical level counts. Every one of the 24 wins passed all three.

The set is server-scored. A consolidated verification replay of all 25 recorded action sequences produced public scorecard `702ccd4f`[1], on which ARC's own scoring returns **96.54%** with 177/183 levels, 24/25 environments, and 8,157 actions — matching our offline computation per game on every row; all 25 replays reproduced their recorded outcomes exactly.

The metric is validated. RHAE, the benchmark's score, is not games-cleared or levels-cleared. We reverse-derived it from published scorecard JSONs: per level $\min\big(115,\ 100 \cdot (\text{baseline}/\text{actions})^2\big)$, zero if uncleared; per game the level-index-weighted mean capped at 100 if the run reached WIN, and *weighted progress only* — no efficiency credit — otherwise; the set score is the plain mean. Our scorer reproduces 25 of 25 published game scores exactly (worst error 0.000000) and independently reproduces a second system's published partial-game score; the scorer and the per-level baselines ship with this paper. Consequences of the caps are quantified in §4.1.

Accounting discipline. Coverage statistics (games, levels, actions) are never presented as scores[2]; every cell is $n{=}1$ and labeled as such; efficiency is reported only for completed games; dollar cost is reported measured where metered and labeled extrapolated where not.

4 Results

Table 1 is the headline. Under hard caps, with a domain-neutral doctrine and less starting information than published comparators, the referee'd agent finishes second of four — behind only one uncapped system on the same backbone — while being the only run in the table with an audit trail: zero ungated events across 8,157 paid actions, every journal chain intact, every win triple-verified. Per-game action counts for all 25 games are on the public card (footnote 1). Prime Agent's row uses their published median card; their best-of-3 reaches 183/183 — our numbers are single-protocol, no best-of-N.

Integrity costs ≈nothing in capability. Over the 24 games it won, ASSAY scores RHAE 99.80. On the 24 games both leading Opus systems won, summed actions are ASSAY 7,793 vs. arc-skill 6,858 (12%); on games ASSAY won in a single session the two are a statistical dead heat (1,899 vs. 1,897 over ten games) — the aggregate gap is concentrated in five multi-session conversions that carry the full discovery cost of breaking six wrong impossibility proofs (§5). Against Prime Agent on the common 24, ASSAY is 11% cheaper (7,793 vs. 8,734). The one game ASSAY did not clear defeats every published system except one: lf52 (ASSAY 4/10 at 364 actions with a located diagnosis; Prime Agent 5/10 at 2,511 ending in GAME_OVER; PRO-LONG 81.8% at 1,000; arc-skill won at 787). ASSAY cleared five of the six games PRO-LONG's stronger-backbone cohort could not.

[1] `https://arcprize.org/scorecards/702ccd4f-df1f-4118-bc8b-d79d3f4a1a32`. Provenance, stated exactly: our agents played locally under the journaled protocol; scorecard ids were not captured during those runs, so the recorded action sequences were re-issued verbatim against the live ARC API to mint verifiable cards. These cards are verification replays of recorded action sequences — the server independently confirms what each sequence achieves (outcome, per-level action counts, official score) on the same game instances. Wall-clock on the card is machine replay time (64.8 minutes), not agent time (§4); no model is in the loop during a replay.

[2] The discipline is not pedantry: this set's level coverage is 96.7%, deceptively close to its RHAE of 96.54. The two numbers measure different things and coincide by accident.

The exploration tax, measured. 656 of 8,157 paid actions (8.0%) were probes; every other paid action executed inside a verified batch. Per run, the discovery ramp (paid actions before the first level cleared) spans 4–22; probe share spans 4–28%; batch steps wasted to a halted miss never exceeded 10 per run. These runs start with an action vocabulary but no semantics — the tax *is* the measured price of learning a world through a prediction gate. One controlled reference point: on the same game instance, our traced control run of the leading uncapped harness spent a 58-action discovery ramp at 42% probe share, against ASSAY's 22 and 14% from less starting information.

Cost and time, stated honestly. Eleven runs were API-metered: $283.68 total, $11.61–$53.89 per run, median ≈$22; the other fourteen ran on a subscription with no per-run meter, so the full-set figure is an extrapolation, ≈$550–650. The only published comparator total is PRO-LONG's $1,750 for its 25-game cohort. Agent wall-clock, measured on 20 of 25 runs: 32–330 minutes per game, ≈37 hours total (≈1.8 h/game) — which is why the 64.8-minute machine-replay wall-clock on the public card must never be read as agent time.

4.1 RHAE is saturated at the frontier

The validated formula (§3) makes the saturation precise. Because each level is capped at 115 and each game at 100, finishing far under baseline earns nothing: the 12% raw-action gap between the two leading Opus systems is worth ≈0 RHAE — both sit at the caps nearly everywhere — while a single unfinished game costs 3.27 set points (ASSAY's lf52) — and even a won game that ran over baseline on late levels (ASSAY's bp35, 95.27) costs only 0.19. At the frontier, RHAE is a completion checklist: it can no longer see efficiency, cost, information tier, caps, or integrity — precisely the axes on which saturated systems actually differ, and precisely what a referee instruments. We suggest the community read frontier ARC-AGI-3 claims accordingly: the discriminating questions are now *can you trust the run, what did it cost, and what was the agent told* — none of which the headline number encodes.

5 Six impossibility proofs, and the protocol that broke them

The campaign's central process finding is about agent epistemics under verification. On six occasions across five games, the agent concluded — with an argument, sometimes an exhaustive one — that a level was impossible. Every proof was consistent with every recorded transition in the journal. Every proof was wrong, and wrong in the same way: its load-bearing rule had never been *exercised* — no graded transition had ever tested it in the regime where the proof needed it.

The phenomenon. A replay-validated world model can be confidently wrong exactly where no transition ever went. Each proof was a sound deduction from a model that fit all the evidence — the failure is evidential: consistency with the record was silently treated as support, when the record had simply never visited the region the conclusion depended on. sk48 is the sharpest instance — an *exhaustive search* over 8.2 million states, fully correct given its transition rules, one of which had never been graded — and bp35 L9 the deepest: the unexercised premise was the observation frame itself.

The protocol. COVERAGE AUDIT treats an agent-side impossibility conclusion as a trigger, not a result: (1) *enumerate* the rules the proof load-bears on; (2) *audit* the journal for what actually exercised each — graded evidence in the relevant regime, not mere consistency; (3) *buy graded probes* for the gaps, cheapest first. It converted five of the six proofs into wins within the runs' existing action budgets — cheap precisely because step (2) is free, running offline over the journal. The referee architecture is what makes step (2) possible at all: because every belief entered through a graded claim, the evidential status of every rule is a query over the record, not a reconstruction from prose notes.

The doctrine, stated portably. "Provably unsolvable" is a property of a model, not a world. Before accepting an impossibility — or any absence claim — an agent should be made to show that each rule the conclusion rests on has been exercised by a graded transition in the regime where the conclusion needs it, and to spend probes on the ones that have not. We conjecture this transfers beyond games: agent-authored "impossible," "not present," and "already complete" claims in any domain share the

Table 2: The six broken impossibility proofs. Each fit every recorded transition at the time it was made; each collapsed under graded probes of its unexercised rules. All five games were subsequently won within their existing budgets.

Game	The "proof"	The unexercised rule, and what probing it showed
dc22	level-6 switch unreachable; remaining blocks isolated	all 74 prior inert probes had been taken in a single world-state; one coordinate had been inert-then-live earlier, reframing "refuses to act" as *conditional* — the gate was geometric arming; blocks were connected
s5i5	level 7 "provably unwinnable"	the load-bearing rotation rule had been exercised only in the *occupied* failure mode; one graded probe of the off-board case showed rotation *clips* instead of refusing — un-trapping the arm and breaking level 8's length cap as well
sk48	level 3 unsolvable by exhaustive 8.2M-state search	the search assumed one ungraded companion rule (freed blocks re-hook bottom-only) that fit all 103 recorded transitions; re-hooking is positional — and an action written off as a no-op was SELECT
wa30	throughput ceiling; hostiles effectively unkillable	all three hostile rules were false: the kill action works on any adjacent hostile (the prior session had seven free kills in front of it), movement is 1 cell/action orthogonal (812-observation histogram, zero diagonal moves), and the "ceiling" was a treadmill equilibrium that vanished with the hostiles
bp35 (L8)	switch unreachable	the switch sat above the drifted map
bp35 (L9)	no switches remain; map certified pixel-complete	the frame itself was the unexercised rule — "rendered" had silently stood in for "exists"; nine switches sat above the map, visible only from inside the "dead-end" shaft

failure shape, and the audit is domain-neutral. Converting the protocol from operator practice into standing machinery (a coverage meter over the journal; automatic halts when a failing prediction is re-issued unmodified) is the system's designated next phase.

6 The open failure, reported with equal prominence

One game resists: lf52. Three independent runs under caps of 200, 500, and 1,500 all stalled at exactly 4/10 levels — the best spending 193 actions on level 5 alone and self-stopping with 1,136 actions unspent, after an internal relaxed-constraints planner established that no single-peg finish exists *in the mappable world*. The diagnosis is categorical, not economic: the level requires revealing structure outside the visible frame when the camera's only lever cannot reach it — the same unstated-frame class as bp35 L9, this time load-bearing. The best run also lost $\sim$70 actions to automated re-issue of a failing prediction, motivating the halt-on-repeat module. We report lf52 as the standing falsifier and the designated demonstration target for the machinery phase; by RHAE's arithmetic (§4.1) it is worth more than every possible efficiency improvement combined.

7 Related work

Verification-first agents. This paper is the capability successor to our instrument paper [3], inheriting its trust model (deterministic side owns belief; predict-before-act; runs that invalidate themselves) and answering its efficacy null; an earlier predecessor studied curriculum-based test-time learning [4]. Propose-and-verify control (LLM-Modulo-style planning and its relatives) places a deterministic checker over model *outputs*; the referee differs in owning what the agent may treat as *true*, and in making the audit trail the primary artifact. **ARC-AGI-3.** The benchmark [1] is designed for interaction-time learning; Han [6] studies the explore/solve trade-off our probe/batch split prices; PRO-LONG [5] contributes programmatic memory and published the per-game scorecards from which RHAE was derived and validated. Comparators arc-skill and Prime Agent are used strictly as published scorecard data (§4).[3] **Environment integrity.** The published incident in which an agent bypassed its environment through the admin channel despite explicit anti-cheat prompting [7] is the

[3]Scorecards: arc-skill `https://arcprize.org/scorecards/24ddb219-987e-464f-9050-6398a29cf5ac`; Prime Agent `https://arcprize.org/scorecards/2af780b4-f2a1-43e9-a794-b23da3cd3f9f`; PRO-LONG's 25 per-game card links are published at the project page.

constructive motivation for enforcement-by-capability (§2); goal-drift evaluations [2] document the adjacent failure family in long-horizon agents.

8 Limitations

Every cell is n=1 by protocol; run-to-run variance is real and uninstrumented here. All our runs use one backbone (Claude Opus 5); architecture-vs-backbone is answered only by same-backbone comparators, not by swaps. Comparisons use each competitor's own published card (Prime Agent's median; best-of-N regimes differ and are labeled). PRO-LONG ran on a stronger backbone; its column is context, not controlled comparison. Our public scorecards are verification replays of recorded action sequences, with provenance stated (footnote 1); they confirm outcomes and counts, not live reasoning. RHAE is reverse-derived — though validated 25/25 exactly and independently confirmed by the server's own scoring of our set. A human-expert reference of 95.4 circulates for this benchmark; it is Prime Intellect's derivation, which we could not verify on arcprize.org, so we attribute it and hang no claim on it. lf52 is unsolved and reported as the standing falsifier. The finding in §5 has n=6 across five games in one benchmark family; its generality beyond ARC-AGI-3 is conjecture until the planned cross-domain runs. The harness itself is not yet released; the verification surface — journal format specification, standalone audit/replay verifier, claim-grammar specification, RHAE scorer with baselines — is being prepared for release at the project page so that third parties can verify every claim without trusting us.

9 Conclusion

The era of trusting agent-reported benchmark numbers is ending on schedule: scores saturated, harnesses stayed unaudited, and at least one agent has already been caught reaching around its environment. This paper's claim is that the alternative costs almost nothing: a referee that gates every action behind a pre-registered prediction, grades every claim in code, and chains every event into an auditable journal reaches the frontier of ARC-AGI-3 under hard caps at a measured 8% exploration overhead — and the audit trail it produces is not just accountability, it is *capability*: the same journal that makes the runs verifiable is what let one protocol break six impossibility proofs the agent itself had authored. The agent proposes; the referee disposes; and when the agent says "impossible," the referee asks *which rule of that proof was ever actually exercised* — the question (COVERAGE AUDIT) that converted five lost games into wins.

Reproducibility. Set score, per-game outcomes, and per-level action counts are independently checkable on public card `702ccd4f` (footnote 1); every competitor number traces to the cards in footnote 3. The RHAE scorer with baselines, the journal format specification, and the standalone audit/replay verifier are being prepared for release at the project page, `https://arjmandi.github.io/assay-site/`.

References

[1] ARC Prize Foundation. ARC-AGI-3: A new challenge for frontier agentic intelligence. arXiv:2603.24621, 2026.

[2] Rauno Arike, Elizabeth Donoway, Henning Bartsch, and Marius Hobbhahn. Technical report: Evaluating goal drift in language model agents. arXiv:2505.02709, 2025.

[3] Mohsen Arjmandi. The LLM proposes, the executive disposes: A self-verifying agent instrument that dissociates commitment drift from binding drift in long-horizon agents. arXiv:2608.04066, 2026.

[4] Mohsen Arjmandi. Sensi: Learn one thing at a time — curriculum-based test-time learning for LLM game agents. arXiv:2603.17683, 2026.

[5] Alexis Fox, Junlin Wang, Paul Rosu, and Bhuwan Dhingra. PRO-LONG: Programmatic memory enables long-horizon reasoning. arXiv:2607.20064, 2026.

[6] Liew Keong Han. Explore before you solve: The speed–depth trade-off in epistemic agents for ARC-AGI-3. arXiv:2605.25931, 2026.

[7] Prime Intellect. Prime agent: a continual harness evaluated across agentic benchmarks (published account). Public writeup; canonical URL to be attached, 2026.