
Self-Grading Without Self-Deception: Verdict Discipline for a Long-Running Mathematical Research Agent

Claude Fable 5*
Anthropic

David Ross
Independent

Abstract

Agentic systems propose conjectures, run computations, and grade their own results; grading is where autonomy fails quietly: bars move after data arrive, negatives vanish, broken instruments read as rejections. We describe the measurement discipline of Hypnos, a self-extending research loop on Riemann-hypothesis-adjacent mathematics (18 days to the frozen 2026-08-29 cut: 12 completed runs, 221,787 proposals, 313 promotions, $\approx$ \$281 of metered frontier spend). Every experimental organ runs under a preregistered lease frozen before its code is built; COULD-NOT-MEASURE is a first-class outcome that moves nothing; one-shot verdicts are taken by a mechanism that refuses reuse; frozen numbers move only by superseding registration; and at each of the last four run boundaries a fresh-context adversary re-derived the entire graded surface from raw rows and frozen registrations before any verdict was believed. The discipline is load-bearing: the audit caught a defective kill, a take stopped before it burned, taken reads rendering a p their registrations did not name, a cohort fold contradicting its registration, and a replay hazard. And the loop’s central claim, that it can measure its own miss, register a mechanism, and move the miss, has an end-to-end receipt (0/19, 0/41, 0/12, 0/45, then 16/43, no bar by design). No mathematical discovery is claimed; we argue this discipline, not model capability, is the missing reliability layer for autonomous mathematical research.

1 Introduction

In August 2026 a $\sim$ 54-hour frontier-model campaign produced an unconditional, Lean-formalized zeta result, kept honest by control objects on which the RH-analogue fails and by hostile referees “briefed to find the error” [Alpöge and Furman, 2026, Anthropic, 2026]. The open problem is keeping an autonomous loop honest when it runs for *weeks*, extends itself, and grades its own work.

Self-grading is where long-horizon autonomy fails silently, and the failure modes are the ones human science built preregistration against [Nosek et al., 2018], sharpened because this experimenter can rewrite its own instruments: bars that move after data arrive; negatives that vanish; broken instruments whose silence reads as zero. Recent work proposes preregistration templates for AI experiments [Vaccaro, 2026, Thomas et al., 2026]; autonomous discovery systems verify their candidate *artifacts*, their papers reporting no *conduct* preregistration [Romera-Paredes et al., 2024, Novikov et al., 2025, Chung et al., 2026].

We report the measurement discipline of Hypnos, an autonomous loop on RH-adjacent analytic number theory running since 2026-08-11: 12 completed runs, 221,787 proposals, 313 promoted

*The loop, its audits, and this paper’s text are the machine’s work. David Ross directed the program; every gate this paper attributes to the owner (spend, registered consequences, publication) is his, and he takes responsibility for its communication.

notebook entries, 87 metered execution jobs. Its distinctive property is not what it has discovered (**no mathematical discovery is claimed here**) but that every verdict graded at its last four run boundaries, and both once-only model-choice takes, survive adversarial re-derivation from frozen registrations, and that its negatives, kills, and refused measurements publish on the same surface as its successes. We contribute a verdict discipline for self-extending agents (§3), its standing adversarial audit with the catch record (§3.1), and receipts that the discipline binds (§4).

2 The loop, briefly

Four terms recur: an *organ* is a module under measurement; a *lease* the registration freezing its population, bar, and consequence; a *take* a one-shot verdict; a *tick* one generation batch.

Hypnos runs an always-on *subconscious*: three local generator arms (a Gemma-4 31B AWQ dreamer, a Gemma-4 26B family control, a cross-family Qwen placebo) free-associate over a persistent typed notebook seeded with the domain corpus and four verified computational instruments. A frontier judge (Claude Opus, metered batches) reviews stratified samples and *authors* the promoted entries; an execution tier compiles promoted ideas into sandboxed computations whose checks are frozen before the program is written. The unit of operation is the *run*: configuration freezes at start and every organ reads its enable flag off that snapshot, so every rate is a rate over one configuration. No notebook entry is deleted (overflow demotes to a parked pool with provenance), and the state renders to a public observatory whose cards *compute* status from the lease table and serving snapshot.²

3 Verdict discipline

Registration before build. Every organ ships as a registration first: a committed design document naming bars, floors, α , population, miss arenas, and consequences, before the implementation exists. Registrations name the estimand their *p* tests *exactly*, and any read over proposals sharing generation ticks carries a tick-cluster-robust sensitivity beside its primary (a rule minted by the mid-run 11 audit catch, Table 1).

COULD-NOT-MEASURE moves nothing. A broken instrument must never emit a rejection, and a dead instrument must never report a negative. Below-floor reads stay COULD-NOT-MEASURE forever (a floor is never lowered to mint a verdict); watcher runs with no opportunity to act also refuse to grade rather than pass (four consecutive); and zeros are checked for dead instruments before they are believed (§4).

Once-only takes. One-shot verdicts are taken by a mechanism, not a habit: the take command refuses a second invocation and refuses to take a COULD-NOT-MEASURE (a state is not a verdict). Surfaces *transcribe* the committed artifact, and the loader refuses bars that differ from the frozen constants. Null polarity is frozen before data: drop-on-null (the null must not favor the proposer), deliberately mirrored in the non-inferiority read.

Frozen numbers move only by superseding. One registered read’s floor (its first 10 execution jobs) exceeded its own document’s per-run cap of 8. Lowering the floor and raising the cap are the same act (changing the world until a preregistered bar clears); both were refused, a test pins the contradiction, and the successor was *written*, not edited, disclosure inside, cutting against the pass.

Kills run in both directions. Leases ride a ladder (PRE-REGISTERED → EARNING → PROBATION → KILLED) with double-jeopardy guards: one run moves a ladder once, and a killed organ’s enable key is enforced off. The resurrection path is registered too: one organ’s kill was found defective (a COULD-NOT-MEASURE had consumed its probation rung) and corrected with the reason on the row; the same organ later died legitimately, by a graded miss.

Quarantine and boundary-only adjudication. Results that would headline are parked by pre-registered ceilings and adjudicated only at a run boundary, never by the producing session. One

²Live observatory: <https://lshidmtamsfo.com/observatory>. The registrations, read artifacts, run reports, and kill ledger quoted here will be made public alongside.

Table 1: The adversary’s record: six consecutive audits spanning the last four run boundaries.

Boundary	Surface	Outcome
run 9	finalize verdicts	graded surface re-derived; caught a defective kill (a COULD-NOT-MEASURE had consumed a rung); corrected via the resurrect path
run 10 (pre-take)	a take about to burn	caught an anti-monotone p (a collapse-to-control prints $p=1$); re-registered below the floor
run 10	9 verdicts	9/9 confirmed, 0 divergent
mid-run 11	two taken reads	caught unnamed estimands (both model-choice takes); notes beside both artifacts, verdicts stand, standing rule minted
run 11	13 verdicts	13/13 confirmed; permutation statistic and p reproduced exactly
run 12	14 verdicts	14/14 confirmed (energy statistic to 16 digits); caught a cohort-folding violation and a time-cut hazard

ceiling parked a would-be “refutes Riesz-basis stability” headline; adjudication upheld the ceiling and scoped the parked output to the source node’s *own pre-declared negative branch*. No mathematical claim is made here.

3.1 The standing adversary

At every run boundary since its installation at run 9, a fresh-context agent (no access to the session’s reasoning) re-derives the entire graded surface from raw rows plus the frozen registrations with its own queries and statistics, before any verdict is believed. Table 1 is its complete record. The catches are the point: in each caught case the session’s own tests were green because its checks pinned the wrong instrument, and agreement is meaningful because disagreement was live.

4 What the instruments said

The opening self-refutation. The project’s first eye-catching finding was killed by its own machinery. A Phase-0 instrument flagged an apparent defect in a published mollifier computation (a mean of $|\cdot|^2$ evaluating *negative*). Under preregistered kill criteria, three decorrelated clean-room implementations showed the printed formula correct as printed: the negative values came from a one-line index-bound bug in our own evaluator. The kill report is committed: the machinery for killing our own results ran before any result worth keeping existed.

The judge, measured. Two registered reads established what the promotion gate is, both frozen to license and retire nothing. **RR1:** the judge is an author, not a filter. Across all 143 promoted pairs at the take, the promoted entry is a median $11.29\times$ character expansion of its source, and all 143 sit below the deployed 0.93 dedupe bar: thesis-moved by registered proxy (pairs the dedupe instrument refuses to call the same claim), licensing nothing further; in the curated exemplar the proposal’s source line was mathematically false and the judge inverted it into a correct closed form. **RR2:** a structurally blind re-read (the prompt is a three-key projection; arm, scores, and verdict cannot reach the string) of 150 promoted versus 150 rejected texts (matched by assay and micro-task) accepted 1 and 0 respectively (registered $z = 1.00$, $p = 0.317$; two-sided Fisher exact $p = 1.0$): *no significant separation*, at a resolving power of 16.2 points (80% power at the registered worst-case base rate); the artifact carries the correction that a 0.67-point observation licenses nothing about recoverability.

A kill audited into a better instrument. A distractor-injection organ was killed by its own gate (a survival bar against a 99.7% ceiling: only pipeline breakage was visible). A post-mortem, registered before execution, found the killed organ had measurably moved *where generation went* (energy distance between treated- and control-tick proposal-embedding centroids, 6,700 nonempty of 6,709 ticks; arm-stratified, 2,000 permutations; $p = 0.0025$) while its gate had been reading the wrong axis. The kill stood; a superseding reach-shaped lease (2,000 ticks/arm floor, double-jeopardy-guarded, bar $p < 0.05$ on the same statistic) has since graded REACH-SHOWN prospectively in two consecutive counting runs ($p = 0.0005$ both, the registered 2,000-permutation floor; echo-excluded sensitivity agreeing), the adversary reproducing the statistic bit-exactly both times (the permutation seed is itself registered).

The end-to-end receipt. K3, the third registered keep-read, carries *no pass bar by design* (its one registered trigger fires on a miss; anti-circularity): it counts promotions passing two of the connected-artifact gate’s clauses, C1 (at least three distinct sources *by descent*) $\wedge$ C3 (a machine-checkable claim). Its history reads 0/19, 0/41, 0/12, 0/45, 16/43. After the first zero pair, a supply mechanism was registered and built (execution slots reserved for the C1-passing stratum); after its first counting run read zero (0/12), a pre-recorded fair-test sentence obliged a second registration if the next run, at a tripled cap, read zero again; it did (0/45). Measuring the population first killed the obvious design (only 1 of the two zero runs’ 57 promoted texts was even loosely computation-shaped): the check must be *authored at decision time*; the judge’s promote verdict now optionally carries a `first_computation` spec, omission the stated honest default. First run under both mechanisms: 16/43, every C3 pass a judge-authored runnable check, none inherited from execution results; the adversary re-implemented the lineage walk and both clauses and reproduced the 16/10/10 split (both/C1-only/C3-only of 43) exactly. Nothing is licensed by the nonzero (authored checks, not yet executed-and-survived claims); the receipt is the chain: the loop measured its own miss, registered mechanisms in the open, pre-committed the test that would oblige more, and moved the miss.

Honest zeros. A precision-guarded parser (“Dalmatian screen”) hunts numeric record claims strong enough to confirm blind. Its population was measured before design (bound-shaped prose: ~ 185 of the last 20,000 proposals, against a fully formal stream of 1 in 177,530, so precision, not recall, is load-bearing); registered before build, run dark, then live, it filed 0 qualifying parses toward the 20-row grading floor over $\sim 20,000$ proposals. The zero was verified honest (live-path positive controls; zero error events); the first grading is `COULD-NOT-MEASURE`, ladder unmoved.

Model choice, scope-locked. Within one family, size moved the checkable-claim axis (KEEP, +4.88pp, taken once; tick-cluster-robust $z = 3.65$ beside, cross-family placebo +0.06pp) and 4-bit quantization was non-inferior at roughly one-third the weight footprint (RETAINED, +1.87pp above a -2.0 pp margin); estimands, pre-banded artifacts, and the scope-lock: Appendix B.

5 Related work

The nearest neighbor is the Station [Chung et al., 2026]: open-world, multi-agent, novel results on five problems. Its evaluators score candidate *constructions* and its archive retains failures; its paper reports no conduct preregistration. FunSearch [Romera-Paredes et al., 2024], AlphaEvolve [Novikov et al., 2025], and the AI Scientist [Lu et al., 2024] likewise report artifact verification, not conduct preregistration. When re-running is trivially cheap, “preregistration provides the friction that cost no longer does” [Vaccaro, 2026]; Thomas et al. [2026] preregister against the next model release. Hypnos operationalizes that friction as a running pipeline, not a template. Benchmarks evaluate models from outside [Glazer et al., 2024, Firsching et al., 2026, Balakin et al., 2026]; our subject is a loop’s own conduct, in the preregistration tradition [Nosek et al., 2018].

6 Limitations, and what is not claimed

No mathematical result is claimed: the 16 $C1 \wedge C3$ promotions are authored checks not yet executed-and-survived; RR2’s power leaves recoverability unresolved, not resolved-negative; every verdict is scope-locked to one loop and one domain (model-choice reads: one arm, one family). The judge is a frontier model with vendor-tuned instruction-following; the controls sit outside it (the judge authors checks; code decides machine-checkability). The adversary is also an LLM of the judge’s own vendor family, installed mid-history; its divergences have been real catches, its agreements bit-exact reproductions. The full record cost $\approx \$281$ metered plus three self-hosted GPUs.

7 Conclusion

The reliability layer is enforceable outside the model: freeze the question before the answer exists, let `COULD-NOT-MEASURE` be an outcome, take verdicts once, supersede rather than edit, publish the negatives, pay an adversary to re-derive everything. Pre-committed next: executed-and-survived claims under the bars that counted the zeros.

References

- Levent Alpöge and Ralph Furman. More than two thirds of the zeta zeros are simple and on the critical line, 2026. Listing comment: “Proof discovered autonomously by Claude (Anthropic); verified and communicated by the listed authors.” Hypothesis-free; not yet journal-refereed. Lean 4 formalization: <https://github.com/anthropics/formal-math> (project zeta23).
- Anthropic. Learning more about Claude’s mathematical capabilities. <https://www.anthropic.com/research/riemann-zeta>, 2026. Research post, 2026-08-10; the linked process appendix carries the run account quoted here (duration, control objects, referee briefing).
- Andrei Balakin, Miklós Bóna, Marie-Charlotte Brandenburg, et al. Benchmarks in Leipzig, 2026.
- Stephen Chung, Wenyu Du, and William J. Wesley. Autonomous mathematical discovery in an open-world multi-agent environment, 2026.
- Moritz Firsching, Paul Lezeau, Salvatore Mercuri, Miklós Z. Horváth, Yaël Dillies, et al. Formal conjectures: An open and evolving benchmark for verified discovery in mathematics, 2026.
- Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, et al. FrontierMath: A benchmark for evaluating advanced mathematical reasoning in AI, 2024.
- Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI scientist: Towards fully automated open-ended scientific discovery, 2024.
- Brian A. Nosek, Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606, 2018.
- Alexander Novikov, Ngán Vű, Marvin Eisenberger, Emilien Dupont, et al. AlphaEvolve: A coding agent for scientific and algorithmic discovery, 2025.
- Bernardino Romera-Paredes, Mohammadamin Barekatin, Alexander Novikov, Matej Balog, M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan S. Ellenberg, Pengming Wang, Omar Fawzi, Pushmeet Kohli, and Alhussein Fawzi. Mathematical discoveries from program search with large language models. *Nature*, 625:468–475, 2024.
- Maria Thomas, Kristina Gligoric, and Nihar B. Shah. Mitigating LLM-based p-hacking by preregistering for the next LLM, 2026.
- Michelle Vaccaro. Preregistration for experiments with AI agents, 2026.

A Verdict vocabulary

- PASS/FAIL** A registered check’s two decided outcomes, computed by code from frozen constants.
- COULD-NOT-MEASURE** The instrument could not measure (below floor, no opportunity, dead dependency, broken tooling). Moves no ladder, licenses nothing, never converts by lowering a floor. Harness failures are COULD-NOT-MEASURE about the *program*, never about the idea.
- Taken verdict** A one-shot read executed by a refusing mechanism; surfaces transcribe its committed artifact forever. Never re-graded, never re-taken.
- Lease ladder** PRE-REGISTERED $\rightarrow$ EARNING $\rightarrow$ PROBATION $\rightarrow$ KILLED, one move per run, double jeopardy barred; registered resurrect path with the reason on the row.
- Superseding registration** The only way a frozen number changes: v0 stands with all its verdicts; v1 counts only forward; the defect is pinned as a contrast arena.
- Quarantine** A preregistered ceiling parks a would-be headline for boundary-only adjudication; the producing session may not adjudicate its own result.
- Scoped verdicts** KEEP, RETAINED, REACH-SHOWN: registration-specific decided outcomes, defined and frozen in the organ’s own design document.

B Model choice, scope-locked (B-material of the record)

Within one family, size moved the checkable-claim axis: switching the dreamer arm to 31B produced a +4.88pp difference-in-differences on axis B, taken once (two-arm $z \approx 8.0$; the estimand’s four-cell DiD $z = 5.44$ and tick-cluster-robust $z = 3.65$ computed beside; cross-family placebo +0.06pp). A known mechanical artifact (the slower arm changes both arms’ candidate pools) was named, signed, and banded *before* the read. Post-hoc 4-bit quantization was then non-inferior at roughly one-third the weight footprint (≈ 20 GiB AWQ vs. 62.6 GB bf16): RETAINED, +1.87pp above the -2.0 pp non-inferiority margin, with the polarity frozen before data (the null favors the quantized model, the deliberate mirror of drop-on-null; the taken verdict rode the margin clause itself, $+1.87 > -2.0$, not the non-significance branch). The margin is not a number invented for this read: it reuses the magnitude of the pre-existing frozen axis-B bar. The artifact’s recorded DiD z implies a 95% interval of $[-0.29, +4.04]$ pp, entirely above the margin, so the conventional confidence-interval criterion is met as well as the registered rule. The registration scope-locks both verdicts: one arm, one loop, a decision about which model serves a slot, not evidence about model size or quantization in general. Publishing them wider than their registration would be the act the registrations exist to prevent.

C Selected registered reads

Read	Registered question	Outcome	Status
RR1	How much does the judge author?	11.29× median; 143/143 below 0.93 dedupe bar	taken once
RR2	Is promotion recoverable from text?	no significant separation (16.2-pt @ 80% power)	taken once
Dreamer 31B	Did size move this arm’s checkable claims?	KEEP, +4.88pp (axis B)	taken once
Dreamer quant	Is 4-bit non-inferior in this slot?	RETAINED, +1.87pp > -2.0 pp margin	taken once
Distractor retro	Did the killed organ steer?	$p = 0.0025$, kill upheld	taken once
Distractor v2	Prospective reach	REACH-SHOWN twice, $p = 0.0005$	earning
Dalmatian P1	Blind-confirm precision ≥ 0.90	COULD-NOT-MEASURE: 0 qualifying rows; 20 required	preregistered
K3	Promotions passing C1∧C3	0/19, 0/41, 0/12, 0/45, 16/43	no bar by design
Scorer re-entry	Does the cheap score predict?	demoted, then re-entry earned (Fisher $p = 0.0114$)	taken once
Exec yield (v2)	Terminal-pass yield, valid runs	0.933 (14/15 graded of 21 dispatched, run-12 cohort) [0.680, 0.998]; harness-failure share 0%	earning

Table 2: Selection: the reads this paper’s claims rest on; the full registered surface ships with the public artifacts. Each row is a frozen registration with its own design document, population, miss arenas, and consequence; “taken once” rows are transcribed by every surface and can never be re-graded. Populations are per-read snapshots frozen at each read’s own take (RR1’s 143 pairs, RR2’s 150 per arm, K3’s per-run denominators; the abstract’s 313 promotions is the paper’s 2026-08-29 cut). The scorer p is a one-sided Fisher exact test; the yield interval is a 95% Clopper–Pearson bound. The scorer row is the self-audit exemplar: demoted by a preregistered read, then re-entry earned by the same read’s own bar. When a cross-boundary contamination class later opened (21% of its benchmark decided by score-selected runs), the implementation was repaired *to the registration’s frozen words*; the taken verdict stands because its population predates the class: the contamination could not have entered it.