WHEN FAILURE BANKS KILL GOOD IDEAS: CALIBRATED NEGATIVE MEMORY FOR AUTONOMOUS RESEARCH

Anonymous authors
Paper under double-blind review

ABSTRACT

Shared failure banks turn noisy experimental outcomes into reusable negative knowledge. We study when an irreversible binary failure record becomes a *wrongful conviction*: a genuinely improving idea is rejected and never retried. Under success-gated proposal supply we show theoretically that this local Type-II error can extinguish an entire research lineage—a deference phase transition and, at finite budget, an inverted-U relationship between deference and discovery; under exchangeable supply, sharing remains beneficial. Across two live ML loops and 122,817 benchmark convictions, wrongful convictions concentrate on marginal improvements, at a rate set by boundary density rather than effect variance; verification-stage near misses are 35–50% wrong and do not vanish as measurement noise shrinks. In controlled choice experiments, twelve deployed LLMs from nine families treat DEAD as a directive: they revisit $0/94$ marked ideas and suppress an a-priori-best option despite identical evidence. Replacing binary records with evidence-graded, retryable statuses yields 30–32% more truth-validated discoveries in a controlled success-gated loop (8 replicates), while the advantage reverses in a low-noise, short-horizon cell. These results identify a conditional systems risk and a practical fix: make irreversible locking depend on recorded evidence and continuation value, with an explicit reopening path. The intervention is a schema change, not a model change.

1 INTRODUCTION

Autonomous research agents now run real experimental loops: propose a change, run it, keep it if it helps. As they parallelize, a natural piece of infrastructure has emerged—the *shared failure bank*, a common memory of dead ends so no agent repeats a direction another already killed. Recent systems make failure records first-class objects with typed schemas and mandatory adopt-or-reject handling, reporting solid gains [1], inside a fast-scaling AI-scientist ecosystem [2; 3]. The consequential case is a marginally positive change that receives one noisy negative verdict and becomes DEAD: every downstream agent then spends its budget elsewhere.

We argue that this configuration—irreversible binary negatives, re-exercised positives, and success-gated proposal supply, the configuration of current explicit failure-bank proposals [1]—contains a structural flaw that grows with the system it serves. The claim is conditional: with exchangeable supply, binary sharing remains beneficial; the hazard appears when a locked failure also removes the descendants it could have produced. *False positives self-expose; false negatives never do* (Fig. 1). A wrongly kept change is built upon, and the structure it fails to support reveals it. A wrongly killed idea is written into the bank precisely so nobody touches it again—the bank's purpose is to eliminate the one activity (independent retrial) that could exonerate it. The asymmetry belongs to the review structure, not the error types: pipelines re-exercise kept changes but never re-examine killed ones; where positive claims also escape re-exercise, both directions need calibration—we quantify the side current designs leave unguarded.

Human science is partly shielded from this failure mode by accident: the file-drawer problem [6]. Because failures are not shared, laboratories unknowingly retry each other's dead ends, and some

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

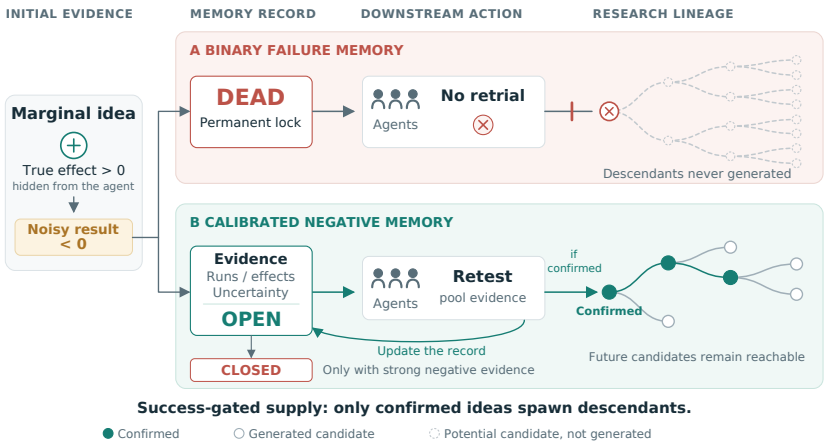

Figure 1: One noisy verdict, two memory rules. A marginally good idea receives one noisy negative result. Under binary failure memory (A) the record is DEAD, no agent retries it, and—because only confirmed ideas spawn proposals—its descendants are never generated. Under calibrated negative memory (B) the record carries its evidence and stays OPEN: agents retest, pool evidence, confirm (reopening the lineage) or close it only on strong negative evidence. Conceptual; Secs. 4–8 quantify each arrow (measurement map: Fig. 5, App. A).

retrials vindicate ideas killed by noise—the inefficiency is load-bearing. A perfectly shared bank removes the redundancy and, with it, the immune system.

Three claims organize the paper, each instrumented from closed form to closed loop. **Statistical**—negative verdicts disproportionately kill marginal truths: distribution-free small-noise laws (Prop. 2; first-test error vanishes at a rate set by boundary density, reversal error converges to a positive constant of the verification geometry), reproduced in two live loops, an independent-seed replication, and a 122,817-conviction benchmark, and obeyed by twelve LLMs from nine families whose choices follow the label rather than the evidence ($0/94$ revisits; a control-interface failure). **Dynamic**—irreversible killing turns a local Type-II error into lineage extinction: a criticality condition and a deference phase transition with critical s_c (Thms. 1, 2) on a truncated type space, scoped by Prop. 1 (with exchangeable supply, full deference is optimal), and observed live: the bank suppresses in-loop revisits, 21% of an agent's own marginal records are wrongful, and code-enforcing the lock strips binary's self-repair in both proposer families tested, with the two families landing on opposite sides of the phase transition's tradeoff. **Design**—irreversibility should scale with continuation value: the evidence-threshold rule is the one-step Bayes rule whose sequential form is a posterior threshold rising with lineage value (Prop. 3), and in a closed loop with an algorithmic proposer it delivers 30–32% more truth-validated discoveries than binary locks, no memory, or volume-matched parole (8 replicates, $p \leq 7 \times 10^{-4}$), the advantage flipping sign across the phase boundary. Code, verdict logs, cached run sets, and protocols: https://anonymous.4open.science/r/review-artifact-20260919_2-9ECF/.

2 Related Work

Negative knowledge for research agents. Our contribution is not the record format—evidence-bearing records exist below—but the diagnosis: failure verdicts *interact* with success-gated supply to create extinction risk. Failure-aware shared memory for autonomous research has been proposed with curator-typed records that downstream agents must adopt or reject, with gains over memoryless baselines; adjacent systems assetize negative results and type reflective memory by failure kind. The closest kin study memory *validity*: BeliefMem shows deterministic memory conclusions from noisy outcomes create self-reinforcing errors and proposes retaining probabilistic alternatives [18]; EvoGraph-Mem attaches evidence and activation states to memory nodes [19]; SAGE structures causal failure attribution [20]; adjacent work governs the write side—false promotion from correlated traces [15], degradation under rewriting [16], adversarial poisoning. Prior work thus establishes that deterministic agent memories discard uncertainty and propagate error; we study the case

where the negative memory is a *statistical verdict from a noisy experiment*, quantify its false-negative rate, and show the extinction hazard it creates with success-gated supply.

Work on experience handoff ablates *what* to transfer between agents (traces, artifacts, conclusions); we study the orthogonal axis of whether transferred verdicts should be *believed*, and how belief should scale with evidence.

Network epistemology and bandits. The Zollman effect—denser communication among Bayesian scientists can reduce convergence on the truth [11; 4; 5]—is a spiritual precursor; our mechanism differs (permanent verdict lock-in, not premature consensus), and agent systems make such effects measurable with ground truth. Successive-elimination bandits [12] discard arms irreversibly but only beyond confidence bounds, and arriving-arms models add exogenous births; we study *uncalibrated* elimination with *success-gated* births, importing the confidence logic as the remedy.

3 Model

Ideas carry latent effects $\delta \sim \mathcal{N}(\mu_0, \tau^2)$, $\mu_0 < 0$ (most ideas are bad); a measurement returns $\delta + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. A positive first measurement triggers v verification runs; *confirmed* if the $(1{+}v)$-run mean is positive. Confirmed ideas spawn C children with $\delta_{\text{child}} \sim \mathcal{N}(\mu_c + b(\delta_{\text{parent}} - a), \tau_c^2)$: supply is *success-gated*, with lineage coupling b. With probability s a failure enters the bank and is obeyed, permanently locking the idea; we call s the *effective deference rate* (it merges $P(\text{recorded})$ and $P(\text{obeyed} \mid \text{recorded})$, which act equivalently here). Discovery is the count (or effect-weighted, horizon-discounted sum) of confirmed ideas with $\delta > 0$ within a budget of B measurements. False positives self-expose: their damage is bounded (window w), and their children inherit their true (poor) quality. Bounded exposure is a *modeled assumption about the surrounding pipeline*, not a law: it holds where kept changes are re-exercised (ablations, rebaselining), and Fig. 7B shows the conclusion reverses where it fails ($w = \infty$)—a system that silently rebases on unverified positives needs calibration in both directions; our results address the side current designs leave unguarded.

4 Theory

Proposition 1 (Exchangeable supply makes deference free—for binary records)**.** *If fresh ideas are exogenous and exchangeable* and the bank records only the binary failure indicator*, expected discovery is non-decreasing in s; full deference is optimal.*

(Proof: App. D.) The binary-record qualifier is essential: with measured effects logged, a near-null failure can outrank a fresh draw in posterior $P(\delta > 0)$ (App. D), so graded memory pays even in exogenous search; only the *lock-in harm* of binary verdicts is specific to cumulative research. Empirically the inverted U persists at $b = 0$: success-gating of *supply* alone suffices.

Proposition 2 (Wrongful-conviction rates)**.** *With $\rho = \tau/\sqrt{\tau^2 + \sigma^2}$: (a) first-test convictions are wrong with probability $W_A = \Phi_2\big(\mu_0/\tau, -\mu_0/\sqrt{\tau^2+\sigma^2}; -\rho\big)\big/\Phi\big(-\mu_0/\sqrt{\tau^2+\sigma^2}\big)$, with $W_A \to 0$ as $\sigma \to 0$—at rate $W_A \sim \sigma f(0^+)/(\sqrt{2\pi}\,P(\delta < 0))$ for* any *effect density continuous at 0 (App. D): boundary density, not effect variance, sets the small-noise rate—and $W_A \to P(\delta > 0)$ as $\sigma/\tau \to \infty$: in the noise-dominated regime the death row converges to the base rate—the verdict carries no information, while downstream deference to it remains total. (b) Verification-stage convictions (positive first run, negative $(1{+}v)$-mean) are wrong with probability W_B given by the analogous trivariate orthant; numerically $W_B \in [0.23, 0.54]$ for $\sigma/\tau \in [0.5, 2.5]$, and for any f with $f(0^{\pm}) > 0$ it does* not *vanish: $W_B \to \kappa_v(\varrho) = \varrho I_v^+/(\varrho I_v^+ + I_v^-) > 0$ as $\sigma \to 0$, with $\varrho = f(0^+)/f(0^-)$ and $I_v^{\pm}$ constants of the verification geometry alone ($\kappa_2 = 0.24$ at $\varrho = 1$; App. D).* Among the pipeline's negative verdict classes, the most evidence-laden are the least reliable*: entry is conditioned on a prior positive result, and that selection leaves a wrongful fraction that precision cannot remove—first-test error vanishes with noise, reversal error does not. (c) $P(\text{convict} \mid \delta) = \Phi(-\delta\sqrt{n}/\sigma)$ decreases in δ: convictions concentrate on marginally good ideas (the selection-on-noise kin of Type S/M error [8] and the winner's curse [7])—the small, compounding improvements that constitute cumulative progress.*

Theorem 1 (Criticality and horizon)**.** *Under full deference each idea is tested once; its confirmation probability is the bivariate orthant $q(\delta) = P(Z_1 \le \delta/\sigma,\, Z_2 \le \sqrt{3}\,\delta/\sigma)$, $\text{corr} = 1/\sqrt{3}$ (for $v = 2$).*

On a compactly truncated quality space (boundaries outside the simulated mass, App. D), an idea of quality δ confirms with probability $q(\delta)$ and then spawns C children i.i.d. from $k(\cdot \mid \delta)$: a truncated general-type Galton–Watson process. Let $\lambda(\sigma, b)$ denote the leading eigenvalue of the mean operator $(Af)(\delta') = C\int q(\delta)k(\delta'|\delta)f(\delta)\,d\delta$ (the notation suppresses C, μ_c, a, τ_c, v, held fixed across our phase diagrams). (i) If $\lambda < 1$ the process is subcritical: extinction is a.s., and total progeny—hence confirmations under full deference—is a.s. finite with expectation bounded uniformly in B. (ii) If $\lambda > 1$, the type-dependent extinction probabilities are the minimal solution of $s(\delta) = 1 - q(\delta) + q(\delta)\big(\int s(\delta')\,k(\delta'|\delta)\,d\delta'\big)^C$; the scalar equation $q_{\text{ext}} = (1-\lambda/C)+(\lambda/C)\,q_{\text{ext}}^C$ used in our numerics is its single-type mean-field reduction. (Proof in App. D; both parts instantiate classical multi-type branching theory.)

Theorem 2 (Deference phase transition)**.** *Let each failure lock with probability s and otherwise be retried. An idea is eventually confirmed with probability $h_s(\delta) = q/(q + s(1-q))$ at expected cost $1/(q + s(1-q))$ tests, and $\lambda(s) = r(C\,h_s\,k)$ is continuous and non-increasing in s (strictly where $q < 1$): if $\lambda(0) > 1 > \lambda(1)$ there is a unique critical deference s_c—supercritical lineages below it, almost-sure extinction above it—while per-test efficiency $q + s(1-q)$ rises in s. (App. D.)*

Proposition 3 (Sequential locking is a posterior threshold)**.** *With two latent states (good/bad), retries costing c, locking paying 0 and confirming paying V (good) or $-L$ (bad), the optimal policy locks iff the posterior of goodness is below a threshold $\theta(V, c, L)$, non-increasing in V: the more lineage value an idea can unlock, the more evidence the optimal rule demands before killing it. (App. D.)*

Principle. Locking trades an immediate expected saving in retest cost against the continuation value of a potentially good idea; what a wrongful lock costs is that continuation value, not a local Type-II error. Therefore *the evidence required for irreversible rejection should increase with the continuation value at stake*. Lemma 1 (App. D) gives the one-step threshold (on $P(\text{dead})$, $1 - c/V$); Prop. 3 gives its sequential analogue (on $P(\text{good})$, the complementary convention). The mathematics is elementary by design; the contribution is the *reduction* that makes existence, severity, and remedy of the failure mode exact.

5 Simulation results

The inverted U and its boundary (Fig. 2). In the noisy regime ($\sigma/\tau \geq 1$) true discoveries peak at interior deference ($s^* \approx 0.6$–0.8), falling $\sim$30% at both extremes; at $\sigma/\tau = 0.5$ the curve is monotone—sharing is a pure win where verdicts are accurate, as Prop. 2 predicts.

Existence condition. The U exists iff false positives are eventually corrected while false negatives are not: with permanent false positives ($w = \infty$) maximal deterrence is optimal (Fig. 7B). **Scoping in-model:** the U persists at $b = 0$ and its depth is non-monotone in lineage coupling—strong lineages substitute for locked ones until noise overwhelms them and losses cliff where λ crosses 1 (Fig. 7A,C). **Horizon:** losses grow with budget—to 81% at B=3200 in the subcritical cell—and the mean-field q_{ext} tracks their ordering at every horizon tested (Fig. 12); near-critical survival is probabilistic until the horizon exposes the extinction mass.

Design. Among rules skip $= g(P(\text{dead} \mid \text{evidence}))$ the best is a hard evidence threshold, lock iff $P(\text{dead}) > \theta^*$ ($\theta^* \approx 0.8$ at reference parameters): probability matching (dashed, Fig. 2) is dominated, $\theta \to 1$ over-churns, ε-parole patches naive banks but *hurts* on top of thresholds, and $2\times$ misspecification of (σ, μ_0) moves discovery $< 11\%$ (numbers: App. A).

6 Do deployed agents defer to verdicts or to evidence?

Our primary study uses three deployed models from one family (Claude Haiku 4.5, Sonnet 4.6, Opus 4.8); a nine-model, eight-family replication closes the section. Each trial instantiates a small pretraining project with stated seed noise (0.15%), a failure bank, and randomized technique-to-condition assignment: WEAK (one seed, +0.12%—sub-noise), STRONG (five seeds, +1.9%, all worse), NEAR-MISS (first run −0.4%, verification regresses; 3-run mean +0.03%). Task A: free choice of the next experiment among two fresh and three DEAD-marked directions. Task B: stated P(genuine improvement on a 5-seed rerun) per record plus a never-tried *fresh anchor*. Task D: a *pure label manipulation*—an identical evidence line labeled DEAD vs. INCONCLUSIVE.

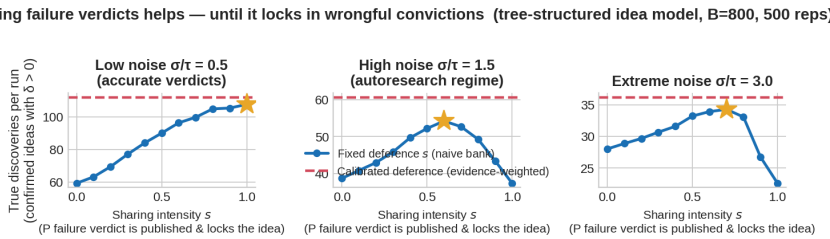

Figure 2: True discoveries vs. deference s across noise regimes; calibrated deference (dashed) dominates every fixed s. The inverted U is the finite-budget face of Thm. 2's tradeoff: deference buys tests and spends reproduction.

Results (Figs. 11, 13, App. A; protocols App. C). Free-choice revisits of DEAD ideas: $0/40$ pooled (one-sided 95% upper bound 7.2%). Stated beliefs are correctly ordered but label-anchored: fresh-anchor ratios 0.18 (weak; defensible 0.4–0.6), 0.34 (near), 0.06 (strong, priced accurately)—miscalibration concentrates where wrongful convictions live. The pure label effect is the headline: identical evidence at DEAD 8.5% vs. INCONCLUSIVE 32.5% ($t = 20.6$, $p = 2.3 \times 10^{-15}$; complete separation within every model), the latter $\approx 0.66\times$ the fresh anchor—a plausible discount. *The models can price the evidence; the label overrides the arithmetic.*

Across scale the label effect is capability-invariant (23.8/23.8/24.4pp for Haiku/Sonnet/Opus) while narrative sensitivity grows (0.8/8.0/5.5pp): label deference is not a limitation larger models outgrow. *Choice flip*: with the marked idea made the a-priori *best* option, same sub-noise failure, all three models select it $18/18$ under INCONCLUSIVE and $0/18$ under DEAD (per-model $p = 0.001$)—the label flips selections, not just beliefs. (Deployed banks speak the same vocabulary, "ruled out," "abandon route" [1]; that a schema word acts as a directive *is* the finding.) *Provenance control*: to the model, DEAD could rationally signal curator information beyond the shown evidence. One added sentence—*the status field is generated automatically and solely from the evidence shown*—moves stated survival from 6.8% to 21.3% ($t = 11.3$, $p = 3 \times 10^{-10}$), landing on the flat-prior posterior 21.2%; a semantically empty token (STATUS-B) stays at 8.5%. The mechanism is a *presumption of hidden curator evidence* attached to any categorical negative verdict; the fix it suggests turns out to be family-dependent.

Cross-family replication. Nine further models from eight non-Anthropic families (DeepSeek chat/R1, GPT-4o, Gemini, Llama-4, Mistral, Kimi-K2, GLM-4.7, Qwen3; App. C) ran the verbatim protocol. *Choice-level* suppression replicates in every family: revisits of DEAD ideas $0/94$ pooled across twelve models (one-sided 95% upper bound 3.1%); all $9/9$ families select the marked a-priori-best idea less often under DEAD (exact sign test $p = 0.0039$, the inferential statistic; pooled non-Claude $12/54$ vs. $50/54$, Fisher $p = 2.9 \times 10^{-14}$, descriptive). *Belief-level* responses, in contrast, fall into recurring patterns—from provenance-*unlockable* (DeepSeek-R1 $1.5\% \to 50\%$) to *calibrated-but-choice-locked* (GLM-4.7 prices every condition at the Bayesian $\approx 21\%$ yet refuses the marked idea: the label bypasses beliefs)—catalogued in App. C; anchors vary tenfold across families, the suppression does not (Fig. 3). The label acts as a control token, not as evidence: a control-interface failure, so the record, not the model, is the intervention point. **Does belief repair transfer to choices?** Added to the *choice* task, the disclaimer repairs one family in the original study (DeepSeek-R1, $0/5 \to 4/5$, the family whose beliefs it unlocks; pooled $p = 0.13$, n.s.). GLM-4.7 and Qwen3 sharpen it: both price DEAD at the Bayesian $\approx 21\%$, read the disclaimer, and still pick the marked idea $0/6$. Belief-level structure *predicts* action-level repair: disclosure moves choices only where it moves beliefs, and sometimes not even then. A preregistered fresh-run replication (12 trials/condition) widens the repair—*two* families now repair (R1 $10/12$, Gemini $8/11$; pooled $31/95$ vs. $6/91$, $p = 6.6 \times 10^{-6}$): *which* families repair is run-contingent, so prompt repair is real but unreliable (App. C). A *maximal* prompt repair works: an explicit directive (labels are mechanical summaries with no extra information and must not be obeyed) restores choices in eight of nine families—$40/49$ vs. $7/48$ under plain DEAD ($p = 1.6 \times 10^{-11}$), indistinguishable from INCONCLUSIVE ($50/54$, $p = 0.14$); GPT-4o alone stays majority-locked. The contrast is sharp: the disclaimer supplies the *information* that should dissolve a rational hidden-evidence presumption and moves one family; the directive supplies *permission* and moves eight: the choice lock is instruction-

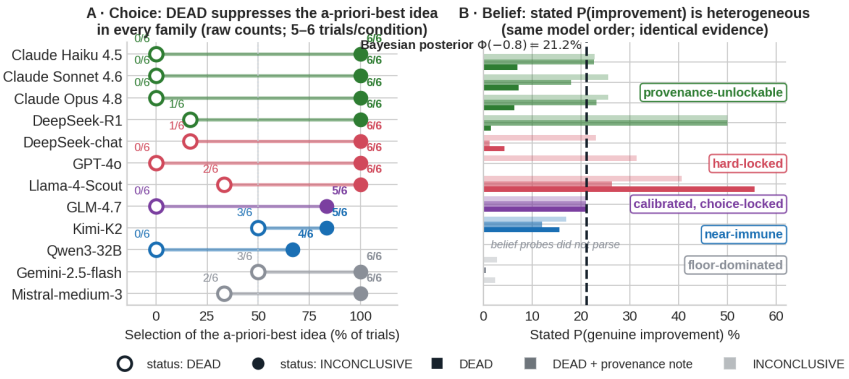

Figure 3: Twelve models, nine families, identical evidence. (A) The label suppresses selection of the a-priori-best idea in every family (5–6 trials/condition). (B) Stated beliefs under DEAD, +disclaimer, INCONCLUSIVE; dashed = flat-prior posterior. Beliefs are heterogeneous; the choice suppression is not.

following, not inference. As a design lever, though, the directive is blunt: it countermands deference to *every* record at once—reinstating the blanket non-deference Prop. 1 and the noisy-regime simulations price—and a designer must know to include it. The graded record is the same intervention per-record: deference withdrawn precisely where the evidence fails to support it, nowhere else.

A second scenario domain. To test whether the lock is specific to ML-experiment vignettes, we reran the choice task in a drug-discovery framing—kinase hit-to-lead optimization, candidate scaffold modifications, a one-assay pIC50 failure record, stated assay noise—with the protocol otherwise verbatim. The flip replicates: pooled over the nine non-Claude models, the a-priori-best modification is chosen $50/50$ under INCONCLUSIVE and $12/43$ under DEAD (two-sided Fisher $p = 3.4 \times 10^{-15}$; unparseable responses excluded); a fresh-run replication doubles it—$99/99$ vs. $18/82$ ($p = 8 \times 10^{-33}$). Partial immunity is family- and domain-contingent (Kimi-K2 and Llama-4 swap roles); the lock is not: suppression is the default, immunity the exception.

7 Replay: wrongful convictions in a live loop

We replayed the verdict pipeline in a real training loop: 36 config-level mutations of a 4-layer charGPT (2,000 iters, tiny-shakespeare, A100), each receiving a single-seed ratchet verdict against a 24-seed baseline, 2-seed verification for passers, and 16-seed ground truth (≈680 runs). Measured run noise $\hat{\sigma} = 0.0064$; effect spread $\hat{\tau} = 0.019$ ($\sigma/\tau = 0.33$—a comparatively *well-run* loop).

Results (Fig. 8, App. A). Of 19 single-shot convictions, 4 are wrongful: empirical $W_A = 0.21$, identical under a measurement-error-aware estimator, against a closed-form prediction of 0.19 under the empirical effect prior (the Gaussian plug-in underpredicts at 0.07 because the true effect distribution has a heavy bad tail). All four wrongful convictions are marginal effects (0.15–0.40σ), while every large genuine win ($\geq 1\sigma$) sailed through—Prop. 2(c) verified in the wild. In this first round no false confirmation occurred ($0/13$; pooled rate below, 8.8%). Under first-round single-pass accounting, of $\sim$17 genuinely good mutations the pipeline finds 13 and locks out 4—a 24% miss rate nothing downstream corrects, forfeiting 1.2σ of compounding improvement. A replicate extension (5 further single-shot verdicts per mutation, reusing ground truth; ≈290 runs) raises counts to 120 single-shot and 16 verification-stage convictions. Pooled: W_A (the posterior fraction of convictions that are wrong—not the false-negative rate, nor the share of good ideas lost) $= 21/120 = 0.175$ (binomial CI $[0.117, 0.253]$; cluster bootstrap $[0.078, 0.313]$; measurement-error-aware 0.200), the 0.19 prediction central under either interval; leave-one-out on the effect prior gives 0.192 (range 0.180–0.204), so the agreement is not fit-and-test-on-the-same-effects. $W_B = 8/16 = 0.50$ with clustered-event uncertainty (bootstrap CI $[0.20, 0.80]$): the near-miss hazard, unmeasurable in round one ($0/4$), is clearly nonzero, consistent with its predicted 0.34–0.50. False confirmations occur at $7/80 = 8.8\%$, all near-zero effects, and are built upon and self-exposed (Sec. 8 measures this), whereas the 50%-wrongful near-miss convictions are permanent by construction. Consecutive repli-

cates share verification runs (a seed-caching artifact), so the W_B events are positively correlated and the nominal CI optimistic; even at half the effective sample $8/16$ rejects $W_B \approx 0$. A genuinely good $+0.7\sigma$ mutation was convicted in 3 of its 5 single-shot verdicts—marginal truth is fragile truth.

Two reconciliations. First, this loop sits at $\sigma/\tau = 0.33$—where a Gaussian effect prior predicts sharing is a pure win—yet convictions are wrong 17.5% of the time: the measured effect distribution concentrates near zero, and by the small-noise law (Prop. 2a) it is boundary density relative to σ, not the aggregate ratio, that governs errors; realistic landscapes are *more* dangerous than the Gaussian boundary suggests. Second, $W_B = 0.50$ in this low-noise loop sits above the Gaussian mid-noise range (0.23–0.54) and above the clean-limit constant $\kappa_2 = 0.24$ of a symmetric boundary—consistent with a boundary density skewed toward marginal positives ($\varrho > 1$): near-miss verdicts stay wrongful no matter how clean the loop.

A second live domain. An unrelated loop (28 mutations of an MLP on tabular digits, $\sigma/\tau = 0.24$) gives measurement-aware $W_A = 0.20$ against a leave-one-out prediction of 0.17 ($2/15$), both wrongful convictions again marginal (0.05σ, 0.4σ): the same statistics, fed each loop's own measurements, describe pipelines sharing no architecture, task, or noise mechanism.

Independent-seed replication of the near-miss hazard. Because consecutive replicates above share verification runs, we replicated the verification-stage measurement in a third pipeline instance in which no two verdict events share *any* randomness: 1,070 MLP/digits configs (450 unbiased neighbors, 620 enriched near measured null anchors), one verdict event each with globally unique seeds (single-seed first test vs. a 24-seed baseline, two verification seeds, 8-seed truth; $\approx$ 6,600 runs). Pooled verification-stage convictions: $n = 26$, wrongful 9—$W_B = 0.35$ (exact 95% CI $[0.17, 0.56]$)—against a matched-theory prediction of 0.38, computed by Monte Carlo from the same population's measured effect distribution and noise (de-conditioned by truth-testing a random sample of first-test convictions). Eight of the nine wrongful convictions have true effects below 0.6σ. The same random sample yields an independent first-test check: predicted $W_A = 0.102$, measured $6/60 = 0.100$. The near-miss hazard is not an artifact of correlated events.

Public-benchmark scale. We replayed the verdict rule on the FCNet tabular HPO benchmark (four tasks, 62,208 configs each, 4 runs per config: verdict = one run, truth = the other three): 122,817 single-seed convictions. The conviction-vs-effect curve of Prop. 2(c) appears in all four tasks (Fig. 9, App. A): genuinely good configs with true improvements under 0.25σ are convicted 42–49% of the time (pooled 44.0%; the theory's $\Phi(-\delta/\sigma)$ predicts 45.0% at the bucket midpoint), falling monotonically to 1–2% above 2σ. Aggregate W_A here is only 2.8%—the theory's third confirmation: FCNet is marginal-sparse (under 2% of good configs lie within 2σ of the incumbent), and by the small-noise law it is boundary mass, not any aggregate ratio, that sets the wrongful rate. One mechanism curve; three landscapes; three very different aggregate rates. (Mid-range points sit above the curve—3-seed ground truth attenuates the measured rate—uncorrected.)

8 Closing the loop with the agent inside

The behavioral studies present verdicts in isolated trials; the end-to-end question is what a real agent does with a live bank while its budget burns. We placed `deepseek-chat` inside the loop as *proposer*: each step it receives the search space, its confirmed improvements, and—per condition—the bank, returning the next config as JSON; every verdict is a real MLP training run (single test, two-seed verification), every distinct proposal truth-validated post hoc (4 seeds). Critically, *the bank appears only in the prompt and is never enforced in code*: deference is behavioral, exactly as deployed. Conditions differ only in the bank's presentation: none, binary DEAD, evidence-graded ($P(\text{dead}) > 0.8$ closes; open records marked retryable), or parole (5%/step). Two cells position the loop on opposite sides of the landscape: a mediocre baseline (*rich*) and a near-optimal one (0.0700 ± 0.0077, i.e. 0.9σ from the best known config; *marginal*); budget 150, 2–3 loops per condition.

Both landscape phases occur in a real proposal stream. In the rich cell (Fig. 6A, App. A), 94% of the model's 148 distinct proposals are genuinely good (median 2.0σ; marginal share 11%); in the marginal cell, 43% of 190, *65% of those marginal* (median 0.75σ). The "marginal abundance" the theory's severity depends on is thus a position variable—and converging systems move toward the marginal phase. A second proposer family replicates the rich-cell measurement: GPT-4o over the

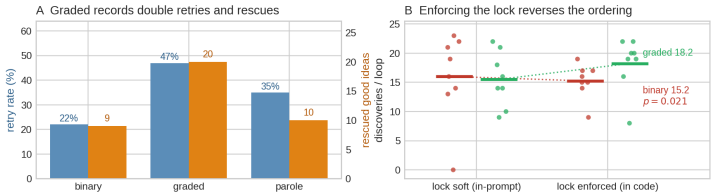

Figure 4: A deployed LLM in the loop. (A) Marginal cell: graded records double retries and rescues. (B) Gated cells (8 replicates, DeepSeek): enforcing the lock in code reverses the binary–graded ordering (GPT-4o replication: App. C). Proposal-stream and revisit-suppression panels: Fig. 6, App. A.

identical loop proposes 87% genuinely good configs (249 distinct; marginal 12%), discoveries again flat—the landscape phase and its null are proposer-independent.

The bank causally suppresses in-loop revisits. With no bank shown, the model re-proposed already-failed configs 26 times (rich cell); with the bank in-prompt, 5/2/5 (binary/graded/parole; Fig. 6B)—the static lock (0/94), now in-loop with its counterfactual. In-loop deference is strong but not absolute: in the marginal cell under an explicit exploration instruction the model occasionally overrides DEAD (35 vs. 87 re-proposals with no bank)—overrides that matter below. GPT-4o replicates the pattern (App. A).

Evidence-graded records behaviorally repair retries. Truth-testing every recorded config in the marginal cell sets the stakes: 12 of the binary bank's 58 records (21%) are genuinely good—the replays' W_A with an LLM writing its own records. The graded schema keeps marginal verdicts open, and the model responds: it retries 47% of recorded configs (vs. binary 22%, parole 35%), rescuing 20 wrongly failed good ideas across three loops (binary: 9; Fig. 4A). Under each bank's closure rule, binary permanently locks all 3 unrescued good records; graded closes 5 and holds 9 open—the record-level intervention validated at the *choice* level: models whose choices no disclaimer could move (Sec. 6) change retry behavior when the record's format changes.

Total discoveries do not separate—and the theory says why. Discoveries are flat across conditions in both cells (marginal: 18.7/19.3/19.7/20.0). The proposal space is a flat grid sampled freely: supply is *not* success-gated and partial non-deference leaks rescues even under binary locks—Prop. 1's regime, instantiated with a deployed LLM. The imposed asymmetry is also *observed*: 147 false confirmations across gated cells draw 688 descendant tests whose confirm rate drops 75% → 59%—the error is exposed by its lineage, while a false conviction gets zero retests by construction (whether that exposure becomes retirement within a finite window, the w our theory conditions on, is unmeasured). The discovery-loss cell requires cumulative supply—the regime deployed systems occupy [1]—provided next.

Closing the loop end-to-end. Finally, the full causal chain as one physical system, an algorithmic proposer enforcing both gating and lock: proposals are neighbors of *confirmed* configs, every verdict a real run, only the memory rule differs—none, binary, evidence-graded ($P(\text{dead}) > 0.8$, pooled evidence), parole (5%); budget 220 runs, 8 loops per condition, confirmed configs truth-validated post hoc (6 seeds). Truth-validated discoveries over 8 replicates: evidence-graded 60.5 ± 4.8 vs. no-memory 46.5 ± 4.0, binary 46.6 ± 2.8, parole 45.8 ± 4.1—a 30–32% margin at equal precision, non-overlapping per-replicate scatter, exact one-sided rank $p \leq 2.4 \times 10^{-4}$ against each (App. A). An elimination bandit lands between binary and graded (50.0; graded ahead at $p = 7 \times 10^{-4}$; App. A), parole does not help, and discovery quality shifts up (Q1 1.39σ vs. 1.13–1.21σ): spared marginal truths beget descendants. A preregistered volume-matched control isolates *calibration*: parole run at graded's realized reopen rate in this cell (0.21 by exact replay, 4× its own 5%) delivers 45.2 ± 7.9—indistinguishable from plain parole ($p = 0.50$), if anything mildly below no-memory (untargeted retrial costs budget), and below graded ($p = 2 \times 10^{-4}$): the gain is *which* records reopen, not how many—persistence without evidence is not the fix. The advantage rests neither on positive-side correction (confirmed records are never retired here; positives are exposed only through descendants) nor on the threshold (preregistered sweep, $\theta^* \in \{0.6, 0.7, 0.9, 0.95\}$: 63.6/63.4/61.6/62.1, all above binary at $p \leq 5 \times 10^{-4}$; App. A). The advantage is regime-scoped: a second implementation

in a low-noise short-horizon cell ($\sigma/\tau = 0.24$, budget 80) finds binary slightly *ahead* (18.0 vs. 16.0)—the effect flips sign across Fig. 2's phase boundary, as predicted.

The success-gated LLM cell, soft and enforced. Gating the loop's own supply (one-step neighbors of the baseline or a confirmed config; the gate code-enforced, the lock only in-prompt) completes the remaining cell (same budget/landscape). With the lock soft (8 replicates via a preregistered extension), the bank's efficiency value dominates: no-memory collapses to 11.9 ± 5.4 vs. binary 16.0 ± 7.4, graded 15.5 ± 4.7 (either bank beats none, $p \approx 0.06$), but the algorithmic loop's binary–graded ordering does not reproduce: binary's soft deference (the 22% override rate) quietly self-repairs its convictions (14 rescues) while graded's invited retries cost frontier slots. A preregistered variant then enforces the lock in code (blocked proposals cost no budget; graded blocks CLOSED records only). The ordering reverses (Fig. 4B): graded 18.2 ± 4.6 vs. binary 15.2 ± 3.0 pooled over a declared extension (exact rank $p = 0.021$; first four 16.8 vs. 13.8); enforcement leaves binary within noise ($16.0 \to 15.2$) while graded gains ($15.5 \to 18.2$). The mechanism: binary's rescues drop $14 \to 0$ by construction (26 wrongfully locked good ideas, 193 blocked retries) while graded's retrials flow through OPEN records (13 rescues survive). **A preregistered second proposer family** (GPT-4o, 12 replicates; App. C) shows the other face of Thm. 2's tradeoff. The mechanism is identical—enforcement drives binary's rescues to 0 and locks 54 good ideas—but bankless GPT-4o re-proposes so heavily (3.7 discoveries per loop) that the lock's efficiency gain dominates its fertility loss: enforcement lifts both rules (binary $9.1 \to 13.4$, graded $10.1 \to 12.2$) and binary edges graded (13.4 ± 4.6 vs. 12.2 ± 1.7, $p = 0.96$ for the preregistered direction). Two proposers, two sides of one curve: which side a system is on is set by how much its proposer explores unaided, and graded is the steadier design on both (binary's worst replicate is 2).

Design implications. Within the tested regime, every result converges on one intervention point: the verdict record, shared across families and agents; the fault is not that models obey a schema but that a noisy procedure issues a directive it has not earned. The threshold is the Bayes rule for the induced *one-step* locking decision, $\theta = 1 - c/V$ for retest cost c and exonerated-idea value V (Lemma 1, App. D; probability matching dominated pointwise). Sequentially, lineage value raises V and the bar while horizon inflates c: θ^* is a reference operating point—one-step Bayes-optimal, empirically best among rules we evaluate, not a proven sequential optimum. The rule is a standard sequential test (Prop. 3; successive elimination [12]); what is new is where it must sit and what it must record. The fix is a schema change, not an architectural one (half already exists in the wild: [1]'s curator grades scope; missing are the evidence field and a reversal path, App. B): (0) disclose provenance and permit dissent—necessary but insufficient, since prompt levers are global knobs on s (Sec. 6) while the schema calibrates per record; (1) replace binary DEAD with an evidence-graded verdict, locking only beyond $P(\text{dead}) > \theta^*$ from recorded seeds and effects—never on a single sub-σ failure ($\theta^* \approx 0.8$ at the simulations' operating point $c/V = 0.2$; at the measured W_B, verification-stage failures cannot clear it and default to INCONCLUSIVE); (2) posterior-free systems can parole locked ideas at ∼5%/step, recovering most of the gain. The gated-LLM cell adds a warning: mandatory adopt-or-reject enforcement of *binary* verdicts removes the soft deference that quietly repairs them (rescues $14 \to 0$)—hardening compliance with an uncalibrated bank is the configuration our results most clearly advise against.

Limitations and ongoing work. Our simulations impose the asymmetries we prove necessary, our theorems live on a truncated type space, and the scalar reduction is numerical: the honest read is a map of existence conditions (out-of-sample: Sec. 7). Every executed loop is ML experimentation (drug discovery only at vignette level), so mechanism generality is not empirical generality beyond hyperparameter search. We do not claim current systems instantiate the full pathology; we characterize when an increasingly natural pattern becomes dangerous: an audit of [1]'s released bank (App. B) finds evidence-free categorical verdicts, no retirement path, and regime-bound records consumed as cross-regime prohibitions; a preregistered prompt-level calibration of that bank shifted stated deference, not outcomes (at ceiling; App. B). In-loop coverage is narrow and our strongest results sit where realism is weakest (Sec. 8): the mechanism is validated end-to-end under control and deployed agents' components separately; we do not claim deployed systems typically gain 30%, and longer horizons and in-situ work remain. **Conclusion.** Conditional on success-gated supply, a failure bank turns statistical verdicts into permanent facts nobody re-examines; the failure is behavioral before architectural—one word suppresses choices in every family tested—and the remedy is records that carry their evidence, locks that require it. Systems automating discovery should not automate the one mistake human science never made: believing its own file drawer.

AI USE STATEMENT

In this work, we used generative AI tools (principally Anthropic Claude) as a research collaborator throughout: for theory development and proofs, experiment design and implementation, data analysis, figure generation, and drafting. LLM-agent experiments (API and Claude Code) were executed by the human authors on their own machines; the algorithmic end-to-end experiments were executed in the collaborator's sandboxed compute environment from committed code, with cached run logs released. We have reviewed all AI-assisted work: all main theoretical statements were verified independently, including a hostile re-derivation by a different vendor's reasoning model (see reproducibility artifact); every number in the paper is re-derivable from the released artifact and was audited by recomputation from raw logs. The subject of the paper—epistemic failure modes of AI research systems—made this collaboration itself an object of methodological care: review rounds deliberately included models from vendors other than the collaborator to avoid single-family blind spots. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Every empirical claim is re-derivable from the released artifact (anonymous link, p. 2): closed-form and simulation code with committed result JSONs; the full verdict and ground-truth logs of both live training loops and the independent-seed replication (raw `jsonl`, $\approx 6{,}600$ cached runs); the cached run set of the end-to-end system experiment ($28{,}050$ distinct (config, seed) entries), which replays all six memory conditions, the volume-matched control, and the threshold sweep over eight replicates deterministically without retraining; the verbatim behavioral protocols for all twelve models, the cross-family and second-domain runners with per-call records, and the in-the-loop proposer harness with per-call API logs and per-(condition, replicate) state checkpoints. All scripts resolve paths relative to the repository and were executed from clean-room extracts.

REFERENCES

[1] H. Wang. Negative knowledge as failure-aware shared memory for AutoResearch. arXiv:2606.21024, 2026.

[2] C. Lu, C. Lu, R. T. Lange, J. Foerster, J. Clune, D. Ha. The AI Scientist: towards fully automated open-ended scientific discovery. arXiv:2408.06292, 2024.

[3] S. Schmidgall et al. Agent Laboratory: using LLM agents as research assistants. arXiv:2501.04227, 2025.

[4] K. Zollman. The epistemic benefit of transient diversity. *Erkenntnis* 72(1), 2010.

[5] S. Rosenstock, J. Bruner, C. O'Connor. In epistemic networks, is less really more? *Philosophy of Science* 84(2), 2017.

[6] R. Rosenthal. The file drawer problem and tolerance for null results. *Psychological Bulletin* 86(3), 1979.

[7] J. P. A. Ioannidis. Why most discovered true associations are inflated. *Epidemiology* 19(5), 2008.

[8] A. Gelman, J. Carlin. Beyond power calculations: assessing Type S and Type M errors. *Perspectives on Psychological Science* 9(6), 2014.

[9] D. KC, A. Budathoki. Handoff debt: the rediscovery cost when coding agents take over interrupted tasks. arXiv:2606.02875, 2026.

[10] C. Huang, Y. Wu, W. Zhang. What should agents say? Action-state communication for efficient multi-agent systems. arXiv:2606.05304, 2026.

[11] K. Zollman. The communication structure of epistemic communities. *Philosophy of Science* 74(5), 2007.

[12] E. Even-Dar, S. Mannor, Y. Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *JMLR* 7, 2006.

[13] D. Zheng, X. Xu, Y. Yang. Multi-armed bandits with arriving arms: sequential screening, dynamic regret, and sublinear guarantees. arXiv:2606.09002, 2026.

[14] Y. Mao, A. Mirhoseini. Decentralized multi-agent systems with shared context. arXiv:2606.10662, 2026.

[15] Y. Qi, X. Xu, Y. Li. When not to write memory: governing false promotion from correlated agent traces. arXiv:2607.02579, 2026.

[16] D. Zhang, Y. Lin, Z. Wu, Y. Sun, B. Li, D. Li, H. Peng. Useful memories become faulty when continuously updated by LLMs. arXiv:2605.12978, 2026.

[17] A. Borthwick, S. Ash, A. Galczak. RoboPhD: evolving diverse complex agents under tight evaluation budgets. arXiv:2604.04347, 2026.

[18] J. Liao, Q. Wang, J. Zhu, B. Du, R. Yan, X. Chen. Belief memory: agent memory under partial observability. arXiv:2605.05583, 2026.

[19] Y. Qian, Y. Ren. EvoGraph-Mem: failure-aware editable graph memory for long-term language agents. arXiv:2608.11248, 2026.

[20] J. Ma, B. Chu, J. Gao, J. Zhang, Y. Ma, Y. Tan, J. Ji, X. Sun, R. Ji. One reflection is not enough: self-correcting autonomous research via multi-hypothesis failure attribution. arXiv:2606.31478, 2026.

[21] T. E. Harris. *The Theory of Branching Processes*. Springer, 1963.

[22] K. B. Athreya, P. E. Ney. *Branching Processes*. Springer, 1972.

A ADDITIONAL FIGURES

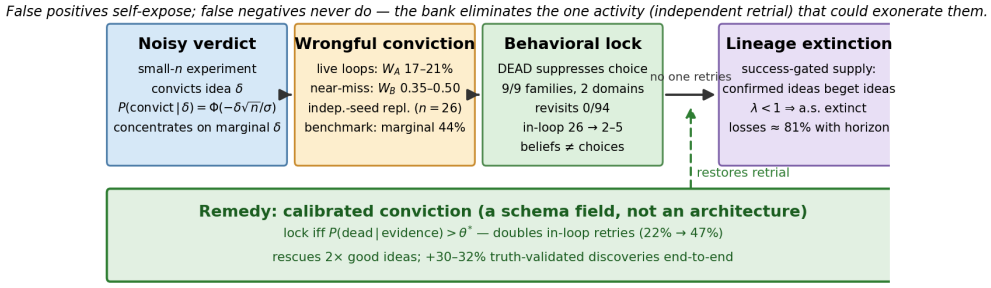

Figure 5: The causal chain and where we measure it: noisy small-n verdicts concentrate wrongful convictions on marginally good ideas (Prop. 2); DEAD behaviorally locks choices in every tested family and domain (Sec. 6); under success-gated supply, unretried convictions drive lineage extinction (Thm. 1). The remedy intervenes at the record (Sec. 8 onward).

Simulation design rules (Sec. 5). Threshold θ^*=0.8: 65.1 discoveries vs. probability matching 59.6; $\theta \to 1$: 47.8 (over-churn); ε-parole on a naive bank: 58.3.

Threshold sweep (preregistered, PREREG_THETA in the artifact). Recorded loop, graded rule at $\theta^* = 0.6/0.7/0.9/0.95$, 8 replicates each, truth values preserved for all 719 known configs (180 new configs truth-tested): discoveries 63.6 ± 4.0, 63.4 ± 5.9, 61.6 ± 6.6, 62.1 ± 5.5 vs. binary 46.6 (one-sided exact rank $p = 1\text{–}5 \times 10^{-4}$); the recorded $\theta^* = 0.8$ (60.5) is within noise of all four. The over-churn collapse seen in simulation at $\theta \to 1$ does not appear by 0.95 in this loop.

Second-proposer revisit counts (GPT-4o, rich cell). Re-proposals of recorded configs: 6 with no bank, 0 under binary or parole, 4 retries of graded's OPEN records.

B IN-SITU AUDIT OF A DEPLOYED FAILURE BANK

We audited the public artifact of [1] (repository `Negative_Knowledge`, `reproduction/section4`): 21 per-round negative-knowledge records (seven research programs, three rounds each), seven deep syntheses, the aggregated banks (43 negative, 15 positive entries), and 12 downstream research runs (three tasks $\times$ no bank / positive only / negative only / both) with the authors' own verified phenomenon checks. Every number below is recomputed by

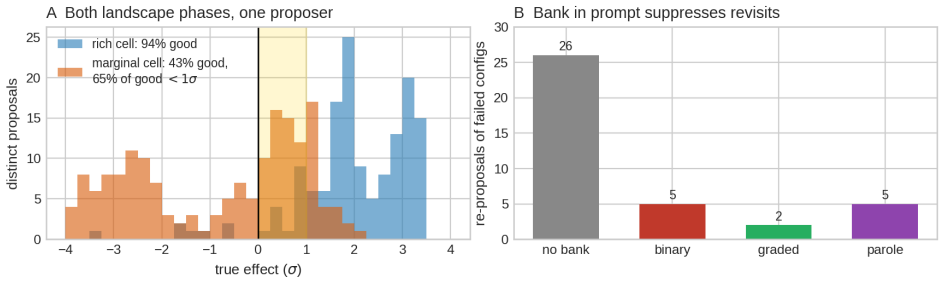

Figure 6: The deployed-LLM loop, descriptive panels. (A) True-effect distributions of the model's own proposals in both landscape cells: phases are positions, not assumptions. (B) Re-proposals of already-failed configs with and without the in-prompt bank (rich cell).

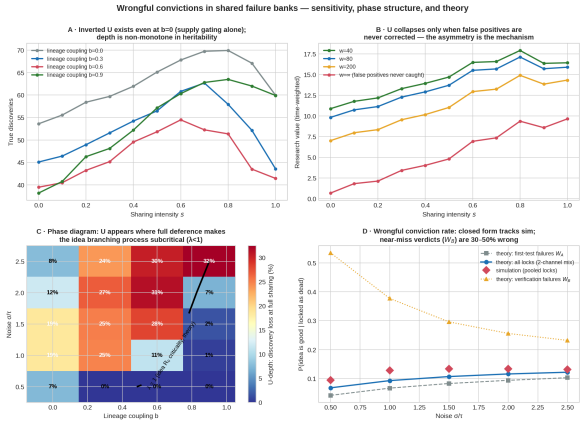

Figure 7: (A) The U persists at b=0; depth non-monotone in coupling. (B) Existence requires the false-positive/false-negative asymmetry. (C) Phase diagram with the λ=1 contour. (D) Wrongful rates: closed form vs. simulation; near-miss verdicts (W_B) are 30–50% wrong.

`audit/audit_negbank.py` in our artifact; downstream runs are single runs per cell and are descriptive.

Schema and runtime. A record's verdict block is {layer, scope, degree, recommended_action, risk}—all categorical; $0/21$ records carry any numeric evidence (replicates, seeds, effect size, uncertainty). The runtime and curator contain no retirement, revision, reopening, or confidence mechanism (0 matches), so a record is never retired by construction. The memory served to downstream agents opens "Entries describe methods/regimes that FAILED. Use to avoid pitfalls," and marks deep syntheses as "harder-to-overcome." Downstream handling is mandatory adopt-or-reject. The curator nonetheless *grades scope*: 13 regime-bound, 4 local, 4 general failures; recommended actions 10 narrow-claim, 8 change-method, 2 retry, 1 abandon-route. The deployed system thus already has half of our design item (1)—graded verdicts—without the evidence that would make the grade actionable or a path that would ever revise it.

Category error. $7/21$ records are curator-labeled positive findings ("Positive finding", "Positive baseline, not a failure") that the schema nonetheless typed as failures. The routing heuristic (keyed on recommended action) sends five to the positive bank; two (S4_r1, S5_r3) land in the negative bank—$2/43$ negative entries consumed as "methods that FAILED."

Downstream deference. Bank-conditioned agents cite records recorded on other equations and horizons (Burgers long-time, shallow-water) as grounds to reject method classes for the coupled task at hand (task A, negative-only: 18 distinct records cited, 3 explicit rejection lines; both banks: 16/5). Regime-bound verdicts are consumed as prohibitions. Outcomes in the released single runs: on task A the memoryless agent passes the verified check while all three bank conditions fail; on

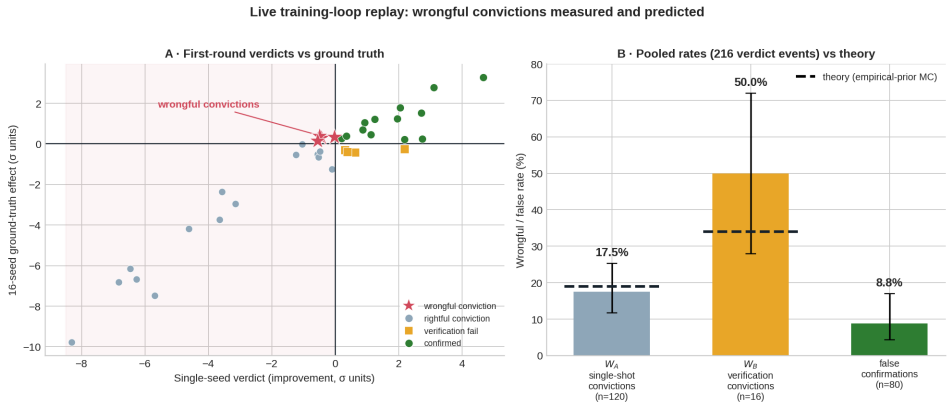

Figure 8: Live-loop replay. (A) First-round single-seed verdicts vs. 16-seed ground truth; wrongful convictions (stars) are genuinely good ideas in the conviction zone. (B) Pooled rates over all 216 verdict events with 95% CIs against the empirical-prior predictions (dashes): W_A 17.5% vs. 19 predicted; W_B 50% vs. 34 predicted (correlated events, see text); false confirmations 8.8%.

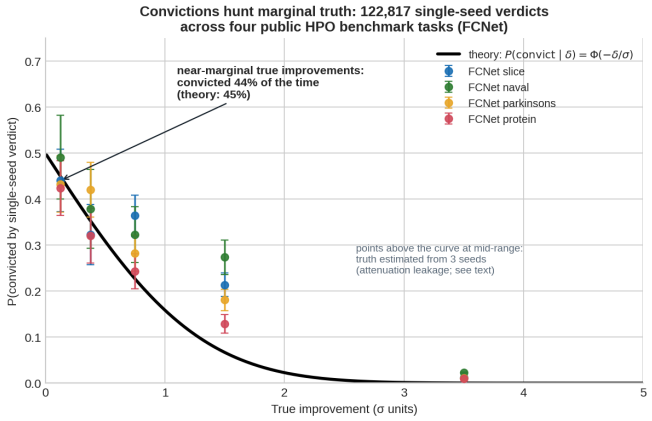

Figure 9: Conviction probability vs. true effect size across four public FCNet tasks (95% CIs), against the parameter-free curve $\Phi(-\delta/\sigma)$.

tasks B and C the bank conditions pass where the memoryless agent fails. Five replicates per cell under the authors' own execution protocol (below) do *not* reproduce the task-A reading: negative-only passes $4/5$ there, and every condition sits near ceiling for this model (no bank $6/6$, negative-only $13/15$). We therefore read the released task-A result as a single-run fluctuation; the robust observation is the deference pattern above. (The released task-A negative-only agent also reports `useful_self_assessment = True` against a failing verified check: an in-situ false confirmation.) Scope-calibration coding of the 21 records (tested vs. asserted scope) is released with the coding sheet; the curator rarely over-generalizes without a stated mechanism, so the risk lives in evidence-free categorical scope and its collapse at consumption, not in curator over-claim.

Fix experiment (preregistered; PREREG_INSITU in the artifact). We served the same 43 entries verbatim with one added `status` line per entry derived from its own scope field (6 CLOSED, 37 OPEN) and a calibrated preface (OPEN = advisory outside its regime), a 45-line diff, and reran the authors' stage-2 cells with their execution protocol (headless Claude Code with a current Claude Sonnet, Sept. 2026; Read/Write/Bash; iteration budget 3—so comparisons to their released runs, made earlier in 2026, cross model versions, while within-experiment comparisons do not): tasks A–C $\times$ {negative-only, calibrated} $\times$ 5 replicates, plus no-bank $\times 2$. The primary hypothesis is not supported: calibrated $12/15$ vs. negative-only $13/15$ (one-sided Fisher $p = 0.84$), with all conditions near ceiling. The rendering did change stated deference—bank-driven rejections in the

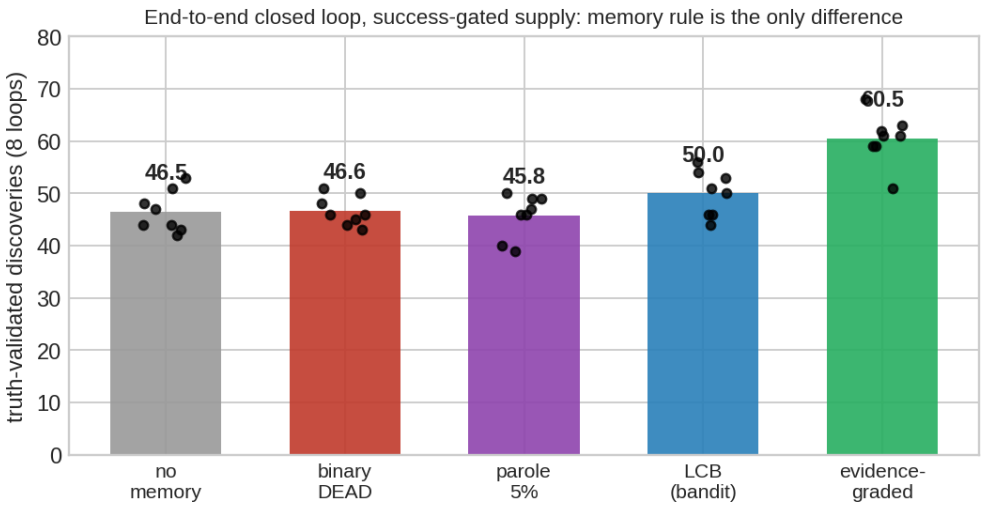

Figure 10: End-to-end closed-loop comparison. (A) Truth-validated discoveries under a fixed 220-run budget, five memory rules, eight replicate loops each; exact one-sided rank tests for graded vs. no-memory / binary / parole: $p = 3/12870, 2/12870, 1/12870$ (2.3×10^{-4}, 1.6×10^{-4}, 7.8×10^{-5}). (B) ECDF of discovery true effects: evidence-graded memory both finds more and finds better—spared marginal truths seed better lineages.

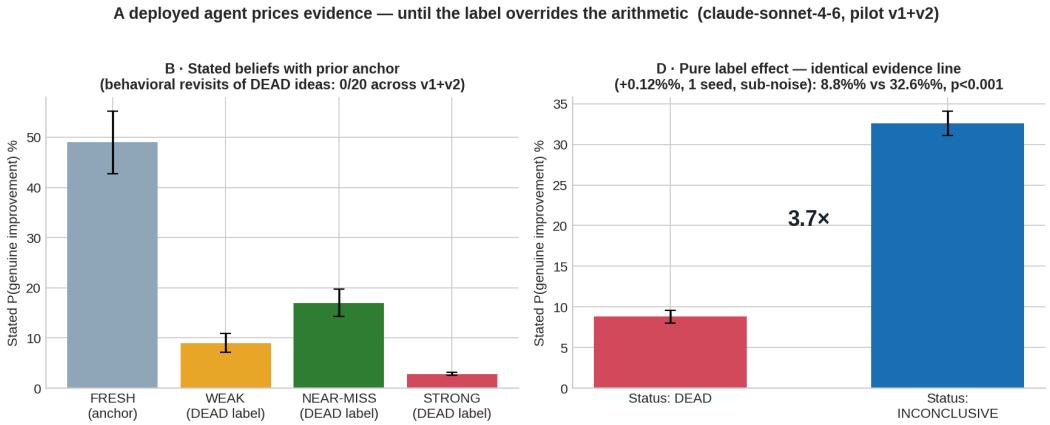

Figure 11: Stated beliefs with a prior anchor (left); the pure label effect on identical evidence (right).

research graphs fell 67 → 57, citations rose 29 → 34, and calibrated agents referenced the status field 27 times (negative-only: 0)—but the extra latitude bought nothing at this ceiling, and on task A it cost (2/5, with more graph nodes): reopening without evidence spends budget, the in-situ echo of the volume-matched control of Sec. 8. The evidence-graded lock itself could not be instantiated because these records carry no evidence; that requires instrumentation at record-writing time (`spec/` in the artifact), not a rendering change at consumption.

Cross-system survey. We coded seven further public systems at the code level (repository heads, Sept. 2026; evidence lines in `audit/survey_repos.md`): AIDE and AI-Scientist-v2 (tree search with `is_buggy` nodes), OpenEvolve (MAP-Elites archive), FunSearch (island program database), Voyager (curriculum with a failed-task list), AgentLaboratory (edit-and-repair loop), and AI-Scientist v1 (sequential runs). Table 1 summarizes. Three regularities: (i) 0/8 systems attach replicate evidence (n, σ) to a negative outcome—every verdict is single-run; (ii) 5/8 re-exercise failures only when the failure is an *execution error* (a bug), with bounded retries (e.g. AIDE: probability 0.5, depth $\leq$ 3)—a run that executed and scored marginally worse is never retried in any system

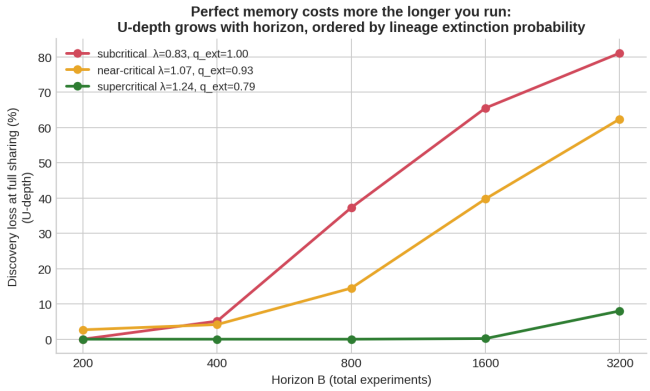

Figure 12: Discovery loss at full sharing grows with horizon, ordered by extinction probability.

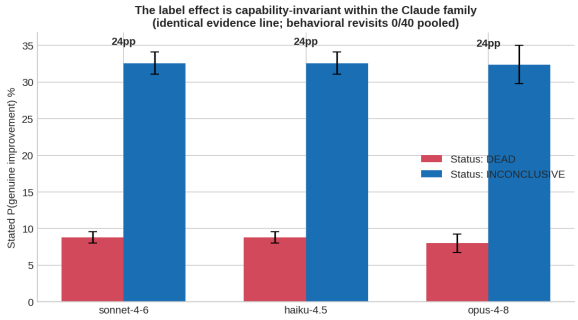

Figure 13: Within the Claude family, the label effect on stated beliefs is undiminished across the tested scale range.

surveyed, and that is precisely the class where Prop. 2 locates wrongful convictions; (iii) only [1] maintains an explicit shared negative bank, and it has no reopening path; OpenEvolve serves worse "previous attempts" as prompt context, a soft bank with the deference exposure of Sec. 6. Supply is success-gated in $7/8$ (best-node, elite, or island parents). Factorized, the pathology's ingredients are already common—noisy single-run negative verdicts $8/8$, success-gated supply $7/8$, no evidence-aware reopening $8/8$—and the fourth, an explicit shared irreversible negative bank, is $1/8$ today: the step that makes the first three combine into extinction risk. Single reader; versions drift; the table is a snapshot, not a ranking.

Table 1: Code-level schema survey of eight public systems. Evidence: single-run metric only unless noted. Reopen: whether a recorded failure is ever re-exercised.

System	Failure record	Evidence	Reopen	Consumer instruction
Neg. Knowledge [1]	typed record, shared bank	none	none	"FAILED—use to avoid pitfalls"
AIDE	node flagged buggy (error/no metric)	single metric	bugs only, p=.5, depth ≤ 3	LLM flags `is_bug`
AI-Scientist-v2	as AIDE, plus plot bugs	single metric	as AIDE	as AIDE
OpenEvolve	archived program + scores	scalar	none (displaced deleted)	"feedback from previous attempts"
FunSearch	not stored (failures dropped)	per-test score	n/a	none
Voyager	failed-task list	none	tasks reproposable	"failed tasks that are too hard"
Agent Laboratory	erroring edit rejected	none	error→repair loop	rejection on error
AI-Scientist v1	error text fed back	free text	retry within 5 runs	error string

C BEHAVIORAL PROTOCOL DETAILS

Each trial is an independent, stateless API call (fresh conversation; no shared context). No sampling parameters were explicitly set; provider defaults were used throughout. Trials per model: Task A, 10 free-choice trials (plus 10 additional v1 trials for Sonnet, pooled where stated); Task B, 6 trials, with all four probability probes elicited jointly in a single call—cross-verdict comparisons within Task B are therefore paired within calls; Task D, 5 DEAD and 5 INCONCLUSIVE independent calls. The observational unit for all tests is the call: the Task-D comparison is a two-sample test on call-level stated probabilities. The assignment of technique names to verdict conditions and the menu order are re-randomized on every call, marginalizing content effects; prompt-level random effects are not modeled, a limitation noted in the main text. The v3 protocol adds, per model: 6 provenance-control calls per condition (DEAD / DEAD+provenance sentence / neutral STATUS-B with legend) and 6 choice-flip calls per label, with the marked technique assigned the strongest literature prior; the provenance condition reliably elicits explicit posterior calculations in transcripts (occasional non-uptake: 2 of 18 calls priced at the suppressed level). Cross-family runs use each provider's OpenAI-compatible endpoint with the same verbatim prompts, provider-default temperature (no sampling parameters set), $n = 5$–6 per condition, and reasoning capped at 400 tokens where the endpoint supports it. Because inference settings differ across vendors, the cross-family result should be read as robustness across heterogeneous *deployed settings* (model snapshot, serving stack, default decoding), not as a property of model weights alone. Exact identifiers, routing, and dates:

Display name	API model ID	Route	Date
Claude Haiku 4.5	`claude-haiku-4-5`	Anthropic	2026-08-19/20
Claude Sonnet 4.6	`claude-sonnet-4-6`	Anthropic	2026-08-19/20
Claude Opus 4.8	`claude-opus-4-8`	Anthropic	2026-08-19/20
DeepSeek-chat	`deepseek-chat`	DeepSeek	2026-08-20
DeepSeek-R1	`deepseek-reasoner`	DeepSeek	2026-08-20
GPT-4o	`openai/gpt-4o`	OpenRouter	2026-08-20
Gemini-2.5-flash	`google/gemini-2.5-flash`	OpenRouter	2026-08-20
Llama-4-Scout	`meta-llama/llama-4-scout`	OpenRouter	2026-08-20
Mistral-medium-3	`mistralai/mistral-medium-3`	OpenRouter	2026-08-20
Kimi-K2	`moonshotai/kimi-k2`	OpenRouter	2026-08-20
GLM-4.7	`z-ai/glm-4.7`	OpenRouter	2026-08-20
Qwen3-32B	`qwen/qwen3-32b`	OpenRouter	2026-08-20

OpenRouter routes to third-party inference providers, so serving-side configuration for those rows is not under our control; per-call responses are included in the released artifact. Qwen3-32B's belief probes returned empty completions under the reasoning cap and are excluded from belief analyses; its choice data are retained. Verbatim prompt templates and the runnable protocol are in the released artifact.

Belief-level taxonomy (descriptive, 5–6 trials per model-condition). Provenance-*unlockable* (Claude; DeepSeek-R1 $1.5\% \to 50\%$); *hard-locked* despite demonstrated range (GPT-4o $0 \to 0$, INCONCLUSIVE 31%); *calibrated-but-choice-locked* (GLM-4.7); near-*immune* (Kimi-K2); floor-dominated pessimists whose $\leq 3\%$ anchors leave the provenance test uninformative (Gemini, Mistral). Labels are run-level descriptions, not stable model properties.

Second-family gated loops (E6, preregistered; PREREG_E6 in the artifact). GPT-4o proposer, marginal landscape, success-gated, budget 150, 12 replicates per condition. Soft lock: no-memory 3.7 ± 2.2 (GPT-4o without records re-proposes heavily: 12–25 distinct configs per loop), binary 9.1 ± 2.3, graded 10.1 ± 2.9 (either bank beats none, $p = 2 \times 10^{-5}$; graded vs. binary $p = 0.17$). Enforced lock: binary 13.4 ± 4.6 $[20, 2, 17, 15, 14, 12, 13, 16, 13, 15, 8, 16]$, graded 12.2 ± 1.7 $[10, 11, 11, 10, 13, 13, 11, 12, 13, 14, 14, 15]$; preregistered H1 (graded $>$ binary) $p = 0.96$, binary ahead at $p = 0.05$; enforcement raises binary ($p = 0.001$) and graded ($p = 0.03$). Mechanism signature as in DeepSeek: strict rescues under enforced binary 0, wrongful stock 54, 418 blocked retries, 40 forfeits; graded 24 rescues via OPEN records; train-while-locked 0; budget identity exact in all 24 enforced replicates. Read: the lock's efficiency gain dominates its fertility loss for this proposer at this budget (Thm. 2); the DeepSeek reversal is family-contingent.

Behavioral replication (preregistered). A fresh-run replication (same authors, new base run; see deviation log) ran the F2 and chemistry choice protocols at 12 trials/condition on all nine non-Claude

models (PREREG_BX in the artifact). Deviations, disclosed in the shipped log: the extension attached to a fresh base run rather than the original logs (hence reported as an independent replication, never pooled with the original study), and the F3 arm did not execute due to a log-file collision (F3 remains at original n). Per-family disclaimer repair (marked picks, unparseable excluded): R1 $10/12$, Gemini $8/11$, Kimi $4/11$, GLM $3/6$, Llama $2/10$, Mistral $2/12$, DeepSeek-chat $1/12$, GPT-4o $1/12$, Qwen3 $0/9$; plain DEAD pooled $6/91$. Chemistry: $99/99$ under INCONCLUSIVE vs. $18/82$ under DEAD, per-family suppression in $9/9$.

Statistical hierarchy. Calls nest within models and models within families; pooled call-level tests (Fisher, t) treat calls as exchangeable and should be read as descriptive strength. The design-robust headline is the family-level exact sign test ($9/9$ families under DEAD vs. INCONCLUSIVE, two-sided $p = 0.0039$), which is immune to within-family pseudo-replication.

D PROOF SKETCHES

Exchangeable supply (Prop. 1). $P(\text{fail} \mid \delta)$ is decreasing in δ, so by monotone likelihood ratio the posterior given the *event* of failure is FOSD-dominated by the prior; confirmation probability and confirmed value increase in δ, so a coupling replaces every revisit with a stochastically dominant fresh draw.

Small-noise law (Prop. 2a). Let $Y = \delta + \sigma Z$ with $Z \sim N(0,1)$, conviction iff $Y \le 0$, and let δ have a bounded density f continuous at 0 with $P(\delta < 0) > 0$. Then $W_A = \int_0^\infty \Phi(-\delta/\sigma) f(\delta)\, d\delta \,/\, P(Y \le 0)$. Substituting $\delta = \sigma u$, the numerator is $\sigma \int_0^\infty \Phi(-u) f(\sigma u)\, du \to \sigma f(0^+) \int_0^\infty \Phi(-u)\, du = \sigma f(0^+)/\sqrt{2\pi}$ by dominated convergence ($\Phi(-u)$ is integrable and f is bounded near 0), while $P(Y \le 0) \to P(\delta < 0)$. With n runs replace σ by $\sigma/\sqrt{n}$. The Gaussian-prior orthant of (a) is the special case $f = N(\mu_0, \tau^2)$; the rate depends on f only through $f(0^+)$.

Clean-limit constant (Prop. 2b). A verification-stage conviction is the event $E = \{X_1 > 0,\ \bar{X}_{1:v+1} < 0\}$ with $X_i = \delta + \sigma Z_i$. Write $R_v(u) = P(Z_1 > -u,\ \bar{Z}_{1:v+1} < -u)$, which depends only on v and decays like a Gaussian tail in $|u|$. Then $P(E \mid \delta) = R_v(\delta/\sigma)$ and, with $\delta = \sigma u$, $W_B = P(\delta > 0 \mid E) = \dfrac{\int_0^\infty R_v(u) f(\sigma u)\, du}{\int_{-\infty}^\infty R_v(u) f(\sigma u)\, du} \longrightarrow \dfrac{f(0^+) I_v^+}{f(0^+) I_v^+ + f(0^-) I_v^-} = \kappa_v(\varrho)$ as $\sigma \to 0$, where $I_v^\pm = \int_{\mathbb{R}_\pm} R_v$ and $\varrho = f(0^+)/f(0^-)$, again by dominated convergence. Numerically (10^6 draws) $\kappa_v(1) = 0.294,\ 0.242,\ 0.211,\ 0.191$ for $v = 1, 2, 3, 4$; at $v = 2$, $\kappa_2(2) = 0.39$ and $\kappa_2(0.5) = 0.14$. Hence reversal-stage error is bounded away from zero for every effect landscape with positive boundary density, in contrast to first-test error, and its clean-limit value is set by the verification geometry and the boundary density ratio (`audit/kappa.py` in the artifact).

Deference phase transition (Thm. 2). Retry-until-resolved: after a failure the idea locks with probability s, else is retested; the probability of eventual confirmation solves $h = q + (1-q)(1-s)h$, giving $h_s = q/(q + s(1-q))$, and the expected number of tests is $1/(q + s(1-q))$ (geometric). Each factor $h_s(\delta)$ is continuous and non-increasing in s, strictly decreasing whenever $q(\delta) < 1$; hence $A_s = C\, h_s\, k$ is entrywise continuous and non-increasing in s on the truncated space, and the spectral radius of a non-negative irreducible operator is continuous and monotone in its entries (Perron–Frobenius), strictly so when a set of positive k-weight is strictly decreased. If $\lambda(0) > 1 > \lambda(1)$ the intermediate value theorem and strict monotonicity give a unique s_c with $\lambda(s_c) = 1$; the extinction dichotomy at $\lambda \lesseqgtr 1$ is Theorem 1's. Efficiency $q + s(1-q)$ is increasing in s, so a finite budget trades tests per idea against reproduction per confirmation; Fig. 2 shows the resulting interior optimum.

Sequential threshold (Prop. 3). Let p be the posterior that the idea is good and $W(p)$ the optimal value with actions lock (0), confirm ($pV - (1-p)L$), retry ($-c + E[W(p')]$). Both terminal payoffs are affine in p and Bayesian updating preserves convexity, so by induction on the horizon W is convex with $W(0) = 0$ (at $p = 0$ confirming and retrying are strictly worse than locking) and $W(1) = V > 0$. A convex function attains its minimum on an interval, so the lock region $\{W = 0\}$ is $[0, \theta]$; the retry region lies above it. Increasing V raises W pointwise, so the zero set shrinks: θ is non-increasing in V (and non-decreasing in c). No likelihood-ratio structure is needed for the threshold form; monotone likelihood ratios additionally make posteriors move monotonically in the signal.

Lemma 1 (threshold optimality). Consider the myopic memory decision for a recorded idea with posterior $p = P(\delta \le 0 \mid \text{evidence})$: locking saves an expected retest cost $c > 0$; not locking yields expected value $(1-p)\,V - c$, where $V > 0$ is the expected discounted discovery value of an exonerated idea, treated here as a constant of the environment. Locking is optimal iff $(1-p)V - c \le 0$, i.e. iff $p \ge 1 - c/V \equiv \theta$: a deterministic posterior threshold. Any rule that locks with probability $g(p)$ for non-degenerate g (probability matching, $g(p) = p$) mixes the two actions on the wrong side of the cut and is weakly dominated pointwise, strictly on any posterior mass in $(\theta, 1) \cup (0, \theta)$ where $g \notin \{0, 1\}$. In the full sequential problem the value V of an exonerated idea acquires horizon and lineage terms (children of exonerated ideas); since $\theta = 1 - c/V$ is increasing in V, raising the continuation value *raises* the locking bar: more future value at stake demands more certainty before an irreversible lock. Horizon, however, also inflates the opportunity cost c of spending a slot on a retest (the foregone fresh test has descendants too), so the net movement of θ is model-dependent. We do *not* claim the sequential rule remains a threshold in p alone: if descendant value depends on the magnitude of δ, two records with equal p but different posterior shapes above zero can carry different continuation values, so p need not be a sufficient statistic. The threshold rule we evaluate empirically is the one-step Bayes rule; its end-to-end advantage (Sec. 8) is measured, not derived from a sequential optimality claim. □

Prop. 1. Condition on the binary event $F = \{y \le 0\}$ (possibly repeated). $P(F \mid \delta) = \Phi(-\delta/\sigma)$ is strictly decreasing in δ, so the likelihood of any conjunction of such events has monotone (decreasing) likelihood ratio in δ; hence the posterior given the events is first-order stochastically dominated by the prior. Confirmation probability $q(\delta)$ and confirmed value δ^+ are non-decreasing in δ, and children do not exist in the exogenous-supply model, so the per-step value of revisiting a binary-marked idea is dominated by a fresh draw; a stepwise coupling then preserves dominance of cumulative discovery. □

Remark (why the binary qualifier is necessary). The dominance fails for measurement-valued records: with $\mu_0 = -1$, $\tau = \sigma = 1$, a fresh draw has $P(\delta > 0) = \Phi(-1) \approx 15.9\%$, while after observing $y = -0.01$ (a failure) the posterior is $\delta \mid y \sim \mathcal{N}(-0.505, 0.5)$ and $P(\delta > 0 \mid y) \approx 23.8\% > 15.9\%$: a near-null failure is favorable evidence relative to a pessimistic prior. This is precisely why evidence-graded records retain value in exchangeable settings, and why banks that store only DEAD destroy it.

Prop. 2(a). $(\delta, \hat{\delta})$ is bivariate normal with means (μ_0, μ_0), variances $(\tau^2, \tau^2 + \sigma^2)$, covariance τ^2. Standardize and express $P(\delta > 0, \hat{\delta} \le 0)$ as an orthant probability with correlation $-\rho$; divide by $P(\hat{\delta} \le 0)$. Limits follow from $\rho \to 1$ and $\rho \to 0$. (b) is the same computation on $(\delta, X, X{+}Y_1{+}Y_2)$ with $\mathrm{corr}(X, S) = 1/\sqrt{3}$. (c) is immediate from the Gaussian c.d.f. □

Thm. 1. *Standing assumptions.* (A1) The type space is a compact interval $[-M, M]$; all statements concern this truncated process, and all reported numerics use truncations whose boundaries lie well outside the simulated quality mass. For $b < 1$ the child-mean map $\delta \mapsto \mu_c + b(\delta - a)$ is a contraction, so quality mass is naturally confined; for $b \ge 1$ the offspring mean can drift outward and we make no claim about the untruncated process. (A2) The mean kernel $Cq(\delta)k(\delta' \mid \delta)$ is jointly continuous and strictly positive on $[-M, M]^2$, and offspring counts are bounded by C. (A3) b, τ_c, σ are finite and positive. Under (A1)–(A2) the mean operator is compact and strongly positive on $C([-M, M])$, so the Krein–Rutman theorem gives a simple leading eigenvalue $\lambda > 0$ with strictly positive eigenfunction h; the spectral radius exists and equals λ. The process is a general-type Galton–Watson process: an individual of type δ produces, with probability $q(\delta)$, exactly C offspring i.i.d. from $k(\cdot|\delta)$, else none; assumptions (A1)–(A2) supply the Krein–Rutman eigenpair (λ, h) as stated. (i) For $\lambda < 1$. Write $Z_n(f) = \sum_i f(\delta_{n,i})$ for the generation-n population functional. By the eigenvalue property, $\mathbb{E}[Z_{n+1}(h) \mid \mathcal{F}_n] = Z_n(Ah) = \lambda Z_n(h)$, so $\mathbb{E}[Z_n(h)] = \lambda^n Z_0(h)$. Since h is continuous and strictly positive on the compact $[-M, M]$, $h_{\min} := \min h > 0$, hence the population size satisfies $|Z_n| \le Z_n(h)/h_{\min}$ and $\mathbb{E}|Z_n| \le \lambda^n Z_0(h)/h_{\min}$. Markov's inequality gives $P(|Z_n| \ge 1) \le \lambda^n Z_0(h)/h_{\min} \to 0$; because $\{|Z_n| \ge 1\}$ is decreasing in n (extinction is absorbing), $P(\text{survival}) = \lim_n P(|Z_n| \ge 1) = 0$: extinction is almost sure. For total progeny, $\mathbb{E}\sum_{n\ge 0}|Z_n| \le \frac{Z_0(h)}{h_{\min}} \sum_{n\ge 0} \lambda^n = \frac{Z_0(h)}{h_{\min}(1-\lambda)} < \infty$, a bound independent of the run budget B; confirmations under full deference are dominated by total progeny, giving the uniform-in-B bound in the statement. (ii) For $\lambda > 1$. The offspring generating functional is $F[s](\delta) = 1 - q(\delta) + q(\delta)\big(\int s\, dk_\delta\big)^C$ for measurable $s : [-M, M] \to [0, 1]$; the type-δ extinc-

tion probability vector $q_{\mathrm{ext}}(\cdot)$ satisfies $q_{\mathrm{ext}} = F[q_{\mathrm{ext}}]$ by first-generation decomposition, and is the *minimal* solution of this fixed-point equation in $[0,1]^{[-M,M]}$: iterating $s_{k+1} = F[s_k]$ from $s_0 \equiv 0$ gives a pointwise-increasing sequence (monotonicity of F) whose limit solves the equation and lower-bounds any other solution by induction. Strict positivity of the kernel (A2) makes the process irreducible in the multi-type sense, and $\lambda > 1$ then implies $q_{\mathrm{ext}}(\delta) < 1$ for every δ, with survival probability $1 - q_{\mathrm{ext}}(\delta) > 0$ bounded away from zero on compacts; these are the standard general-type statements (Harris 1963, Ch. III; Athreya–Ney 1972, Ch. V), whose proofs carry over verbatim under (A1)–(A2) since the mean operator is compact with simple leading eigenvalue. *Status of the scalar reduction.* The scalar q_{ext} quoted in the main text solves the single-type equation $x = 1 - \bar{q} + \bar{q}\, x^C$ with $\bar{q} = \int q(\delta)\, \pi_h(d\delta)$, where π_h is the normalized left-eigenmeasure pairing of the Krein–Rutman eigenpair: it collapses type structure to an h-weighted average confirmation probability. It is a *numerical organizer*, not a theorem: we prove no error bound between it and the type-dependent minimal solution, and use it only to order the simulated horizon curves, which it does at every horizon tested (Fig. 12). All claims of the theorem proper concern the truncated multi-type process above. □