WHEN STATISTICAL DISCOVERIES BECOME STATE: RELIABILITY IN CONTINUAL RESEARCH AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

When scientific discovery agents operate continually, statistical discoveries can become operational state that guides subsequent search rather than remain terminal outputs. This creates a Decision-to-State Gap: guarantees for local admission decisions do not characterize the reliability of the persistent state that drives future search. We study this gap in a controlled benchmark with exact ground truth, 128 hypotheses, and a frozen stateful proposer, distinguishing currently usable discoveries (CURRENT) from immutable admission history (EVER). Across three preregistered positive-dependence stress conditions, permanent append yielded terminal Global-Null active-state FDRs of 0.410–0.422, compared with 0.044–0.064 under cumulative recertification; the approximate 95% simultaneous bands for all three paired gaps lay entirely above zero. Controlled interventions on proposal-visible state changed the next candidate in 20% of selected planted-signal snapshots, while equal-size states with different membership yielded different next candidates in 37.5% of planted pairs. A frozen real-LLM transport pilot further found state-conditioned proposal-distribution shifts beyond same-state repeat instability ($\Delta^{CE} = 0.279$, 95% stability interval [0.158, 0.413]). An exploratory downstream-error signal motivated an independently seeded, preregistered 320-pair confirmation, but its primary hypothesis, an increase in distinct future false admissions, was not confirmed. Together, these results identify reusable statistical state as a distinct reliability object and motivate a governance principle: preserve admission history while separately governing future action capability.

1 INTRODUCTION

Continual scientific agents reuse discoveries as parents, retrieval items and transformation sources for later hypotheses (Lu et al., 2024; Yamada et al., 2025). StatefulDiscovery coordinates frontier selection with investigation state (Chen et al., 2026); HypoExplore explores candidate parent hypotheses while experimental evidence updates a hypothesis memory bank (Koo et al., 2026). A statistical decision can therefore become future action capability.

We call the resulting mismatch the *Decision-to-State Gap*. We identify admission-to-capability composition as a distinct reliability object: statistical admissions become persistent proposal capability that future search consumes. Permanent retention makes this object mismatch visible under repeated exposure, while direct state interventions test whether operational state changes search. In the controlled proposer, visible identities determine legal mutation neighborhoods, so retirement, retention and membership substitution can change later candidates even when raw Evidence remains hidden.

Classical BH controls a fixed rejection family, while online and batched procedures target specified decision streams under their assumptions (Benjamini & Hochberg, 1995; Zrnic et al., 2020; Ramdas et al., 2018). State governance is the next compositional layer: it maps admissions into CURRENT capability, retirement/readmission and EVER provenance. The question is whether the controlled decision object coincides with the persistent operational object on which the system acts.

An exact-truth synthetic catalog identifies these objects directly. Across three preregistered dependence conditions, per-batch BH with permanent append yields terminal Global-Null active-state FDR 0.410–0.422, versus 0.044–0.064 for cumulative recertification. In prospectively selected planted retired states, CURRENT/EVER permission changes the next candidate in 8/40 snapshots and produces divergence across three matched cycles. A truth-blind equal-size intervention changes

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.

15/40 EVER/size-matched proposals, showing that membership can matter beyond capacity. A separate frozen 40-snapshot real-LLM pilot tests the corresponding state-capability contrast under one provider/model.

Mechanism identification and consequence confirmation are distinct. A separately seeded 320-pair primary test evaluates the exploratory increase in distinct future false admissions and is **NOT CONFIRMED**; the result leaves both amplification and equivalence unresolved. Our contributions are:

1. **Reliability-object identification.** We formalize and measure the distinction among local admission, CURRENT capability and EVER history.
2. **State-to-search mechanism and transport.** Controlled membership interventions alter search, and the corresponding contrast transports to the tested real-LLM proposer beyond same-state repeat instability.
3. **Mechanism–consequence separation.** The independent primary is **NOT CONFIRMED**, showing that search divergence and downstream statistical harm require separate empirical tests.

2 Continual Research as a Closed Statistical Loop

2.1 Decision, CURRENT and EVER

Let $S_t = (L_t, P_t, C_t, E_t, J_t)$ denote committed research state: ordered hypotheses/provenance L_t, once-observed statistics P_t, CURRENT identities C_t, EVER identities E_t, and lifecycle/consumption ledger J_t. A local statistical decision returns a rejection set A_t. Governance updates CURRENT and admission history separately:

$$E_t = E_{t-1} \cup A_t, \qquad C_t = G_t(C_{t-1}, P_t, A_t). \tag{1}$$

Retirement removes an identity from CURRENT without deleting it from EVER; readmission reverses that permission change. For candidate set M, $\mathrm{FDP}(M) = |M \cap \mathcal{H}_0| / \max(|M|, 1)$ and $\mathrm{FDR}(M) = \mathbb{E}[\mathrm{FDP}(M)]$. The three reliability objects are

$$\begin{aligned} \mathrm{FDR}_{\mathrm{local}}(t) &= \mathbb{E}[\mathrm{FDP}(A_t)], \\ \mathrm{FDR}_{\mathrm{CURRENT}}(T) &= \mathbb{E}[\mathrm{FDP}(C_T)], \\ \mathrm{FDR}_{\mathrm{EVER}}(T) &= \mathbb{E}[\mathrm{FDP}(E_T)]. \end{aligned} \tag{2}$$

These objects represent new admissions, capability available now and complete admission history, respectively; below we refer to $\mathrm{FDR}_{\mathrm{CURRENT}}$ as active-state FDR. Group power uses all 16 nonnull equivalence groups; candidate yield uses a candidate-budget denominator.

2.2 The Decision-to-State Gap

An unchanged testing rule can induce different reliability objects through its decision-to-state mapping: CURRENT may improve while EVER retains a past false admission. Under the Global Null, each set's FDP is its nonempty indicator, separating present from historical exposure.

2.3 Governance and proposal capability

The proposal interface receives permitted capability $V_t = \Gamma_t(S_t)$ from the full committed state. With keyed randomness U_{t+1}, the loop is

$$V_t \xrightarrow{\pi, U_{t+1}} H_{t+1} \xrightarrow{\text{freeze, Evidence}} Z_{t+1} \xrightarrow{D} A_{t+1} \xrightarrow{G} S_{t+1}. \tag{3}$$

Figure 1 separates operational permission from provenance. Raw Evidence remains outside the proposer, while decision-derived identities affect later capability. Statistical dependence must therefore be assessed in a global filtration containing hypotheses, Evidence, certification, CURRENT/EVER and stopping; payload isolation governs access rather than global independence (Wang et al., 2025).

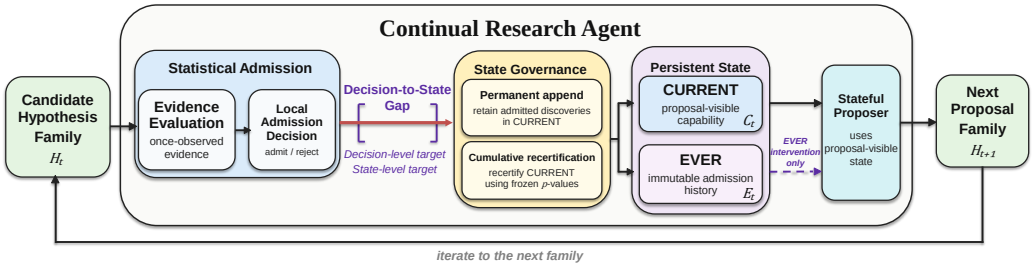

Figure 1: **Decision-to-state composition in a continual research agent.** Local statistical decisions are mapped by state governance into proposal-visible CURRENT capability and immutable EVER history, highlighting the Decision-to-State Gap between the decision object and the persistent state later consumed by search. Permanent append retains admissions in CURRENT, whereas cumulative recertification updates CURRENT using frozen p-values. The stateful proposer then generates the next hypothesis family and closes the loop; the dashed EVER edge appears only in the controlled EVER intervention.

3 Controlled Benchmark and Experimental Design

3.1 Synthetic ground-truth environment

The frozen catalog combines 16 features with eight unary templates, yielding 128 candidates and 64 RankIC-equivalence groups. Four hidden signal coordinates define 32 nonnull candidates and 16 nonnull groups in the planted world. Complete panels contain 60 assets and 180 days; Gaussian features generate $r_{d+1,a} = .16 \sum_{j \in \mathcal{I}} x_{d,a,j} + \epsilon_{d+1,a}$ in the planted world and noise-only returns under the Global Null. Truth is never proposer-visible; Appendix A gives the construction.

3.2 Evidence and tested-null semantics

Evidence tests $H_0 : \mathbb{E}[IC_d] = 0$ against a positive alternative using 35 daily finite-asset sample RankICs and a Student-t_{34} reference. The benchmark uses a 60/20/20 research/evidence/final-holdout split with lookback 5; the final holdout remains locked. Under the restricted Gaussian construction, null-coordinate fields are independent of signal/noise/returns, aligning structural truth with the tested mean null (Appendix A); exact finite-sample validity of the t reference remains a separate obligation.

The benchmark mechanically enforces Evidence access, hypothesis freeze and once-only evaluation. Statistical validity is a separate layer: candidate-level p-value validity, dependence from shared panels, rolling overlap and rank partners, and adaptive-family conditioning determine which classical BH/BY guarantees apply (Benjamini & Hochberg, 1995; Benjamini & Yekutieli, 2001). The preregistered calibration screen provides an empirical anti-conservatism check for the implemented Evidence engine; it does not replace theorem-level validity analysis.

3.3 Governance policies: permanent append and cumulative recertification

Permanent append applies BH($q = .05$) to each new batch of ten and permanently retains admitted identities in CURRENT. Cumulative recertification reapplies BH($q = .05$) to all frozen once-observed p-values and defines the resulting rejection set as CURRENT, allowing retirement/readmission without reevaluating Evidence. EVER is the union of admissions:

$$C_t^{\text{Append}} = C_{t-1}^{\text{Append}} \cup A_t, \qquad C_t^{\text{Recert}} = \text{BH}_{.05}(P_{1:t}). \tag{4}$$

We use cumulative recertification as an empirical governance baseline built from existing BH: it determines CURRENT while reusing the same once-observed Evidence. Its role is to measure the resulting operational-state object rather than introduce a new testing procedure.

3.4 CONTROLLED STATE INTERVENTIONS

The proposer mutates with probability 0.8 by drawing a visible parent and then a legal unselected catalog neighbor; empty capability or no neighbor triggers a parentless uniform restart. Its proposal-visible fields are identity, expression and generation, with no retirement label. Prospectively selected sources are the earliest exactly reconstructible retired states at cycles 2–9 with EVER≠CURRENT and at least 30 candidates remaining. Forty Global-Null and 40 planted sources share history, catalog, settings and RNG across arms. A size-matched state augments CURRENT with truth-blind, distinct, never-admitted historical identities until its cardinality matches EVER, with no new Evidence.

Each rollout arm proposes 30 hypotheses over three ten-item cycles. Slot-keyed arm-invariant seeds prevent draw-count drift; both batches freeze before a fresh shared panel, each hypothesis is evaluated once, and decisions commit before the next cycle. The sole treatment is CURRENT versus EVER capability. An earlier unmatched comparison jointly varied memory feedback and panel generation, so we treat it only as a joint-system stress test.

3.5 REAL-LLM TRANSPORT PROTOCOL

The transport pilot uses the same 40 planted retired-state snapshots, CURRENT, EVER, the existing size-matched state, and fixed finite eligible next-candidate catalogs. Neutral prompts expose only canonical proposal-visible identities/descriptions and the allowed set; truth, Evidence, admission status and outcomes remain hidden. One DeepSeek `deepseek-v4-pro` proposer receives $K = 3$ calls per arm and snapshot (360 total); all responses are valid, with no provider failures or repairs. Provider and serialization details are in Appendix H.

Let Y_{iar} be the candidate selected for snapshot i, arm $a \in \{C, E, P\}$ and repeat $r \in \{1, 2, 3\}$, where C denotes CURRENT, E EVER and P the size-matched state. We used the frozen cross-arm disagreement and same-arm repeat-instability summaries

$$\widehat{B}_i^{ab} = K^{-2} \sum_{r,s} \mathbf{1}\{Y_{iar} \neq Y_{ibs}\}, \qquad \widehat{W}_i^{a} = \frac{1}{\binom{K}{2}} \sum_{r<s} \mathbf{1}\{Y_{iar} \neq Y_{ias}\}, \tag{5}$$

The primary score is $\widehat{\Delta}^{CE} = N^{-1} \sum_i [\widehat{B}_i^{CE} - (\widehat{W}_i^{C} + \widehat{W}_i^{E})/2]$ for $N = 40$; Δ^{EP} is analogous. Under independent categorical repeats its population analogue is $\frac{1}{2}\|p_i^a - p_i^b\|_2^2$, so the score isolates state-conditioned proposal-distribution change after accounting for same-state repeat instability. The pilot tests transport of state-to-search sensitivity under the frozen model and configuration; FDR transport lies outside this experiment.

3.6 EVALUATION PROTOCOL AND INFERENTIAL FAMILIES

The dependence experiment preregisters $N = 500$, ten cycles, $\rho = 0/.3/.7$ and both worlds. A score-blind shared 100-candidate family fixes selection before outcomes; fresh panels are paired across policies, and the three rho conditions share innovations. Dependence affects only nominal-null feature blocks. Six absolute terminal Global-Null active-state FDRs and three paired permanent-append-minus-recertification gaps form one approximate simultaneous 95% bootstrap family (Appendix D); the three rho conditions are paired rather than independent replications.

The original downstream analysis is exploratory; its 380-cell family and the 60-cell size-matched family remain separate. Independent confirmation freezes one primary, $D_i = Y_i(\mathrm{EVER}) - Y_i(\mathrm{CURRENT})$, where Y_i counts distinct future truth-null hypotheses admitted at least once among the 30 proposals generated during the three-cycle rollout. We refer to this endpoint as distinct future false admissions. The first 320 eligible sources from a separately seeded prospective frame enter a one-sided paired source-mean t-type test of $H_0 : \mathbb{E}[D] \leq 0$ at $\alpha = .05$; this discrete-count reference is asymptotic, not exact. The sole lineage secondary is descriptive (Appendix F).

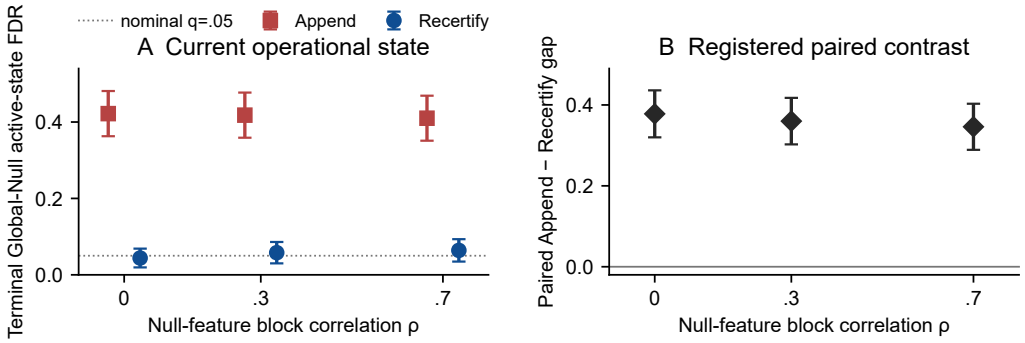

Figure 2: Primary state-reliability result ($N = 500$). Left: terminal Global-Null active-state FDR for per-batch BH with permanent append and cumulative recertification. Right: paired permanent-append-minus-recertification gaps. Error bars are the frozen approximate simultaneous 95% bands from one nine-coordinate family. The three rho conditions share innovations; they are not independent replications or a tested rho trend. The dotted nominal q=.05 reference is not a theoretical state-control guarantee.

Table 1: Registered paired dependence results; intervals share a single nine-coordinate approximate simultaneous family. All absolute bands appear at full stored precision in Appendix D.1.

ρ	Append FDR	Recertify FDR	Gap	Approx. simultaneous 95% band
0	0.422	0.044	0.378	[0.3199, 0.4361]
0.3	0.418	0.058	0.360	[0.3025, 0.4175]
0.7	0.410	0.064	0.346	[0.2890, 0.4030]

4 Results

4.1 Persistent-state reliability can diverge from local control

The baseline makes the reliability-object mismatch concrete: terminal Global-Null active-state FDR is 0.444 under permanent append versus 0.060 under cumulative recertification, with planted equivalence-group power 0.9306 versus 0.9251 (99.4% observed retention; utility equivalence was not tested). Under the Global Null, permanent retention accumulates exposure in the persistent state, whereas BH's fixed-family theorem governs its specified rejection family (Benjamini & Hochberg, 1995). The registered dependence experiment below is the principal quantitative reliability result; Appendix J records historical active/EVER and stopping diagnostics.

4.2 The reliability gap persists under dependence stress

The preregistered controlled positive-dependence experiment yields active-state FDR 0.410–0.422 under permanent append versus 0.044–0.064 under cumulative recertification (Figure 2); all three paired gaps have approximate simultaneous lower bands above zero (Table 1). Its $\rho = 0$ condition uses this protocol's score-blind family and fresh paired panels, distinct from the earlier baseline. These experiments establish robustness over the registered null-block dependence family; broader dependence structures and a rho trend remain outside the tested design.

Before execution, 768,000 calibration evaluations produced 2,688 valid diagnostics and zero Holm rejections (Holm, 1979). The preregistered coarse empirical screen detected no violation under its prespecified criteria; exact p-value validity, BH assumptions and anytime control remain theorem-level questions.

The reliability objects therefore diverge under the registered stress conditions; we next ask whether changing the operational state changes what the system searches.

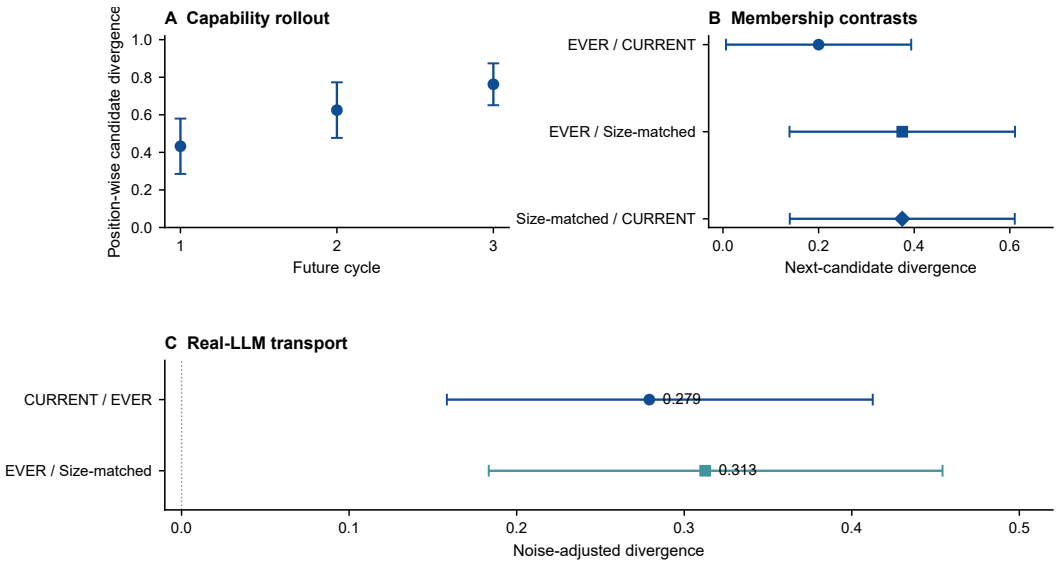

Figure 3: State-to-search mechanism evidence. A: proposal-visible state interventions changed position-wise next-candidate behavior across three cycles in selected planted-signal snapshots; whiskers are the 380-cell exploratory approximate simultaneous bands. B: EVER and a size-matched state with different membership yielded different next proposals in 15/40 planted pairs, showing that membership can matter beyond capacity; whiskers use the separate 60-cell exploratory approximate simultaneous family and one frozen truth-blind padding draw. C: the frozen real-LLM transport pilot (DeepSeek `deepseek-v4-pro`; 40 frozen snapshots; three repeats/arm) shows positive noise-adjusted divergence for CURRENT versus EVER and EVER versus a size-matched state; whiskers are 95% stability intervals from 10,000 whole-snapshot bootstrap resamples. Panel C provides narrow, mechanism-consistent transport evidence rather than general external validity.

4.3 Controlled proposal-visible state interventions change search

Changing proposal-visible membership changes search. In 40 planted retired-state snapshots, EVER versus CURRENT selects a different next candidate in 8/40 pairs: 0.200, marginal Wilson 95% interval [.105,.348]. Shared history, catalog, proposer and RNG identify a capability intervention conditional on this selected population.

Position-wise planted candidate divergence across three matched cycles is 0.4325, 0.6250 and 0.7625 (Figure 3A). Because the proposer receives capability identities but no retirement labels, the contrast operates through proposal-visible membership. Downstream harm is tested separately in Section 4.6.

4.4 Equal-size states with different membership remain behaviorally distinct

The size-matched state retains CURRENT and adds truth-blind, distinct, never-admitted historical identities until its cardinality matches EVER. Planted-pair divergence is 8/40 for EVER/CURRENT, 15/40 for EVER/size-matched and 15/40 for size-matched/CURRENT (Figure 3B). Thus

$$|V_1| = |V_2| \not\Rightarrow \pi(\cdot \mid V_1) = \pi(\cdot \mid V_2). \tag{6}$$

Equal capacity need not yield equal behavior. The single size-matched-state realization establishes an identity-sensitive equal-size contrast; averaging over padding draws and additive decomposition remain separate questions.

4.5 Real-LLM transport of state-to-search sensitivity

The controlled proposer identifies the mechanism; the transport pilot asks whether corresponding state sensitivity appears in a learned proposer. Across 40 planted snapshots, $W_C = 0.125$,

$W_E = 0.0333$ and $B_{CE} = 0.3583$, giving $\Delta_{CE} = 0.2792$ (10,000 whole-snapshot bootstrap 95% stability interval [0.1583,0.4125]). The membership contrast is $\Delta_{EP} = 0.3125$, with corresponding 95% stability interval [0.1833,0.4542]; all 360 calls are valid with zero provider failure. The positive noise-adjusted contrasts show that proposal-visible state shifts the tested model's candidate distribution beyond repeat instability. This transports state-to-search sensitivity for the frozen snapshots, prompts and `deepseek-v4-pro`; FDR and downstream-harm transport remain open. Appendix H reports the remaining diagnostics.

4.6 Search divergence and downstream harm require separate empirical tests

A. Exploratory observation. Proposal sensitivity alone does not determine downstream statistical harm, which requires a separate outcome test. The original 40-pair planted rollout generated this hypothesis: EVER–CURRENT distinct future false admissions differ by +0.200, but the complete exploratory simultaneous band [-0.136,0.536] spans zero. All other displayed harm and utility bands also span zero (Appendix F), motivating independent confirmation.

B. Independent preregistered confirmation. We froze one primary endpoint and tested it on 320 prospectively selected sources from an independent 10,000-trajectory frame; all matched three-cycle pairs completed without reusing original sources. Mean $Y_{\mathrm{CURRENT}} = .3594$ and $Y_{\mathrm{EVER}} = .3781$, so mean $D = .0188$, one-sided $p = .280$, with conventional two-sided 95% interval [-.045,.082]. The primary is **NOT CONFIRMED** (Figure 4). The independent result leaves downstream amplification unresolved rather than supporting either amplification or equivalence, and it provides no universal CURRENT-safety conclusion. Appendix F preserves full precision, execution counts and the prespecified minimum effect of interest (.10).

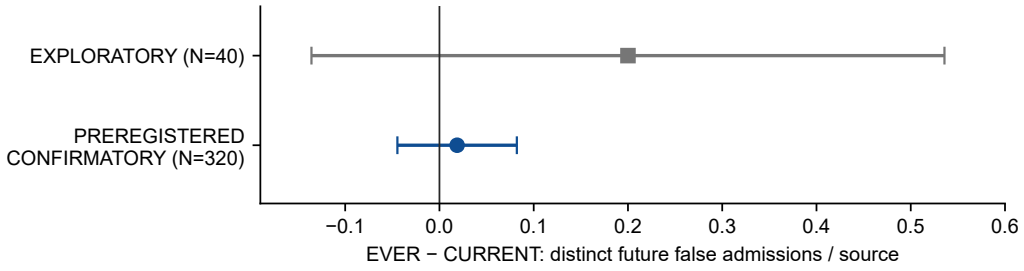

Figure 4: Distinct future false admissions per source: exploratory versus independent confirmation. The exploratory 40-pair band is approximate simultaneous 95% from the full 380-cell audit; the 320-pair confirmation uses a conventional two-sided 95% t interval alongside a preregistered one-sided test ($p = .2804$). Different interval constructions are explicitly separated; neither stage confirms the original +0.200 downstream increase. Intervals spanning zero do not establish equivalence.

Prespecified descriptive lineage result. The prespecified truth-null parent-consumption secondary remains descriptively different but non-confirmatory; it cannot alter the primary status. Appendix F gives its full definition and precision.

5 Discussion

5.1 Reliability objects in continual agents

Reliability should be attached to the state the system actually uses. Local admission describes new decisions, CURRENT describes present action capability, and EVER preserves complete admission history. A valid local procedure can therefore coexist with a different error rate for a differently composed persistent object. Recertification improves CURRENT empirically here while EVER preserves provenance. Utility equivalence was not tested, and Evidence validity and dependence assumptions remain separate from the access contract.

5.2 Preserve history, govern capability

Immutable provenance is compatible with restricted permission: retirement can remove a mutation parent without deleting why it was admitted or retired. Because members determine available transformations, capacity limits alone do not preserve behavior. CURRENT can implement the proposal-visible pool while EVER remains audit history; retired evidence may still serve other explicit roles. We treat this as an engineering principle for separating provenance from action permission; optimal memory policies and production-agent validation remain open.

5.3 Behavioral mechanisms versus downstream consequences

The gap between search divergence and admitted error motivates several possible attenuation mechanisms. The lineage secondary changes descriptively, whereas the independent primary is **NOT CONFIRMED** for more distinct future false admissions. Candidate paths may enter dead-end or low-admissibility regions, reconverge on high-signal regions, or encounter an Evidence/recertification bottleneck. The present experiments leave these attenuation mechanisms unresolved. Appendix I discusses cumulative recertification's computational trade-off.

5.4 Scope and external validity

The strongest identification comes from a finite, unary exact-truth benchmark and a selected retired-state population. Dependence robustness covers the registered null-block family, with calibration serving as an empirical screen. The size-matched intervention provides one frozen equal-size realization. Real-LLM transport covers one model, finite candidate catalogs, 40 planted snapshots and neutral prompts. The independent downstream-harm primary leaves the hypothesized increase unconfirmed. Open-ended agents, broader model families, real-market settings and downstream reliability transport remain future tests.

6 Related Work

Scientific-agent state and validation. StatefulDiscovery and HypoExplore connect investigation or hypothesis memory to exploration (Chen et al., 2026; Koo et al., 2026); EvoScientist distills research histories and AutoSci separates active from long-term memory (Lyu et al., 2026; Qian et al., 2026). ScienceAgentBench evaluates executable components (Chen et al., 2025), while XScientist and fuzz testing emphasize provenance, re-execution and protected validation (Luo, 2026; He et al., 2026). POPPER uses sequential falsification (Huang et al., 2025). We instead measure the transition from statistical admission identity to proposal-visible capability, intervening on CURRENT/EVER permission and equal-size membership under exact truth.

Statistical decision control. BH/BY, LORD++, SAFFRON, Batch-BH and e-BH/stopped e-BH govern specified rejection sets or streams under their assumptions (Benjamini & Hochberg, 1995; Benjamini & Yekutieli, 2001; Ramdas et al., 2017; 2018; Zrnic et al., 2020; Wang & Ramdas, 2022; Wang et al., 2025). A continual system additionally maps admissions into CURRENT capability, retirement/readmission and EVER provenance. Our permanent-append policy uses ordinary fixed-level per-batch BH followed by persistent state retention, distinct from Zrnic et al. (2020)'s level-allocating Batch-BH. Our contribution is empirical measurement of the admission-to-capability composition that connects these two layers.

7 Conclusion

Continual agents act on what they retain, not only on what they once tested. Local decisions and persistent operational state are distinct reliability objects, and controlled changes to proposal-visible state alter future search, including under equal-size membership interventions. A frozen pilot transports this state-sensitive proposal contrast to one specified real LLM. The downstream-harm primary remains **NOT CONFIRMED**, separating mechanism identification from consequence confirmation. Preserve history; govern capability.

AI USE STATEMENT

Generative AI assisted framing, methods planning, implementation, analysis, interpretation, literature triage, debugging, figure planning and drafting beyond copy editing. Benchmark data come from specified generators, not model-authored observations; numerical claims and citations are checked against frozen artifacts and primary records. AI outputs are not scientific evidence, and human authors remain responsible for review before submission.

REPRODUCIBILITY STATEMENT

Appendices document tested-null semantics, covariance, calibration, inferential families, source selection, endpoints, recovery and pre-treatment timestamps. A versioned artifact map binds immutable results, ledgers and deterministic plot/table scripts, including transport records. Manuscript integration reruns no experiment or inference; a sanitized anonymous supplement has been prepared as a separate release artifact.

REFERENCES

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001. doi: 10.1214/aos/1013699998.

Jiayao Chen, Shi Liu, and Linyi Yang. StatefulDiscovery: Evidence-calibrated claim formation in open-ended scientific discovery. *arXiv preprint arXiv:2606.11851*, 2026. URL https://arxiv.org/abs/2606.11851.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, Vishal Dey, Mingyi Xue, Frazier N. Baker, Benjamin Burns, Daniel Adu-Ampratwum, Xuhui Huang, Xia Ning, Song Gao, Yu Su, and Huan Sun. ScienceAgentBench: Toward rigorous assessment of language agents for data-driven scientific discovery. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/f12b4df26344f3be803c06b555252efe-Abstract-Conference.html.

Yifeng He, Jicheng Wang, Yinzhe Zhao, Chengyang Shi, Jiachen Liu, and Hao Chen. The greatness of science cannot be planned: Agentic auto-research is fuzz testing. *arXiv preprint arXiv:2608.09855*, 2026. URL https://arxiv.org/abs/2608.09855.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Kexin Huang, Ying Jin, Ryan Li, Michael Y. Li, Emmanuel Candes, and Jure Leskovec. Automated hypothesis validation with agentic sequential falsifications. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 25372–25437, 2025. URL https://proceedings.mlr.press/v267/huang25n.html.

Jaywon Koo, Jefferson Hernandez, Ruozhen He, Hanjie Chen, Chen Wei, and Vicente Ordonez. Agentic discovery with active hypothesis exploration for visual recognition. *arXiv preprint arXiv:2604.12999*, 2026. URL https://arxiv.org/abs/2604.12999.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024. URL https://arxiv.org/abs/2408.06292.

Jixiang Luo. XScientist: A Git-like research protocol for long-running autonomous scientific discovery. *arXiv preprint arXiv:2607.12301*, 2026. URL https://arxiv.org/abs/2607.12301.

Yougang Lyu, Xi Zhang, Xinhao Yi, Yuyue Zhao, Shuyu Guo, et al. EvoScientist: Towards multi-agent evolving AI scientists for end-to-end scientific discovery. *arXiv preprint arXiv:2603.08127*, 2026. URL https://arxiv.org/abs/2603.08127.

Weitong Qian, Beicheng Xu, Zhongao Xie, Bowen Fan, Guozheng Tang, et al. AutoSci: A memory-centric agentic system for the full scientific research lifecycle. *arXiv preprint arXiv:2605.31468*, 2026. URL https://arxiv.org/abs/2605.31468.

Aaditya Ramdas, Fanny Yang, Martin J. Wainwright, and Michael I. Jordan. Online control of the false discovery rate with decaying memory. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/7f018eb7b301a66658931cb8a93fd6e8-Abstract.html.

Aaditya Ramdas, Tijana Zrnic, Martin Wainwright, and Michael Jordan. SAFFRON: An adaptive algorithm for online control of the false discovery rate. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4286–4294, 2018. URL https://proceedings.mlr.press/v80/ramdas18a.html.

Hongjian Wang, Sanjit Dandapanthula, and Aaditya Ramdas. Anytime-valid FDR control with the stopped e-BH procedure. *Statistics & Probability Letters*, 226:110512, 2025. doi: 10.1016/j.spl.2025.110512.

Ruodu Wang and Aaditya Ramdas. False discovery rate control with e-values. *Journal of the Royal Statistical Society: Series B*, 84(3):822–852, 2022. doi: 10.1111/rssb.12489.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The AI scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*, 2025. URL https://arxiv.org/abs/2504.08066.

Tijana Zrnic, Daniel Jiang, Aaditya Ramdas, and Michael Jordan. The power of batching in multiple hypothesis testing. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3806–3815, 2020. URL https://proceedings.mlr.press/v108/zrnic20a.html.

A Tested Null, Structural Truth and Evidence Boundary

The implemented test concerns $H_0 : \mathbb{E}[IC_d] = 0$, positive one-sided, where IC_d is daily Spearman correlation across 60 complete assets after the candidate transformation. It is not a regression coefficient test, raw Pearson correlation, or a single-day Spearman significance test. Evidence aggregates 35 usable days by sample mean/SE with a Student-t_{34} reference; minimum finite-day count 20 remains a guard, not a changed target.

Under the frozen Gaussian DGP, a nominal-null coordinate's entire field is independent of signal coordinates, Gaussian noise and returns. Raw/rolling transformations preserve that independence. Conditional on a transformed factor's ranks, the return ranks form an independent uniformly distributed permutation across assets, whose expected centered rank inner product is zero. This bridges structural assignment to the *finite-asset sample* RankIC mean-null; population Spearman and finite-asset expected Spearman are not interchangeable numerical formulas. In the planted world 96 candidates are mean-null and 32 have positive signal; the Global Null has 128 mean-null. Independence of fields aligns labels but does not make the finite 35-day t-reference exact. Rolling overlap and shared returns can couple daily/cross-hypothesis statistics.

Mechanical isolation, candidate-level null validity, cross-hypothesis dependence, BH independence/PRDS and adaptive-family conditional validity must be checked separately. The global filtration includes candidate identities, Evidence observations, certification outputs, active-memory state and stopping. Raw payload hiding is only an access boundary: Evidence→decision→certified identity→future capability still carries feedback. None of the empirical p-value experiments closes arbitrary-stopping proof obligations; deterministic Evidence infrastructure validates normalization/event order, not new e-value power or a theorem.

B Dependence Construction and Preregistered Execution

Historical artifact labels C0 and C2 denote the permanent-append and cumulative-recertification policies, respectively.

Source order is $x_0, \ldots, x_{15}$. Signal coordinates are (x_9, x_{10}, x_6, x_5), with covariance I_4. Null blocks are (x_{15}, x_3, x_0, x_2), $(x_{12}, x_8, x_{14}, x_1)$ and $(x_{13}, x_{11}, x_4, x_7)$. Each block has $E_4(\rho) = (1-\rho)I_4 + \rho\mathbf{1}\mathbf{1}^\top$, giving a permuted $\mathrm{diag}(I_4, E_4, E_4, E_4)$ at $\rho = 0, .3, .7$. All signal–null covariance is exactly zero. Null-only Cholesky transformation keeps signal features and returns unchanged; block eigenvalues are $1-\rho$ (multiplicity 3) and $1+3\rho$ (multiplicity 1). The tested-null argument in Appendix A therefore survives this restricted construction. Mixing signal and null blocks would not have this property and is not included.

Main execution uses 500 seeds, two worlds, three rho conditions, $T = 10$, $K = 10$, and a score-blind common 100-candidate order. Fresh base innovations are paired across conditions/worlds for each seed/cycle; C0 and C2 replay the same frozen observations without additional Evidence. Main Evidence evaluations are 300,000; calibration evaluations 768,000; combined 1,068,000. C0 applies per-batch BH(.05) append; C2 applies cumulative BH(.05) to once-observed statistics with retirement. No rho trend, downstream-harm endpoint or outcome-dependent expansion is introduced. Package/environment hashes, target/access/identity/seed, shards, journal attempts and atomic cycle commits are audited before complete global inference.

C Calibration Screen and Diagnostic Criteria

Calibration uses $N = 2000$ Global-Null panels per candidate/rho and all 128 distinct hypotheses, including rank partners. Seven thresholds are .0005/.001/.0025/.005/.010/.025/.050. The C0/C2 operational set contains 320 unique rational thresholds from $i/(200t)$ for $t = 1, \ldots, 10$ and $i = 1, \ldots, 10t$; only seven representative points are screened, not all 320. The 2688-coordinate diagnostic family is $128 \times 3 \times 7$, with each coordinate gate-critical.

At threshold α, the diagnostic null is $\Pr(p \le \alpha) \le \alpha + \max(.005, .2\alpha)$, tested by an upper-tail binomial reference and joint Holm FWER .01 (Holm, 1979). Boundaries are .0055/.006/.0075/.010/.015/.030/.060, respectively: boundary/alpha ratios 11/6/3/2/1.5/1.2/1.2. At extreme tails this is a coarse catastrophic screen, not near-exact calibration. The frozen protocol defines three execution statuses: `PASS` for complete valid data with no diagnostic rejection, `FAIL` for at least one diagnostic rejection, and `INVALID` when integrity, resource or pending failures take precedence. The 768,000 complete valid records and 2688 diagnostics with zero Holm rejections meet the `PASS` status. No invalid rows are dropped, and this status does not prove superuniformity, PRDS, BH conditions or anytime validity.

D Inferential Families and Registered Dependence Results

The registered dependence experiment corresponds to frozen artifact label `A1`. Each resample draws 500 *whole* seed vectors jointly over the six absolute Global-Null coordinates and three paired gaps. With original mean SE s_k, the bootstrap maximum is $\max_k |\widehat{\theta}_k^* - \widehat{\theta}_k|/s_k$; 10,000 draws and the 95% higher quantile define bands $\widehat{\theta}_k \pm cs_k$. Bands are not truncated to feasible ranges. Zero-SE coordinates retain descriptive estimates with undefined inferential bands. All nine reported coordinates are evaluable. This is an approximate simultaneous bootstrap construction, not exact finite-sample coverage.

The 380-cell exploratory rollout family uses all prespecified summary cells across endpoints/cycles/strata; paired source IDs resample within world-by-CURRENT-empty strata. 334 finite nondegenerate cells enter one maximum-standardized bootstrap calibration, 44 constant and two undefined cells remain disclosed. The size-matched follow-up has its own 60-cell family, with constant cells kept descriptive. These inferential families are kept separate and are not retrospectively pooled into a confirmatory analysis.

D.1 ALL NINE DEPENDENCE COORDINATES AT FULL STORED PRECISION

Coordinate	Estimate	Approximate simultaneous 95% band
theta_C0_0	0.422	[0.36283856114472385, 0.4811614388552761]
theta_C2_0	0.044	[0.01943177009255993, 0.06856822990744006]
theta_C0_0.3	0.418	[0.3589162276169167, 0.47708377238308325]
theta_C2_0.3	0.058	[0.030000000000000013, 0.086]
theta_C0_0.7	0.41	[0.3510835572277715, 0.46891644277222844]
theta_C2_0.7	0.064	[0.03468117950480411, 0.0933188204951959]
delta_0	0.378	[0.3199155775462888, 0.4360844224537112]
delta_0.3	0.36	[0.30250106237672036, 0.4174989376232796]
delta_0.7	0.346	[0.2890169741790126, 0.40298302582098733]

E ORIGINAL SNAPSHOT, SIZE-MATCHED AND CONSEQUENCE CONTRACTS

Original source selection keeps the earliest eligible retired state per trajectory, then 40 sources/world by the frozen source-ID hash rule. Eligibility requires completed cycle 2–9, retirement, EVER≠CURRENT, at least 30 remaining candidates and exact history reconstruction. CURRENT may be empty. The artifact condition `NO_STATE` denotes the no-state proposal condition. The snapshot comparison includes this no-state condition, EVER and CURRENT, which share source/catalog/RNG; one proposal per condition gives 240 controlled invocations and zero new Evidence. The full-proposal divergence trigger of at least 8/80 is an engineering magnitude threshold, not a statistical success criterion. All 80 then enter the prespecified three-cycle rollout; source-specific divergence is not a filter. Three matched cycles give 4800 Evidence evaluations, shared fresh panels only after both batches freeze, no raw same-cycle feedback.

The pooled EVER/CURRENT next-proposal divergence is 46/80 but is secondary because it combines nonempty capability changes with CURRENT-empty on/off cases; among nonempty CURRENT sources across worlds, 11/45 differ. The main text therefore reports the prespecified 8/40 planted contrast and the three-cycle planted trajectory.

The artifact label `PADDED` denotes the size-matched state used in the main text. The size-matched state is implemented by the frozen artifact construction `CURRENT_RANDOM_HISTORY`: canonical CURRENT plus distinct never-admitted historical identities, chosen by frozen truth-blind hashes without replacement until EVER item count. Generation/expression metadata are original, not synthesized to match neighbors. The original 80 sources and 240 snapshot outcomes remain immutable. All source, seed and arm comparisons satisfy their checks; no new p-values are evaluated. One size-matched draw tests an identity-sensitive composition contrast, not an average over all random histories. Two 15/40 contrasts can involve different pairs and are neither independent replications nor an additive decomposition. Exact 64-bit seeds are retained as decimal strings in a versioned display projection after a concatenated float export; actual generators used the frozen integers. No scientific rerun or outcome change occurred.

All four displayed original exploratory planted-harm contrasts have approximate simultaneous bands spanning zero: distinct future false admissions +.200[-.136,.536], truth-null recorded-parent events +.675[-.199,1.549], active FDP +.0106[-.0105,.0317], EVER FDP +.0079[-.0063,.0222]. Utility contrasts for newly explored/discovered true groups and candidate yield span zero; there is no equivalence or noninferiority test. Under the Global Null, the distinct future false admissions and active-FDP differences are $-.05$ and $-.025$ with zero-spanning bands, preserving the counterexample to behavioral divergence automatically implying statistical harm. The original timing diagnostic is the earliest proposal position of a finally ever-discovered true group, not chronological time of certification. Labels are corrected without rescaling or rewriting results.

F INDEPENDENT DOWNSTREAM-HARM CONFIRMATION: PLANNING, SELECTION AND PRIMARY ENDPOINT

Exploratory source SD .6076436202502 is used only as a nuisance planning estimate; the previously observed exploratory mean +.200 is not assumed true. Analytical and centered-empirical residual

calculations considered multiple effect sizes; these planning simulations are not scientific replication and do not preserve every discrete count-DGP property. Human choices freeze one-sided source-mean t-type testing, a minimum effect of interest (MEI) of .10, target power .90 and $N = 320$ before new source/treatment observations. Planning power does not guarantee realized power or establish an effect above the MEI.

The separate source frame is fixed at indices $j = 0, \ldots, 9999$ with new innovation/seed namespaces, the canonical planted assignment and the original eligibility semantics. One state per trajectory uses earliest qualifying cycle; all 10,000 source trajectories complete and are globally validated before treatment. Prospective order hashes the *index*, not source outcomes; the first 320 eligible of 2928 are frozen globally, then split by selection ordinal into 160 whole pairs/host. Original source IDs are excluded; original source panels are not reused as new confirmation observations. Seed/hash separation enforces nonreuse and collision checks, not a mathematical IID proof; the independent fresh-innovation model and conditional selected-source population motivate the asymptotic test.

In each source/arm, Y counts DISTINCT truth-null newly generated hypotheses among 30 future proposals generated during the three-cycle rollout which enter EVER at least once by the end of three cycles. Source-prefix history is excluded. Repeated admission/readmission does not increase Y. This endpoint is called distinct future false admissions. The paired contrast is $D_i = Y_i(\mathrm{EVER}) - Y_i(\mathrm{CURRENT})$, not terminal false-active size, a lineage event count or historical false-memory count. Full batches freeze in both arms before each shared fresh Evidence panel; historical and newly observed p-values enter identical cumulative BH(.05) recertification, with capability intervention only. Every new hypothesis is evaluated once; atomic cycle commit precedes subsequent proposal.

For complete valid 320 pairs with SD> 0, $T = \sqrt{320}\,\overline{D}/s_D$ and $p = \Pr(t_{319} \geq T)$; rejection is strictly $p < .05$ for the positive direction. SD 0 predefines $p = 1$, no rejection and undefined T/CI. No permutation/random assignment makes sign-flip inference exact under a mean-null; this is an asymptotic t reference for independent finite-variance source contrasts. Negative point mean gives REVERSE, not a post-hoc opposite-direction success; non-rejection is NOT CONFIRMED, not no effect or equivalence. INVALID is integrity failure, never a scientific negative. No interim effect inspection, replacement, retry, sample extension or endpoint change occurs.

F.1 Full stored confirmatory result

Field	Frozen result
n	320
record_count	19200
mean_Y_CURRENT	0.359375
mean_Y_EVER	0.378125
mean_D	0.01875
SD_D	0.5761386436241853
T	0.5821692405756351
df	319
p_one_sided	0.28043195263952453
ci_95	[-0.0446152204261772, 0.08211522042617721]
median_D	0.0
range_D	[-2.0, 2.0]
support_D	-2: 4, -1: 32, 0: 244, 1: 34, 2: 6
status	NOT CONFIRMED
mei_planning_only	0.1

F.2 Prespecified descriptive secondary

The only secondary is false-parent consumption rate: F counts successful recorded-parent slots whose parent is truth-null, and P counts any successful recorded-parent slot. Each source/arm rate is F/P if $P > 0$, otherwise 0 with a zero-exposure flag retained. Source-weighted arm means/paired differences are not aggregate $\sum F / \sum P$ ratios. Failed unlogged no-neighbor parent draws are not

inferred. Retirement and truth are distinct. 10,000 whole-pair resamples give a descriptive percentile interval; no secondary test or primary rescue is authorized.

Field	Descriptive frozen result
mean_CURRENT	0.04093177749225744
mean_EVER	0.058370962691622806
paired_mean_difference	0.01743918519936537
percentile_ci_95	[0.010578716447152603, 0.024676678560712127]
bootstrap_resamples	10000

G Versioned Package Recovery and Pre-Treatment Freeze

The original source merge stopped after generation with a missing administrative dependency: a package whitelist omitted a module imported by the seed-audit helper. The original package and `INVALID` markers, along with the valid registered-dependence results, were preserved. Separately authorized recovery restored the exact pre-science module bytes, verified from successful timestamped patch history, not a fabricated Git commit. All scientific code remained byte-identical; administrative import restores serialization/closure only. Isolated imports and fake lifecycle tests preceded real source reuse.

All 10,000 native sources passed reconstruction/journal/receipt/identity/truth/access/seed/state checks without regeneration. 75,000 files in the imported odd half matched host originals. The recovery package is a separate 174-file namespace; its effective runtime hash differs from the retained scientific design identity. Neither the registered dependence experiment nor any source unit was rerun. Merge/seed/global selection completed before treatment. Selected 320 freeze is 2026-09-17T12:30:14.566022Z; first treatment starts at 12:31:03.054262Z(A) and 12:31:03.923314Z(B). Hash-bound freeze files reached both hosts first. The incident yields no alternative endpoint, excluded row or scientific retry; it is a reproducibility event, not a statistical result. Complete 19200-treatment observations and final global 320 inference were verified with historical `INVALID` records intact.

H Real-LLM Transport Record

The transport pilot is separate from the earlier broader provider campaign. DeepSeek `deepseek-v4-pro` used reasoning disabled, temperature 0, top-p 1, no seed, no tools, nonstreaming strict JSON-schema output with exactly one required string field `candidate_id` and no additional properties. It followed the frozen protocol package, run configuration and ascending 360-call schedule. Forty snapshots, three arms and three repeats completed without replacement calls, interim adaptation, output repair or size-matched-state redrawing. The underlying response ledger and parsed output table are immutable scientific artifacts with recorded hashes. The anonymous supplement releases the sanitized scientific ledger and parsed choices while omitting raw provider envelopes and provider request/response identifiers.

Full diagnostics are $W_C = 0.125$, $W_E = 0.0333$, $W_P = 0.0917$, $B_{CE} = 0.3583$, $B_{EP} = 0.3750$ and $B_{CP} = 0.2889$. These give $\Delta_{CE} = 0.2792$ with a 95% stability interval [0.1583,0.4125] and $\Delta_{EP} = 0.3125$ with a 95% stability interval [0.1833,0.4542] under 10,000 whole-snapshot bootstrap draws. Across snapshot-arm cells, 88.3% have three identical selections, 10.0% have two distinct selections and 1.7% have three; invalid and provider-failure rates are zero. The pilot supports only provider/model- and snapshot-conditional state sensitivity. It does not pool the historical 162 campaign attempts, validate an open-ended agent, or reproduce synthetic FDR.

I Computational Scope of Recertification

Cumulative recertification reuses once-observed Evidence: each hypothesis is evaluated once and historical p-values are reused, not recomputed. Governance updates a growing rejection family, increasing naive ranking/sorting, storage/indexing, state-maintenance costs and update latency. In-

cremental ranking, periodic or batched recertification, checkpointed governance and hierarchical or bounded active-state policies are future systems directions whose reliability requires separate evaluation. Scaling governance is distinct from repeated Evidence evaluation.

J Stopping Identity, Historical Corrections and Negative Evidence

Let τ be the first $t \leq T$ with $C_t \neq \varnothing$, using $\tau = T$ and empty state when no such time exists. Under the Global Null, every item is null and EVER is the union of admissions. Therefore

$$\mathrm{FDP}(C_\tau) = \mathbf{1}\{\exists t \leq T : C_t \neq \varnothing\} = \mathbf{1}\{E_T \neq \varnothing\} = \mathrm{FDP}(E_T). \tag{7}$$

The reported τ-FDR = .101 and EVER-FDR = .101 are the same event. Its Wilson interval [.0823,.1197] includes .10; a point planning-threshold miss is not a demonstrated population violation. Nominal $q = .05$ is not that operational threshold. Segmentation $20\times5/10\times10/5\times20$ with the same terminal cumulative p-family deterministically yields the same cumulative-recertification terminal BH set; this is definition consistency, not three independent robustness results. Intermediate/path/EVER behavior can still differ. Historical artifact labels `OPEN`/`CLOSED` refer to the earlier unmatched joint-system comparison. That comparison changes both memory feedback and panel-generation protocol, so it is joint-system stress, not feedback-alone causality.

The historical phrase "32 nonnull equivalence groups" is a label error: implementations use 16 groups and 32 nonnull candidates. No historical power is rescaled. The earlier broader real-provider diagnostic remains incomplete under its completeness/repeat-instability gates; it is not benchmark replication and is not pooled with the completed transport pilot documented in Appendix H. The native bridge feasibility audit did not identify a defensible CURRENT/EVER trace mapping, and no native treatment was run. These limitations do not refute the controlled mechanism and do not confirm general LLM behavior.

K Artifact Map and Exact Claim Boundaries

The versioned supplement map identifies immutable calibration/main results, exploratory 380-cell and size-matched 60-cell audits, pre-treatment source/seed freezes, 320 paired primary counts, the descriptive lineage interval, runtime/environment summaries and recovery history. Deterministic manuscript plotting reads existing intervals; it performs no benchmark generation, hypothesis proposal, Evidence evaluation, bootstrap recalculation or provider call. The claim contract links each abstract/introduction/contribution/result/discussion/conclusion claim to named sources and inferential status. Scientific truth, family definitions and historical reports remain read-only. Manuscript source uses local asset paths and an anonymous template; operational credentials/host addresses are not part of the paper. Supplement sanitization, author approval and OpenReview submission are handled as distinct release gates.