ACTION-CONDITIONED EVIDENCE CONTRACTS FOR PROSPECTIVE EVALUATION OF SCIENTIFIC AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Final-answer correctness alone does not reveal whether a scientific agent used admissible evidence or revised its reported belief appropriately. We introduce a prospective evaluation protocol in which an agent commits to a belief and an action before receiving evidence, then reports an updated belief and proposes releasing or withholding its conclusion. A Prospective Scientific Episode links these records; an Action-Conditioned Evidence Contract supplies admissibility constraints. On 160 synthetic episodes, runtime filtering adds holds for a non-checking rule policy but none for aligned policies. Across 480 runs of the non-checking policy, full enforcement blocks 33 releases admitted by provenance-and-quality checks; in 32 of these, the posterior mode matches the oracle label. At matched coverage, contract risk falls from 10.0% to zero, while oracle error risk rises from 14.55% to 15.82%. Reanalysis of 313 historical language-model trajectories shows only one additional hold from runtime filtering; supplied likelihoods and a domain–condition confound restrict these results to diagnostics of recorded responses. A live register and an eight-task executable pilot illustrate the workflow without establishing a capability gain. The linked episode record exposes disagreements between admissibility, outcome correctness, and release decisions that a single "verified" label would hide.

1 INTRODUCTION

Scientific-agent evaluations increasingly test executable research tasks, tool use, and complete workflows (Chen et al., 2025; Bragg et al., 2026). Yet a correct-looking answer and a replayable trace do not establish that an agent tested its belief with admissible new evidence. A source can be authentic but premature; a measurement can be reproducible but obtained by an unauthorized action; an admissible observation can contradict the belief that motivated its collection. A single "verified" label obscures these differences.

Consider an agent predicting whether a public environmental series will cross a threshold during a future week. Before that week, it chooses between an authoritative archival product, a provisional feed, and an inadmissible shortcut, under a fixed budget. After the week, it should update only on the product it selected. An expired rolling feed should produce a hold without directional updating; an admissible measurement that contradicts the prior should produce revision. A retrospective forecast, a valid but uninformative action, and a scientifically correct update are different evaluation outcomes.

We formulate this unit of evaluation as a *Prospective Scientific Episode* (PSE), and its action-bound commitments and constraints as an *Action-Conditioned Evidence Contract* (ACEC). The agent commits before the evidence is revealed, observes the chosen action's evidence, revises its belief, and proposes release or hold. The evaluator scores action quality, belief reports, and outcome correctness. The runtime checks observable integrity conditions and constrains release; it neither reads private reasoning nor certifies the truth of every accepted conclusion.

Our question is: *When does prospective enforcement add value beyond an agent's own checking, and what does a passed contract fail to establish?* We answer with paired interventions on frozen trajectories, reanalysis of two-stage model responses, and two real-world instantiations.

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.

We contribute a *prospective episode-level evaluation design* with a *joint estimand*: a vector of diagnostic targets linked to the same trajectory. PSE binds pre-evidence commitment and action to received evidence, reported update, and release proposal. ACEC defines admissibility; paired runtime-off/on replay measures incremental filtering and coverage cost on fixed trajectories. Endpoint-specific eligibility and denominators distinguish action information, admissibility, updating, oracle error, and selective release without reducing them to one score. Historical responses supply action and update records; the live register probes prospective timing, and the pilot tests executable tasks.

2 Related Work

Scientific agents and prospective evaluation. ScienceAgentBench and SciCode evaluate scientific programming, while AstaBench provides research tasks, tools, and controlled agent comparisons (Chen et al., 2025; Tian et al., 2024; Bragg et al., 2026). ForecastBench scores predictions of unresolved events (Karger et al., 2025); Belief-R and Reflection-Bench evaluate belief revision and epistemic agency (Wilie et al., 2024; Li et al., 2025). Our target is to align action, admissibility, update, outcome, and release assessments along the same prospective record, with explicit endpoint-specific denominators. The real-data pilot uses prescribed estimators; it is not a reproduction of either scientific coding benchmark.

Contracts and provenance. Preregistered Belief Revision Contracts specify evidence triggers, admissible revision, and externally verifiable evidence for enforcement (Alqithami, 2026). ResearchLoop represents task contracts, evidence objects, claim ledgers, and admission rules (Xia & Wang, 2026). PROV-O and RO-Crate standardize provenance and research artifacts (World Wide Web Consortium, 2013; Soiland-Reyes et al., 2022); preregistration constrains outcome-dependent redesign (Chambers & Tzavella, 2022; Vaccaro, 2026). We claim no new hashing, timestamping, preregistration, or reference-monitor primitive. ACEC supplies the binding and admission mechanism; our contribution centers on the linked evaluation and its empirical diagnostics. Gate-subset comparisons ablate our implementation, not implementations of these neighboring systems.

Information and selective release. Expected information gain (EIG) is a classical objective for Bayesian experimental design (Lindley, 1956; Chaloner & Verdinelli, 1995). We use it to score actions under supplied priors and likelihoods, not as a new acquisition rule. Selective prediction makes abstention costs explicit through risk and coverage (Geifman & El-Yaniv, 2017). We similarly distinguish contract risk from outcome error at matched coverage. Agent-security evaluations such as AgentDojo study prompt injection (Debenedetti et al., 2024); our bounded implementation probes do not establish comparable adversarial robustness.

3 Prospective Episodes and Action-Conditioned Contracts

3.1 Commit, observe, revise

Let I_0 be the initial information, $\mathcal{H}$ the hypothesis set, $\mathcal{A}$ the action menu, and B the budget. Before directional evidence is revealed, the agent submits

$$m_0 = (b_0, \widehat{y}, \kappa, a, Q_0), \qquad a \in \mathcal{A}_{\text{allowed}}, \quad c(a) \le B, \tag{1}$$

where b_0 is a belief distribution, $\widehat{y}$ a prediction, κ a declared falsifier and stopping or revision condition, and Q_0 a frozen quality rule. The implementation records κ without validating its semantic adequacy (Section 7). The chosen action produces $e = \Omega_a(a, \xi)$, with environmental uncertainty ξ. The agent returns $m_1 = (b_1, d, z)$: a revised belief, release/hold proposal, and structured reasons linking the update to evidence. Only the committed action's outcome enters the second-stage packet (Figure 1). An inadmissible observation can justify holding without being evidence against the hypothesis.

A SHA-256 digest of a canonicalized commitment binds its initial information and selected action. This detects later modification; it is not a signature and proves neither absence of prior exposure nor honest execution. A separation record discloses shared models, data, methods, operators, instruments, and timing; in this implementation it is a disclosure, not an enforced independence test.

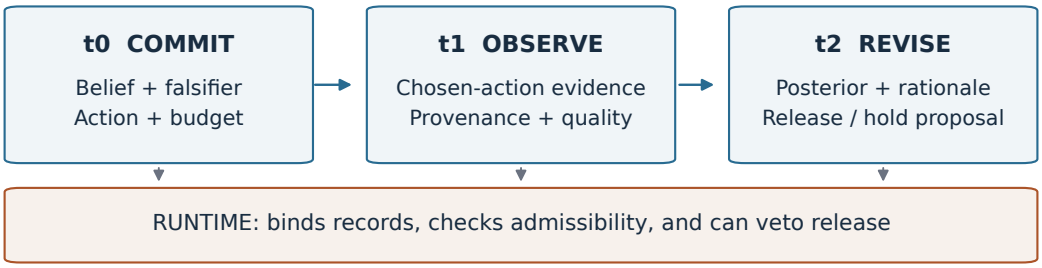

Figure 1: **Agent judgment and runtime enforcement are separate.** The first packet contains no outcome; the second contains only chosen-action evidence. The evaluator decomposes this record into action, admissibility, update, outcome, and release assessments. The runtime can veto release, but cannot certify the revised belief. Live release recommendations remain pending human authorization.

3.2 What gates enforce, and what evaluators measure

Five gate groups check temporal integrity, provenance, quality, authorization, and commitment integrity. In the synthetic evaluator, these mean access no earlier than commitment; source-declaration, hash-validity, and receipt flags; quality at or above a frozen threshold; an authorized or provisional selected action; and an intact commitment. Fixture provenance flags are trusted inputs. Commitment integrity is set true in the rule sweep and is therefore untested there. Budget compliance is measured separately. These checks do not classify a plausible but wrong conclusion as false.

Let r_i denote a proposed release and $G_i = \prod_j g_{ij}$ the conjunction of gates. For the same frozen trajectory, the disposition intervention is

$$r_i^{\text{off}} = r_i, \qquad r_i^{\text{on}} = r_i G_i. \tag{2}$$

With N trajectories, coverage is $C = N^{-1} \sum_i r_i$, contract-violating release rate is $F = N^{-1} \sum_i r_i(1 - G_i)$, and intervention rate is $V = N^{-1} \sum_i \mathbf{1}[r_i^{\text{off}} \neq r_i^{\text{on}}]$. Complete mediation under the same gate definition implies

$$F_{\text{on}} = 0, \qquad V = C_{\text{off}} - C_{\text{on}} = F_{\text{off}}. \tag{3}$$

These are operator properties, not evidence of better scientific reasoning. Empirical questions concern proposed violations, the incremental restriction of each gate group, and errors among retained releases. Holding actions, evidence, beliefs, and random draws fixed isolates disposition; it does not estimate how agents adapt to a deployed runtime.

The joint estimand is a vector of diagnostic targets on linked prospective records, each with its specified analysis population, conditioning, and denominator. "Joint" denotes this evaluation design, not a scalar score, full joint-distribution estimate, or causal decomposition of scientific ability. Action quality is EIG $I(H; O \mid a)$ under the provided prior and likelihood, constrained by budget and authorization. Reported posteriors are compared with Bayes on gate-passing evidence; preserving the prior on gate failure is the contract's chosen reference, not a universal Bayesian theorem. Oracle outcome error compares the posterior argmax with synthetic truth. For a release set S, contract risk and oracle error risk divide by $|S|$, whereas F divides by all N trajectories. Zero contract risk therefore does not imply zero outcome error risk.

3.3 Time and the trust boundary

The schema distinguishes the interval described by evidence, its generation, public availability, governed access, commitment, and reveal. The live implementation records the outcome window, commitment, planned reveal, and gateway-call time; public availability is a placeholder, and generation and actual subject delivery are not independently recorded. Only an external anchor before the outcome window supports the live prospective label. RFC 3161 timestamps (Adams et al., 2001) attest that a digest existed by a time, not that no alternative batch was run; local sealing is not public preregistration.

Guarantees assume mediated evidence access and trusted sources, timestamp authority, and operator. The gateway has unauthenticated operator endpoints, and direct store writes or out-of-band export remain outside the demonstrated boundary; a small author-run attack suite does not establish deployment security (Appendix E). Epistemic status and publication authority remain separate: the live system records recommendations while actual release requires human authorization.

4 Evaluation Design

Synthetic fixture and paired intervention. PSE-Overlap has 160 episodes: 80 clean/fault construction pairs, 16 pairs in each of five families (temporal leakage, source corruption, evidence degradation, authorization conflict, and contradictory evidence). Each offers three actions with costs, authorization states, and hypothesis-conditioned likelihoods. Seed 20260810 fixes the generator. Construction-pair members draw different priors, actions, and outcomes; they are not one-field counterfactuals. Injected faults and observables overlap: some faults pass gates, and 17/80 condition-clean episodes fail them. Hence hidden condition, gate validity, and oracle correctness define different endpoints.

We retain four rule policies: non-checking, permissive, balanced, and conservative. They vary both observable thresholds and action filtering (Appendix A); cross-policy differences are not a pure causal intervention on checking strength. Within each policy, runtime off/on uses identical actions, evidence, Bayesian updates, and seeds (17, 29, 43). The $160 \times 4 \times 2 \times 3 = 3{,}840$ records are paired traces, not independent tasks. Internal ablations compare provenance plus quality (P+Q) with the full contract, scoring both against the full gate definition.

Historical two-stage model responses. The recorded API subjects are `deepseek-v4-flash` and `deepseek-v4-pro`, at temperature 0 and a 16,000-token output limit. Stage A receives a prior, hypotheses, action likelihoods, costs, risks, authorization, and budget, but no outcome. Stage B receives only the committed action's evidence and retained commitment fields, in a fresh call without chat history. Oracle truth, hidden fault labels, and runtime verdicts are withheld. Opaque identifiers withhold direct condition labels; action menus are shuffled and assigned neutral IDs. However, Stage-A domain metadata remains perfectly confounded with clean/fault status: astronomy/genomics are clean and materials/Earth-system are fault. Rule policies ignore domain; historical responses cannot be deconfounded by deleting it afterward. Their reanalysis supplies diagnostic evidence for the evaluation decomposition, not a scientific-capability estimate.

Five output-limit failures and two abstentions on the sole episode with no feasible action leave 156 Flash and 157 Pro mapped commitments. All seven stand-in Stage-B packets are excluded from posterior and release endpoints. Action scores retain all 160 scheduled episodes, giving zero utility to missing/infeasible selections and to the empty feasible set; a 159-episode sensitivity removes only the latter. Abstention on that episode respects the budget, although the old harness lacks a supported no-action branch. These responses come from one provider and scaffold, without a same-task prompt control.

Reanalysis and uncertainty. EIG scoring, tie correction, normative-posterior and matched-coverage analyses are post hoc reanalyses of recorded trajectories. We independently re-scored local rows and reconstructed the issued packets, without new subject calls; this is local verification, not independent experimental replication. Figure 2 shows exact counts, replacing older episode-bootstrap intervals. New synthetic intervals resample 80 construction-pair clusters, retaining members, nested seeds, and paired arms (5,000 resamples for EIG; 2,000 for posterior and release metrics; seed 20260925). They describe this fixed fixture, not population risk. Live intervals resample predicates and retain repeated commitments. No clean-task non-inferiority margin was preregistered. Artifact limits are stated alongside each real-world result and in Appendix E.

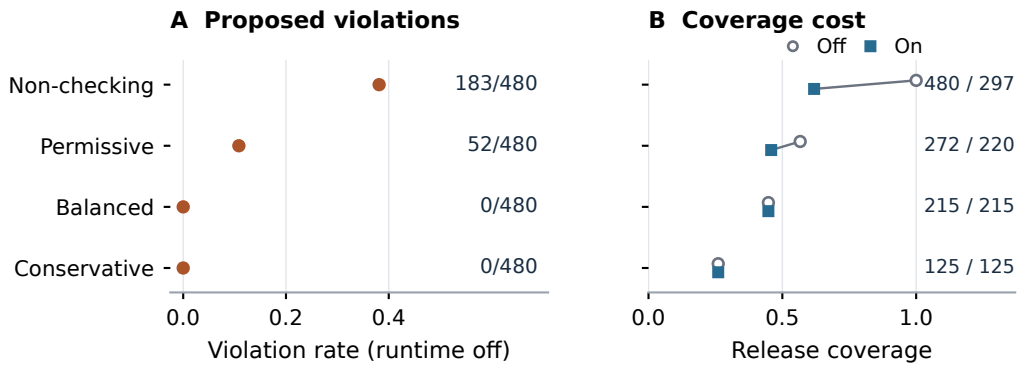

Figure 2: **Enforcement benefit and coverage cost across four rule policies.** Each policy has 480 decisions per arm (160 episodes, three nested seeds). Left: proposed violations without runtime; every full-runtime count is zero by Equation 3. Right: off/on release coverage. Exact counts and operating points are shown without inferred continuous curves. Only within-policy off/on comparisons isolate disposition enforcement.

Table 1: **Compliance and correctness at matched coverage.** Non-checking policy, 480 trajectories. Oracle error is posterior argmax disagreement with synthetic truth, not an independently verified natural-language error. Random thinning uniformly retains 297 of the 330 P+Q releases; entries are analytic expectations on this fixed pool, not another model run. Risks divide by releases.

Runtime	Releases	Violations	Oracle errors	Contract risk	Oracle risk
None	480	183	76	38.13%	15.83%
P+Q	330	33	48	10.00%	14.55%
P+Q, random thinning	297	29.7	43.2	10.00%	14.55%
Full	297	0	47	0.00%	15.82%

5 What Enforcement Adds—and What It Does Not

5.1 The marginal benefit depends on existing checking

For the non-checking policy, full enforcement prevents 183/480 contract-violating releases and lowers coverage from 1.000 to .619 (Figure 2). For permissive it prevents 52/480, reducing coverage from .567 to .458. Balanced and conservative already hold every gate-invalid trajectory: enforcement changes no decisions, with coverage .448 and .260. Action and belief-update scores are unchanged within each off/on pair. Equation 3 guarantees zero violations after enforcement. The informative quantity is how often each policy proposes a release that the contract rejects.

5.2 Additional compliance need not improve outcome correctness

P+Q removes 150 of the non-checking policy's 183 violations. The remaining gates remove another 33 releases, lowering coverage from 330/480 to 297/480 (Table 1). These are distinct comparisons: 183 measures no-runtime to full, while 33 measures P+Q to full. Among those additional 33 holds, 32 posterior argmaxes agree with the oracle and one does not. Removing temporal, provenance, quality, or authorization checks from the full contract separately reintroduces 24, 84, 54, or 9 violations; commitment removal changes none because it is unexercised.

At the same 297/480 coverage, full enforcement has zero contract risk versus 10.0% for random thinning of P+Q releases. Oracle error risk instead rises from 14.55% to 15.82%: a paired difference of +1.28 percentage points (95% pair-cluster interval [+0.35,+2.49]). Thus selective gate enforcement is better than arbitrary withholding at removing contract violations, but does not improve outcome correctness in this fixture. The difference concerns synthetic oracle labels under gates designed to test integrity rather than truth.

Table 2: **Action quality on all 160 scheduled episodes.** EIG is in bits. Max EIG is an optimization reference with disclosed distributions. The ceiling is 159/160 because one episode has no feasible action. Results are post hoc scores of the same responses.

Policy	Feasible choice	EIG-optimal	Mean EIG	Attainment
Uniform all	82.3%	35.6%	.2584	62.0%
Uniform feasible	99.4%	48.1%	.3017	77.3%
Cheapest feasible	99.4%	13.8%	.2018	55.9%
Most-expensive feasible	99.4%	99.4%	.4047	99.4%
Max EIG/cost	99.4%	64.4%	.3659	90.5%
Max EIG (reference)	99.4%	99.4%	.4047	99.4%
Flash	97.5%	83.1%	.3737	92.4%
Pro	98.1%	86.3%	.3837	94.6%

Full enforcement retains all 297 gate-passing non-checking trajectories, including all 250 that are also oracle-correct. Across both valid and invalid trajectories it retains 250/404 correct outcomes; the difference is precisely why correctness and admissibility need separate denominators. For permissive, P+Q and full release the same 220 trajectories with the same 19 oracle errors. That policy retains 201/257 jointly correct, gate-passing trajectories; its other holds originate in the proposal, not additional runtime vetoes. Added gates have no measured incremental benefit when the policy and existing checks already cover the relevant failures.

5.3 Action-information diagnostics under disclosed likelihoods

We compute $I(H;O \mid a)$ using the supplied prior and action likelihoods, and compare against six baselines (Table 2). Feasible actions are authorized or provisional and affordable. All optima within 10^{-12} receive credit. Normalized attainment divides selected EIG by that episode's maximum feasible EIG; it is not the optimal-choice hit rate. Missing/infeasible outputs receive zero, and the one empty feasible set has zero utility and no optimal hit for every policy. Random-policy entries integrate over choices exactly; deterministic ties follow displayed order.

Most-expensive feasible uses only cost and eligibility, yet matches max-EIG utility on all 160 episodes, attaining the optimum on all 159 nonempty menus. Seven action IDs differ because of EIG ties. Pairing costs and likelihood peaks by rank creates a shortcut: these records cannot distinguish probability-based selection from cost ordering or establish actual model use of the shortcut.

Flash and Pro select an EIG-optimal action on 133/160 and 138/160 episodes. Their attainments exceed uniform feasible by 15.1 points ([11.9,18.2]) and 17.3 ([14.3,19.9]), but exceed EIG/cost by only 2.0 ([$-2.0, 6.0$]) and 4.1 ([0.5,7.8]). Mean selected cost is .880 for both versus .766 for EIG/cost. These diagnostics retain the domain confound and do not establish cost efficiency; Appendix B reports the 159-episode sensitivity.

In this symmetric fixture, EIG and the former concentration score induce the same action ranking. Crediting all tied optima changes the old 127/131 model hit counts to 133/138; it does not reflect new responses or a demonstrated ranking advantage for EIG. Eleven feasible menus have tied maxima.

5.4 Reported updating and release diagnostics

On gate-passing evidence, both models approximate the normative posterior to mean absolute error about 2×10^{-5}, with the correct direction for every nonzero reference change (Table 3). This measures agreement of recorded reports with a supplied finite-model calculation within the confounded fixture, not scientific discovery or an effect of ACEC wording. Synthetic Brier sums squared errors over both hypotheses (range $[0, 2]$); live Brier below uses a single binary probability (range $[0, 1]$).

Flash changes its belief on four gate-failing episodes; Pro preserves the prior on all 56. Across all eligible episodes, prior-to-posterior Brier improvement is .1058 ([.0402,.1674]) and .1247 ([.0681,.1799]). These estimates use synthetic truth independently of gate definitions but remain conditional on model-selected actions and supplied evidence. Neither a concern keyword nor a contradiction-family label establishes whether the observed evidence semantically refutes the original belief.

Table 3: **Posterior reports, coverage, and outcome errors.** Stage B excludes missing action commitments. MAE averages absolute error over two belief coordinates and gate-passing episodes. Direction excludes 14 zero-reference-change episodes per model. No-update uses evaluator-gate failures, not only self-judged inadmissibility. Off/on keeps reports fixed.

Endpoint	Flash	Pro
Eligible / gate-passing trajectories	156 / 100	157 / 101
Gate-passing posterior MAE	1.94×10^{-5}	1.85×10^{-5}
Correct direction / nonzero reference changes	86/86	87/87
Gate-passing prior $\rightarrow$ posterior Brier	.4439 $\rightarrow$ .2947	.4440 $\rightarrow$ .2502
Oracle-correct gate-passing reports	80/100	84/101
Prior preserved on gate failure (10^{-6} tolerance)	52/56	56/56
Releases, runtime off $\rightarrow$ on	81 $\rightarrow$ 80	79 $\rightarrow$ 79
Contract-violating releases, off $\rightarrow$ on	1 $\rightarrow$ 0	0 $\rightarrow$ 0
Oracle-incorrect releases, off $\rightarrow$ on	13 $\rightarrow$ 13	10 $\rightarrow$ 10

The runtime blocks one Flash release and no Pro releases, across 313 scorable trajectories. The blocked Flash report was oracle-correct, leaving 13 oracle errors among its 80 retained releases. Full enforcement retains 67/80 and 69/84 jointly correct, gate-passing reports; the remaining holds were already proposed by the subjects. Near-zero incremental enforcement thus coexists with accurate numerical updating, incomplete coverage, and residual outcome error.

6 Real-World Instantiations

The live register tests prospective timing and evidence delivery; the pilot tests prescribed executable tasks.

6.1 A live register with future outcomes

Two batches each specify 27 predicates across discharge, geomagnetic activity, solar activity, sea ice, waves, and CO_2. Agents commit to threshold forecasts and budgeted evidence products using frozen history. The gateway delivers only the selected product, obtaining an authoritative outcome separately for scoring. Actions select products, not environmental interventions. There are up to three commitments per model and predicate; repetitions are not additional independent outcomes.

Batch A targets August 21–28, 2026 and has 79 Flash and 66 Pro commitments. Its first external anchor is 12.71 hours after window opening, so it remains *late anchored*. Batch B targets August 22–29, has 52 Flash and 81 Pro commitments, and is anchored 10.04 hours before opening. A credit failure prevented 29 scheduled Flash commitments in B; none was filled after opening. The windows overlap by six days and share definitions, including the same CO_2 target week. They are not independent replications. Each batch resolves 26/27 predicates: delayed reveal let the authoritative rolling proton product expire. Secondary alerts do not replace that product in primary scoring.

On properly anchored B, Flash scores binary Brier .2340 versus .2991 for the frozen comparator on 50 commitments; Pro scores .2546 versus .2943 on 78. Agent-minus-comparator differences are $-.0651$ (95% predicate-cluster interval $[-.1368, .0022]$) and $-.0397$ ($[-.1049, .0238]$). Both intervals include zero. Always predicting no scores .2200 and .2308 on the respective resolved sets, outperforming both subjects. Accordingly the register does not establish forecasting superiority. Comparator definitions, A's results, and time-field limitations are in Appendix C; paired Brier emphasis and authoritative-only scoring were chosen after reveal.

All 11 unresolved evidence packets produce a hold without updating within .01. Among 267 gate-passing packets, 248 recommend release (132/139 in A; 116/128 in B); all 278 actual publication states remain pending human authorization. Live validity checks source admissibility, coverage, and quality, not every synthetic gate. Small posterior losses on resolved packets primarily measure uptake of a product that states the answer. Local audit verifies stored commitments, packet hashes, and records, but retained source extracts and receipts do not contain complete HTTP response bodies; source parsing cannot be fully replayed.

6.2 A FROZEN EXECUTABLE REAL-DATA PILOT

A pilot uses eight prescribed tasks on penguin morphology, wine chemistry, airfoil acoustics, and atmospheric CO_2 (Gorman et al., 2014; Aeberhard & Forina, 1992; Brooks et al., 1989; Lan & Keeling, 2026). It compares a *shared-interface control* with *explicit ACEC framing*: both have structured plans, failure criteria, confidence, two-stage access, and an output schema. The contrast is additional framing. Requests specify `gpt-6-sol`, medium reasoning, through native Codex CLI; the same model family assisted harness development, limiting author independence.

Both arms meet the specified estimators and propose release on 8/8 tasks; held-out metrics agree at reported precision. We verified the recovered 91-file freeze and 16 two-stage chains, and replayed all 16 programs in the original sandbox, matching every archived output. Independent implementations confirm all eight references. The execution-validity filter changes no releases; this is not an interactive five-gate 2×2 experiment. Confidence predicts implementation success. The ceiling and zero difference establish neither capability improvement nor equivalence. Appendix D gives complete results and replay boundaries.

7 EVALUATION VALIDITY AND LIMITATIONS

The strongest controlled evidence concerns a disposition mechanism on one synthetic generator. Disclosure of likelihoods makes action and posterior endpoints tractable, but does not test acquiring a scientific model from ambiguous evidence. Gates inherit trusted fields and thresholds, miss observationally overlapping faults, and do not assess scientific relevance or latent common causes. There is no clean-task non-inferiority result, human baseline, external red team, or autonomous laboratory evaluation.

Falsifier semantics remain unvalidated. In the full-160 synthetic evaluation, the falsifier is carried into Stage B as text; its scorer checks only for at least three whitespace-separated tokens. We did not validate scientific meaning, testability with the chosen action, or consistent application to revision and stopping. Falsifier meaning enters neither eligibility nor runtime gates. Recording a falsifier therefore does not establish that a falsifiable hypothesis was tested.

Evaluator defects materially affect interpretation. Earlier runs leaked condition labels or unselected outcomes, favored native action position, and conflated concerns, revisions, and releases. Those runs do not support the retained historical-model diagnostics. The issued-menu reconstruction repairs a documented salt mismatch but does not remove the domain confound or recreate hosted-model responses. EIG, normative-posterior, and matched-coverage endpoints are disclosed post hoc analyses. The same-task synthetic prompt control remains missing; the separate eight-task real-data framing comparison does not supply it. Appendix E records these corrections without treating archive recovery as independent replication.

Live evidence has one properly anchored weekly batch, unequal completion, correlated predicates, delayed retrieval, and incomplete source and timing records. The pilot has prescribed estimators, eight tasks from four historical datasets, an all-success ceiling, and local rather than externally anchored freezing. Offline program replay verifies archived computations; it does not repeat model sampling or establish external replication. Neither instance supports general forecasting skill, calibration, or autonomous scientific capability. Complete mediation is an assumption beyond the evaluated boundary; actual publication authority remains separate from a passed contract.

8 CONCLUSION

Prospective episodes place action, evidence, update, and release decisions on one record. In this fixture, extra gates hold 33 more releases than provenance-and-quality checks, including 32 with oracle-correct posterior argmaxes; choosing the most expensive feasible action also attains maximum EIG on every nonempty menu. The first result separates admissibility from correctness; the second exposes a cost shortcut in action scoring.

AI USE STATEMENT

Language models served as evaluation subjects. AI assistants also supported protocol and sandbox development, pilot task and reference-program construction, analysis code, results interpretation, and manuscript drafting. AI assistance also included literature search and reference verification. The synthetic and live API subjects are identified in Section 4; the real-data pilot's recorded requested CLI model and author-assistance relationship are stated in Section 6.2. Earlier non-blind development results involving an Anthropic-family subject sharing a model family with the harness-authoring assistant are excluded. OpenAI Codex assisted this manuscript rewrite and offline reanalysis of existing records; no new model-subject experiment was run for the rewrite. Model-generated suggestions were not empirical observations. The human authors retain responsibility for claims, artifacts, and disclosures.

ETHICS STATEMENT

The evaluations use synthetic episodes and public environmental and scientific data, without human-subject experiments or autonomous instrument control. Authorization fields express declared policy, not substantive biosafety, legal, or institutional approval. Gate compliance does not establish safety for unrestricted scientific action. Implementation probes concern a limited research threat model.

REPRODUCIBILITY STATEMENT

The accompanying anonymous evaluation archive runs offline with Python 3.9+ and its standard library. Its entry point, `reproduce.py`, re-scores stored trajectories and raw responses for Tables 1–3, including the main bootstrap intervals and cost-only baseline; archived pilot programs, inputs, labels, references, and execution outputs support rescoring of Appendix D. The README maps outputs to paper tables, and per-file provenance distinguishes original hashes from anonymized publication copies. The default run makes no model calls and does not execute pilot programs; their original macOS Seatbelt execution is described separately. Complete live HTTP bodies were not archived, so live source retrieval cannot be exactly replayed from this archive. The supplementary ZIP also includes the manuscript source.

REFERENCES

Carlisle Adams, Pat Cain, Denis Pinkas, and Robert Zuccherato. Internet X.509 public key infrastructure time-stamp protocol (TSP). Technical Report RFC 3161, IETF, 2001.

S. Aeberhard and M. Forina. Wine. UCI Machine Learning Repository, 1992. URL https://archive.ics.uci.edu/dataset/109/wine. Dataset.

Saad Alqithami. Preregistered belief revision contracts. arXiv:2604.15558 [cs.AI], 2026. URL https://arxiv.org/abs/2604.15558.

Jonathan Bragg, Mike D'Arcy, Nishant Balepur, Dan Bareket, Bhavana Dalvi, Sergey Feldman, et al. AstaBench: Rigorous benchmarking of AI agents with a scientific research suite. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.21652.

T. Brooks, D. Pope, and M. Marcolini. Airfoil self-noise. UCI Machine Learning Repository, 1989. URL https://archive.ics.uci.edu/dataset/291/airfoil+self+noise. Dataset.

Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304, 1995. doi: 10.1214/ss/1177009939. URL https://doi.org/10.1214/ss/1177009939.

Christopher D. Chambers and Loukia Tzavella. The past, present and future of registered reports. *Nature Human Behaviour*, 6:29–42, 2022. doi: 10.1038/s41562-021-01193-7.

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.

Edoardo Debenedetti, Jie Zhang, Mislav Balunović, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. In *Advances in Neural Information Processing Systems 37 (Datasets and Benchmarks Track)*, 2024.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, 2017. URL https://papers.neurips.cc/paper_files/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Kristen B. Gorman, Tony D. Williams, and William R. Fraser. Ecological sexual dimorphism and environmental variability within a community of Antarctic penguins (genus Pygoscelis). *PLOS ONE*, 9(3):e90081, 2014. doi: 10.1371/journal.pone.0090081. URL https://doi.org/10.1371/journal.pone.0090081.

Ezra Karger, Houtan Bastani, Chen Yueh-Han, Zachary Jacobs, Danny Halawi, Fred Zhang, and Philip E. Tetlock. ForecastBench: A dynamic benchmark of AI forecasting capabilities. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2409.19839.

Xin Lan and Ralph Keeling. Trends in atmospheric carbon dioxide: Mauna Loa monthly means. NOAA Global Monitoring Laboratory and Scripps Institution of Oceanography, 2026. URL https://gml.noaa.gov/ccgg/trends/data.html. Dataset; monthly means.

Lingyu Li, Yixu Wang, Haiquan Zhao, Shuqi Kong, Yan Teng, Chunbo Li, and Yingchun Wang. Reflection-Bench: Evaluating epistemic agency in large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 36236–36264. PMLR, 2025. URL https://proceedings.mlr.press/v267/li25cu.html.

D. V. Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956. doi: 10.1214/aoms/1177728069. URL https://doi.org/10.1214/aoms/1177728069.

Stian Soiland-Reyes, Peter Sefton, Mercè Crosas, Leyla Jael Castro, Frederik Coppens, José M. Fernández, Daniel Garijo, Björn Grüning, Marco La Rosa, Simone Leo, et al. Packaging research artefacts with RO-Crate. *Data Science*, 5(2):97–138, 2022. doi: 10.3233/DS-210053. URL https://doi.org/10.3233/DS-210053.

Minyang Tian, Luyu Gao, Shizhuo Dylan Zhang, Xinan Chen, Cunwei Fan, Xuefei Guo, Roland Haas, Pan Ji, Kittithat Krongchon, Yao Li, Shengyan Liu, Di Luo, Yutao Ma, Hao Tong, Kha Trinh, Chenyu Tian, Zihan Wang, Bohao Wu, Yanyu Xiong, Shengzhu Yin, Minhui Zhu, Kilian Lieret, Yanxin Lu, Genglin Liu, Yufeng Du, Tianhua Tao, Ofir Press, Jamie Callan, Eliu Huerta, and Hao Peng. SciCode: A research coding benchmark curated by scientists. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0963. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/36850592258c8c41cecdaa3dea5ff7de-Abstract-Datasets_and_Benchmarks_Track.html. Datasets and Benchmarks Track.

Michelle Vaccaro. Preregistration for experiments with AI agents, 2026. URL https://arxiv.org/abs/2606.11217. arXiv:2606.11217.

Bryan Wilie, Samuel Cahyawijaya, Etsuko Ishii, Junxian He, and Pascale Fung. Belief revision: The adaptability of large language models reasoning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 10480–10496. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.586. URL https://aclanthology.org/2024.emnlp-main.586/.

World Wide Web Consortium. PROV-O: The PROV ontology. W3c recommendation, World Wide Web Consortium, 2013. URL https://www.w3.org/TR/prov-o/. Edited by Timothy Lebo, Satya Sahoo, and Deborah McGuinness.

Yihan Xia and Taotao Wang. ResearchLoop: An evidence-gated control plane for AI-assisted research. *arXiv preprint arXiv:2605.28282*, 2026. URL https://arxiv.org/abs/2605.28282.

A Synthetic Specification and Endpoint Details

Generator and observables. Seed 20260810 generates 80 construction pairs with two hypotheses and three actions per episode. Members share a nominal family, not identical covariates or outcomes. The prior on H_1 is drawn from $[.35, .65]$, costs from $[.35, 1.45]$, budget from $[.9, 1.6]$, likelihood peaks s_a from $[.55, .90]$, and risk from $[.01, .14]$. Costs and peaks are separately sorted in descending order, then paired by rank. For outcomes (O_1, O_2, O_0) the likelihood is (s_a, u_a, u_a) under H_1 and (u_a, s_a, u_a) under H_2, with $u_a = (1 - s_a)/2$. The realized prior range is .354–.643. All 480 channels are symmetric; the legacy concentration and EIG rankings coincide on all 160 menus.

Contradiction-family faults set truth opposite the prior favorite; this occurs with probability .25 in that family's clean cases and .15 in other families. Outcomes still sample from the specified likelihood, so a family label alone does not prove that the selected observation contradicts the prior. Quality has mean .71 for degradation faults and .86 otherwise, standard deviation .07, clipped to $[.01, .999]$, with gate .75. Source faults expose a failed flag with probability .85 versus .12 otherwise; failed hash, receipt, and source declaration are drawn with probabilities .45, .35, and .20. Temporal faults precede commitment with probability .70 (20–900 seconds early); otherwise they are near-misses (1–120 seconds after). Legitimate near-misses occur with probability .15; ordinary lead is 180–3,600 seconds. Authorization faults deny or expire the most diagnostic action. These rules create observable overlap, including 17/80 gate-invalid clean episodes.

Rule policies. Let ℓ be access lead, q quality, and d the posterior drop of the prior favorite. Non-checking never flags. Permissive flags $\ell < -60$, invalid hash, $q < .70$, denied authorization, or $d \geq .20$. Balanced flags $\ell < 0$, any failed provenance flag, $q < .75$, expired/denied or missing authorization, or $d \geq .15$. Conservative flags $\ell < 180$, any failed provenance flag, $q < .80$, a status other than authorized, or $d \geq .10$. Any flag proposes a hold. All policies update by Bayes; contradiction is not a runtime gate. They rank affordable actions by $s_a - .25c(a) - .15\,\mathrm{risk}(a)$, choosing the top with probability .90 and a lower rank uniformly otherwise. Checking policies first filter denied/expired actions; non-checking does not. If no permitted action remains, they fall back to affordable actions; if none exist, no action is selected and authorization fails. Seeds 17, 29, and 43 are nested within the fixed episodes, not independent task samples.

Information gain and references. With $p_a(o) = \sum_h b_0(h)p(o \mid h, a)$, we compute

$$I(H; O \mid a) = \sum_{h,o} b_0(h)p(o \mid h, a) \log_2 \frac{p(o \mid h, a)}{p_a(o)}, \qquad U_i = \frac{I(H; O \mid a_i)}{\max_{a \in \mathcal{F}_i} I(H; O \mid a)}, \tag{4}$$

where $\mathcal{F}_i$ is the affordable authorized/provisional set. Missing or infeasible selections and empty $\mathcal{F}_i$ receive $U_i = 0$. Uniform-all samples all three actions; uniform-feasible samples $\mathcal{F}_i$; cheapest and most-expensive minimize/maximize cost without likelihoods; EIG/cost maximizes the ratio; max-EIG is the disclosed-distribution optimum. Uniform policies are exact expectations; deterministic ties use displayed order. EIG ties within 10^{-12} all receive hit credit. The numbers of menus with 0, 1, 2, and 3 feasible actions are 1, 17, 48, and 94; ten eligible menus have two optimal actions and one has three.

For gate-passing evidence, the reference is $b^*(h) = b_0(h)p(o \mid h, a)/p_a(o)$. On gate failure it is $b^* = b_0$ by contract convention. Posterior MAE averages $|b_1(h) - b^*(h)|$ over coordinates, then episodes. Direction compares the prior favorite's signed change when $|b^*(h) - b_0(h)| > 10^{-8}$; 14 gate-passing zero-change episodes per subject remain in MAE and outcome scores but not direction. Synthetic Brier is $\sum_h (b(h) - \mathbf{1}[h = H])^2$. Oracle error compares posterior argmax with stored truth. Neither metric evaluates free-text scientific claims.

B Model Denominators and Reconstruction

Eligibility and sensitivity. Stage-A JSON extraction succeeds on 157 Flash and 158 Pro outputs; three and two hit the output limit. Each subject also abstains on the same episode with budget 1.02 and costs 1.42, 1.30, 1.23. This leaves 156/157 mapped commitments and excludes 4/3 Stage-B stand-ins. Extraction decodes the first brace-balanced JSON object after optional code-fence removal; it is not schema validation. Every mapped commitment satisfies authorization and budget.

Table 4: **Sensitivity restricted to 159 episodes with a feasible action.** Missing outputs on eligible episodes remain failures. Attainment and regret use per-episode feasible EIG normalization.

Policy	EIG-optimal	Attainment	Normalized regret
Uniform all	35.8%	62.4%	37.6%
Uniform feasible	48.4%	77.8%	22.2%
Cheapest feasible	13.8%	56.2%	43.8%
Most-expensive feasible	100.0%	100.0%	0.0%
Max EIG/cost	64.8%	91.0%	9.0%
Max EIG	100.0%	100.0%	0.0%
Flash	83.6%	93.0%	7.0%
Pro	86.8%	95.2%	4.8%

Posterior eligibility. Maximum per-episode MAE is .00021049 (Flash) and .00010615 (Pro). For prior preservation, self-judged inadmissibility gives Flash 52/55, whereas all evaluator-gate failures give 52/56 in Table 3. The 10/12 and 11/12 contradiction-family scores measure a .15 probability drop rather than semantic refutation by selected evidence; the main revision analysis therefore uses normative posterior comparisons.

Exact packet reconstruction. The archived, unmodified two-stage builder with salt `full160` regenerates all 160 shuffled maps, 160 Stage-A packets, and 320 subject-specific Stage-B packets byte-for-byte in a scratch directory. The previously documented `full160-v1` yields zero byte matches in each category and is retained only as a negative control. The original command line was not recovered; reproducing its outputs restores the parameter provenance but does not recreate model sampling. Input-file hashes were unchanged after local checks. The exact domain association remains: astronomy/genomics each have 40 clean episodes, materials/Earth-system each have 40 fault episodes. Stage B omits domain but retains the action and falsifier; reconstruction does not remove information already available in Stage A.

Intervals and matched coverage. Pair-cluster bootstrap samples 80 construction-pair IDs with replacement, retaining members, seeds, and compared arms. EIG uses 5,000 draws; posterior and release endpoints use 2,000, with seed 20260925 and sorted IDs. We use interpolated percentile limits: matched oracle-risk and Flash Brier-improvement upper bounds are 2.49 percentage points and .1674. Recovered posterior/release scripts use the same sampling order but discrete limits, giving 2.50 and .1675 instead. The original EIG script interpolates; its Pro-minus-uniform lower bound is 14.3477 percentage points, which rounds directly to 14.3 at one decimal place (V3.1 displayed 14.4). Random thinning from 330 P+Q releases to 297 gives expected violations $33(.9) = 29.7$ and oracle errors $48(.9) = 43.2$. This analytic comparator adds no trials or continuous risk–coverage curve.

C Live Register: Products, Scoring, and Artifacts

Products and comparators. The budget and reference-product cost are two. Alternatives generally cost one and provide provisional or nonresolving products; NSIDC offers a monthly product (one) and no collection (zero). Menus shuffle deterministically per episode, subject, and sample. Frozen-comparator probabilities and authoritative scoring outcomes are hidden from subject packets. Discharge uses nominal 61-day history ending three days before opening; Kp, flares, and waves use approximately 90, 30, and 45 days. Frequency comparators use overlapping seven-record windows, not necessarily seven consecutive dates. NSIDC uses same-span declines in 2000–2025, CO_2 the last 12 annual increments, and protons an external third-day forecast. The flare history builder uses M counts, but both frozen histories contain zero X counts.

Sampling, time, and missingness. Live temperature is .3 with three scheduled samples per predicate. Commitment limits are generally 16,000 tokens in A (one completed Flash response retained from a 4,000-token pass) and 40,000 in B; Stage B uses 40,000. B's missing Flash commitments are all third samples plus the second CO_2 samples. Its 58 failed attempts concern 29 missing episode/sample pairs, not 58 independent episodes. Scores weight commitments equally;

Table 5: **Live predicates and reference acceptance rules.** Counts total 27 per batch. Acceptance does not imply complete observation or final measurements.

Source	n	Predicate	Reference and acceptance
USGS	8	Max daily discharge reaches historical 80th percentile.	NWIS daily means (00060), ≥ 6 days; provisional allowed.
GFZ Kp	2	Max three-hourly Kp ≥ 5 or ≥ 6.	GFZ, ≥ 50 intervals; nowcast or definitive.
SWPC flares	2	M-plus-X total ≥ 1 or ≥ 3.	Daily Solar Data, all seven days.
SWPC protons	1	Any five-minute ≥ 10 MeV flux ≥ 10 pfu.	Rolling seven-day product, ≥ 200 target-day points.
NSIDC	2	Extent below historical-decline thresholds.	Sea Ice Index v4 target daily row; missing area $\leq .05$.
NDBC	10	Max wave height reaches historical 80th/90th percentile.	Realtime2 WVHT, ≥ 6 days.
NOAA GML	2	Weekly annual increment < 2.0 or < 1.5 ppm.	Weekly CSV, annual comparator and ≥ 3 contributing days.

Table 6: **Live forecast scores rechecked from recorded outcomes.** n is resolved predicates / scored commitments; Δ is agent minus frozen-comparator binary Brier. Intervals resample predicates. Always-no is a descriptive baseline. A is late anchored.

Batch	Subject	n	Agent	Comparator	Always-no	Δ [95% interval]
A	Flash	26/76	.2545	.2617	.2632	$-.0072$ $[-.0754, .0551]$
A	Pro	24/63	.2558	.2628	.2857	$-.0070$ $[-.0835, .0650]$
B	Flash	26/50	.2340	.2991	.2200	$-.0651$ $[-.1368, .0022]$
B	Pro	26/78	.2546	.2943	.2308	$-.0397$ $[-.1049, .0238]$

2,000 predicate-cluster bootstrap draws (seed 7) retain all repetitions, without removing dependence among related predicates.

A's first external anchor is August 21 at 12:42:42 UTC; B's is August 21 at 13:57:48 UTC. The proton windows are one day, and both batches' CO_2 windows are August 23–30. Twenty-five predicates per batch were revealed 4–13 days late. CO_2 was retrieved September 6 at 07:36–07:37 UTC, after its 00:00 planned gate. The recorded reveal time denotes that gate, agent-availability time denotes gateway-call start, public time is `source-defined`, and event time denotes the outcome window. These are not independently verified six-time records.

What is replayable. Local checks reproduce scores, commitment and packet digests, source-extract hashes, log chains, and anchor-file bindings from retained records. They do not newly authenticate TSA certificate chains. All 476 inspected receipt instances mark complete raw source bodies unarchived: extract hashes can be recomputed, raw-body hashes cannot. Offline scoring therefore checks stored outcomes, not original source parsing or recovery of expired products.

D Frozen Executable Real-Data Pilot

Evidence status and contrast. The original study archive contains requests, raw CLI events, programs, inputs, execution receipts, references, and scoring code. Its 329 distribution hashes and 91 frozen-file hashes match. We checked 16 Stage-A–execution–Stage-B bindings, rebuilt 19 task/input/label files byte-for-byte from four archived datasets, and independently implemented eight reference estimators. Sandboxed offline reexecution reproduces all 16 archived result objects exactly; rescoring reproduces the complete summary. This verifies existing records without new model sampling or external replication. The shared interface is described in Section 6.2; ACEC additionally emphasizes prospective commitment, ordered evidence checks, and release conditions. Each task has one chain per arm, with hash-fixed arm order.

Wine and Airfoil use SHA-256 row-ID splits independent of response values; Penguins and CO_2 use chronological splits. Airfoil rows are not separated by experimental configuration. C01 is

Table 7: **All eight prescribed tasks (verified frozen protocol).** Four datasets support eight tasks; rows are not eight independent scientific domains. P01, W01, and C02 have numerical-reference checks but no held-out prediction metric.

ID	Specified estimator and required output	Data and split
P01	OLS body mass on flipper length, sex, species; male coefficient, classical SE, normal 95% interval.	Penguins: 344 observations, 333 complete cases.
P02	OLS body-mass prediction from morphology, sex, species.	216 birds (2007–08) train; 117 (2009) test.
W01	Standardized PCA of 13 chemical variables; first three variance ratios, absolute PC1 loadings.	Wine: 178 assays.
W02	Nearest-centroid cultivar classifier with training-only standardization.	Wine: 151 train; 27 test.
A01	Sound-pressure OLS, prescribed log physical inputs and raw angle.	Airfoil: 1,244 train; 259 test.
A02	Choose raw/log OLS by pooled five-fold out-of-fold RMSE, then refit.	Same Airfoil split; folds from row IDs.
C01	Linear trend plus annual sine/cosine.	CO_2: 180 months (2000–14) train; 60 (2015–19) test.
C02	Difference in seasonally adjusted linear growth rates between 1980–99 and 2000–19.	CO_2: 480 historical months.

retrospective held-out prediction, not live forecasting. Withheld labels on public historical data do not exclude prior model exposure. This is not a ScienceAgentBench or SciCode replication.

Execution and hidden evaluation. The frozen environment records Codex CLI 0.155.0-alpha.9.2, Python 3.12.14, NumPy 2.3.5, and pandas 2.2.3. Requests specify `gpt-6-sol` and medium reasoning; returned model identifiers are null, so provider weights are not independently identified. Temperature and native context were not independently controlled; raw-API equivalence is not assumed. The author-assistance relationship is disclosed in Section 6.2.

In the original collection, each Stage-A program executes once under a macOS Seatbelt filesystem/network sandbox, with 30-second CPU and 35-second wall limits, Python/NumPy/pandas and allowed standard libraries. Network, shell, installation, private reference files, and other trial records are forbidden. The archived chains contain no program repair or replacement. Stage B receives the commitment and execution receipt, including bounded output, code/input digests, process status, input integrity, result presence, and its own result summary. Hidden labels, reference values, held-out scores, and evaluator pass/fail are not revealed; the program cannot be replaced. The later offline audit uses the same execution boundary without model calls.

Success requires zero exit status, a JSON object, unchanged inputs, exact required keys and row order, finite outputs, and numerical agreement with the specified estimator under `math.isclose` with both absolute and relative tolerance 10^{-5}. The replayed release filter only tests completion, zero exit status, JSON-object output, and unchanged inputs; it does not inspect hidden numerical correctness or every required field. Thus execution-filtered releases are not a full five-gate runtime experiment.

Results and collection accounting. Both arms have 8/8 valid programs, executions, estimator successes, proposed releases, and filtered releases. Averaging $(q_i - S_i)^2$ using raw confidence reports and verified $S_i = 1$ gives Control .0001000 before and .0000625 after; ACEC .0002500 and .0002125. This ceiling sample does not establish calibration. Table 9 preserves separate task scales; A02 differs between arms only by 3×10^{-15} RMSE. Coefficients, loadings, slopes, and predictions were all checked against prescribed estimators, without establishing new scientific findings.

The archive has 16 responses per stage, no parse failures or recorded subject-tool violations, 128 reconnect notices, and HTTPS fallback on all 32 requests; server-side attempt counts are unknown. Each planned study ID has one recorded start and terminal response, with no recorded collector-level answer repair or replacement; this does not prove unique backend sampling. A startup failure produced no captured study answer and was misclassified as a tool violation. After three nonstudy

Table 8: **Complete per-task outcomes and confidence, locally rechecked.** Both arms have verified estimator success and recorded release on every row; the execution filter changes none. q is read from raw responses and predicts implementation success, not a scientific hypothesis.

	Success / decision		Control q		ACEC q	
Task	Control	ACEC	Before	After	Before	After
P01	Pass / release	Pass / release	.990	.995	.990	.990
P02	Pass / release	Pass / release	.990	.990	.980	.990
W01	Pass / release	Pass / release	.990	.995	.990	.990
W02	Pass / release	Pass / release	.990	.995	.980	.980
A01	Pass / release	Pass / release	.990	.990	.990	.990
A02	Pass / release	Pass / release	.990	.990	.980	.980
C01	Pass / release	Pass / release	.990	.995	.980	.980
C02	Pass / release	Pass / release	.990	.990	.990	.990

Table 9: **All held-out metrics, recomputed from program outputs and hidden labels.** Both arms implement the same prescribed task. The training-mean comparator applies to regression only.

Task	Metric	Test n	Control	ACEC	Training mean
P02	RMSE (g)	117	280.5040	280.5040	822.7593
W02	Accuracy	27	.9259	.9259	—
A01	RMSE (dB)	259	5.1711	5.1711	6.7633
A02	RMSE (dB)	259	4.6096	4.6096	6.7633
C01	RMSE (ppm)	60	2.1665	2.1665	22.9266

connection probes, the eight-task protocol was refrozen before valid answers; 80 prepared synthetic tasks remained uncollected. Local freezing is not an external anchor or public preregistration.

E Corrections and Verification Boundaries

Historical endpoint corrections. The retained non-checking summary records 108 releases among 240 fault-conditioned trajectories. Of these, 33 meet the fixed revision-drop threshold and 36 select a permitted alternative to authorization bait. The residual 39 comprises 18 quality, 15 temporal, three source-overlap cases, and three below-threshold revisions. It is a family-conditioned behavioral residual, not a scientific-error rate or the oracle error endpoint in Table 1; the family label does not certify semantic contradiction. It is therefore excluded from headline claims. Raw records and original aggregate values remain unchanged.

Leaked IDs or supplement filenames, packets exposing unselected outcomes, operator-transcribed or non-blind outputs, and fixed-order menus do not support the retained model diagnostics. Scoring changed from condition-clean false alarms to gate-relative denominators, keywords to belief movement, and action ranking to feasible, tie-aware credit. These were post-inspection corrections; they did not repair late anchoring, delayed reveal, or domain confounding.

Attack endpoints. The 16 author-designed scripts test commitment absence, mutation, duplication and timing; policy changes; evidence substitution; log deletion; replay; and direct writes or export. Fourteen meet the script-defined blocking endpoint; this includes log deletion, which occurs but is subsequently detected. Direct in-process writes and assumed out-of-band export succeed. A raw 16/16 "detectable" counter counted an ordinary failed-gate hold as detecting export; excluding that case yields 15/16 attack-effect visibility. These are different definitions, not three interchangeable success rates. The tampering test modifies evidence before a decision, not during a concurrent race; the source-assertion test does not execute prompt injection. A replayed hold does not independently test monotonic disposition. The simulated time authority accepts a caller-supplied clock, provenance flags are trusted, and the live runner invokes gateway code in process. Separate gateway probes demonstrate absence of tested subject-side routes only. No malicious-operator or general prompt-injection defense is established.