SCORES ALONE DO NOT PROVE DISCOVERY: THE DISCOVERY CERTIFICATION PROTOCOL FOR AUDITING AI RESEARCH AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

AI research agents combine public information and experimental feedback to produce measurable results. The Discovery Certification Protocol (DCP) turns an outcome claim into an executable audit under a registered model, information boundary, and budget. Gate 1 validates useful improvement. Gate 2 tests recovery by matched agents given the starting information and observed Web content, with run history and new measurements withheld. Core requires adequate registered controls, zero recoveries, and a finite-sample recovery bound. Optional Gate 3 compares truthful and neutral feedback from a shared checkpoint; Evidence adds a supported effect and a null-policy equivalence check. Controlled SQLite and virtual catalyst audits pass both decision kernels. On real-data response surfaces, Yacht and Ionosphere pass the Core kernel after zero recoveries in 96 attempts, with an upper bound of 0.0468. Each target combines ten observed utilities and six predictions into a 16-entry data product. Yacht scores 0.7677 on reconstruction of all 32 switch effects, with utility-prediction MAE 0.0315 on its six unmeasured configurations. Fresh truthful continuations recover the target level in $9/30$ and $16/30$ trials, respectively, separating achieved utility from process repeatability. A deterministic verifier reproduces these local decisions from frozen records.

1 INTRODUCTION

AI research agents choose actions, run experiments, inspect measurements, and revise their proposals. Their outputs include programs, models, data products, and experimental recipes. Recent systems have found useful algorithms and mathematical constructions and have begun to automate broader research workflows (Fawzi et al., 2022; Mankowitz et al., 2023; Romera-Paredes et al., 2024; Novikov et al., 2025; Lu et al., 2026). As these systems become research collaborators, their reports need evidence connecting measured results to the research that produced them.

Consider an agent that runs 87 experiments and reports a program with score x. A sealed test establishes the program's utility. A matched agent may also reach x using initial data, public material, and its own reasoning. Separately, truthful experimental observations may improve the chance of reaching x. These are three measurable questions about one outcome. Implementation and baseline choices shape measured gains (Melis et al., 2018; Musgrave et al., 2020). In automated research, independent implementations of a fixed idea can change its ranking (Ning et al., 2026e). DCP extends this measurement discipline to the information and feedback available to a research agent.

The central design choice is to admit every valid route to the same numerical outcome. A challenger may combine familiar components, transfer a technique across domains, or produce a different implementation. It receives the registered background, initial observations, tools, budget, and captured Web content. Run history and new scientific measurements are withheld. An executable validity check and a score threshold determine recovery, making the rule portable across research domains.

Recovery witnesses and recovery probabilities play complementary roles. A witness establishes an available route within the registered scope and closes the audit as recovered. A positive Core decision requires an adequate audit with zero recoveries and a bound on a fresh episode's recovery probability.

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

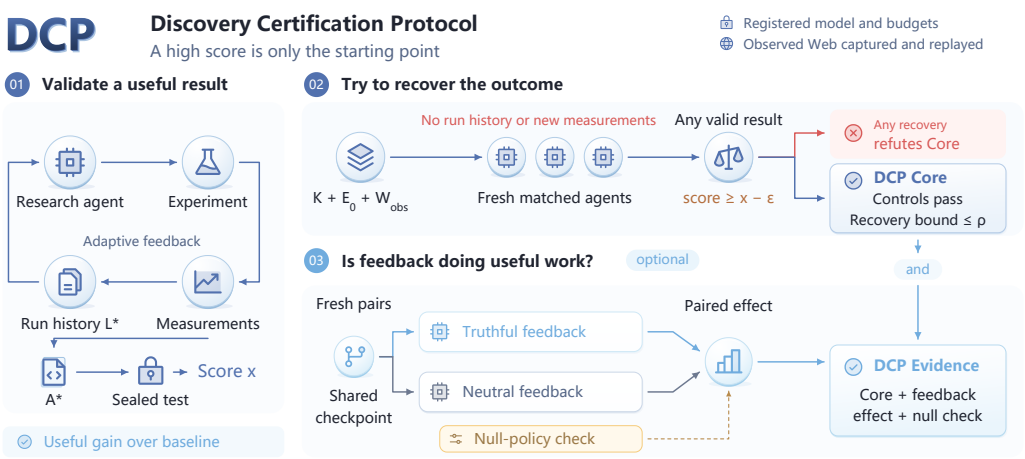

Figure 1: DCP asks three questions about one numerical outcome. Gate 1 validates result A^* on sealed data. Gate 2 gives fresh agents K, E_0, W_{obs} with run history and new measurements withheld. A valid score $\geq x - \varepsilon$ refutes Core; a Core pass requires adequate controls, zero recoveries, and a recovery bound below ρ. Gate 3 adds Evidence through a paired feedback effect and a null-policy equivalence check.

Evidence adds a randomized comparison of truthful and neutral feedback from a shared checkpoint. This separation gives a successful target run, a recoverable outcome, and a beneficial feedback policy distinct evidential meanings.

The paper makes three contributions.

- We define a general evidence interface for measurable AI research outputs. Its numerical recovery rule accepts alternative methods under a registered information and resource boundary.
- We combine qualified recovery witnesses, finite-sample recovery bounds, and randomized feedback effects in two explicit decisions, Core and Evidence. Their definitions separate certificate eligibility from statistical and causal conclusions.
- We demonstrate local Core and Evidence decisions in controlled workflows and two real-data Core passes with Evidence Inconclusive for distinct reasons. Recovery witnesses and incomplete records exercise the remaining decision paths. Portable bundles support deterministic replay.

2 Related Work

AI research agents and their evaluation. AlphaTensor, AlphaDev, FunSearch, and AlphaEvolve combine search with objective evaluation (Fawzi et al., 2022; Mankowitz et al., 2023; Romera-Paredes et al., 2024; Novikov et al., 2025). The AI Scientist family and AI co-scientist automate broader research and hypothesis development (Lu et al., 2026; Yamada et al., 2025; Gottweis et al., 2026). A specialist-agent ablation removes prior-trial history while retaining current-best code and score (Ning et al., 2026a). Molecular and materials agents freeze selected interventions for held-out evaluation (Ning et al., 2026b;c). MLE-bench, MLGym, MLR-Bench, and PaperBench evaluate engineering, open-ended research, and replication (Chan et al., 2025; Nathani et al., 2025; Chen et al., 2025a; Starace et al., 2025). DiscoveryBench and ScienceAgentBench assess data-driven scientific workflows and their outputs (Majumder et al., 2025; Chen et al., 2025b). DiscoveryWorld measures task success, scientific actions, and explanatory knowledge (Jansen et al., 2024). FIRE-Bench evaluates rediscovery of published scientific insights (Wang et al., 2026), while Bhushan et al. (2026) distinguish psychological novelty, historical novelty, and usefulness. DCP adds outcome recovery and feedback interventions to these evaluations.

Iterative feedback and experimental design. ReAct, Tree of Thoughts, Reflexion, and Self-Refine develop tool use, search, reflection, and feedback (Yao et al., 2023b;a; Shinn et al., 2023; Madaan et al., 2023). Controlled revision studies separate additional solving, prompt structure, and draft content (Ning et al., 2026d). SkillLearnBench compares execution-derived and teacher-guided skill refinement (Zhong et al., 2026). DCP tests feedback through randomized pairs (Rubin, 1974), an equivalence region for independent null-policy checks (Schuirmann, 1987), and preregistered confirmatory choices (Nosek et al., 2018).

Evaluation integrity and process attestation. Broad evaluation, contamination studies, and memorization research examine the influence of prior exposure (Hendrycks et al., 2021; Srivastava et al., 2023; Liang et al., 2023; Golchin & Surdeanu, 2025; Carlini et al., 2021). DeepResearchGym provides stable retrieval from fixed Web corpora (Coelho et al., 2026). SWE-bench emphasizes executable outputs (Jimenez et al., 2024). Proof-of-Learning records training states to attest a training procedure (Jia et al., 2021). ScientistOne traces claims, scores, references, and code to supporting evidence (Meng et al., 2026). DCP adds recovery and feedback interventions to research audits.

3 METHOD

3.1 THE OUTCOME AND ITS INFORMATION BOUNDARY

An operational discovery claim specifies a useful outcome, a registered model and budget, a starting information packet, and a numerical recovery rule. Core bounds matched recovery with run history and new measurements withheld. Evidence estimates a feedback-policy effect from a registered checkpoint.

K contains fixed background, the model, tools, and known methods. E_0 contains starting observations fixed independently of research actions. L^* is the target run's history used by later actions. A^* is the final output selected before sealed scoring, and P is the executable validity and recovery rule.

The E_0 and $L^\star$ boundary follows information provenance. An example present at the start belongs to E_0. A measurement returned because the agent chose an action belongs to $L^\star$. Fixed human guidance enters E_0; guidance chosen after intermediate results enters a human-agent lineage. Core targets adaptive research with at least one new observation consumed by a later research action. A pool of independent candidates followed by final score-based selection is a search-and-selection baseline.

The validity predicate and numerical threshold define recovery across all admissible artifacts. Historical priority is a separate scholarly assessment. The same outcome audit applies to programs, models, data products, and recipes.

3.2 PRECOMMITMENT AND SEALED EVALUATION

Registration precedes the target run and fixes the task, model, information, interface, tools, budgets, baseline, validity rules, selection and stopping procedures, and statistical analysis. It specifies the complete production procedure, including authorized measurement queries and their disclosure policy. Returned task measurements enter $L^\star$; the evaluator protects the remaining private reference data and final outcome scores. Every started confirmatory audit consumes a preallocated error budget, including recovered and incomplete audits. The registered selection rule freezes $A^\star$ before final scores are revealed. Committed baseline, target, and control artifacts share the same validity and scoring rules.

3.3 GATE 1 ESTABLISHES USEFUL IMPROVEMENT

Let $\mathrm{Valid}(a)$ denote compliance with the registered artifact constraints. Let b denote the baseline artifact and x the mean sealed score of $A^\star$, on a registered $[0, 1]$ scale. Write LCB for the registered lower confidence bound. For the smallest useful gain $\delta_{\min} > 0$, Gate 1 requires

$$\mathrm{LCB}[\mu(A^\star) - \mu(b)] \geq \delta_{\min}, \tag{1}$$

where μ is mean utility over the registered evaluation population. Sampled populations use a finite-sample confidence interval; exhaustive finite evaluations use an exact mean. Gate 1 also requires valid

baseline and target artifacts and a baseline below the recovery region. Confirmed target invalidity or insufficient utility fails the audit. Unresolved checks produce an inconclusive decision.

3.4 GATE 2 TESTS RECOVERY UNDER A REGISTERED CHALLENGER

Gate 2 gives a fresh matched agent the same K, complete E_0, model, interface, known components, and registered production opportunities. It withholds $L^\star$ and gives the challenger all Web bytes observed by the target run, denoted W_{obs}. The agent may reason, compile, and perform engineering checks. Responses to its actions follow a frozen policy that supplies non-directional messages and withholds new task scores and scientific measurements.

Every valid method is eligible. For a registered tolerance ε, the recovery rule is

$$P(a) = \text{Valid}(a) \ \wedge \ \text{score}(a) \geq x - \varepsilon. \tag{2}$$

The tolerance expresses a small substantively equivalent score difference. Registration requires $0 \leq \varepsilon < \delta_{\min}$. Gate 1 checks the baseline's position after sealed scoring. Repeated paired evaluation and simultaneous confidence intervals handle measurement uncertainty. Generators receive the metric and selection rule; the verifier computes x after all artifacts are committed.

A qualified recovery witness satisfies both Equation 2 and the registered execution, information, and provenance conditions. It provides an executable alternative route under the claimed resources and triggers the Core veto, recorded as *recovered* with Core *refuted*. The veto tests the existence of an audited alternative route. The probability bound below quantifies how often a fresh registered episode finds one. Both quantities remain in the report when a rare route is found.

Let Q_B be the registered distribution of complete challenger episodes at budget B. An episode includes its model calls, candidate opportunities, and selection rule. A best-of-k procedure counts as one k-candidate episode. Let $H = 1$ when any admissible candidate recovers the target, and define $p_B = \Pr_{Q_B}(H = 1)$. Registration fixes audit size n, recovery-bound threshold ρ, and one-sided error allocation α_{recovery}. With zero recoveries in n independent episodes, the exact Clopper–Pearson upper bound is (Clopper & Pearson, 1934)

$$p_{\text{upper}} = 1 - \alpha_{\text{recovery}}^{1/n}. \tag{3}$$

Core requires zero qualified witnesses in the registered n episodes, $p_{\text{upper}} \leq \rho$, and complete audit checks. This supports $p_B \leq \rho$ at confidence $1 - \alpha_{\text{recovery}}$ under the stated sampling model. Checks cover opportunity budgets, information access, independent draws, valid-output rates, and the registered control task. Controls establish the operation they exercise; challenger scores measure search performance. Candidate score tests retain a separate error allocation. Core concerns the achieved level x; fresh truthful continuations measure process repeatability. Figure 2 shows the distinct effects of audit size on precision and witness detection.

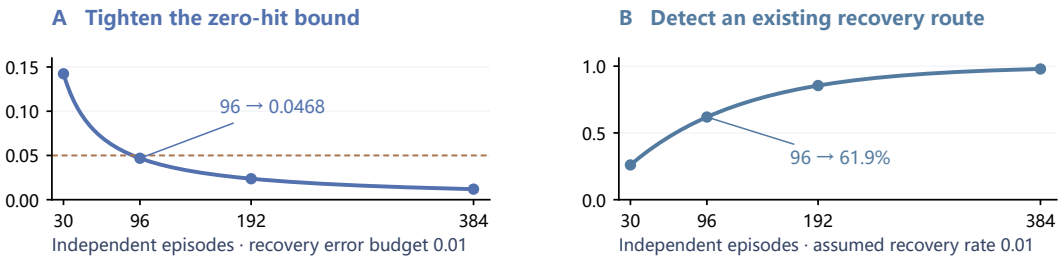

Figure 2: Two effects of audit size. The zero-hit upper bound uses recovery error budget 0.01; the orange line marks $\rho = 0.05$. Witness detection assumes an illustrative recovery rate $p_B = 0.01$. At 96 episodes, the corresponding values are 0.0468 and 61.9%.

3.5 GATE 3 ESTIMATES A FEEDBACK-POLICY EFFECT

Gate 2 measures recovery from the starting information. Gate 3 measures how subsequent feedback changes fresh outcomes from a shared checkpoint c. A rule registered before the run selects c, including its files and available history. Fresh paired branches receive either truthful feedback from

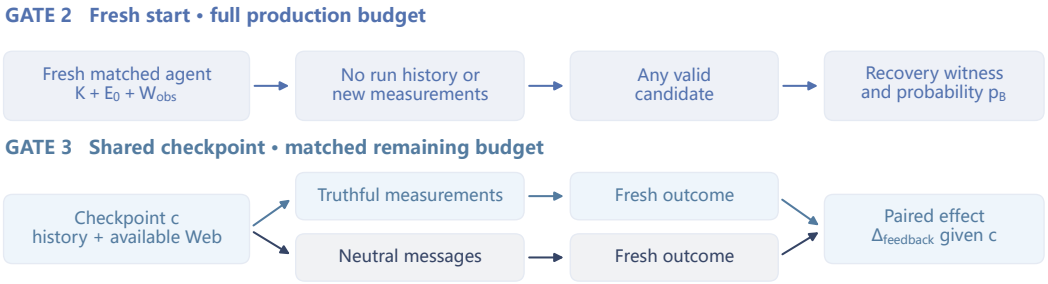

Figure 3: Information flows in the two tests. Gate 2 starts with K, E_0, W_{obs}, withholding run history and new measurements. Gate 3 branches from one checkpoint containing its available history and Web, holds remaining resources fixed, and compares truthful and neutral feedback.

their own actions or messages from a specified neutral policy. Both arms share the model, tools, remaining budget, starting state, and output checks. Figure 3 summarizes both tests.

Registration specifies the timing, schema, and content of both feedback policies. The neutral policy withholds directional task measurements, and later actions may diverge in response to the assigned messages. Arm assignment, pair count, selection, and stopping are frozen before the study. Fresh replicates estimate the average policy effect conditional on the checkpoint.

For registered utility u, the estimand is

$$\Delta_{\text{feedback}} = \mathbb{E}[u(A_{\text{truthful}}) - u(A_{\text{neutral}}) \mid c]\,. \tag{4}$$

A binary utility compares recovery probabilities. This is the total effect of the two registered feedback policies, conditional on c, including their downstream influence on the agent's actions.

Independent null tasks compare the neutral policy with a second non-directional policy on problems whose answers are determined by E_0. This null-policy equivalence check requires its interval to lie inside $\pm\delta_{\text{sham}}$. Its scope is the known-answer family. The target-task effect in Equation 4 includes the behavior induced by the chosen neutral policy. The registered buffer δ_{sham} is a decision margin, with no assumed bound on target-task neutral-policy bias. Evidence requires $\text{LCB}(\Delta_{\text{feedback}}) \geq \delta_{\text{evidence}} + \delta_{\text{sham}}$ and the equivalence check. A frozen rule retains every start and governs pre-action pair replacements.

Appendix Figure 8 traces the SQLite-Web audit through all three gates.

3.6 Web access

A recording gateway stores each query, response time, and exact model-visible bytes. Gate 2 discloses the complete observed packet at the start, so its recovery bound conditions on W_{obs}. Gate 3 pairs inherit the pages available at their checkpoint and share a frozen environment for later Web requests. A claim about the contribution of Web access uses an additional Web-withholding intervention.

3.7 Threat model and trust boundary

DCP treats the claimant and evidence producer as potentially strategic. In a formal deployment, an independent audit authority approves the question and resource match, holds the sealed test, and anchors records before execution. Independently anchored ledgers and provider-bound records expose missing starts, substituted models, early private-score access, and altered evidence. The checker evaluates numerical and record consistency under the accepted registration. Independent provenance verification and registry countersigning complete formal issuance.

Evidence production and decision checking have separate interfaces. Task adapters specify validity, scoring, and feedback; a shared deterministic verifier evaluates the resulting records. The reported audits were produced by task-specific runners. A lightweight reusable harness supports collection and control execution, and the offline verifier recomputes decisions from the resulting bundles. This separation lets independent readers check the same evidence across agent implementations.

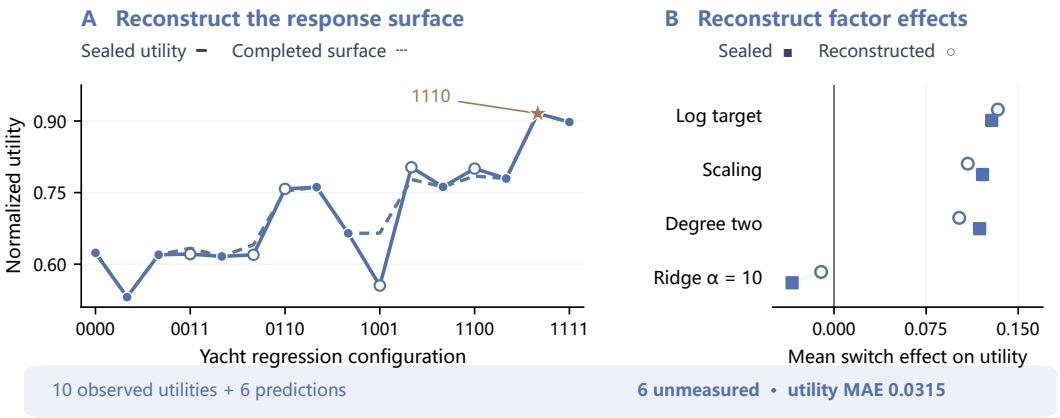

Figure 4: The Yacht response surface combines ten observed utilities with six predictions. (A) Filled points mark the two public anchors and eight queried utilities inserted by the evaluator; open points mark undisclosed values. The star marks the accompanying pipeline selection. (B) Each mean effect averages eight matched pairs. The completed surface scores 0.7677 on all 32 effects; unmeasured-six utility-prediction MAE is 0.0315.

3.8 Decisions and their meaning

Core and Evidence report *Pass*, *Refuted*, *Inconclusive*, or *Not evaluated* (Table 3). Inconclusive carries *audit incomplete* or *statistical uncertainty*. Prospective decisions use their pre-execution classification version. Table 7 identifies source policies and later amendments, with their timing and analysis scope. Appendix B defines these versioned failure rules.

4 Experimental Audit Cases

4.1 Study design

The primary suite combines Yacht and Ionosphere response-surface audits with controlled cases. Authors collected and replayed these records locally; every reported pass denotes a kernel decision with `formal_certificate_issued=false`. Independent authority review and countersigning define formal issuance. Artifacts were committed before final scoring under preallocated audit slots. Table 7 records version bases and original decisions; Appendix D.1 presents an exploratory Dry Bean analysis.

Gate 1 enumerated each finite sealed workload. Gate 2 sampled fresh episodes from each registered Q_B. SQLite-Web randomized equivalent presentations, catalyst randomized public starting recipes, and real-data tasks repeated fixed initial materials with provider-default sampling. Table 4 specifies these distributions and per-episode budgets B; $n = 96$ is the audit sample size. Gate 3 sampled fresh paired continuations of a frozen checkpoint. Development cases guided neutral-channel design; policies, thresholds, equivalence bands, and pair distributions were frozen before confirmatory branches opened. Provider-reported identities were checked through the registered CLI. SQLite-Web used `deepseek-v4-flash`, catalyst used `deepseek-v4-pro`, and real-data sessions used `deepseek-flash` with matched interfaces and budgets.

4.2 Real-data research workflows

Yacht uses hull and velocity data for regression (Gerritsma et al., 1981); Ionosphere uses radar returns for classification (Sigillito et al., 1989). These measurement-dependent profiles share device calibration's controlled information-access design, with utility surfaces obtained from pipelines on real data. Each scored artifact $A^\star$ combines two public anchors, eight measured utilities, and six predictions over four binary interventions. The evaluator inserts authorized observations into the submitted table. Baseline, challenger, and neutral observations contain only the two anchors; the baseline fills its other 14 entries with their mean. Gate 1 scores all 32 single-switch effects as

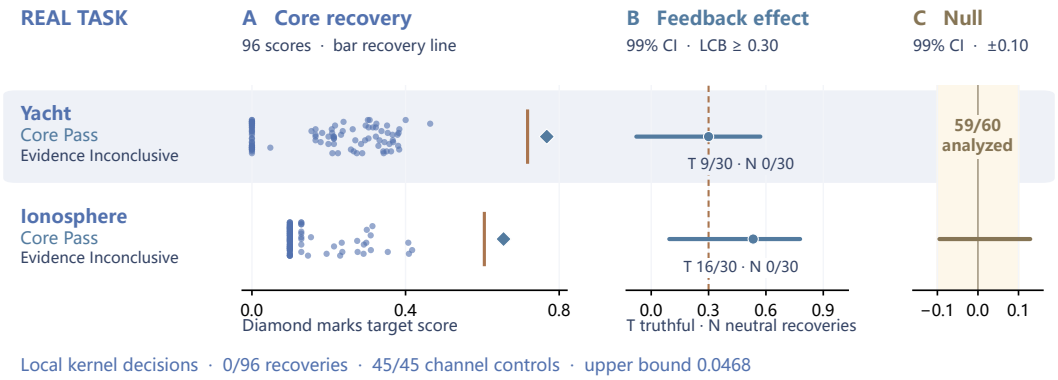

Figure 5: Local Core passes and separate feedback decisions. (A) Challenger scores lie below the recovery lines; diamonds mark target scores. (B) T/N reports truthful/neutral recovery counts beside 99% effect intervals; the required lower bound is 0.30. (C) The null band is ± 0.10. Yacht analyzes 59 of 60 pairs; Ionosphere analyzes all 60. Both report Evidence Inconclusive, with reasons audit incomplete and statistical uncertainty, respectively.

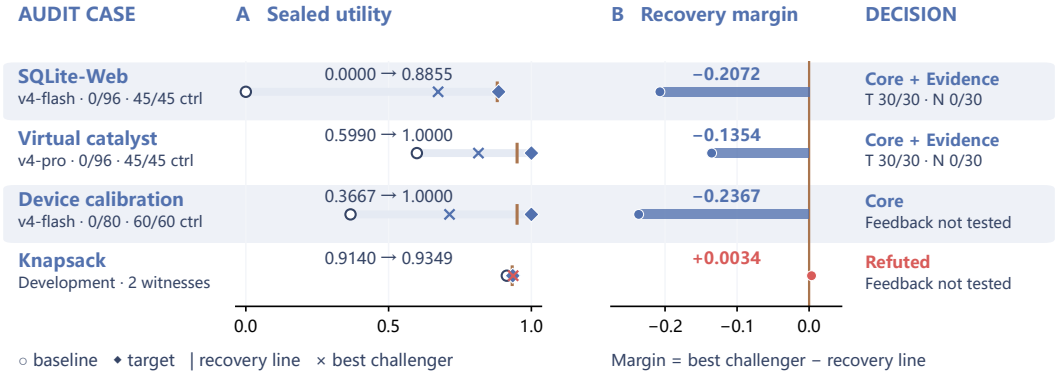

Figure 6: Controlled audit outcomes. (A) Sealed baseline, target, recovery line, and best challenger. (B) Best-challenger score minus the recovery line; a positive margin supplies a witness. Recovery and control counts accompany each case. Knapsack supplies two developmental witnesses.

$\max(0, 1 - \mathrm{MAE}_{\mathrm{effect}} / \operatorname{mean} |\mathrm{effect}|)$. This evaluates the total acquisition-and-completion gain. Core recovery concerns this data product; selected-pipeline utility is an accompanying readout.

Yacht and Ionosphere vary transformations, scaling, complexity, and regularization. Each audit registered 96 challengers, 45 file-channel controls, 30 feedback pairs, and 60 null pairs. Controls copy a supplied marker and submit a finish action, testing the read/write channel. Gate 3 starts at checkpoint 0 before each branch selects eight masks. Truthful replies return their utilities; neutral replies contain `status=no-measurement` and an empty map. The end-to-end effect includes direct insertion and subsequent predictions. The neutral arm shares Gate 2's task state, prompt, budget, and response policy; Gate 2 additionally carries an empty-Web manifest. Separate sampling records preserve this implementation distinction.

4.3 Controlled Calibration Profiles

The SQLite-Web audit in Figure 6 presents 16 query families, four high-traffic families, and a four-index budget. Its score measures the reduction in traffic-weighted SQLite virtual-machine work while preserving every query result. The agent captured two documentation pages and used measured per-family work to select indexes. Gate 2 received the pages; Gate 3 compared measured traffic feedback with schema-matched uniform weights from a checkpoint after Web capture.

Virtual catalyst control offers $8^5 = 32{,}768$ recipes. An agent selects an anchor, receives a 36-well one-factor plate, and commits a recipe evaluated on 256 sealed conditions. Truthful and neutral continuations share the anchor checkpoint. Device calibration supplies a Core control with known

hidden gains. Multidimensional knapsack supplies two distinct legal recovery witnesses from one developmental challenger episode. An undersampled affine audit exercises incomplete control adequacy. Appendix D and the controlled audit records give the exact task and execution details.

5 RESULTS

The main results connect response-surface reconstruction, matched recovery attempts, and feedback effects. Figures 4 and 5 show the real-data evidence; Figure 6 shows distinct controlled-case decisions.

5.1 THE REAL WORKFLOW RECONSTRUCTS AN INFORMATIVE RESPONSE SURFACE

Yacht's completed surface scores 0.7677, against baseline 0, with utility-prediction MAE 0.0315 on its six unmeasured configurations. Its reconstructed mean utility effect of log-transforming the target is $+0.1334$, against the sealed $+0.1284$; scaling gives $+0.1089$ against $+0.1210$. Figure 4 also shows polynomial-feature and regularization effects. The surface selects measured configuration `1110`, combining target transformation, scaling, degree-two features, and Ridge $\alpha = 1$. Its utility 0.9159 accompanies the audited data product as a selection readout. The Core recovery line 0.7177 applies to effect reconstruction across the complete surface.

5.2 REAL-DATA AUDITS SEPARATE RECOVERY AND FEEDBACK EVIDENCE

Yacht and Ionosphere pass the Core kernel at outcome levels 0.7677 and 0.6551, with zero recoveries in 96 episodes and $45/45$ channel controls. Their best challengers score 0.4641 and 0.4171, below recovery lines 0.7177 and 0.6051; the recovery upper bound is 0.0468. Fresh truthful continuations recover these levels in $9/30$ and $16/30$ trials; both neutral arms recover $0/30$. Core evaluates the achieved level, while these rates quantify process repeatability.

Both Evidence decisions are Inconclusive. Yacht's feedback interval is $[-0.078, 0.571]$, and a changed protected observation leaves its null-policy check audit incomplete. Ionosphere's feedback interval is $[0.094, 0.779]$, and its null interval $[-0.095, 0.129]$ crosses the equivalence band, yielding statistical uncertainty. Figure 5 presents these separate decision fields. The version table records the Ionosphere timeout classification and its unchanged final verdict.

5.3 CONTROLLED AUDITS EXERCISE THE DECISION SPACE

SQLite-Web reduced traffic-weighted work by 88.55%, and virtual catalyst control reached its sealed optimum. Both pass the Core and Evidence kernels (Figure 7). SQLite-Web's Evidence pass uses the corrected implementation of its registered read-before-first-write rule; Table 7 records the cross-bundle replay and version basis. Device calibration adds a Core pass, and knapsack supplies a recovery witness (Figure 6). The affine case and a separate SQLite addendum remain Inconclusive. Appendix Table 5 gives exact values.

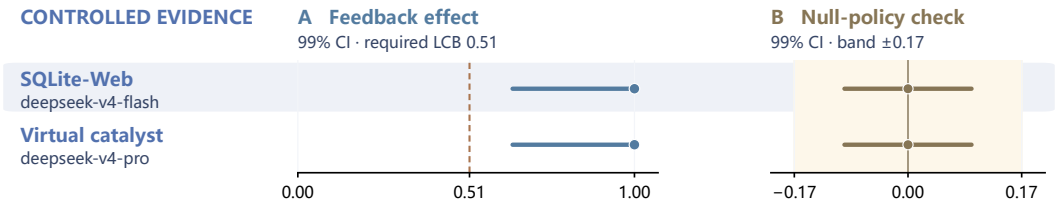

Figure 7: Controlled Gate 3 evidence. Each audit has 30 target pairs and 60 null pairs. Registered 99% feedback intervals clear the 0.51 lower-bound requirement; null intervals lie within ± 0.17.

6 DISCUSSION

6.1 THE RECOVERY RULE ADMITS ALTERNATIVE METHODS

A component ablation evaluates an intervention within a chosen implementation. DCP evaluates recovery across every admissible implementation offered by the registered challenger. A recombina-

tion of known modules, a transfer from another domain, or a different program can all supply the same numerical witness. Two distinct legal knapsack solutions cross the recovery line under this rule. Task adapters specify validity, sealed utility, and a starting information packet. Program search (Romera-Paredes et al., 2024; Novikov et al., 2025), AI scientist workflows (Lu et al., 2026; Yamada et al., 2025), and ML engineering benchmarks (Chan et al., 2025; Nathani et al., 2025) can instantiate the same audit interface when their candidate histories and budgets are registered.

6.2 Reporting a research contribution

A reusable report gives the outcome artifact, threshold, model and information scope, episode budget, recovery witness or bound, truthful repeatability, and checkpoint-conditional feedback interval. Authors submit the record, and auditors replay decisions and examine alternative routes. Yacht and Ionosphere illustrate local Core passes with distinct reasons for Evidence Inconclusive. Multi-task deployments register each task instance and report the decision distribution. A separately registered no-measurement arm can assess target-task sensitivity to the neutral policy.

6.3 Deployment and retrospective recovery

Prospective use starts with an approved registration and sealed evaluator, followed by evidence collection, control execution, offline replay, and independent countersigning. Auditors can stop at the first qualified recovery and reserve paired studies for feedback claims. Confirmatory collection used 419 catalyst sessions costing 54.27 USD and 504 SQLite-Web sessions costing 61.13 USD. These totals cover target, challengers, controls, feedback pairs, and null pairs. Generation incurs model cost; a reader replays the saved evidence offline.

Retrospective use reconstructs an outcome target and seeks a qualified recovery witness. An audit with pre-publication-equivalent model and information access can test the original recovery scope. An audit using post-publication knowledge measures present-day recoverability. Records identify this temporal scope and their analysis-policy version. Formal issuance requires prospective registration and complete evidence collection.

7 Conclusion

The Discovery Certification Protocol gives measurable AI research outputs an executable, layered evidence interface under a registered model, information boundary, and budget. Gate 1 establishes sealed utility. Gate 2 accepts any valid route to the numerical target with run history and new measurements withheld; Core requires adequate registered controls, zero witnesses, and a finite-sample bound at a fixed audit size. Gate 3 compares truthful and neutral feedback from a shared checkpoint; Evidence requires a positive effect above a registered margin and a null-policy equivalence check. A deterministic verifier replays local decisions; independent authority review and countersigning complete formal issuance.

The evidence covers controlled software optimization, virtual experimental control, and response-surface reconstruction on real datasets. SQLite-Web reduced traffic-weighted work by 88.55%, and virtual catalyst control reached its sealed optimum; both pass the Core and Evidence kernels. Yacht and Ionosphere pass Core with zero recovery in 96 attempts and $45/45$ channel controls. Their truthful repeatability rates of $9/30$ and $16/30$ accompany Evidence Inconclusive decisions. Yacht's completed surface combines ten observed utilities with six predictions, scores 0.7677 on effect reconstruction, and achieves unmeasured-six utility-prediction MAE 0.0315. Knapsack supplies an executable recovery witness. These records make achieved outputs, alternative routes, process repeatability, and feedback effects inspectable through one evidence interface.

Reproducibility Statement

Frozen machine-readable bundles and evidence records support the numerical claims. The appendices report task construction, registrations, model and interface contracts, evaluation units, thresholds, statistics, budgets, costs, and identifiers. The anonymous supplement contains the controlled examples, three real-data decision bundles, versioned provenance receipts, and figure sources. It also contains the installable offline verifier, task-specific evidence producers, and reusable general-purpose harness. Decision replay is deterministic and offline. Fresh evidence production uses the model and environment specified in each registration. The harness exposes capture, challenge, paired feedback testing, finalization, and bundle generation as separate operations.

Ethics Statement

AI research agents can accelerate useful discovery while increasing the need for clear and contestable evidence. DCP supports accountable reporting by binding each decision to a registered model, information boundary, budget, probability bound, and replayable record. Certificates should be presented with this scope, and captured Web data should follow applicable privacy, licensing, and security requirements. Deterministic replay gives authors, reviewers, and independent auditors a common record on which to examine a research claim.

AI Use Statement

Generative AI tools supported ideation, protocol and statistical design, adversarial review, experiment design, implementation, testing, evidence inspection, result interpretation, literature work, and manuscript editing. GPT and Codex assisted these activities. Claude interfaces supported review and provided the registered CLI execution layer. `deepseek-flash`, `deepseek-v4-flash`, and `deepseek-v4-pro` served as the audited agents and matched challengers in the respective studies. Image generation informed early teaser drafts; the publication teaser is native SVG artwork with vector export. Quantitative plots were generated with Matplotlib from frozen evidence. Executable evaluators and statistical code determine scores and decisions. The authors review the scientific claims, check numerical evidence and primary sources, test the code, and take responsibility for the final content and artifacts.

References

Shitanshu Bhushan, Yunxiang Zhang, and Lu Wang. Can LLM agents discover? evaluating creativity on ML engineering tasks. In *Conference on Language Modeling*, 2026. URL https://arxiv.org/abs/2608.30047.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *30th USENIX Security Symposium*, pp. 2633–2650, 2021. URL https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting.

Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, Lilian Weng, and Aleksander Madry. MLE-bench: Evaluating machine learning agents on machine learning engineering. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.07095.

Hui Chen, Miao Xiong, Yujie Lu, Wei Han, Ailin Deng, Yufei He, Jiaying Wu, Yibo Li, Yue Liu, and Bryan Hooi. MLR-Bench: Evaluating AI agents on open-ended machine learning research. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2025a. doi: 10.52202/085713-3945.

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*, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/f12b4df26344f3be803c06b555252efe-Abstract-Conference.html.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404. URL https://academic.oup.com/biomet/article/26/4/404/291538.

João Coelho, Jingjie Ning, Jingyuan He, Kangrui Mao, Abhijay Sai Paladugu, Pranav Setlur, Jiahe Jin, Jamie Callan, João Magalhães, Bruno Martins, and Chenyan Xiong. DeepResearchGym: A free, transparent, and reproducible sandbox for deep research. In *Proceedings of the 2026 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval*, pp. 34–43. ACM, 2026. doi: 10.1145/3805713.3820409. URL https://doi.org/10.1145/3805713.3820409.

Alhussein Fawzi, Matej Balog, Aja Huang, Thomas Hubert, Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Francisco J. R. Ruiz, Julian Schrittwieser, Grzegorz Swirszcz, et al. Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature*, 610:47–53, 2022. doi: 10.1038/s41586-022-05172-4.

J. Gerritsma, R. Onnink, and A. Versluis. Yacht hydrodynamics, 1981. URL https://archive.ics.uci.edu/dataset/243/yacht+hydrodynamics.

Shahriar Golchin and Mihai Surdeanu. Data contamination quiz: A tool to detect and estimate contamination in large language models. *Transactions of the Association for Computational Linguistics*, 13:809–830, 2025. doi: 10.1162/tacl.a.20. URL https://aclanthology.org/2025.tacl-1.37/.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Petar Sirkovic, Artiom Myaskovsky, Grzegorz Glowaty, Felix Weissenberger, Alessio Orlandi, et al. Accelerating scientific discovery with co-scientist. *Nature*, 655:487–496, 2026. doi: 10.1038/s41586-026-10644-y.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Peter Jansen, Marc-Alexandre Côté, Tushar Khot, Erin Bransom, Bhavana Dalvi Mishra, Bodhisattwa Prasad Majumder, Oyvind Tafjord, and Peter Clark. DiscoveryWorld: A virtual environment for developing and evaluating automated scientific discovery agents. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 37, pp. 10088–10116, 2024. doi: 10.52202/079017-0324. URL https://arxiv.org/abs/2406.06769.

Hengrui Jia, Mohammad Yaghini, Christopher A. Choquette-Choo, Natalie Dullerud, Anvith Thudi, Varun Chandrasekaran, and Nicolas Papernot. Proof-of-learning: Definitions and practice. In *2021 IEEE Symposium on Security and Privacy*, pp. 1039–1056, 2021. doi: 10.1109/SP40001.2021.00106.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can language models resolve real-world github issues? In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=VTF8yNQM66.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=iO4LZibEqW.

Chris Lu, Cong Lu, Robert Tjarko Lange, Yutaro Yamada, Shengran Hu, Jakob Foerster, David Ha, and Jeff Clune. Towards end-to-end automation of AI research. *Nature*, 651:914–919, 2026. doi: 10.1038/s41586-026-10265-5.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, 2023. doi: 10.52202/075280-2019.

Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. DiscoveryBench: Towards data-driven discovery with large language models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/0d70af566e69f1dfb687791ecf955e28-Abstract-Conference.html.

Daniel J. Mankowitz, Andrea Michi, Anton Zhernov, Marco Gelmi, Marco Selvi, Cosmin Paduraru, Edouard Leurent, Shariq Iqbal, Jean-Baptiste Lespiau, Alex Ahern, et al. Faster sorting algorithms discovered using deep reinforcement learning. *Nature*, 618:257–263, 2023. doi: 10.1038/s41586-023-06004-9.

Gábor Melis, Chris Dyer, and Phil Blunsom. On the state of the art of evaluation in neural language models. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=ByJHuTgA-.

Rui Meng, Bhavana Dalvi Mishra, Jiefeng Chen, Chun-Liang Li, Palash Goyal, Mihir Parmar, Yiwen Song, Yale Song, Rajarishi Sinha, Parthasarathy Ranganathan, Burak Gokturk, Jinsung Yoon, and Tomas Pfister. ScientistOne: Towards human-level autonomous research via chain-of-evidence. *arXiv preprint arXiv:2605.26340*, 2026. URL https://arxiv.org/abs/2605.26340.

Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *European Conference on Computer Vision*, pp. 681–699, 2020. doi: 10.1007/978-3-030-58595-2_41.

Deepak Nathani, Lovish Madaan, Nicholas Roberts, Nikolay Bashlykov, Ajay Menon, Vincent Moens, Mikhail Plekhanov, Amar Budhiraja, Despoina Magka, Vladislav Vorotilov, Gaurav Chaurasia, Dieuwke Hupkes, Ricardo Silveira Cabral, Tatiana Shavrina, Jakob Foerster, Yoram Bachrach, William Yang Wang, and Roberta Raileanu. MLGym: A new framework and benchmark for advancing AI research agents. In *Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=ryTr83DxRq.

Jingjie Ning, Xiaochuan Li, Ji Zeng, Hao Kang, and Chenyan Xiong. Auto research with specialist agents develops effective and non-trivial training recipes. *arXiv preprint arXiv:2605.05724*, 2026a. URL https://arxiv.org/abs/2605.05724.

Jingjie Ning, Xiaochuan Li, Ji Zeng, Chenyan Xiong, and Guolin Ke. Closed-loop auto research for molecular property prediction: Discovering and certifying generalizable improvements. *arXiv preprint arXiv:2606.22731*, 2026b. URL https://arxiv.org/abs/2606.22731.

Jingjie Ning, Xiaochuan Li, Shanshan Zhong, Ji Zeng, and Guolin Ke. Auto research for materials: Auditable AI-scientist workflows with held-out transfer. *arXiv preprint arXiv:2607.17100*, 2026c. URL https://arxiv.org/abs/2607.17100.

Jingjie Ning, Xueqi Li, and Chengyu Yu. Revision or re-solving? decomposing second-pass gains in Multi-LLM pipelines. In *Conference on Language Modeling*, 2026d. URL https://arxiv.org/abs/2604.01029.

Jingjie Ning, Shanshan Zhong, Xiaochuan Li, Ji Zeng, and Chenyan Xiong. One run is not an idea: The implementation lottery in automated research. *arXiv preprint arXiv:2607.26587*, 2026e. URL https://arxiv.org/abs/2607.26587.

Brian A. Nosek, Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606, 2018. doi: 10.1073/pnas.1708274114.

Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, et al. AlphaEvolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*, 2025. URL https://arxiv.org/abs/2506.13131.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan S. Ellenberg, Pengming Wang, Omar Fawzi, et al. Mathematical discoveries from program search with large language models. *Nature*, 625:468–475, 2024. doi: 10.1038/s41586-023-06924-6.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974. doi: 10.1037/h0037350.

Donald J. Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6):657–680, 1987. doi: 10.1007/BF01068419.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, 2023. doi: 10.52202/075280-0377.

V. Sigillito, S. Wing, L. Hutton, and K. Baker. Ionosphere, 1989. URL https://archive.ics.uci.edu/dataset/52/ionosphere.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=uyTL5Bvosj.

Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, Johannes Heidecke, Amelia Glaese, and Tejal Patwardhan. PaperBench: Evaluating AI's ability to replicate AI research. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 56843–56873. PMLR, 2025. URL https://proceedings.mlr.press/v267/starace25a.html.

UCI Machine Learning Repository. Dry bean, 2020. URL https://archive.ics.uci.edu/dataset/602/dry+bean.

Zhen Wang, Fan Bai, Zhongyan Luo, Jinyan Su, Kaiser Sun, Xinle Yu, Jieyuan Liu, Kun Zhou, Claire Cardie, Mark Dredze, Zhiting Hu, and Eric P. Xing. FIRE-Bench: Evaluating AI agents on the rediscovery of scientific insights. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.02905.

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.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, 2023a. doi: 10.52202/075280-0517.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=WE_vluYUL-X.

Shanshan Zhong, Yi Lu, Jingjie Ning, Yibing Wan, Lihan Feng, Yuyi Ao, Leonardo F. R. Ribeiro, Markus Dreyer, Sean Ammirati, and Chenyan Xiong. SkillLearnBench: Benchmarking continual learning methods for agent skill generation on real-world tasks. In *Conference on Language Modeling*, 2026. URL https://arxiv.org/abs/2604.20087.



Figure 8: One recorded SQLite-Web audit. The target captured two Web pages before its first experiment and used traffic feedback to select A^*. Gate 2 received those pages while L^* and new task measurements were withheld. Gate 3 compared fresh feedback policies from a shared checkpoint after Web capture. The bottom strip reports controls and null-policy checks.

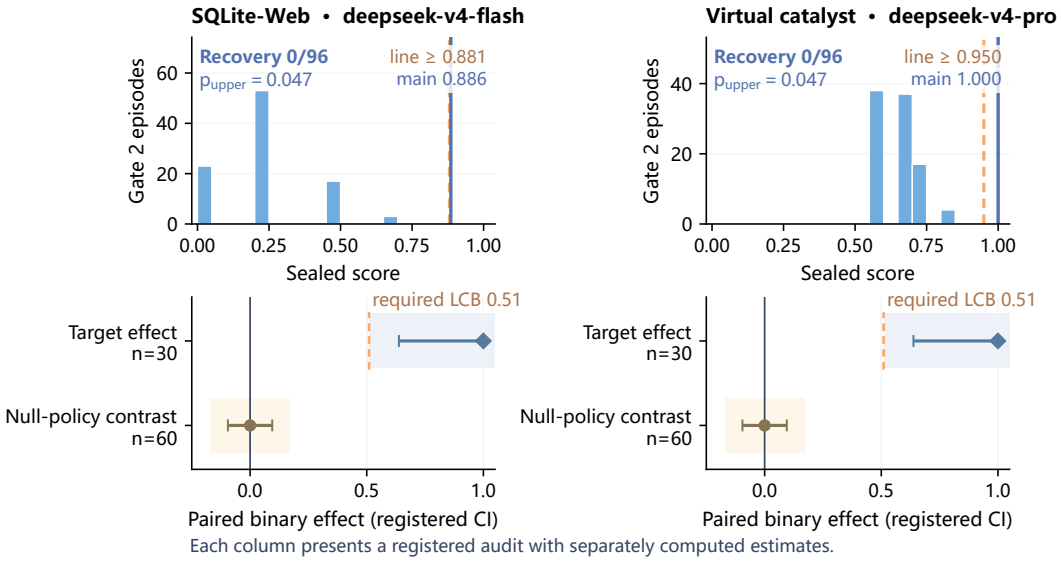

Figure 9: Controlled three-gate audits in two domains and model scopes. The upper row shows all 96 Gate 2 scores per audit. The lower row shows registered 99% intervals for the checkpoint-conditional feedback effect and independent null-policy checks.

A Statistical Decision Details

A.1 Main Result

For paired private evaluation units, DCP computes differences between $A^\star$ and the baseline. The sampled-population profile uses a bounded empirical-Bernstein interval on the registered $[0, 1]$ score scale. A claim over a fully enumerated finite set uses its exact mean. Each uncertainty calculation uses its registered sampling unit.

The value x is the registered evaluation mean of $A^\star$, selected before sealed scoring. The baseline must be valid and lie below the recovery region under its adjusted uncertainty check.

A.2 Candidate Decisions

Each challenger candidate is scored on the same private units as $A^\star$. Let d_i be its paired score difference from $A^\star$. Candidate i is a confirmed hit when its adjusted lower confidence bound is at

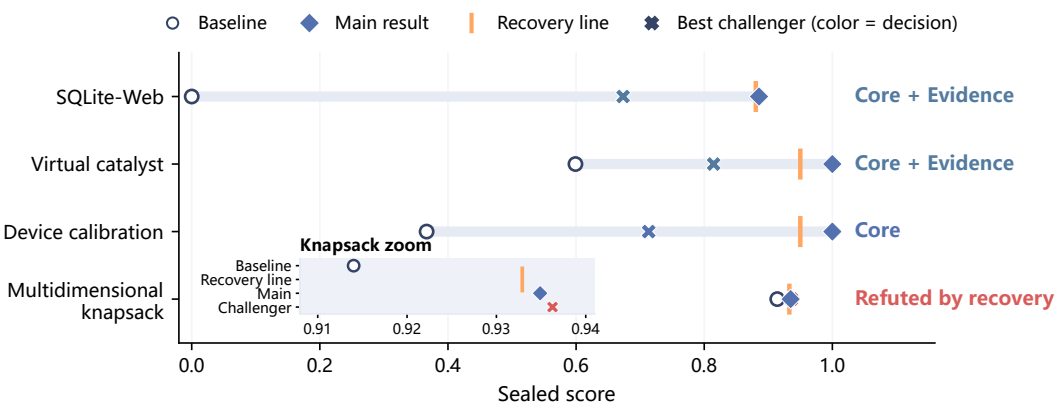

Figure 10: The sealed-score geometry across four artifacts. The best no-lineage challenger stays below the registered recovery line in the three Core cases and crosses it in knapsack. The inset resolves the knapsack scores. The challenger may use any valid method.

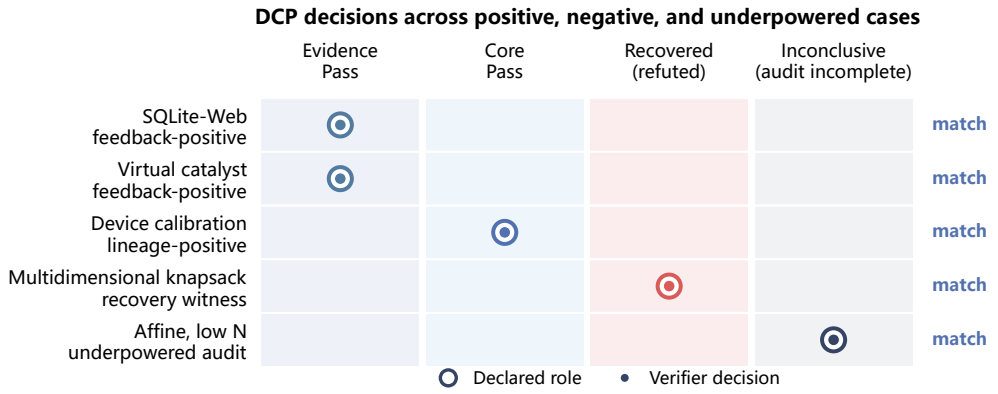

Figure 11: The controlled calibration suite spans two Evidence positives, one constructed information-deficit Core positive, one developmental recovery refutation, and one intentionally undersampled audit. Rings show the declared calibration role and filled dots show the deterministic verifier decision.

least $-\varepsilon$. It is a confirmed miss when its adjusted upper confidence bound is below $-\varepsilon$. Every other candidate is unresolved.

The family of candidate intervals shares a registered error budget. A simple profile assigns Bonferroni corrected error to every possible candidate slot, controlling false recovery across candidates that can change the decision. Equation 3 receives a separate error budget. Its prerequisites are zero qualified hits, resolved candidates, and a complete adequate control ledger.

The SQLite-Web and virtual catalyst audits allocated 0.005 each to main validation, candidate classification, positive-control adequacy, and neutral recovery, and 0.01 each to challenger recovery, the target feedback effect, and null-policy checks. The allocations total 0.05. Both used $\rho = 0.05$, 45 positive controls, and minimum control recall 0.8. Their feedback threshold was $\delta_{\text{evidence}} = 0.34$ and decision buffer was $\delta_{\text{sham}} = 0.17$. The real-data audits used the same per-gate error allocations with $\delta_{\text{evidence}} = 0.20$ and $\delta_{\text{sham}} = 0.10$.

When $\alpha_{\text{recovery}} = 0.05$, 59 independent misses put the one-sided upper bound near 0.05. With $\alpha_{\text{recovery}} = 0.025$, 72 misses are required. Device calibration allocated 0.02, required at least 77 misses, and registered 80 episodes. Each bound concerns one fresh episode from its registered model, policy, information, and budget distribution.

A.3 Recovery witnesses and repeated search

For independent Bernoulli episode outcomes with recovery probability p_B, the probability of zero recoveries in n episodes is $(1 - p_B)^n$. Thus a fixed-sample zero-hit audit reaches the registered bound when $n \geq \lceil \log(\alpha_{\text{recovery}}) / \log(1 - \rho) \rceil$. For the two controlled three-gate audits and all three real-data audits, $\alpha_{\text{recovery}} = 0.01$ and $\rho = 0.05$ require 90 episodes; the registered count was 96. Candidate score classification and the other statistical gates retain their separate error allocations.

A witness establishes an available route, while the binomial bound quantifies the probability that a fresh episode produces any admissible recovery. DCP assigns them different decision roles. A campaign of k independent episodes from the same Q_B has recovery probability $1 - (1 - p_B)^k$. A best-of-k production procedure therefore includes its full search and selection budget in one registered episode. Fresh Gate 3 pairs estimate the checkpoint-conditional feedback effect separately from selection of a successful target.

Additional episodes tighten a zero-hit bound and increase the chance of finding an existing rare route. For $p_B = 0.01$, 96 independent episodes find at least one witness with probability 0.6190; their probability of satisfying the zero-witness condition is 0.3810. Core uses this fixed-sample eligibility rule together with the bound threshold. A witness remains visible beside the bound and registered audit size. Table 1 gives the exact-rule sensitivity. At $\alpha_{\text{recovery}} = 0.01$, zero-hit audits need at least 44, 90, 228, and 459 episodes for $\rho = 0.10, 0.05, 0.02, 0.01$, respectively.

Table 1: Audit size and the two Gate 2 quantities. With $\alpha_{\text{recovery}} = 0.01$, the upper bound conditions on zero hits. Witness detection assumes an illustrative fresh-episode recovery probability $p_B = 0.01$.

Independent episodes	**30**	**96**	**192**	**384**
Zero-hit recovery upper bound	0.1423	0.0468	0.0237	0.0119
Witness detection at $p_B = 0.01$	0.2603	0.6190	0.8548	0.9789

A.4 Feedback decisions

The paired feedback estimator uses normalized utility differences or paired differences in recovery success, conditional on the frozen checkpoint. The utility and both policies are registered before the pairs start. The binary interval combines four exact one-sided Clopper–Pearson limits (Clopper & Pearson, 1934). Let $\delta_{\text{evidence}} > 0$ be the minimum effect and $\delta_{\text{sham}} > 0$ the registered decision buffer. Support requires passing null-policy checks and

$$\text{LCB}(\Delta_{\text{feedback}}) \geq \delta_{\text{evidence}} + \delta_{\text{sham}}.$$

The effect-below-threshold decision requires

$$\text{UCB}(\Delta_{\text{feedback}}) + \delta_{\text{sham}} < \delta_{\text{evidence}}.$$

Remaining interval positions are inconclusive. The null interval describes equivalence on the registered known-answer family. The protocol quantifies neutral recovery with separate candidate checks and a finite-sample probability bound. Table 2 gives the minimum supportive counts for the controlled audits' registered binary effect condition.

Table 2: Registered binary feedback-effect decision boundary. Entries give the minimum truthful-only wins with zero neutral-only wins and all other pairs tied, using $\alpha_{\text{evidence}} = 0.01$ and effect threshold $\delta_{\text{evidence}} + \delta_{\text{sham}} = 0.51$. "No pass" means that even all truthful-only wins fall short. Core, null-policy checks, and audit integrity have separate checks.

Paired episodes	**20**	**30**	**40**	**60**	**100**
Minimum truthful-only wins	No pass	28	35	48	71

B Audit Validity

Valid numerical interpretation requires both branches to implement the registered comparison. DCP records the following evidence.

- The registration is sealed before any private result or confirmatory null case is opened.
- Model identity, interface, tools, information packet, and resource budget match their registered values.
- Every started attempt remains in the ledger, including timeouts, invalid outputs, and infrastructure failures.
- Candidate outputs are committed before private scores are returned.
- The main output, baseline, and controls use one symmetric evaluator.
- Each independent episode identifier corresponds to a committed draw record; its internal candidate opportunities share one statistical unit.
- All Web responses given to the model are recorded byte for byte when Web access is enabled.

Classification versions. We use reporting identifiers for the real-data policies. F_0 denotes each run's own frozen source policy, including its completion and reserved-channel checks. F_1 names the recorded Ionosphere timeout amendment, and F_2 names the Dry Bean regular-file amendment. Yacht retains F_0; Ionosphere reports F_1 alongside its F_0 record; exploratory Dry Bean reports F_2 alongside F_0. These labels index archived policies and receipts. Table 7 gives amendment timing. Controlled tasks retain their task-specific registered rules.

Outcome failures and comparison integrity. Under F_1, a timeout with verified accepted-model activity and intact protected channels receives the registered failed-attempt utility. Under F_2, an additional regular file containing public material leaves an existing valid candidate eligible for scoring; absent or invalid candidates receive failure utility. Wrong or unverified model identity, an executed forbidden tool, unauthorized information, and modified protected feedback remain incomplete-comparison conditions. Each source F_0 retains its original checks and decision. Pre-action infrastructure faults follow the registered attrition rule, and every start remains in the ledger. Independent authority review and signatures govern formal issuance.

Table 3: Decision vocabulary used throughout the paper. A pass denotes a local kernel decision. The reason field qualifies an Inconclusive decision.

Label	Meaning and recorded value
Pass	Registered numerical and audit checks pass; kernel value `certified`.
Refuted	A qualified recovery or confirmed failed claim condition; `refuted`. For Evidence, a confirmed effect-below-threshold finding maps here after Core, integrity, and null-policy prerequisites pass.
Inconclusive	Decision remains unresolved; `inconclusive`, with reason `audit_incomplete` or `statistical_uncertainty`.
Not evaluated	Optional grade omitted or blocked by a prerequisite; `not_tested`.

B.1 Episode budgets and sampling distributions

Table 4 specifies the recorded 96-episode audits, including exploratory Dry Bean. Budget B covers one challenger episode with one final-candidate opportunity; n counts audit repetitions. Each CLI session can contain several provider requests and file-tool calls, governed by its cost cap and timeout. Decoding uses provider defaults with effort `low`; commands impose no separate request-count or token cap. Target allowances are three \$0.50/360 s sessions for SQLite-Web, two \$0.50/360 s sessions for catalyst, and the same three-session caps as challengers for real-data tasks. Table entries distinguish these allowances from actual target use.

Fixed opportunity and stopping rules. All episodes use the registered model, public task, initial evidence, evaluator, thresholds, and stateless CLI interface. SQLite challengers receive all target-observed Web bytes and commit one index plan in one session. Catalyst challengers choose an assay anchor in the first session and commit a recipe in the second, with a registered no-signal assay between sessions. The nonce maps each starting control independently to $\{0, 1, 2, 4, 5, 6, 7\}$. Real-data challengers may select eight masks and receive an empty measurement map; the evaluator uses their final submitted table with the two public anchors. Their action loop ends on `finish`, a session

Table 4: Challenger allowances, target consumption, and sampling scope. Every task registered $n = 96$ episodes with one final-candidate opportunity each. Target use lists sessions, reported USD, and summed CLI wall seconds, excluding evaluator and gateway time.

Task	Challenger budget B	Target use	Variation in Q_B
SQLite-Web	1 session, $1.50, 1,080 s; 2 CPUs, 3 GiB	3 / 0.297 / 85.1	Candidate and family aliases, row order, four prompt variants
Virtual catalyst	2 sessions, each $0.50 and 360 s; 2 CPUs, 3 GiB	2 / 0.229 / 93.0	Public nonce and five-control starting recipe
Yacht	Up to 3 sessions, each $1.00 and 240 s; 2 CPUs, 2 GiB	2 / 0.352 / 91.6	Fixed task; fresh provider sampling
Ionosphere	Up to 3 sessions, each $1.00 and 240 s; 2 CPUs, 2 GiB	2 / 0.280 / 78.4	Fixed task; fresh provider sampling
Dry Bean Web (R)	Up to 3 sessions, each $1.00 and 360 s; 2 CPUs, 2 GiB	3 / 0.483 / 121.8	Fixed task and Web packet; fresh provider sampling

failure, or the three-session cap. Cost exhaustion and timeout terminate each session. Intermediate edits remain within the same episode, and final scoring opens after all candidate commitments.

Provider model identifiers. Assistant payloads report `deepseek-v4-flash` for SQLite-Web, device calibration, and knapsack, `deepseek-v4-pro` for catalyst, and `deepseek-flash` for Yacht, Ionosphere, and Dry Bean. The configured DeepSeek endpoint routes the CLI model slot `claude-sonnet-4-6` to the first and third backends; this slot is a client-side alias. Catalyst requests `deepseek-v4-pro` directly. Identity checks use the provider's assistant field. The identifiers, frozen endpoints, and collection windows define the model scope; the provider supplies no weight-checkpoint revision for `deepseek-flash`. Short figure labels use v4-flash, v4-pro, and flash.

Collection windows. All real-data model sessions were collected on 25 September 2026. Recorded session timestamps, rounded to seconds in UTC, span 10:58:42 to 15:26:16 for Yacht, 11:13:45 to 14:01:20 for Ionosphere, and 16:49:28 to 21:10:18 for Dry Bean. These windows run from the first target-session start through the last audit-session completion, including all started branches.

Positive-control inputs. SQLite-Web controls receive the known optimal plan, an alias-mapping test, and a captured-Web content marker. Catalyst controls receive a complete informative assay for a fresh reaction and must submit its optimum. Device controls receive every hidden gain in episode-local coordinates. Yacht, Ionosphere, and Dry Bean controls receive a task-neutral marker in `CANARY.txt`, copy it exactly, and submit a finish action. These checks respectively assess supplied-solution submission, assay-to-recipe execution, calibration submission, and file-channel operation. Challenger search performance is evaluated through the recorded recovery attempts and their stated Q_B.

Information and sampling scope. Controlled sessions expose Read and Edit tools; real-data sessions expose Read, Write, Edit, Glob, and Grep. Shell and native Web tools are disabled. Yacht, Ionosphere, and catalyst use no Web. Dry Bean permits one host-mediated UCI request with a 20 s timeout, a 2,000,000-byte raw limit, and a 300,000-byte delivery limit; Gate 2 uses the recorded response. All 96 initial workspace manifests and first-prompt receipts match within each real-data task. Their independent, identically distributed episode model concerns repeated provider sampling conditional on that fixed configuration. Draw receipts identify started episodes; the recorded real-data commands leave the provider's random seed unspecified. The resulting p_B bound applies to each task's stated generation process and budget.

C AUDIT CASE DETAILS

SQLite-Web denotes the event-traffic profile with two captured documentation pages; SQLite with Web disabled is a separate earlier Core audit, and the e-commerce profile supplies the separate parent and Gate 3 addendum in Appendix E.

Table 5: Controlled audit decisions. Gate 1 shows baseline-to-target utility; T/N gives truthful/neutral recoveries. Zero-hit bounds are 0.0468 for 96 episodes and 0.0477 for 80, with respective recovery error budgets 0.01 and 0.02. Knapsack supplies two developmental witnesses.

Case	Gate 1 score	Best challenger	Recovery line	Recovery	Controls	Gate 3 T/N	Decision
SQLite-Web v4-flash	0.0000 → 0.8855	0.6734	0.8805	0/96	45/45	30/30, 0/30 null 60, pass	**Core + Evidence**
Virtual catalyst v4-pro	0.5990 → 1.0000	0.8146	0.9500	0/96	45/45	30/30, 0/30 null 60, pass	**Core + Evidence**
Device calibration v4-flash	0.3667 → 1.0000	0.7133	0.9500	0/80	60/60	Not run	**Core**
Knapsack v4-flash, dev.	0.9140 → 0.9349	0.9363	0.9329	Found	n/a	Not run	**Refuted**

C.1 SQLITE WITH CAPTURED WEB ACCESS

The registered assistant payload was `deepseek-v4-flash`. The main score was 0.8855470, the baseline score was 0, and the recovery line was 0.8805470. All 96 matched challenger episodes were valid confirmed misses. Their scores took four values between 0.0368978 and 0.6733847. The zero-hit upper bound was $1 - 0.01^{1/96} = 0.0468381$. All 45 positive controls succeeded, giving a one-sided recall lower bound of 0.8889268. Gate 1 used the exact enumerated finite workload and required a gain of $\delta_{\min} = 0.5$.

The gateway recorded two HTTPS responses from `sqlite.org`. Gate 2 received the normalized model-visible text and content hash for every response. The target feedback test used 30 fresh pairs. Truthful and neutral recovery counts were $30/30$ and $0/30$. The paired exact 99% interval was $[0.6379275, 1]$. In 60 fresh null pairs, both arms succeeded in every pair, giving a difference of 0 and an exact 99% interval of $[-0.0950339, 0.0950339]$.

The machine-readable record is `artifacts/dcp_sqlite_event_web_fullflow_confirmatory_v1_certificate/certificate.json`. Its identifier begins with `b3f0af47968ba2cb`. The self-contained publication Bundle containing the same certificate is `examples/audits/sqlite-web` and its identifier begins with `eff6cf27afa58da5`. All SQLite-Web decisions and Gate 3 numbers in this paper are read from this certificate. Appendix E records its verifier-versioned provenance.

Confirmatory collection used 504 model sessions and 61.13 USD. Including three interface preflight sessions gives 507 sessions and 61.17 USD.

C.2 VIRTUAL CATALYST OPTIMIZATION

The registered model identity was `deepseek-v4-pro`. The main score was 1.0, the baseline score was 0.5989989, and the recovery line was 0.95. Gate 1 required $\delta_{\min} = 0.35$, with recovery tolerance $\varepsilon = 0.05$. All 96 matched challenger episodes were valid confirmed misses, and the best score was 0.814649. The zero-hit upper bound was $1 - 0.01^{1/96} = 0.0468381$. All 45 positive controls succeeded, giving a one-sided recall lower bound of 0.8889268.

The target feedback study used 30 fresh pairs. Truthful and neutral recovery counts were $30/30$ and $0/30$. The paired exact 99% interval was $[0.6379275, 1]$. In 60 fresh null pairs, both arms had equal outcomes in every pair, giving a difference of 0 and an exact 99% interval of $[-0.0950339, 0.0950339]$.

Replay uses `examples/audits/virtual-catalyst`. Its identifier begins with `fa17603f2184bb7b`.

Confirmatory collection used 419 model sessions and 54.27 USD. Development and confirmatory collection together used 435 sessions and 56.40 USD. Both source and replay records accompany the supplement.

C.3 SQLite with Web disabled

This secondary Core audit predates the Web-enabled case and is retained for direct replay at `examples/audits/sqlite-core`. The main score was 0.9432886, the baseline was 0, and the recovery line was 0.9382886. All 80 challenger episodes were valid confirmed misses. The best challenger scored 0.7947267. The zero-hit upper bound was $1 - 0.02^{1/80} = 0.0477239$. All 45 positive controls succeeded, giving a one-sided recall lower bound of 0.8889268. Gate 1 used the exact enumerated finite workload and required $\delta_{\min} = 0.5$. The evidence bundle identifier is `3690bebc8b3ac882`. The 130 recorded model sessions cost 25.56 USD.

C.4 Device calibration

The public replay Bundle is `examples/audits/device-calibration-core`. The main score was 1.0, the baseline score was $110/300 = 0.3667$, and the recovery line was 0.95. All 80 challenger episodes were valid confirmed misses. The best score was $214/300 = 0.7133$. The zero hit upper bound was $1 - 0.02^{1/80} = 0.0477$. All 60 positive controls succeeded, producing a one sided Clopper Pearson lower bound of 0.9155. Gate 1 used the exact enumerated 300-item set and required $\delta_{\min} = 0.5$. Its identifier begins with `dff7b57b062362e0`. The recorded provider cost was 19.08 USD.

C.5 Multidimensional knapsack

The main score was 0.9348968, the baseline score was 0.9140206, and the recovery line was 0.9328968. Two candidate opportunities inside one challenger episode produced scores 0.9362833 and 0.9355607. These qualified recovery witnesses trigger the Core veto. The anonymous supplement provides a six-file challenger input manifest whose reconstructed digest matches both prestart records in the frozen event chain. The packet marks task feedback unavailable and contains no earlier candidate score. The record supports a constructive recovery decision. On the deterministic finite sealed set, the main gain was 0.0208762, above the registered $\delta_{\min} = 0.01$.

C.6 Development checks retained as inconclusive

An affine parity run exercised all three gates with two challenger episodes, two target pairs, and two null pairs. The observed main score and feedback contrast were both 1.0, while the positive-control recall lower bound was 0.0707. The verifier returned an incomplete audit at control adequacy. An active automaton run showed a 1.0 main score against 0.5 no-feedback outputs and remained incomplete in challenger adequacy and paired interface records. Earlier single-response induction tests measure E_0-dependence because their labeled examples were available at the start. These developmental records are reported separately from the primary audits.

D Real-Data Workflow Records

The audited artifact in these workflows is a completed response surface for 16 executable configurations. For a submitted prediction table $\hat{u}$ and an authorized observation map O, the evaluator applies the fixed rule

$$A^{\star}_m = C(\hat{u}, O)_m = \begin{cases} O(m), & m \in \mathrm{dom}(O), \\ \hat{u}_m, & m \notin \mathrm{dom}(O). \end{cases} \tag{5}$$

Its registered outcome measures recovery of 32 adjacent-configuration utility effects. Clustering utility is NMI against private labels, Yacht uses $1/(1 + \mathrm{MAE}/\mathrm{IQR}_{\mathrm{train}})$, and Ionosphere uses balanced accuracy. The two public anchors are E_0. Eight utilities returned by the truthful measurement action enter $L^{\star}$ and expand O to ten entries; the other six remain private through artifact commitment. Challenger and neutral observations contain only the two anchors. Final effect-recovery scores are

revealed after candidate generation closes. Unmeasured-six MAE summarizes prediction error on configurations whose utilities stayed private.

Baseline and insertion-only comparison. The registered baseline preserves the two public anchors and assigns their mean to the other 14 entries. The offline insertion-only diagnostic additionally inserts the target's eight frozen measurements and leaves its six remaining entries at that same mean. Table 6 reports this fixed-batch comparison. Insertion alone clears the registered $\delta_{\min} = 0.20$ gain for Yacht and Ionosphere. Agent predictions further improve their effect-reconstruction scores by 0.4691 and 0.2289. Gate 1 measures the full acquisition-and-completion gain. The tolerance remains $\varepsilon = 0.05$, with $\alpha_{\text{recovery}} = 0.01$ and $\rho = 0.05$. Gate 3 requires a feedback lower bound of 0.30 and a null interval inside $[-0.10, 0.10]$.

Table 6: Offline completion diagnostic on the frozen target measurement sets. Insertion-only adds eight observed utilities to the two-anchor baseline and uses the anchor mean for the remaining six entries. The final column measures the additional score from the submitted predictions at fixed observations. The registered Gate 1 baseline remains the two-anchor column.

Task	Two anchors	Insertion-only	Completed target	Added score
Yacht	0.0000	0.2986	0.7677	+0.4691
Ionosphere	0.0988	0.4263	0.6551	+0.2289
Dry Bean (R)	0.0000	0.1439	0.7124	+0.5685

Registered intervention. The adapter selects checkpoint 0, containing the public task, two anchors, and initial prediction table, before any eight-mask selection. In Dry Bean this state also precedes Web capture; both fresh arms first replay the same frozen UCI response. Each arm then chooses its own eight-mask batch. Truthful replies use schema `wwb_feedback_v1`, status `measured`, and an eight-entry measurement map. Neutral replies use the same schema, status `no-measurement`, and an empty map. The agent subsequently submits its prediction table, and Equation 5 inserts the authorized values. The Gate 3 estimand is the end-to-end recovery effect on this completed data product. It includes direct insertion of eight measured values and any improvement in predictions for the remaining six configurations. Even with unchanged residual predictions, inserting measured values can improve the score. The known-answer null family compares empty-map policies labelled `no-measurement` and `reference-null`.

D.1 Exploratory Dry Bean Web response surface

Dry Bean uses bean-image features and seven private labels for clustering (UCI Machine Learning Repository, 2020). Its four switches control signed-log features, scaling, 95% principal-component projection, and one versus ten KMeans initializations. Utility is normalized mutual information (NMI). This exploratory record complements the primary Yacht and Ionosphere audits.

The captured UCI page has snapshot identifier `54e2ea93` and is replayed byte for byte. Figure 12 compares the completed surface with all 16 reference utilities. Full-grid MAE is 0.0100, and unmeasured-six MAE is 0.0268. Reconstructed mean effects for signed log, scaling, PCA, and ten restarts are $(+0.0291, +0.1782, -0.0237, +0.0259)$; sealed values are $(+0.0304, +0.1747, -0.0271, +0.0075)$. Mean absolute true edge effect 0.0605176846 and edge error 0.0174069010 give reconstruction score 0.7124. Measured configuration `1101` attains the highest pipeline NMI, 0.7222, as an accompanying selection result. The final analysis bundle at `examples/audits/dry-bean-web`, identifier `d5c04462cdd1ffb9`, returns Core Pass and Evidence Pass under the revised analysis policy. Reported model cost is 48.39 USD.

Exploratory analysis scope. The source record has an incomplete challenger-contract attestation because episode 78 created an extra regular file in the reserved channel directory. The file contained a public request identifier; recorded protected observations and instructions stayed unchanged. The revised policy treats this file addition as an ordinary workspace error and scores the existing candidate, preserving all 96 slots. That policy was recorded after Gate 2 scoring, so its Core and Evidence passes describe an exploratory reanalysis. The original Core Inconclusive and dependent Evidence Not evaluated decisions remain in Table 7. This case supplies response-surface and feedback observations within that stated scope.

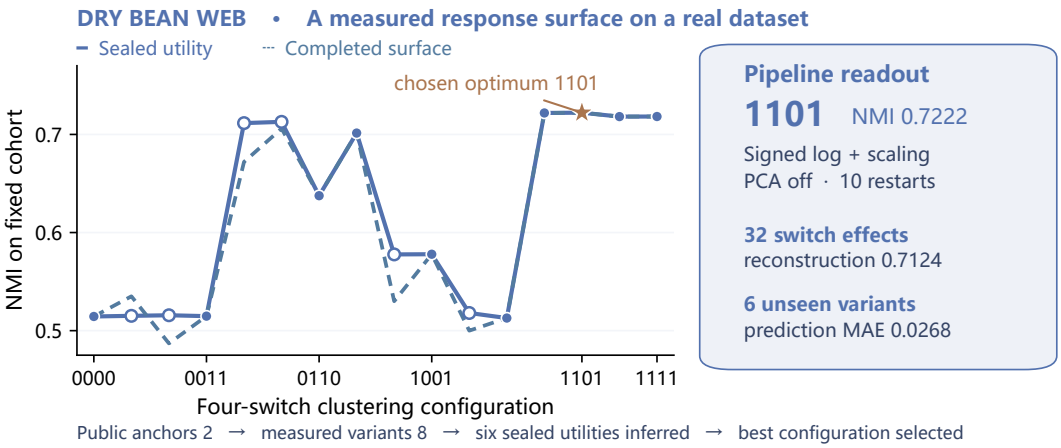

Figure 12: Exploratory Dry Bean response surface. Ten observed utilities and six predictions form the scored data product. Pipeline selection is an accompanying readout. Table 7 records the original and revised-policy kernel decisions.

Paired feedback statistics. The separately collected Gate 3 record contains 30 target pairs and 60 null pairs. Its 27 truthful-only recoveries give effect 0.90 and 99% interval $[0.4772, 0.9910]$; all null contrasts are zero, with interval $[-0.0950, 0.0950]$. These observations are present in both bundles. They quantify the end-to-end feedback effect independently of the original certificate's short-circuited Evidence field. The revised-policy Evidence decision inherits the reanalysis scope of Core.

D.2 Yacht and Ionosphere

Yacht's target score was 0.7677, against a baseline of 0, and its best challenger scored 0.4641 below the 0.7177 recovery line. Its full-audit bundle at `examples/audits/yacht` has identifier prefix `72e6c50143839ba8`. The target feedback effect interval included zero. One null branch modified the protected observation during calibration pair 25, leaving 59 analyzable null pairs and Evidence Inconclusive with an incomplete-audit label. The recorded model cost was 39.23 USD.

Ionosphere's target score was 0.6551, and its best challenger scored 0.4171 below the 0.6051 recovery line. Its versioned full-audit bundle at `examples/audits/ionosphere` has identifier prefix `017a4fd527476dbd`. All 60 null pairs were analyzed, with one nonzero contrast. Their equivalence interval $[-0.0950, 0.1289]$ crossed the registered ± 0.10 band; the target effect lower bound 0.0944 stayed below 0.30. An accepted-model timeout was classified as a failed branch in the versioned record, and the Evidence verdict remained inconclusive. The recorded model cost was 32.98 USD. Both retain local Core Pass with zero recoveries in 96 attempts and $45/45$ file-channel controls.

E Audit Record

Each audit record binds the registered claim and information boundary to its final output, sealed scores, challenger contract, attempt ledger, controls, intervals, and decision. Feedback records add paired branches and null-policy checks. The verifier replays these records offline. Local results retain `formal_certificate_issued=false`; independent provenance verification and countersigning govern formal issuance. Historical priority remains `not_assessed`.

Cross-bundle replay. The current kernel replays all 11 supplied audit bundles and the linked e-commerce addendum, including the archived source versions, with every saved decision reproduced. Raw SQLite checks cover the event-traffic audit, the e-commerce parent, and its addendum under their recorded path rules. The read-ordering implementation correction changes only the event-traffic Evidence decision from Inconclusive to Pass. Ionosphere retains its decisions under F_1, and Dry Bean's exploratory changes are separately attributed to policy F_2.

Table 7: Decision provenance and classification versions. P denotes Pass, I denotes Inconclusive, and N denotes Not evaluated. Decision pairs list Core/Evidence. Policy entries give source-to-report versions defined in Appendix B. Appendix E specifies cross-bundle replay coverage.

Record	Source	Reported	Policy	Version basis
SQLite-Web	P/I	P/P	Task policy retained	Corrected read-before-first-write implementation; all 120 saved null branches replayed. Source Evidence reason was audit incomplete.
Virtual catalyst	P/P	P/P	Task policy retained	Original registration and decision retained.
Yacht	P/I	P/I	$F_0 \to F_0$	Original decision retained. A changed protected observation leaves the null-policy comparison audit incomplete.
Ionosphere	P/I	P/I	$F_0 \to F_1$	Accepted-model timeout amendment recorded after start and before Gate 3 scoring. Both versions retain statistical uncertainty; Core is unchanged.
Dry Bean (R)	I/N	P/P (R)	$F_0 \to F_2$	Regular-file amendment recorded after Gate 2 scoring. Revised passes are exploratory; source Core remains audit incomplete.

Separate registered SQLite feedback record. The earlier e-commerce parent `1f9d3142` yielded Core Pass and Evidence Inconclusive. Its registered Gate 3 addendum `210b680c` started 22 target pairs and 45 null pairs with $\alpha_{\text{evidence}} = \alpha_{\text{sham}} = 0.01$. Twenty-one truthful branches reached the target, one had a model failure, and all 22 neutral outputs stayed below the recovery line. Raw replay identifies an unfinished, unverified experiment session in reference-null pair 28 and relative-path reads outside the frozen exact-path convention in pair 32. The ordering correction leaves both conditions unresolved, preserving Evidence Inconclusive. The parent and addendum use the recorded exact-path convention; event-traffic uses its separately frozen workspace-normalized path policy.

Verifier-versioned event-traffic certificate. An earlier ordering verifier marked two sequential-write sham branches incomplete in frozen event-traffic bundle `2889e78f`. Offline recertification applied the registered exact-read-before-first-write rule, replayed subsequent writes in order, and validated all 120 sham branches. Raw sessions, registration, Gate 1 and Gate 2 inputs, target pairs, thresholds, and error allocations stayed fixed. The canonical certificate `b3f0af47` is packaged in publication bundle `eff6cf27`; its recertification receipt is in the supplement.