# WHEN SUCCESS IS INVALID: LATENT VALIDITY IN AI-GUIDED MACHINE LEARNING RESEARCH

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

A machine-learning artifact can satisfy its nominal evaluator without supporting the scientific claim attached to it. *Latent validity* is the satisfaction of claim-critical requirements left uncertified by a particular evaluation interface. Our central contribution is a claim-to-control compilation path: structure the residual from the claim and interface, map each requirement to a complete mediation cut or detection boundary, and compile it into a guard or evidence rule, a channel-relative completeness test, a trivalent witness, and a fail-closed release decision. We evaluate components of this lifecycle. A diagnostic archive finds that 113/824 artifact passes fail a separate process proxy; a frozen historically informed roster shows human-adjudicated candidate failures recurring within and across requested routes. In blinded archived packets, the human labeler resolves process evidence for 33/34 complete traces but only 5/116 incomplete traces, and every resolved incomplete-trace verdict is fail. Only 84/135 nominal intervention pairs contain an application receipt for the requested executable control. Every constructed integrity case matches the frozen oracle: intact compliant or allowed-control streams release, intact violations fail, and corruptions block. These results characterize the claim-to-control lifecycle rather than estimating prevalence, ranking agents, or establishing causal guard efficacy: whether computational success supports a scientific claim depends partly on admissible process evidence, not on the final artifact alone.

## 1 INTRODUCTION

AI research agents now run empirical studies end to end: they read data, write code, fit models, and report results. Each run here is given a task contract that bounds what the agent may use and a test that accepts or rejects its model. On one constructed hospital-escalation task, whose contract sealed a deployment benchmark, an agent submitted a model that passed that test. Its trace records a different process: before freezing the model, the agent read that benchmark and scored every candidate feature on it, then reported validation-based selection (Section 6). The score cannot separate that run from a clean one; the retained trace can. The same gap opens when preprocessing is fit on held-out data or selection adapts to test feedback (Kapoor & Narayanan, 2023), when local checks miss a global constraint, or when a metric loses the lineage of the state that produced it. The artifact passes; the evidence does not support the scientific claim.

*Latent validity* is the satisfaction of claim-critical requirements that a nominal evaluation leaves uncertified. It is defined relative to both the scientific claim and the evaluation interface. Here, *latent* means uncertified by that interface, not inherently unobservable. The same artifact may support a descriptive claim but not forward generalization. A test score is untouched-sample evidence only if test data remain outside adaptation; adaptive reuse instead requires a mechanism designed to preserve statistical validity (Dwork et al., 2015; Blum & Hardt, 2015). An *invalid success* passes the visible acceptance criterion while violating at least one latent-validity requirement.

AI research agents turn this blind spot into a systems problem. They optimize visible success signals while inspecting data, revising decisions, and mutating state across long workflows. Two runs can produce the same accepted artifact yet carry different scientific meaning because their information,



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.

decision, or state histories differ. Output-only evaluation cannot recover that difference. Natural-language instructions state intent; they neither enforce the required boundary nor prove that it held. The key object is the residual between what the claim requires and what the interface certifies. Our framework turns that residual into executable rules across the four failure surfaces of Table 1.

Table 1: **Four common latent-validity failure surfaces.** Each row shows what a passing output can hide and how a claim-preserving system should respond. The set is claim-dependent, not exhaustive.

| Failure surface | What a passing output can hide | Rule action |
|---|---|---|
| **Information flow** | Held-out or restricted data influenced a claim-critical decision. | Block prohibited access before it can affect that decision; record the access path. |
| **Decision protocol** | Evaluation feedback selected the model, seed, or stopping rule. | Freeze claim-critical decisions before evaluation feedback; record later changes. |
| **Joint feasibility** | Components pass separately, but their combination violates a shared budget, capacity, or precedence constraint. | Enforce global constraints before execution or release; retain a joint certificate. |
| **Evidence lineage** | A metric and checkpoint exist but are not bound to the code, data, configuration, and state that produced them. | Bind them to the state that produced them; block unsupported release. |

Leakage, provenance, and runtime enforcement are ingredients of our system, not its conceptual novelty. The novelty is the *claim-relative residual construction*: start from a scientific claim and its nominal evaluator, determine what acceptance leaves uncertified, and compile those residual requirements into frozen prevention, detection, evidence, and release semantics.

We make four contributions. **Conceptually**, we formalize interface-relative latent validity. **Methodologically**, we give a claim-to-rule protocol that admits a rule only on a claim-necessity argument and a non-entailment certificate, then freezes it before evaluation on a new lineage. **Architecturally**, we build a fail-closed release system that executes those frozen rules as mediation cut sets, trivalent witnesses, and channel-relative completeness tests, binding each verdict to a run-bound receipt. **Empirically**, we separate questions that are often conflated: whether output and process instruments disagree, whether retained evidence permits certification, whether failures recur on a frozen roster, whether requested controls were applied, and whether the evidence architecture behaves as specified. These studies form an evidence ladder. Our unit of analysis is the claim-bearing research process and its rule design, not an agent leaderboard.

## 2 Latent Validity

**Claim-relative definition.** Let $c$ specify a scientific claim, including its estimand when applicable, target setting, and stated assumptions. Let $\tau$ denote a research process and $z$ its artifact. Let $\Omega_c$ contain every $(\tau, z)$ that could be offered as evidence for $c$, including processes that violate its requirements. Each requirement is a predicate $r : \Omega_c \to \{0, 1\}$, and the declared requirement set $\mathcal{C}(c)$ contains the information, decision, feasibility, and lineage conditions needed to report $c$ under its assumptions. We decompose $c$ at the claim level into conditions that can be certified independently of one another: preprocessing isolation and split isolation are separate requirements, because checking split hashes certifies only the second. Coarser bundling keeps already-certified components inside the residual. This neither guarantees completeness nor proves $c$ true.

Evaluation interface $e$ exposes only $\pi_e(\tau, z)$, on which its nominal predicate $A_e$ operates. Define certification over a nonempty acceptance region $\mathsf{Acc}_e(c)$:

$$
\begin{aligned}
\mathsf{Acc}_e(c) &= \{(\tau, z) \in \Omega_c : A_e(\pi_e(\tau, z)) = 1\}, \\
A_e \models_c r &\iff \mathsf{Acc}_e(c) \neq \varnothing \ \wedge\ \forall (\tau, z) \in \mathsf{Acc}_e(c),\ r(\tau, z) = 1, \\
\mathcal{L}(c, e) &= \{r \in \mathcal{C}(c) : A_e \not\models_c r\}, \\
Y^{\mathrm{latent}}_{c,e}(\tau, z) &= \bigwedge_{r \in \mathcal{L}(c,e)} r(\tau, z).
\end{aligned}
\tag{1}
$$

A reject-all interface thus certifies nothing rather than everything vacuously. Here $A_e$ is the nominal acceptance rule, not ground truth; its certification envelope is the object under study. If two accepted processes expose the same view but differ on $r$, that view cannot certify $r$. For a fixed claim, a nonempty acceptance refinement can only shrink the residual:

$$\varnothing \neq \mathsf{Acc}_{e_2}(c) \subseteq \mathsf{Acc}_{e_1}(c) \quad \Longrightarrow \quad \mathcal{L}(c, e_2) \subseteq \mathcal{L}(c, e_1). \tag{2}$$

Merely exposing more information does not suffice; the acceptance rule must use it. The blind spot is induced by the interface, not an agent family.

Latent-validity requirements are not a new sibling of construct, internal, or external validity. They describe the claim-critical part of any such category that the chosen interface leaves uncertified; latent validity is their joint satisfaction.

The *declared joint endpoint* combines nominal success with every residual requirement:

$$Y_{c,e}^{\mathrm{declared}}(\tau, z) = A_e(\pi_e(\tau, z)) \wedge Y_{c,e}^{\mathrm{latent}}(\tau, z). \tag{3}$$

An *invalid success* has $A_e(\pi_e(\tau, z)) = 1$ and $Y_{c,e}^{\mathrm{latent}}(\tau, z) = 0$. Both $c$ and $e$ matter: exploration and final estimation can authorize different actions, and a requirement stops being latent once the interface certifies it. "Declared" matters as well: the framework cannot prove that $\mathcal{C}(c)$ is complete.

**Trivalent operational semantics.** For each $r$, a frozen channel model $\chi_r$ names the channels and claim-critical interval to be evidenced. Writing $E_r$ for the evidence retained for $r$, $Q_r(E_r; \chi_r) = 1$ certifies capture relative to that declaration; it says nothing about omitted channels. Let $w_r(E_r) \in \{\mathsf{P}, \mathsf{F}, \mathsf{U}\}$ and define

$$\widehat{r}_\chi = \begin{cases} \mathsf{F}, & w_r(E_r) = \mathsf{F}, \\ \mathsf{P}, & Q_r(E_r; \chi_r) = 1 \ \wedge \ w_r(E_r) = \mathsf{P}, \\ \mathsf{U}, & \text{otherwise.} \end{cases} \tag{4}$$

A witness must be specified to be fail-sound, $w_r(E_r) = \mathsf{F} \Rightarrow r(\tau, z) = 0$, and monotone on append-only prefixes $E \preceq E'$ of one run, $w_r(E) = \mathsf{F} \Rightarrow w_r(E') = \mathsf{F}$; both are design requirements, not measured properties. A concrete violation certificate therefore establishes fail from a partial trace and survives to the final evidence, whereas positive certification requires complete declared-channel evidence. Let $\widehat{A}_e$ be the evidence-grounded verdict of the nominal interface and set $\widehat{Y}_{c,e} = \widehat{A}_e \wedge_K \bigwedge_{r \in \mathcal{L}(c,e)} \widehat{r}_\chi$. Strong-Kleene conjunction (Kleene, 1952) fails on any $\mathsf{F}$, passes only on all $\mathsf{P}$, and is unresolved otherwise; $\phi$ releases only pass.

**Evidence follows the claim.** Evidence supports progressively stronger statements. An artifact check supports nominal success. A sound witness supports its named requirement. Verified control application, verified control-arm absence, and a comparable design are needed to attribute a change to a control. A transfer claim would additionally require independently authored lineages; this paper makes none. The first unmet condition limits the claim, regardless of whether the observed result is favorable, null, or harmful.

## 3 Designing Latent-Validity Rules

The framework transforms the residual set $\mathcal{L}(c, e)$ into a frozen rule system $\Gamma_{c,e} = \{q_r : r \in \mathcal{L}(c, e)\}$. Each record $q_r = (n_r, \xi_r, \chi_r, B_r, g_r, a_r)$ binds six decisions: a claim-necessity argument $n_r$; a non-entailment certificate $\xi_r$; the declared channel model $\chi_r$; a mediation cut $B_r$; a preventive guard $g_r$ or detection-only choice; and an audit specification $a_r = (Q_r, w_r, v_r)$ containing the completeness test, witness, and rule version. The fail-closed map $\phi$ is architecture-wide. Thus $\mathcal{C}(c)$ states what the claim requires, whereas $\Gamma_{c,e}$ states how the system controls and audits what the interface leaves uncertified. Table 2 works one requirement through all six decisions.

For a preventive rule, $B_r$ is a complete mediation cut relative to $\chi_r$, adapting the complete-mediation principle of Saltzer & Schroeder (1975) to the declared channel model, when every trajectory that reaches an $r$-violating completion from an allowed initial state through a channel declared in $\chi_r$ crosses at least one mediated transition in $B_r$; it is minimal when no proper subset has that property. Concurrent workflows may admit several incomparable minimal cuts, so the earliest mediated transition is a design preference, not a unique event. If no enforceable complete cut exists, the rule is detection-only.

**Derive requirements from the claim.** Rule design starts from the claim, not from a fixed failure list. Decompose $c$ into its target, assumptions, information boundaries, decision protocol, joint constraints, and required provenance. For each resulting requirement, ask what observation through $e$ would certify it. If nominal acceptance can occur without that observation, the requirement enters $\mathcal{L}(c, e)$. The four surfaces in Table 1 guide this search; they are non-exclusive mechanisms, not task genres or a closed taxonomy. Appendix Table 17 refines them into ten operational families.

**Admit rules with non-entailment certificates.** An observed failure can suggest a requirement but cannot establish one. Admission requires $n_r$ and $\xi_r$: the former connects rule to claim; the latter establishes $A_e \not\models_c r$. It is either an accepted counterexample violating $r$, or an accepted view $u$ whose fiber $\pi_e^{-1}(u) \cap \Omega_c$ contains processes differing on $r$. This structural certificate needs no engineered exploit. A usable rule then fixes its boundary, control, evidence, and witness before evaluation. Prompts may explain a rule; they neither mediate its boundary nor prove it held.

**Freeze before evaluation.** The development protocol is **specify, discover, admit, compile, freeze, evaluate, revise**. A counterexample may motivate a revision but cannot validate that revision. Version $v_r + 1$ is frozen before evaluation on a new, independently authored lineage. This separation lets the rule set improve without using the same example to define a failure, tune its detector, and certify its repair.

## 4 A Claim-Preserving Runtime Architecture

The runtime architecture executes a frozen $\Gamma_{c,e}$ across three planes (Figure 1). The agent acts on the execution plane. The control plane mediates enforceable boundaries such as data access, phase transitions, artifact commits, and global constraints. The evidence plane runs witnesses and stores receipts outside the evaluated agent's write authority. A rule either prevents a violation at every applicable member of $B_r$, or detects it afterward and blocks release. In both cases, $w_r$ issues a trivalent verdict on the retained evidence for the requirement.

**Threat model.** The evaluated agent may modify its permitted workspace, call permitted tools, and adapt to exposed signals. We assume it cannot modify the frozen rule implementation, witness, or receipt store or bypass a mediated control boundary. The architecture does not defend against host compromise, sandbox escape, dishonest providers, malicious witnesses, or an incorrect requirement specification.

A boundary must cover every relevant channel: arbitrary shell access requires OS-enforced file isolation, services require capability brokers, and phase or global rules require tokens, transactions, or solvers. Our architecture types and records a denial, never silently substituting data. A blocked request is evidence but does not itself falsify an information-flow requirement; a separate protocol may forbid the attempt itself.

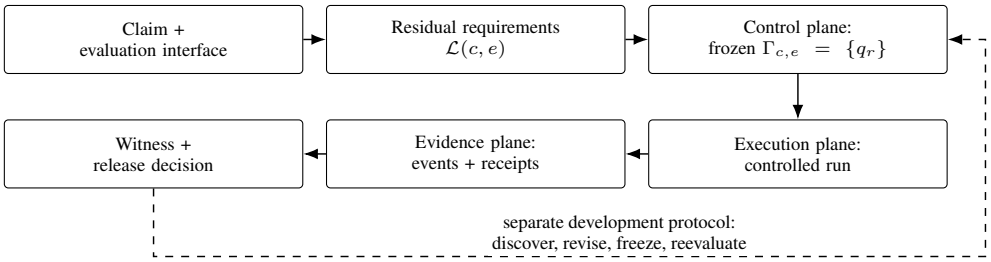


Figure 1: **Claim-to-control architecture.** The claim and interface define what nominal success leaves uncertified. The control plane governs execution; the write-isolated evidence plane supports witnessed release. The dashed loop is the separate rule-development protocol, not online rule tuning.

**Evidence-bound release.** The system returns an evidence-bearing run bundle: artifact, claim and interface identifiers, frozen rule versions, coverage manifests, guard and witness receipts, verdicts, unresolved set, and provenance root. Write $E_r = E_r(\tau, z)$. Claim release requires nominal success,

complete evidence, a passing witnessed verdict and authentic receipt for every residual requirement, and an application receipt for every preventive guard:

$$\begin{aligned} R_{c,e}(\tau, z) = A_e(\pi_e(\tau, z)) \\ \wedge \bigwedge_{r \in \mathcal{L}(c,e)} \big([Q_r(E_r; \chi_r) = 1] \wedge [\phi(w_r(E_r)) = 1] \wedge [\eta_r = 1]\big) \\ \wedge \bigwedge_{\substack{r \in \mathcal{L}(c,e) \\ g_r \neq \varnothing}} [\rho_r = 1]. \end{aligned} \tag{5}$$

Here $[Q_r = 1] \wedge [\phi(w_r) = 1]$ is exactly $[\widehat{r}_\chi = \mathsf{P}]$, so release applies $\phi$ to the witnessed residual of (4) and adds receipts. $\eta_r$ binds the evidence, completeness test, witness, outcomes, rule version, and provenance root. $\rho_r$ binds the applied guard, configuration, version, and the same provenance root; it establishes recorded treatment fidelity, not correctness or efficacy. Incomplete evidence, fail, unresolved, or a missing receipt blocks release. Isolation is an architectural assumption, not a sandbox-security result.

**Conditional safety proposition.** Assume for every residual requirement that (i) $\chi_r$ is adequate, so $r(\tau, z) = 0$ only if $\tau$ realizes a violating trajectory and every such trajectory crosses a declared channel; (ii) $Q_r$ soundly verifies capture on those channels over the claim-critical interval; and (iii) $w_r$ is pass-sound relative to $\chi_r$, so on such complete evidence $w_r(E_r) = \mathsf{P}$ rules out every declared-channel path to $r = 0$. Together (i)–(iii) give $Q_r(E_r; \chi_r) = 1 \wedge w_r(E_r) = \mathsf{P} \Rightarrow r(\tau, z) = 1$. If receipts are authentic and run-bound, rule versions match, and the evidence plane is isolated, then $R_{c,e} = 1 \Rightarrow Y_{c,e}^{\text{declared}} = 1$. Completeness of $\mathcal{C}(c)$ is additionally required to license the entire claim. A sound guard that mediates every member of a complete $B_r$ before irreversible effect makes $r = 0$ unreachable within the declared channel model.

Table 2: **Illustrative frozen rule card: preprocessing leakage.** Completeness is relative to the declared channel model.

| Element | Frozen content |
|---|---|
| Claim $c$ | Under stated sampling assumptions, the sealed-test score is untouched-sample evidence for a named population. |
| Requirement $r$ | No test-derived information influences preprocessing fit before fitted state is committed. |
| Necessity $n_r$ | Test-dependent fitting makes the test sample part of adaptation, so its score loses the claimed untouched-sample interpretation. |
| Non-entailment $\xi_r$ | Identical fitted parameters can arise from training-only and test-informed fits, so the same accepted artifact/score view has process preimages with $r = 1$ and $r = 0$. |
| Channel model $\chi_r$ | Declared ingress paths from test bytes or derived statistics to fit state: data handles and aliases, subprocess/IPC, network/cache imports, and fitted-state commit. |
| Cut and guard $B_r, g_r$ | Mediate every declared ingress before commit with train-scoped capabilities; deny and record unauthorized requests; commit freezes the fitted-state hash. |
| Evidence and $Q_r$ | Bind split hashes, capability events, denials, subprocess inheritance, fit commit, and stream closure. $Q_r = 1$ only when the coverage manifest accounts for every $\chi_r$ channel over the critical interval. |
| Witness and release | Fail on a recorded successful forbidden flow; pass only with complete training-only provenance; otherwise unresolved. A blocked attempt is evidence, but does not falsify $r$ unless attempt-prohibition is separately declared. Only nominal pass plus witnessed pass releases the claim. |

## 5 Empirical Evaluation

The experiments follow an evidence ladder, reported in the same order in Section 6: diagnostic disagreement, evidence sufficiency, frozen-roster recurrence, treatment fidelity, and architecture behavior. Unless stated otherwise, retrospective findings are conditional on their archived instruments and collection process and license neither prevalence, agent ranking, causal efficacy, nor transfer. They test components of a claim-to-control lifecycle, not a benchmark of agents. Appendix Tables 3 and 18 record each study's licensed conclusion and the validation status of each framework component.

### 5.1 Measures and adjudication

The archived study keeps two operational axes separate. *Artifact acceptance* asks whether the final deliverable satisfies its visible task contract. *Process evidence* asks whether the retained record

supports the named information, decision, feasibility, and lineage requirements. Their conjunction is the operational joint endpoint. In the 905-run discovery snapshot, the process axis is approximated by a narrow programmatic transcript-scope label. The discovery stage asks whether nominal artifact success can disagree with a separate process instrument; the artifact checker is the nominal interface $e$, and neither instrument is ground truth. Empirical "valid/invalid/unresolved" labels denote operational pass/fail/unresolved verdicts, not direct observation of the underlying Boolean requirement.

For cross-provider adjudication, we request three `gpt-5.6-sol[1m]` judgments at ultra reasoning and three `claude-opus-5[1m]` judgments at maximum reasoning for each of 150 opaque evidence packets, judging the two axes separately. An axis resolves only with at least 5/6 identical votes, which necessarily entails at least 2/3 matching votes within each provider. The joint verdict passes only when both axes pass, fails when either fails, and is unresolved otherwise, matching the strong-Kleene rule above. This strict multi-run consensus serves as a fallible operational adjudication reference, not human ground truth (Zheng et al., 2023). Only 92 rows overlap the discovery snapshot, and the label-enriched frame is not representative. It measures evidence sufficiency, not prevalence or agent quality. Appendix C gives the blinding, packet, and consensus protocol.

**Secondary human audit.** After freezing the model panel, one blinded human labeler labels the 150 archived packets and all 120 current-route packets. Frozen panel labels are joined afterward only for comparison. Joint verdicts use the same trivalent rule; this single reading is not consensus or ground truth (Appendix C).

### 5.2 Frozen prospective observational recurrence study

The recurrence study is prospective and protocol-frozen. Its frozen protocol covers a completed 120-endpoint source collection over 20 historically informed task-template lineages, two requested current routes, and three isolated sessions per lineage–route cell. Before panel labeling, eight lineages (48 endpoints) were assigned to rubric development and excluded from outcome analysis. The primary recurrence denominator is therefore 12 lineages and 72 endpoints, 36 per route. Recurrence is defined as at least two of three sessions within a lineage–route cell. Cross-route recurrence requires that criterion under both routes for the same lineage. The study can diagnose repeated failure on this frozen roster; it cannot estimate population prevalence or rank agents. Every source endpoint, development lineages included, carries six blinded judgment slots: 716/720 judgments parse. Four terminal-missing slots leave one primary endpoint ambiguous but still in the denominator, so 71/72 are panel-adjudicable.

### 5.3 Receipt-grounded controls

The control audit uses a separate 421-row intervention ledger. Fixed pair and duplicate rules leave 135 nominal pairs. Only 84 pairs record `applied=true` for the requested executable control; the other 51 are excluded from control contrasts. Only the 84 applied pairs enter descriptive contrasts; their receipts do not independently verify control-arm absence. The 12-case interface suite separately tests positive, targeted-negative, null-agent, and malformed-evidence behavior. Collection was adaptive, arm order was not randomized, and repeated executions share construction lineages. Appendix Sections A and B give the complete provenance flow.

**Architecture studies.** We freeze 16 constructed traces and apply six stream corruptions to each, yielding 112 trivalent-witness cells. A separate phase-boundary follow-up uses eight fresh generated selection lineages, two routes, and monitor-only versus typed-denial arms (32 endpoints); identical arm-blind instructions state the authorized split protocol. Its prespecified primary estimand averages paired joint-release differences across routes within lineage. Appendix D gives both protocols and a sensitivity-limited 72-endpoint precursor.

## 6 Results

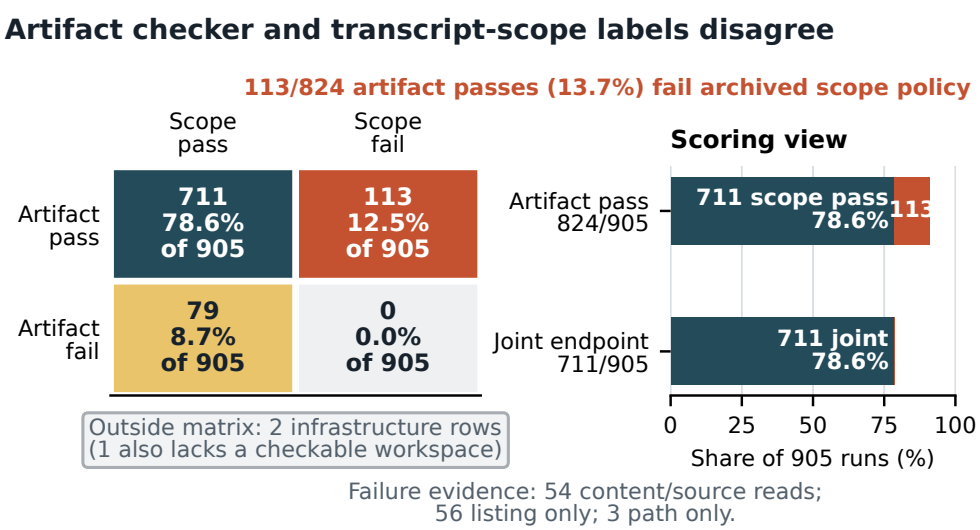


Figure 2: **Artifact success and recorded process evidence diverge.** The artifact checker accepts 824/905 runs, but only 711 also pass the transcript-scope instrument. The 113 artifact-only passes are candidate invalid successes under this narrow, partial operational proxy, not independently adjudicated process violations. The four matrix cells contain 903 runs; two infrastructure rows remain outside the matrix.

### 6.1 Diagnostic: artifact acceptance and process proxy disagree

Of 824 artifact-checker passes, 113 fail the separate process-scope proxy (Figure 2). This establishes a diagnostic disagreement: nominal acceptance did not certify the additional proxy requirement. It is not an estimate of the true invalid-success rate. The disagreements contain 54 content/source-read records, 56 listing-level records, and three path-only records. The empty artifact-fail/scope-fail cell is not hard-coded: route-dependent observability and the proxy's binary treatment of missing evidence are analyzed in Appendix A.

The supported result is narrow: artifact acceptance does not certify the recorded process requirement in this archive.

### 6.2 Evidence sufficiency: incomplete traces prevent positive certification

The panel produces 895/900 parsed judgments and resolves artifact acceptance for 117/150 runs. The panel rubric mirrors (4): a process-pass verdict is prohibited for an incomplete structured trace. The model panel resolves process evidence for 21/34 complete traces and 7/116 incomplete traces (Figure 3); the human labeler resolves 33/34 and 5/116. Every resolved incomplete-trace process verdict is fail: exactly the asymmetry (4) specifies, in which incomplete evidence cannot establish pass while a concrete violation certificate still establishes fail. The human and model panel agree on 133/150 process verdicts and on 24/25 where both resolve. This is the operational consequence of the frozen semantics, not evidence that trace completeness causally improves judging.

### 6.3 Frozen-roster recurrence and a complete-trace case

On this frozen roster, route labels are requested, not authenticated served-model identities. Appendix Table 10 separates endpoint classification from prespecified $\geq 2/3$ within-lineage recurrence; cross-route recurrence requires the same lineage under both requested routes.

The 72 endpoints include 23 substantive-invalid outcomes and three panel-adjudicated invalid successes. The human labeler also labels all three as candidates and flags four more (one panel-ambiguous and three not classified as invalid success by the panel). The human candidates recur in one requested-`claude` and two requested-`codex` lineage–route cells, including one lineage under both routes; the panel counterpart is zero, one, and zero.

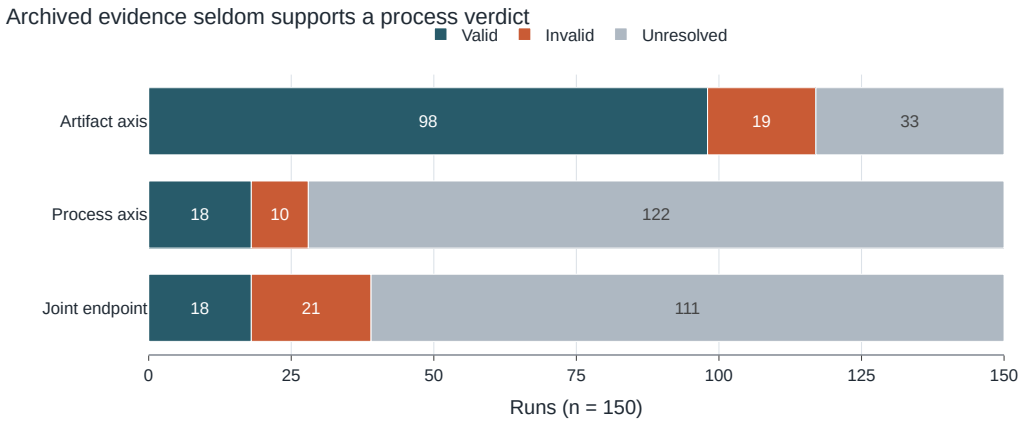


Figure 3: **Archived evidence supports far fewer process verdicts than artifact verdicts.** Each bar partitions all 150 runs under the frozen 5/6 cross-provider consensus rule. Unresolved is an explicit outcome, not a pass. Artifact and joint verdicts resolve for 117 and 39 runs. Process evidence resolves for 21/34 complete and 7/116 incomplete traces (28/150 overall).

**Concrete case: accepted artifact, contradicted process.** In one complete-trace hospital-escalation run (packet A207), the opening example of this paper, the final manifest reported validation-based selection and the visible artifact test passed. Before freezing the model, however, the agent loaded the sealed deployment benchmark and printed benchmark accuracy for every candidate feature. The artifact interface cannot distinguish this run from a clean one; the retained trace can. Both the panel and the human labeler mark the process requirement fail, the panel unanimously across all six slots. A claim-preserving implementation would withhold the benchmark capability until a model-selection commit is receipted.

### 6.4 Treatment fidelity: requested controls are not applied controls

Only 84/135 nominal intervention pairs contain a receipt showing application of the requested executable control; the other 51 cannot enter executable-control contrasts. Receipt-qualified outcomes are mixed and lineage-sparse, with each large favorable cell confined to one baseline-first construction lineage. Receipts establish recorded treatment application, not correctness or efficacy; Appendix B gives full accounting.

### 6.5 Corrupted evidence blocks; clean-protocol runs remain releasable

All 112 constructed integrity cells match the frozen oracle: 8/8 intact compliant or allowed-control streams release, 8/8 intact violations fail, and all 96 corruptions become unresolved and block; the 12-case executable interface suite likewise matches its oracle in all 12 cases. Separately, all 32 clean-protocol prospective endpoints remain releasable in both arms, and no endpoint attempts prohibited precommit evaluation, so the prospective study establishes clean-path compatibility, not targeted guard efficacy or equivalence. Both arms stated the split protocol explicitly, so the null bears on compliance with a stated instruction, not on behavior under an unstated one. A one-shot audit budget present in both arms did act on agent-initiated behavior, denying two repeat postcommit audit requests; both endpoints still released.

## 7 Related Work

**Scientific validity, leakage, and agent evaluation.** Classical validity links observations to claims; ML leakage shows how accepted results can violate those links (Cronbach & Meehl, 1955; Shadish et al., 2002; Kapoor & Narayanan, 2023). Autonomous systems now execute multi-step scientific workflows (Boiko et al., 2023), and AI Scientist extends an ML pipeline through experimentation

to manuscript generation (Lu et al., 2024). Capability benchmarks measure completion (Huang et al., 2024; Chan et al., 2025; Jing et al., 2025; Jimenez et al., 2024). Complementary work targets engineering-process discipline (Madiraju & Madiraju, 2026), experimental fidelity (Yu et al., 2026), step-level process supervision (Lightman et al., 2024), contamination mitigation (Jacovi et al., 2023), and discovery certification (Ning et al., 2026). We derive the residual scientific requirements that task acceptance cannot certify.

**Provenance and evidence systems.** Scientific-workflow provenance records derivation and execution history (Davidson & Freire, 2008), which PROV-DM formalizes as entities, activities, agents, and relations (Moreau & Missier, 2013). A recent survey organizes evidence tracing in LLM agents (Wang et al., 2026); agent systems maintain trajectory and claim–evidence structures (Zhuang et al., 2026; Ren et al., 2026), ControlA adds provenance-aware safeguards (Gueroudji et al., 2025), and a conceptual framework argues that evidence licenses a claim (Li, 2026). These record what happened, constrain workflows, or argue the principle; our upstream step derives which records and controls a particular claim and evaluator require.

**Runtime enforcement.** Design by contract makes a specification a checkable assertion (Meyer, 1992), while reference-monitor principles, information-flow control, and runtime monitors motivate complete mediation, information-flow restriction, and execution-time policy enforcement (Saltzer & Schroeder, 1975; Sabelfeld & Myers, 2003; Schneider, 2000). We claim neither a new reference monitor nor a host-security theorem: sandboxes constrain host-level execution, whereas our guards mediate the methodological boundaries a claim requires at a complete cut set. The claim-relative path $(c, e) \to \mathcal{L}(c, e) \to \Gamma_{c,e}$—not provenance or monitoring alone—is the contribution.

## 8 Limitations

The synthetic or fictionalized empirical frames are not representative of natural science. Claim-to-rule derivation is author-built and AI-assisted, not independently replicated on unseen claims, so requirement coverage remains unvalidated. Historical controls are nonrandomized and lineage-sparse; 116 archive traces are incomplete; the human audit is one unreplicated reading; routes are unauthenticated; and repeated sessions are dependent. LLM-as-judge verdicts carry documented systematic biases (Zheng et al., 2023), and cross-provider consensus cannot detect those the two providers share: 5/6 agreement reduces sensitivity to single-run stochastic variation, not inductive bias common to frontier LLMs. That single human reading, agreeing on 24/25 jointly resolved process verdicts, is a non-model comparison but cannot rule such bias out. No prevalence, ranking, causal, or transfer claim follows.

Fault injection is constructed, and the clean-path follow-up contains no target attempt; neither establishes efficacy or natural fault rates. No witness repairs an omitted requirement or channel: $Q_r = 1$ certifies only frozen $\chi_r$. No rule here has been revised, frozen at $v_r + 1$, and re-evaluated, so the development protocol is specified but not exercised.

## 9 Discussion and Conclusion

A model and score do not by themselves determine the scientific claim they support: the relevant requirements depend on the claim, while what remains latent depends on the interface. The framework is executable when a requirement maps to a declared channel model and complete mediation cut; target choice and scientific assumptions remain semantic commitments that receipts cannot guarantee.

Fail-closed release makes a deliberate trade: missing evidence cannot certify compliance, although concrete evidence can contradict it. Retrospective traces cannot recover events never recorded; the remedy is prospective instrumentation or a narrower claim. Our evidence supports output–process divergence, evidence-sufficiency and treatment-fidelity gaps, constructed stream integrity, frozen-roster recurrence, and clean-path compatibility. Each names a different link between claim and release, which is why the architecture combines mediation, completeness testing, control receipts and abstention. Independent rule derivation, known-truth witness calibration, natural workflows, and target-boundary interventions remain the decisive tests.

## ETHICS STATEMENT

The evaluated tasks are synthetic or fictionalized. The human annotation is a methodological research instrument: the audit retains categorical human judgments, not annotator identity, demographics, or behavioral outcomes. Raw execution traces can nevertheless contain credentials, private paths, or protected material, so any reviewer artifact requires privacy, licensing, binary-file, anonymity, and redaction reviews. Publishing guard details may also enable gaming. The experiments establish neither deployment safety, sandbox security, nor human impact.

## AI USE STATEMENT

OpenAI Codex was used as an authoring and research-support tool for research questions and hypotheses, the conceptual framework and mathematical formulation, methodology and experimental design, software implementation and testing, generation and refinement of synthetic research tasks, data cleaning and reformatting, data analysis and interpretation, figure and table generation, literature discovery and reference support, brainstorming, and manuscript structure, drafting, editing, title development, and keyword selection. It also assisted the formal statement and revision of the conditional safety proposition. It was not used for translation, transcription, survey or interview questions, or human-subject analysis; those tasks were absent or not applicable. Separately, configured OpenAI and Anthropic models were used as blinded model-panel judges; their consensus is treated as an operational reference label, not human ground truth. References to Claude, Codex, and Gemini as evaluated agents concern study subjects. OpenAI and Anthropic models critiqued interim drafts and informed their organization, terminology, and evidence presentation. The authors reviewed all AI-assisted text and code, regenerated numerical claims from bound inputs, tested executable analyses, and checked citations against source records. The authors retain responsibility for the paper's content, including its claims, evidence, citations, and compliance.

## REPRODUCIBILITY STATEMENT

The appendices document the evidence inventory, row accounting, lineage groups, retrospective analyses, model and human adjudication, current-route recurrence, constructed stream faults, and the two frozen phase-boundary studies. Hash-bound inputs and deterministic scripts reconstruct the locked snapshot, model-panel and human-labeler analyses, tables, figures, and sealed prospective outcomes. An anonymous reviewer artifact remains contingent on privacy, licensing, anonymity, and binary-file review; independent reproduction is not yet claimed.

## REFERENCES

Avrim Blum and Moritz Hardt. The ladder: A reliable leaderboard for machine learning competitions. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1006–1014, 2015. URL `https://proceedings.mlr.press/v37/blum15.html`.

Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624:570–578, 2023. doi: 10.1038/s41586-023-06792-0.

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.

Lee J. Cronbach and Paul E. Meehl. Construct validity in psychological tests. *Psychological Bulletin*, 52(4):281–302, 1955.

Susan B. Davidson and Juliana Freire. Provenance and scientific workflows: Challenges and opportunities. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, pp. 1345–1350, 2008. doi: 10.1145/1376616.1376772.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015. doi: 10.1126/science.aaa9375.

Amal Gueroudji, Tanwi Mallick, Renan Souza, Rafael Ferreira da Silva, Robert Ross, Matthieu Dorier, Philip Carns, Kyle Chard, and Ian T. Foster. ControlA: Agentic workflow control mechanisms for reliable science. In *2025 IEEE International Conference on e-Science (e-Science)*, pp. 415–426, 2025. doi: 10.1109/eScience65000.2025.00086.

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

Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. MLAgentBench: Evaluating language agents on machine learning experimentation. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav Goldberg. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5075–5084. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.308. URL https://aclanthology.org/2023.emnlp-main.308/.

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.

Liqiang Jing, Zhehui Huang, Xiaoyang Wang, Wenlin Yao, Wenhao Yu, Kaixin Ma, Hongming Zhang, Xinya Du, and Dong Yu. DSBench: How far are data science agents from becoming data science experts? In *International Conference on Learning Representations*, 2025.

Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, 2023.

Stephen Cole Kleene. *Introduction to Metamathematics*. North-Holland, Amsterdam, 1952.

Hongmin Li. The calibration turn in AI-assisted research: A conceptual and methodological framework for evidence-licensed claims. *arXiv preprint arXiv:2606.31273*, June 2026. URL https://arxiv.org/abs/2606.31273v1. Version 1, submitted 30 June 2026 (2026-06-30T07:46:54Z).

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024. URL https://iclr.cc/virtual/2024/poster/17549.

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

Meher Bhaskar Madiraju and Meher Sai Preetam Madiraju. RigorBench: Benchmarking engineering process discipline in autonomous AI coding agents. *arXiv preprint arXiv:2606.22678*, 2026.

Bertrand Meyer. Applying "design by contract". *Computer*, 25(10):40–51, 1992. doi: 10.1109/2.161279.

Luc Moreau and Paolo Missier. PROV-DM: The PROV data model. W3C recommendation, World Wide Web Consortium, April 2013. URL https://www.w3.org/TR/2013/REC-prov-dm-20130430/.

Jingjie Ning, Shanshan Zhong, Xiaochuan Li, and Ji Zeng. Scores alone do not prove discovery: The discovery certification protocol for auditing AI research agents. *arXiv preprint arXiv:2609.09219*, September 2026. URL https://arxiv.org/abs/2609.09219v1. Version 1, submitted 7 September 2026 (2026-09-07T00:16:28Z).

Zhenjiang Ren, Ruiji Li, Xujing Zhang, Ziliang Pang, Shuo Ren, and Jiajun Zhang. EviGraph: Evidence-guided autonomous research agents. *arXiv preprint arXiv:2608.04738*, August 2026. URL https://arxiv.org/abs/2608.04738v2. Version 2, revised 6 August 2026; version 1 submitted 5 August 2026 (v1: 2026-08-05T12:02:55Z; v2: 2026-08-06T07:15:59Z).

Andrei Sabelfeld and Andrew C. Myers. Language-based information-flow security. *IEEE Journal on Selected Areas in Communications*, 21(1):5–19, 2003. doi: 10.1109/JSAC.2002.806121.

Jerome H. Saltzer and Michael D. Schroeder. The protection of information in computer systems. *Proceedings of the IEEE*, 63(9):1278–1308, 1975. doi: 10.1109/PROC.1975.9939.

Fred B. Schneider. Enforceable security policies. *ACM Transactions on Information and System Security*, 3(1):30–50, 2000. doi: 10.1145/353323.353382.

William R. Shadish, Thomas D. Cook, and Donald T. Campbell. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin, 2002.

Yiqi Wang, Jiaqi Zhang, Taotao Cai, Zirui Liu, Qingqiang Sun, Zequn Sun, Zhangkai Wu, Manqing Dong, Mingkai Zheng, Xuefei Yin, and Yanming Zhu. From agent traces to trust: A survey of evidence tracing and execution provenance in LLM agents. *arXiv preprint arXiv:2606.04990*, 2026. URL https://arxiv.org/abs/2606.04990v4. Version 4.

Lezhi Yu, Xiaogang Xu, Yuhua Zhou, Shuibing He, and Aimin Pan. Beyond execution: Auditing experimental fidelity in LLM-driven scientific research. *arXiv preprint arXiv:2608.26753*, 2026.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623, 2023.

Zijie Zhuang, Changxin Lao, Pengbo Xu, Hanwen Xu, Ruochen Yang, Yingzhi He, Peng Zhang, Jiangxia Cao, Yusheng Huang, Guohong Mu, Jian Liang, Ruiming Tang, Shuang Yang, Zhaojie Liu, Wenwu Ou, and Kun Gai. From trajectories to evidence: Auditable experimental records for industrial research agents. *arXiv preprint arXiv:2608.05235*, August 2026. URL https://arxiv.org/abs/2608.05235v1. Version 1, submitted 5 August 2026 (2026-08-05T13:37:38Z).

# A EVIDENCE ACCOUNTING

Table 3: **Core empirical evidence.** Each analysis has a distinct unit and licensed conclusion; one result cannot upgrade another analysis's labels, independence, or claim scope.

| Stage | Material and unit | Primary readout | Licensed conclusion |
|---|---|---|---|
| **Discover** | 905 archived runs over 57 tasks; 903 jointly scored | Artifact–scope contingency | Whether nominal success and a separate narrow process instrument can disagree |
| **Adjudicate** | 150 archived runs; six blinded high-reasoning judgment slots per run, 895/900 parsed | Trivalent artifact-acceptance, process-evidence, and joint verdicts | What the retained evidence lets a strict cross-provider consensus resolve |
| **Human audit** | 270 unique opaque packets; one blinded human labeler | Direct trivalent verdicts and agreement with the frozen panel | Secondary evidence-sufficiency and fixed-roster sensitivity; no ground-truth, inter-rater, prevalence, or ranking claim |
| **Enforce** | 421 intervention records; 135 nominal pairs; 84 receipt-qualified applied-control pairs | Receipt fidelity and joint artifact–witness success | Which requested controls have qualifying application receipts and how three rule interfaces behave in their audited settings |
| **Diagnose recurrence** | 72 endpoints; 12 historically informed lineages; two routes; three sessions per cell | Endpoint classes; prespecified $\geq 2/3$ recurrence | Fixed-roster recurrence only; no prevalence, ranking, or causal claim |

Table 4: Evidence objects have distinct licensed uses but are not necessarily disjoint: all 905 locked run IDs are nested in the later 2,012-run archive. Nested counts do not create replication or upgrade another row's claim status.

| Evidence object | Unit and scope | Inclusion and labels | Licensed use |
|---|---|---|---|
| Locked diagnostic snapshot | 905 run rows; 57 tasks; 298 Claude, 305 Codex, 302 Gemini rows | Archived task-checker and programmatic transcript-scope labels; 903 matrix-eligible rows plus two infrastructure rows outside the matrix, one without a checkable workspace | Archive-conditional artifact/policy contingency only |
| Later mechanism archive | 2,012 run rows across 57 tasks | Adaptive collection; programmatic graders; contains all 905 locked run IDs | Mechanism discovery, not an independent replication or inferential sample |
| Historical intervention audit | 421 primary ledger rows; 135 nominal complete pairs; task IDs collapsed into post-hoc audit lineages | 84 applied executable pairs, 42 paired non-applied registry requests, and 9 prompt/schema pairs; nonrandom arm order (106/28/1) across all 135 nominal pairs, and 55/28/1 among the 84 applied pairs | Treatment-fidelity and descriptive lineage/partition contrasts only |
| Earlier frozen, provenance-ineligible acquisition attempt (v1) | 180 planned endpoint slots; 120 usable Claude/Codex endpoints; 60 Gemini endpoints technically missing | Frozen wrapper used incompatible command-line flags; no post-hoc repair | Executed-calls and missingness diagnostics only; not pooled with the corrective cohort |
| Gemini corrective cohort | 60 endpoints / 30 pairs | Separately hash-bound cohort with programmatic artifact-and-scope labels | Technical completion and descriptive contrast audit only |
| Model-panel agreement audit | 150 archived runs; 10 families $\times$ 3 agent labels $\times$ 5 rows | 900 planned slots, all terminal: 895 parsed votes and 5 exhausted missing slots across 4 rows; artifact/process/panel joint verdict resolves for 117/28/39 rows | Evidence-sufficiency and archived-label agreement only; no prevalence, ground-truth accuracy, or cross-agent claim |
| Frozen prospective observational recurrence study | Frozen 120-endpoint source collection over 20 historically informed task lineages, two requested routes, and three fresh sessions per cell | Eight lineages (48 endpoints) assigned to rubric development before panel labeling; primary analysis fixes 12 lineages, 72 endpoints, and 36 matched task–session pairs; 716/720 panel judgments parsed, four terminal-missing slots affect one endpoint, and 71 endpoints are adjudicable | Descriptive endpoint, paired, and recurrence results on the frozen roster only; no prevalence, agent-ranking, or causal claim |
| Architecture validation studies | 112 constructed stream-integrity cells; 72-endpoint phase-gate precursor; 32-endpoint corrected follow-up | Frozen and sealed outcomes, including the precursor's literal-prose sensitivity | Declared verifier behavior and explicit-protocol compliance only; no pooled guard-efficacy, equivalence, ranking, or broad-transfer claim |

Table 5: Retrospective provenance flow. Each stage records what was recovered from the archive and the strongest boundary that survives. The sequence is an audit reconstruction, not a prospective collection protocol.

| Stage | Recorded operation | Evidential boundary |
|---|---|---|
| Task construction | Authors built synthetic or fictionalized contracts with declared inputs, outputs, procedures, and forbidden scope. | Constraint-first design is documented, but construction was adaptive and neither the task portfolio nor ten-family coding received independent construct validation. |
| Historical execution | Agent calls were accumulated across tasks, regimes, and follow-ups; repeated calls share tasks and settings. | Assignment, stopping, and arm order were not randomized; exposure is unbalanced and exact requested-model identity is absent for 338/905 snapshot rows. |
| Recorded scoring | Task-specific artifact checks and an archived programmatic transcript-scope policy supplied separate labels. | The scope label is a partial operational instrument, not a complete process measure; raw transcripts were not independently re-annotated for this paper. |
| Snapshot binding | A fixed list of 905 run IDs was recorded with a pre-existing summary and content digests. | The list defines the analysis denominator and the digests make later byte drift detectable; absent external attestation, neither proves acquisition time, immutability, or preregistration. |
| Intervention re-audit | Raw run receipts determined whether the requested executable control recorded `applied=true`; fixed pair and duplicate rules yielded 135 nominal pairs and 84 applied-control pairs. | Non-applied and prompt-only arms remain outside executable-control contrasts; order is confounded. |
| Dependence audit | Task variants were collapsed by a fixed post-hoc lookup, then task contrasts were averaged within lineages. | Shared construction ancestry and outcome-blind grouping were not established, so lineage summaries remain descriptive and cannot license transfer. |

**Row-to-claim trace.** For a snapshot row, an accepted task-specific artifact check sets $Y^{\text{artifact}} = 1$; the archived transcript policy separately supplies $Y^{\text{scope}}$. A $(1, 0)$ row contributes once to the 113 artifact-only cell, and its stored trace category contributes once to the 54/56/3 evidence split; neither step turns the policy label into independently adjudicated exposure. A historical intervention record enters a contrast only when its baseline and intervention keys match and the requested guard's raw receipt records `applied=true`. The resulting paired difference is averaged within its post-hoc task lineage before any equal-lineage summary. Thus an artifact pass, a policy disagreement, an applied-control pair, and a transfer unit are four different evidential objects rather than interchangeable successes.

**Evidence-category provenance.** The 54/56/3 split is copied from the locked snapshot's row-level `transcript_failure_evidence` records, not newly assigned for this paper: 54 contain content-read or source-inspection evidence, 56 contain listing evidence only, and three contain path references only. Its three categories are:

- `content_read_or_source_inspection`: an affirmative recorded content-read or source-inspection signal;
- `inventory_or_listing`: listing-level evidence without that signal; and
- `path_reference`: path-reference evidence only.

Each of the 113 policy-failure run IDs appears exactly once in those records, and the stored evidence flags agree with the three categories. These are programmatic evidence categories under the archived

policy, not independent human adjudications of exposure or process invalidity. The low-level source contains 904 rows with checker output; 903 non-infrastructure rows are eligible for the two-axis matrix, while both infrastructure rows remain outside it.

**Strong-evidence sensitivity.** An explicitly post-hoc nested-threshold analysis retains 113/824 artifact passes under any recorded evidence, 110/824 after excluding path-only evidence, and 54/824 when requiring content-read or source-inspection evidence. The strongest subset spans 12 tasks and three recorded axes; its largest task contributes 7/54 rows, but all 54 carry the Codex route label. Tightening the recorded-evidence threshold therefore preserves the existence of the artifact/process measurement gap while narrowing its scope; against the 79 artifact-fail/scope-pass rows, the signed direction reverses under the strict 54-row threshold and is therefore not robust. These filters do not validate the archived proxy, identify causal exposure, or support prevalence or agent comparisons.

**Concentration.** This discrepancy is highly concentrated in the adaptive archive: by recorded agent label, 111/113 artifact-accepted policy failures occur among 305 Codex rows, 2 among 302 Gemini rows, and 0 among 298 Claude rows. By recorded family, data scope (47) and preprocessing (38) account for 85/113. These are composition facts, not agent or family comparisons: exposure was neither balanced nor randomized, and all Codex rows lack a requested model ID.

**Trace observability and the empty cell.** A post-hoc source audit finds provider-agnostic checker code but provider-dependent evidence: all 305 Codex logs contain rendered tool-call and result markers, versus 0/298 Claude and 0/302 Gemini logs. The checker scans a merged CLI text stream without a completeness gate, so missing or unrendered actions become "no detected violation," not unresolved. This makes the 111/2/0 route concentration inseparable from detection opportunity.

The implementation permits an artifact-fail/scope-fail outcome; the observed zero is not hard-coded. Of 79 eligible artifact failures, 48 have no applicable archived scope policy and 31 have no recorded policy match; only 12/79 retain the structured tool-call/result format. Current-source replay reproduces 109/113 archived policy failures and still no double failures. The four label differences cannot be attributed with retained provenance, so the locked labels remain unchanged. These audits explain the recorded matrix but cannot infer unobserved behavior or process compliance.

Table 6: Primary-axis breakdown of the 903 eligible rows. Artifact-only means artifact-accepted/scope-invalid; scope-only is the converse. Data scope contributes the largest event count (62/192), whereas label leakage has the highest observed within-axis rate (11/30). Unequal axis denominators (15–241 rows) and adaptive composition preclude prevalence or risk comparisons.

| Primary task axis | Eligible | Artifact-only | Scope-only | Any disagreement |
|---|---|---|---|---|
| Artifact integrity | 149 | 1 | 13 | 14/149 (9.4%) |
| Category generalization | 15 | 0 | 0 | 0/15 (0.0%) |
| Data scope | 241 | 47 | 15 | 62/241 (25.7%) |
| Global planning | 78 | 0 | 17 | 17/78 (21.8%) |
| Label leakage | 30 | 0 | 11 | 11/30 (36.7%) |
| Preprocessing leakage | 171 | 38 | 7 | 45/171 (26.3%) |
| Selection bias | 30 | 0 | 7 | 7/30 (23.3%) |
| Split validity | 60 | 0 | 0 | 0/60 (0.0%) |
| Temporal validity | 47 | 12 | 2 | 14/47 (29.8%) |
| Validation protocol | 82 | 15 | 7 | 22/82 (26.8%) |
| Total | 903 | 113 | 79 | 192/903 (21.3%) |

**Task- and axis-level sensitivity.** Run weighting gives $192/903$ (21.26%) disagreement; the unweighted mean of 57 task rates is 20.16%. Omitting each task in turn yields 20.36–21.62% and 19.53–20.52%, respectively, with at most 13/192 events from one task. Thus no task drives the aggregate under this archive mixture; these post-hoc ranges are neither inferential intervals nor evidence of stability in another task population.

Every single-axis deletion leaves a nonzero disagreement rate, but the magnitude of disagreement and, under equal-task weighting, the direction of the signed artifact-minus-scope gap depend on the

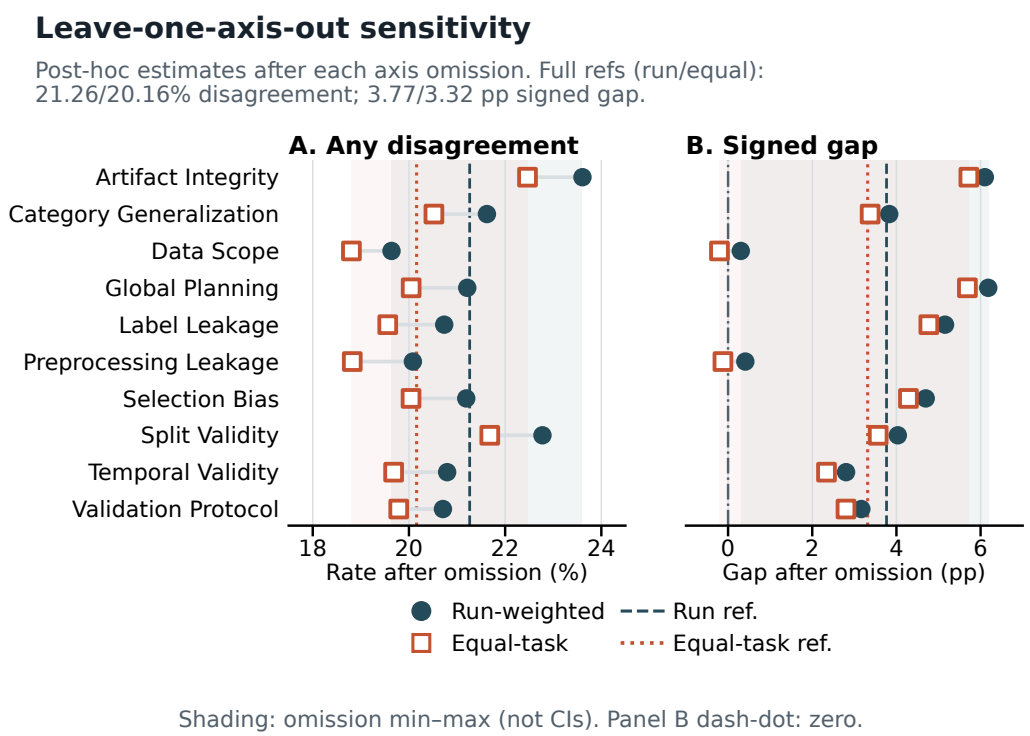


Figure 4: Exhaustive leave-one-recorded-primary-axis-out sensitivity for the fixed 903-row eligible archive. Each row deletes all eligible runs assigned to the named primary axis, then recomputes (A) any artifact/scope disagreement and (B) the signed artifact-minus-scope gap (artifact-only minus scope-only shares, in percentage points). Filled circles and open squares use run and equal-task weighting; dashed and dotted lines mark the run-weighted and equal-task full-archive values, respectively. Any-disagreement rates span 19.64–23.61% run-weighted and 18.80–22.47% equal-task. The equal-task signed gap becomes slightly negative only after omitting data scope or preprocessing leakage. These exhaustive post-hoc deletion ranges cover the ten recorded primary-axis labels and diagnose composition; they are not uncertainty intervals, independent replications, population estimates, causal effects, or model comparisons.

observed axis mixture (Figure 4). These are deterministic reweightings of the same adaptive archive, not inferential intervals or external stability evidence.

**Label-flip sensitivity.** Figure 5 complements the fixed-label task/axis deletion analysis by holding the artifact label fixed and perturbing only the archived programmatic transcript-scope label. A 2.5% budget cannot eliminate disagreement or reverse its signed direction; a 5% budget cannot eliminate disagreement but can reverse direction. Separately, 34 one-sided invalid-to-valid flips make the signed gap zero and 35 reverse it. Under equal-task weighting, fixed-label disagreement is 20.1591%; at the same 45-flip budget, its exact sharp marginal bounds, shown rounded, are 14.8542–25.6145%, while the equal-task mean signed gap spans $-1.8444$ to 8.8135 percentage points. For the lower and upper endpoints separately, optimizing over the choice of at most one task (or primary-axis) cluster and three actions—leave unchanged, set every matrix-eligible scope label valid, or set every such label invalid—gives run-weighted disagreement bounds of 184–210 out of 903 for tasks and 145–356 out of 903 for axes. The task endpoints recode 8 and 24 labels in different tasks; both axis endpoints select data scope and recode 47 and 194 labels, so these whole-cluster stresses are not comparable to a fixed 45-row budget. These are separate sharp finite-archive assumption bounds, not sampling uncertainty, estimated error rates, or human-validated corrections; they do not identify which labels, if any, are wrong.

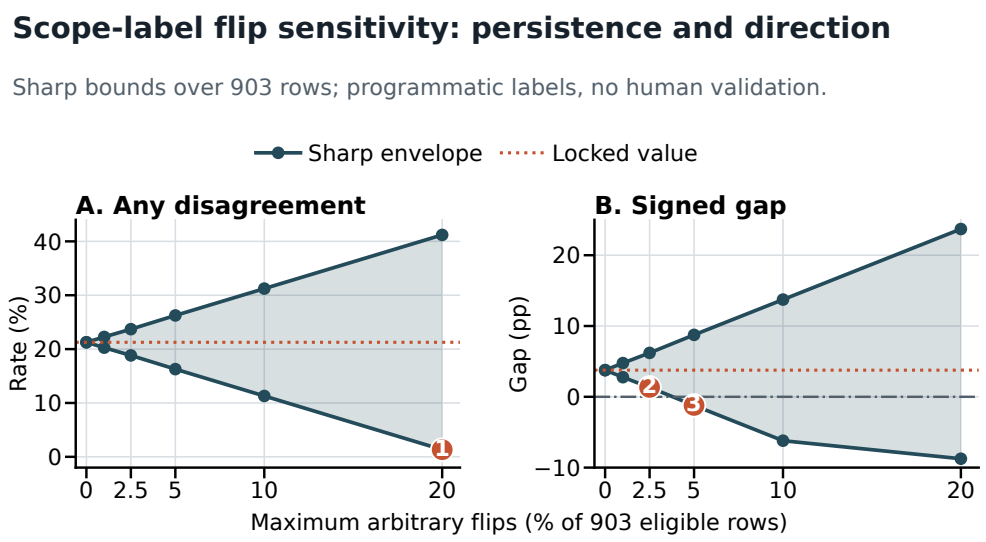


Figure 5: Run-weighted deterministic sensitivity to arbitrary flips of the 903 archived programmatic transcript-scope labels, with artifact acceptance held fixed. Panel A bounds the any-disagreement rate; Panel B bounds the signed artifact-minus-scope gap. Numbered anchors separate two conclusions: at **[1]** 20%/180 flips the sharp minimum disagreement remains 12/903 (1.33%); at **[2]** 2.5%/22 flips the signed-gap lower bound is $+12/903$, whereas at **[3]** 5%/45 flips its envelope first includes zero on the evaluated grid ($-11/903$ to $+79/903$). Thus disagreement remains nonzero on this grid while its direction can reverse under assumed recodings; this does not assert that any label is wrong or that reversal occurred. Points are evaluated budgets and shading joins separate sharp marginal extrema, which need not be jointly attainable. These are post-hoc assumption envelopes, not confidence intervals, calibrated error rates, human validation, population estimates, causal effects, or model comparisons.

### A.1 EXECUTABLE INTERFACE SMOKE SUITE

The executable interface suite contains 12 scripted cases across access, registry, and global-solution checker interfaces. All 12 match expectations: three positive controls pass, three targeted-negative controls fail, and three null-agent plus three malformed-evidence controls fail closed. This validates the scripted interfaces only; it is not evidence about historical graders or agent behavior.

## B DETAILED HISTORICAL COHORT ACCOUNTING

Each intervention-ledger row has a unique `row_id` and `run_id`; the reporting-unit key is the tuple of `source_phase`, `task_id`, `family`, `agent`, `seed`, `regime`, and `guard_id`. Baseline and nominal-intervention rows pair after dropping `regime` and `guard_id`. The five ledgers' 421 rows comprise 163 Claude, 129 Codex, and 129 Gemini records spanning 18 task IDs, 14 post-hoc audit lineages, and four families. Rows record the reported client/model string, timestamps, outcomes, hashes, and planned settings. The historical runner commands did not transmit the recorded temperature, top-$p$, or token-limit fields to provider CLIs; the `seed` was a repeat label, not an exposed provider RNG seed. All 135 complete pairs agree on task contract, reported client/model string, plan fields, and package hash.

The flow closes exactly. Of 421 rows, 147 are outside the baseline/designated-intervention pool: 57 prompt-only checklist rows, 45 checker-only rows, and 45 post-feedback reruns whose runtime records also show no applied registry guard. The sole duplicated reporting unit is resolved to its earliest record, and both distinct runs passed. Under that rule, the one-lineage applied registry-guard diagnostic is $1/6 \to 5/6$; excluding the duplicated logical cell instead gives $0/5 \to 4/5$, without changing the one-lineage limitation. The remaining 274 records include that duplicate; discarding it leaves 273: 135 baselines and 138 nominal intervention endpoints—84 applied executable guards, 45 non-applied registry requests, and nine prompt/schema interventions. Of these, 270 form 135 pairs; three non-applied artifact-integrity endpoints are unpaired. Hence $421 = 147 + 1 + 270 + 3$, where the 1 is the discarded duplicate, with no baseline-only key and no diagnostic regime silently contributing to an effect.

The 36 audit-held-out registry requests all record `applied=false`; their observed $91.7\% \to 80.6\%$ arm-label contrast is therefore not a guard effect. This is precisely the phantom intervention that receipt-gated accounting excludes.

Table 7: Runtime-receipt audit. Numeric contrasts appear only for applied executable controls; non-applied and prompt-only records are separate non-estimable or non-executable findings. Rates are pair-weighted descriptive summaries, with no lineage-level inference or pair-level $p$-values. "Held-out" denotes an inherited audit partition, not independent authorship or guard-construction holdout. No row is confirmatory.

| Family | Partition | Pairs / lineages | Outcome and boundary |
|---|---|---|---|
| *Applied executable-control contrasts* | | | |
| Artifact integrity | Applied construction | 6 / 1 | 16.7% → 83.3% (+66.7 pp); stated duplicate rule; one-lineage diagnostic. |
| Data scope | Construction | 36 / 3 | 63.9% → 52.8% (−11.1 pp). |
| Data scope | Audit-designated held-out | 36 / 4 | 66.7% → 72.2% (+5.6 pp); pooled data-scope difference −2.8 pp. Inherited task-specific partition, not guard-construction holdout. |
| Global planning | Construction | 6 / 1 | 0.0% → 100.0% (+100 pp); one lineage, so no transfer claim. |
| *Non-estimable or non-executable findings* | | | |
| Artifact integrity | Requested; not applied | 42 / 4 | 0/42 requested registry guards applied (6 construction; 36 audit-held-out); no executable effect is estimable. |
| Split validity | Prompt arm | 9 / 1 | Nine paired prompt/schema records; no executable-control receipt or effect. |

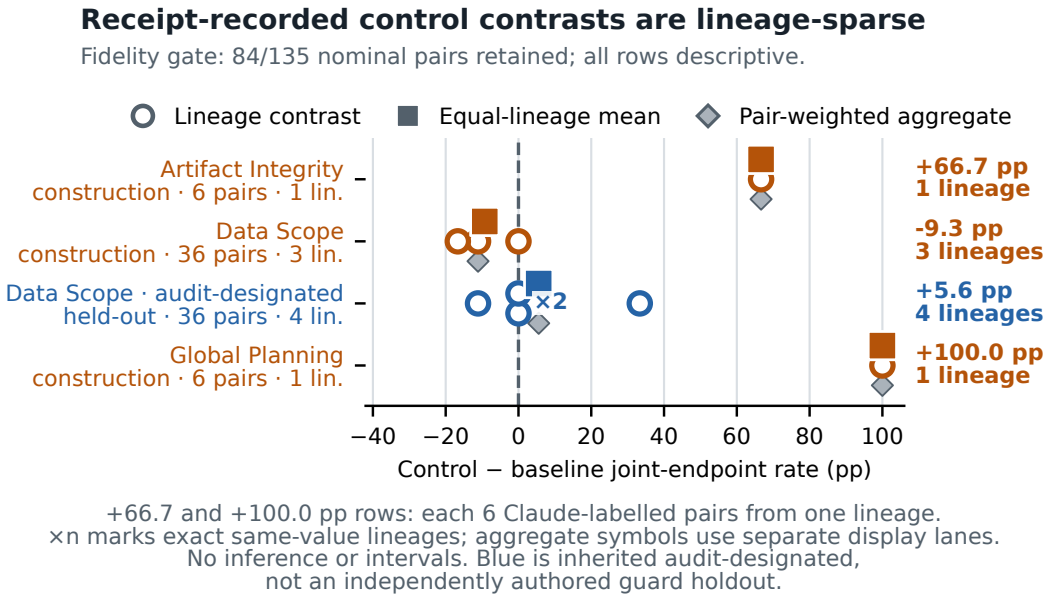


Figure 6: **Receipt-qualified control outcomes are mixed and lineage-sparse.** Open circles are post-hoc lineage contrasts; squares are equal-lineage means; diamonds are pair-weighted summaries. Artifact integrity and global planning have large favorable descriptive contrasts but one lineage each. For data-scope construction, the annotated $-9.3$ pp is the equal-lineage mean; its pair-weighted diamond is $-11.1$ pp, as reported in the main text. Data-scope contrasts differ across partitions. The 51 pairs without an executable-control receipt are excluded.

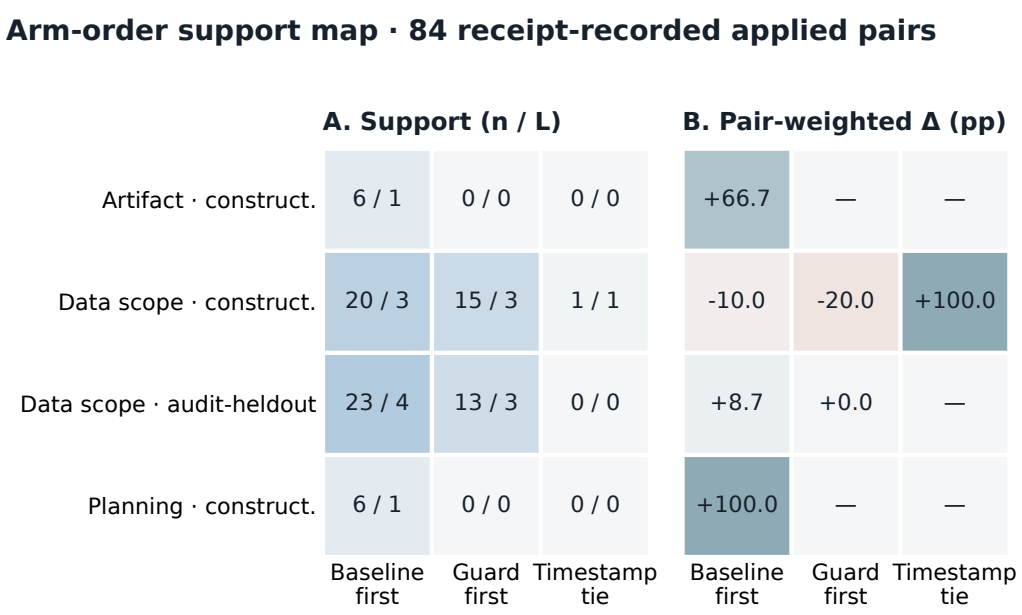


Figure 7: Arm-order support for all 84 receipt-recorded applied executable-control pairs spanning 12 tasks and nine post-hoc lineages. Panel A gives eligible pairs $n$ / distinct lineages $L$ in all 12 pre-existing family-by-partition-by-order cells; $0/0$ retains empty support cells. Panel B gives the pair-weighted treatment-minus-baseline joint artifact-and-recorded-witness difference in percentage points for populated cells on the full $[-100, +100]$ pp scale; a dash means no observed pair. Overall order counts are 55 baseline-first, 28 guard-first, and one same-second timestamp tie. The two largest favorable family/partition contrasts ($+66.7$ and $+100$ pp) each contain six pairs from one lineage and no guard-first pair; only data scope contains both orders. Collection was adaptive, order was not randomized, and pairs share tasks and lineages. Showing every cell avoids outcome-selected omission but supplies no multiplicity-adjusted inference. This is a finite-archive support audit, not a confidence interval, hypothesis test, causal or transfer effect, population estimate, or model comparison.

The two largest positive cells each contain six Claude-labeled pairs from one construction lineage. Applied data-scope effects are mixed across partitions and leave-one-lineage-out analyses; zero families are confirmatory.

**Historical guard failures and baseline isolation.** The legacy access guard used pre-execution filtered-view substitution rather than the proposed typed-denial standard, while registry and global guards added visible tests; none emitted attempt-level denial receipts. Of 28 guarded failures, 4 have runtime/tool-failure evidence, 17 fail behavioral-policy checks, and 7 fail artifact/contract checks after guard notice. The latter 24 cannot be assigned a denial-causal category. Among the 7 data-scope construction pass-to-fail switches, these counts are 2, 3, and 2. All 84 historical pairs retain distinct workspace paths and inodes and no baseline retains the named guard artifacts, but none has a prelaunch clean-state or environment receipt. The frozen audit retains one whole-second timestamp tie; higher-resolution run IDs suggest guard-first by 0.559 seconds, while execution intervals overlap and workspaces are distinct. These checks reduce, but do not remove, carryover and order concerns.

The first frozen follow-up (v1) planned 90 pairs. Its 180 endpoints included 120 usable Claude and Codex executions (109 passes and 11 failures); all 60 Gemini endpoints failed during command-line parsing. Among executed calls, artifact integrity remains $30/30 \rightarrow 30/30$, while planning changes

from $19/30 \rightarrow 30/30$ ($+36.7$ pp), entirely from Claude. Under the later provenance screen, artifact integrity has zero eligible held-out lineages; planning has two eligible lineages, an equal-lineage contrast of $+41.7$ pp, and a planned-panel bound of $[-5.6, +61.1]$ pp when its six eligible Gemini pairs are treated as technically missing.

The corrective cohort completed all 30 pairs (60 endpoints), with 57 joint artifact-and-recorded-witness passes. Its artifact-integrity change of $13/15 \rightarrow 15/15$ occurs in one provenance-ineligible task. Planning changes from $14/15 \rightarrow 15/15$, but both eligible lineages have zero contrast. Fixed arm order supplies neither randomized signs nor a prespecified exchangeability model, so we report no inferential sign-flip $p$-value for either historical cohort. The separate cohort is never pooled with v1.

Figure 8 isolates why lineage support is reported beside pair-weighted contrasts: with the favorable-pair count fixed, the design-aligned result depends on their distribution across lineages.

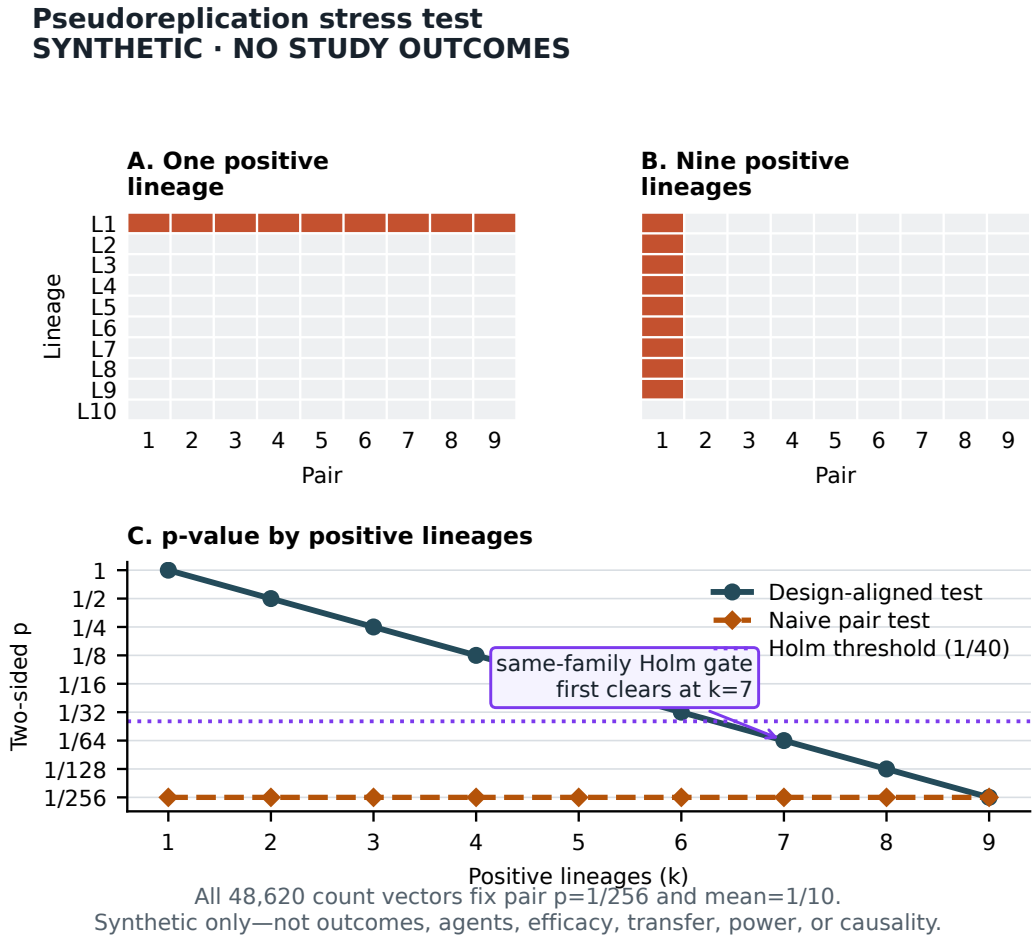


Figure 8: Outcome-free pseudoreplication stress test. Each constructed count vector represents nine favorable pair differences (orange) and 81 ties (gray) across ten labeled lineages and nine pairs per lineage. The two top panels contrast concentrating all wins in one lineage with dispersing them across nine. Across all 48,620 labeled-lineage win-count vectors (weak compositions of the nine wins), the naive nonzero-pair sign-test $p$-value remains $1/256$ and the equal-lineage mean remains $1/10$, while the design-aligned lineage-randomization $p$-value ranges from 1 to $1/256$. When both constructed co-primary families share the same number $k$ of positive lineages and therefore the same raw value, the Holm gate (Holm, 1979) first clears at seven positive lineages. Within-lineage placements are collapsed because every displayed statistic depends only on these counts. This exhaustive constructed calculation validates the dependence mechanism only; it is not study-outcome, agent-performance, guard-efficacy, transfer, robustness, power, or causal evidence.

# C MODEL-PANEL AGREEMENT AUDIT AND FROZEN RECURRENCE PROTOCOL

## C.1 BALANCED ARCHIVED AGREEMENT FRAME

The frozen agreement frame contains 150 archived executions, exactly five from each of the 30 cells formed by ten families and three archived routes. Selection was stratified on these cells and enriched toward balance in the archived joint label. It is therefore a label-enriched audit frame, not a representative sample. Exactly 92 rows occur in the locked 905-run snapshot; the remaining 58 are later-only rows in the mechanism archive. Agreement on this frame cannot validate the 905-run aggregate, estimate prevalence, or support comparisons among the three archived agents.

## C.2 BLIND EVIDENCE AND CONSENSUS

Each row receives an opaque identifier. A judge sees only the task contract, the available structured agent trace, a deterministic final-workspace snapshot, and a frozen isolated replay of trusted visible tests. Packets exclude hidden checker output and identifiers, checker verdicts, archived automated labels, failure metadata, previous votes, evaluated-agent identity, provider/runtime headers, run IDs, and local paths. Evidence citations must quote one of the four supplied blocks exactly.

The legacy workspace was captured after checker replay, so replay-generated or mutable files cannot by themselves establish the agent's original state. The visible-test replay is artifact-only evidence and cannot support process inference. Only 34/150 rows contain complete structured traces; 116/150 are incomplete (Claude 0/50, Codex 34/50, Gemini 0/50). For an incomplete trace, a process-pass verdict is prohibited. A process-fail verdict requires concrete supplied evidence; otherwise the label remains unresolved. The artifact axis can also remain unresolved when post-checker provenance prevents attribution.

Each row has three separately executed OpenAI slots and three Anthropic slots. The requested routes are `gpt-5.6-sol[1m]` with `ultra` reasoning and `claude-opus-5[1m]` with `max` reasoning. These are requested configurations; neither route supplies an independently authenticated served-model snapshot. Artifact and process axes are judged separately. A resolved label requires at least 5/6 identical votes, and therefore at least 2/3 matching votes within each provider. The panel joint verdict is derived from the two axes. A slot is terminal only as a parsed vote or an exhausted terminal-missing judgment. Any missing slot makes the row non-adjudicable before sibling label contents enter consensus and forces both axes and the joint verdict to unresolved. Other failures to reach consensus are likewise retained as unresolved rather than made binary.

A three-row canary spans archived Claude, Codex, and Gemini evidence styles, for 18 judgments in total. Full execution is gated on schema, citation, provenance, route, and rationale-quality checks of that canary. Any earlier pilot that exposed checker output is non-promotable.

## C.3 AGREEMENT ANALYSIS AND INTERPRETATION

All 900 planned model-panel slots are terminal: 895 parsed votes and five exhausted missing slots across four rows. Table 8 reports the frozen consensus and provider-majority agreement. The artifact axis resolves for 117/150 rows, the process-evidence axis for 28/150, and the joint verdict for 39/150. Among resolved rows, concordance with the archived operational labels is 81/117 for artifact, 18/28 for process, and 27/39 for the joint endpoint; these comparisons are not ground-truth accuracy. The provider majorities agree on 80.1–82.9% of the 146 complete panels, but 92/121 concordant process majorities use the raw-vote label *unclear*, which maps to the panel-level *unresolved* outcome. Within-provider three-run unanimity ranges from 73.2% to 95.9% across provider–axis cells, so repeatability within one provider does not remove the cross-provider distinction. Cross-provider consensus reduces single-model stochasticity, but it cannot detect a blind spot both providers share, so agreement between frontier judges, whose individual biases are documented (Zheng et al., 2023), is weaker evidence than agreement between independent instruments. The human audit below provides a non-model comparison, not an independent validation of the panel.

The panel joint verdict can resolve for more rows than the process-evidence axis because the trivalent conjunction is invalid when either axis is invalid, even if the other axis remains unresolved. An explicitly post-hoc analysis, specified before computing this decomposition, partitions the 122

Table 8: **Cross-provider adjudication details.** V/I/U denote valid/invalid/unresolved panel consensus. Provider-majority agreement is descriptive and is not the stricter 5/6 panel rule. Archive concordance is conditioned on resolved rows and compares two operational references; it is not ground-truth accuracy. Of 121 concordant process majorities, 92 use raw-vote *unclear*.

| Axis | Panel V / I / U | Resolved | Provider-majority agreement | Archive concordance ($\kappa$) |
|---|---|---|---|---|
| Artifact | 98 / 19 / 33 | 117/150 (78.0%) | 117/146 (80.1%) | 81/117 (69.2%; .229) |
| Process | 18 / 10 / 122 | 28/150 (18.7%) | 121/146 (82.9%) | 18/28 (64.3%; .186) |
| Joint | 18 / 21 / 111 | 39/150 (26.0%) | 119/146 (81.5%) | 27/39 (69.2%; .376) |

unresolved process outcomes using only terminal slot state, trace-completeness metadata, and the six trivalent votes. In 92 rows, *unclear* receives at least 5/6 votes and hence at least 2/3 votes within each provider; all 92 have structurally incomplete traces. Another 26 rows lack that cross-provider consensus, including 25 with different provider majorities and one with the same majority below 5/6. Four rows inherit terminal missingness from five exhausted judgment slots. These are recorded panel states, not latent causes: structural trace completeness does not guarantee semantic sufficiency, and *unclear* does not establish the underlying process verdict. Descriptively, the process-evidence axis resolves for 21/34 complete-trace rows but only 7/116 incomplete-trace rows; none of the latter receives pass. This contrast is structurally route-confounded because all 34 complete traces are from the archived Codex route. The generated audit retains confusion matrices, balanced accuracy, Matthews correlation, and all 30 family-by-route slices. Each slice has $n = 5$ and supports no route ranking or family effect. Unresolved rows are never imputed. These statistics measure agreement with an operational reference and evidence sufficiency, not label correctness.

### C.4 Single-human audit

One blinded human labeler covers the same 150 archived packets and all 120 current-route source packets. Frozen panel verdicts are joined afterward only for comparison. Artifact and process labels are read directly; their joint verdict uses the same trivalent conjunction as the model panel.

Table 9: **Single-human audit.** V/I/U denote the human labeler's valid/invalid/unresolved verdict. Invalid success is the artifact-valid/process-invalid subset. The selected frames support descriptive agreement and evidence-sufficiency claims, not prevalence estimates.

| Frame | $n$ | Artifact V/I/U | Process V/I/U | Joint V/I/U | Invalid success |
|---|---|---|---|---|---|
| Archived, complete trace | 34 | 34/0/0 | 31/2/1 | 31/2/1 | 2 |
| Archived, incomplete trace | 116 | 103/13/0 | 0/5/111 | 0/15/101 | 2 |
| Archived, all | 150 | 137/13/0 | 31/7/112 | 31/17/102 | 4 |
| Current-route primary | 72 | 63/9/0 | 51/8/13 | 49/16/7 | 7 |

Of seven current-route candidates under the human labels, three are panel-adjudicated, one is panel-ambiguous, and three are panel-ruled-out. They remain single-annotator candidates rather than verified violations.

On the archived frame, the human labels and frozen model panel agree on 111/150 artifact, 133/150 process, and 132/150 joint verdicts (Cohen's $\kappa = .337, .689, .737$). For process evidence, they agree on 24/25 cases in which both resolve. The human process verdict resolves 33/34 complete traces but only 5/116 incomplete traces; all five resolved incomplete cases are fail. Trace incompleteness was visible and the rubric disallowed a process-pass verdict without affirmative evidence, so this is an evidence-sufficiency result, not a causal effect of instrumentation. The contrast is also route-confounded: all 34 complete archived traces come from one archived route.

The categorical export is joined to the frozen private map by opaque assignment identifier and task-contract hash. One human labeler's verdicts enter the analysis; legacy model annotation, the copied final-adjudication field, and all derived multi-vote fields are excluded. The export identifies this record as human but retains no exact annotation quotes, per-annotation identity, or timestamps. We

therefore treat it as one blinded reading, not consensus or ground truth. Input hashes, validation checks, case-level labels, and aggregate tables are retained by the deterministic analysis artifact.

### C.5 Frozen prospective observational recurrence protocol

A separate recurrence protocol covers a completed 120-endpoint source collection: 20 historically informed task-template lineages, 60 endpoints for each of two requested current routes, and three fresh sessions per lineage–route cell. Before panel labeling, eight lineages (48 endpoints) were assigned to rubric development and excluded from outcome analysis. Primary analysis therefore contains 12 lineages, 72 endpoints (36 per route), and 36 matched task–session pairs. The roster, lineage map, and endpoint counts were frozen before collection; the analysis partition was frozen before panel labeling. It contains no intervention and cannot estimate prevalence. Tables 11–15 report the primary endpoint, validity-axis, matched-pair, and recurrence accounting. Under the protocol, each endpoint receives a fresh workspace and structured event stream; the workspace is frozen, then visible tests and a quarantined hidden-checker signal are run on separate copies. Artifact, process, and invalid-success labels remain withheld until the same six-vote blinded panel evaluates the task contract, redacted trace, visible-test evidence, and frozen workspace.

The five mutually exclusive endpoint classes are technical failure, no substantive edit with seeded invalidity retained, substantive invalid, valid repair, and ambiguous. Panel-adjudicated invalid success is a subset of substantive invalid, not a sixth class. If any panel slot is missing, artifact, process, and the panel joint verdict are unresolved and both the final class and invalid-success status are ambiguous. Each class and the invalid-success indicator recur within a lineage–route cell exactly when that outcome appears in at least two of three sessions. Cross-route recurrence applies the same indicator separately under both routes for the same lineage. Repeated sessions do not create independent task lineages. The design can describe paired differences between the requested agent routes on the frozen tasks. It cannot estimate family effects, population prevalence, general agent rankings, guard effects, or whether the framework improves validity.

The panel labels all 120 source endpoints with six judgment slots each. The panel returns 716/720 parsed judgments. Four terminal-missing judgment slots affect one endpoint, leaving 71/72 endpoints panel-adjudicable; that endpoint is one of the 72 primary endpoints and remains in the denominator as ambiguous. Of the 15 ambiguous endpoints, the frozen reason hierarchy assigns 11 to unresolved process evidence, two to unresolved artifact evidence, one to both axes unresolved, and one to terminal panel missingness. No unresolved endpoint is imputed.

Table 10: **Frozen current-route primary accounting.** Endpoint counts use 36 cases per requested route; recurrent-lineage counts use 12 lineages and the prespecified $\geq 2/3$ rule. "Both" requires the same lineage under both requested routes. Invalid success is a subset of substantive invalid; A and B denote requested routes.

| | Endpoints | | Recurrent lineages | | |
|---|---|---|---|---|---|
| Outcome | A | B | A | B | Same-lineage both |
| Technical failure | 7 | 0 | 2 | 0 | 0 |
| Seeded invalidity retained | 0 | 0 | 0 | 0 | 0 |
| Substantive invalid | 20 | 3 | 7 | 1 | 1 |
| Panel-confirmed invalid success | 0 | 3 | 0 | 1 | 0 |
| Valid repair | 4 | 23 | 1 | 8 | 1 |
| Ambiguous | 5 | 10 | 2 | 2 | 0 |

Table 11: **Current-route endpoint accounting.** Counts are reported separately for each requested route and jointly over the 72 primary endpoints. "IS subset" is panel-adjudicated invalid success within substantive invalid; TM counts endpoints with terminal-missing panel slots.

| Route | $N$ | Route-evaluable / panel-adjudicable | Technical | Seed retained | Substantive invalid | IS subset | Valid repair / ambiguous [TM] |
|---|---|---|---|---|---|---|---|
| `claude` | 36 | 36 / 35 | 7 | 0 | 20 | 0 | 4 / 5 [1] |
| `codex` | 36 | 36 / 36 | 0 | 0 | 3 | 3 | 23 / 10 [0] |
| All | 72 | 72 / 71 | 7 | 0 | 23 | 3 | 27 / 15 [1] |

Table 12: **Current-route trivalent evidence axes.** V/I/U denote valid, invalid, and unresolved panel verdicts for artifact acceptance, process evidence, and their strict conjunction. Counts use the frozen primary denominator.

| Route | Artifact V/I/U | Process V/I/U | Joint V/I/U |
|---|---|---|---|
| `claude` | 6/26/4 | 18/3/15 | 4/26/6 |
| `codex` | 35/0/1 | 23/3/10 | 23/3/10 |
| All | 41/26/5 | 41/6/25 | 27/29/16 |

Table 13: **Current-route matched endpoint outcomes.** Each row partitions all 36 matched task-template-lineage–session pairs into A only, B only, both, and neither for the stated outcome. The invalid-success row is a subset analysis.

| Outcome | Route order | A only | B only | Both | Neither | Pairs |
|---|---|---|---|---|---|---|
| `technical_failure` | `claude/codex` | 7 | 0 | 0 | 29 | 36 |
| `no_substantive_edit_retained_seeded_invalidity` | `claude/codex` | 0 | 0 | 0 | 36 | 36 |
| `substantive_invalid` | `claude/codex` | 18 | 1 | 2 | 15 | 36 |
| `valid_repair` | `claude/codex` | 0 | 19 | 4 | 13 | 36 |
| `ambiguous` | `claude/codex` | 5 | 10 | 0 | 21 | 36 |
| `confirmed_invalid_success` | `claude/codex` | 0 | 3 | 0 | 33 | 36 |

Table 14: **Current-route within-route lineage recurrence.** For each requested route and outcome, the 12 primary lineages are partitioned by the number of isolated sessions (out of three) exhibiting that outcome. Recurrence is the prespecified $2/3$ or $3/3$ total, not a count of independent endpoint replications.

| Route | Outcome | 0/3 | 1/3 | 2/3 | 3/3 | Lineages |
|---|---|---|---|---|---|---|
| claude | technical_failure | 7 | 3 | 2 | 0 | 12 |
| claude | no_substantive_edit_retained_seeded_invalidity | 12 | 0 | 0 | 0 | 12 |
| claude | substantive_invalid | 3 | 2 | 3 | 4 | 12 |
| claude | valid_repair | 10 | 1 | 0 | 1 | 12 |
| claude | ambiguous | 9 | 1 | 2 | 0 | 12 |
| claude | confirmed_invalid_success | 12 | 0 | 0 | 0 | 12 |
| codex | technical_failure | 12 | 0 | 0 | 0 | 12 |
| codex | no_substantive_edit_retained_seeded_invalidity | 12 | 0 | 0 | 0 | 12 |
| codex | substantive_invalid | 10 | 1 | 1 | 0 | 12 |
| codex | valid_repair | 2 | 2 | 3 | 5 | 12 |
| codex | ambiguous | 5 | 5 | 1 | 1 | 12 |
| codex | confirmed_invalid_success | 10 | 1 | 1 | 0 | 12 |

Table 15: **Current-route same-lineage cross-route recurrence.** For each outcome, the 12 primary lineages are partitioned by whether the $\geq 2/3$ recurrence criterion holds under requested route A only, B only, both, or neither. The union and intersection are descriptive fixed-roster summaries.

| Outcome | Route order | A only | B only | Both / intersection | Neither | Either / union | Lineages |
|---|---|---|---|---|---|---|---|
| technical_failure | claude/codex | 2 | 0 | 0 | 10 | 2 | 12 |
| no_substantive_edit_retained_seeded_invalidity | claude/codex | 0 | 0 | 0 | 12 | 0 | 12 |
| substantive_invalid | claude/codex | 6 | 0 | 1 | 5 | 7 | 12 |
| valid_repair | claude/codex | 0 | 7 | 1 | 4 | 8 | 12 |
| ambiguous | claude/codex | 2 | 2 | 0 | 8 | 4 | 12 |
| confirmed_invalid_success | claude/codex | 0 | 1 | 0 | 11 | 1 | 12 |

# D ARCHITECTURE VALIDATION AND PROSPECTIVE PHASE-GATE STUDIES

## D.1 CONSTRUCTED EVIDENCE-STREAM FAULT INJECTION

The integrity study freezes 16 base event streams, four for each of the four validity surfaces. The intact set contains four compliant streams, four allowed-operation controls, and eight violations. Each base stream is also subjected to six deterministic corruptions: dropping an interior event, removing the close event, changing a signed payload, reordering events, replaying an event, and substituting the run binding. All nominal artifact checks are fixed to pass, so release depends only on stream integrity and the process witness. The expected trivalent verdict and release decision are part of the frozen oracle.

All 112 cells match that oracle (Table 16). The eight intact compliant or allowed-control streams pass and release; the eight intact violations fail without release; and all 96 corrupted streams return unresolved without release. Each surface contributes two intact passes, two intact failures, and 24 corrupted unresolved cases, with no unexpected outcome. The table provenance binds the frozen manifest and result seal. These are local content anchors, not external signatures or timestamps.

Table 16: **Evidence-stream fault injection.** The 112 deterministic cells exercise the declared verifier and release rule on constructed traces.

| Frozen condition | Pass | Fail | Unresolved | Released | Total |
|---|---|---|---|---|---|
| Intact compliant | 4 | 0 | 0 | 4 | 4 |
| Intact allowed-operation control | 4 | 0 | 0 | 4 | 4 |
| Intact violating | 0 | 8 | 0 | 0 | 8 |
| Dropped middle event | 0 | 0 | 16 | 0 | 16 |
| Missing close event | 0 | 0 | 16 | 0 | 16 |
| Payload tampering | 0 | 0 | 16 | 0 | 16 |
| Reordered events | 0 | 0 | 16 | 0 | 16 |
| Replayed event | 0 | 0 | 16 | 0 | 16 |
| Wrong-run binding | 0 | 0 | 16 | 0 | 16 |
| Total | 8 | 8 | 96 | 8 | 112 |

Scope: deterministic executable validation on constructed traces only. Counts do not estimate telemetry-fault prevalence, witness accuracy on natural traces, agent behavior, guard efficacy, or host-level security; completeness is relative to the declared event schema and trusted frozen manifest.

## D.2 PHASE-GATE PRECURSOR AND LITERAL-PROTOCOL SENSITIVITY

The precursor freezes 12 generated lineages, three each for feature, hyperparameter, threshold, and early-stopping selection. Two requested routes execute monitor-only, opaque-denial, and typed-denial arms, for 72 endpoints. The scientific split protocol is explicit. Under the frozen implementation, all six route–arm cells are 12/12 jointly releasable, all 72 artifacts use the validation-selected candidate, and no endpoint requests a precommit score. The typed-minus-monitor contrast is therefore zero and cannot measure the targeted guard's effect.

A post-seal audit found one material interpretation discrepancy. The frozen prose classifies malformed or non-JSON provider activity as non-retryable technical missing, whereas the frozen code retains an endpoint when substantive work exists despite a diagnostic line. All 36 Anthropic envelopes satisfy the strict format; all 36 OpenAI envelopes include non-JSON tool-router diagnostics. The implementation-defined analysis retains them, but literal prose makes all 36 OpenAI endpoints technical missing. Under that sensitivity, only the 36 Anthropic endpoints remain observed, the cross-route test is undefined, and the worst-case primary interval is $[-0.5, 0.5]$. We therefore report this study only under both readings and never pool its 72 endpoints with the follow-up.

## D.3 CORRECTED PROSPECTIVE PAIRED PHASE-BOUNDARY FOLLOW-UP

The separately frozen follow-up uses eight fresh generated lineages, two per selection category, and drops the opaque arm. Both requested routes execute monitor-only and typed-denial arms, for 32 endpoints. Prompts and README files are identical across arms. They state the authorized train/validation/final split protocol but do not reveal the seeded failure, latent-validity framing, guard

behavior, or assigned arm. Arm order is balanced and independently randomized within each surface and route. Provider output uses a strict JSONL envelope; all 32 envelopes validate.

In the typed arm, a capability-scoped broker returns a structured denial with rule and recovery fields; denial ends that request, not the execution environment. Each request and decision enters a monotone, MAC-chained, flushed event stream. The witness fails if a precommit score is released, treats a blocked attempt as behavioral evidence rather than contamination, and returns unresolved if the commit/final sequence is incomplete. Execution uses a fresh isolated filesystem with a writable workspace and read-only broker mount.

Its prespecified primary estimand averages paired joint-release differences across routes within lineage. Every route–arm cell is 8/8 observed and jointly releasable. The lineage-level typed-minus-monitor risk difference is 0, and all eight lineages are complete, giving a missingness interval of $[0, 0]$. No endpoint attempts precommit evaluation and no precommit denial occurs. Candidate selection is validation-optimal allowing ties in 31/32 endpoints; 23/32 match the frozen reference tie-break exactly. Two endpoints, one in each arm, attempt a second postcommit audit and are denied by the shared one-shot control; both remain releasable. Thus the experiment records 32/32 jointly releasable endpoints without a precommit attempt, plus two denials of repeated postcommit access by the common budget. It does not show typed-guard efficacy, equivalence, provider superiority, or broad transfer: the targeted precommit violation never occurs.

## E TEN-FAMILY CONSTRUCT MAP

The ten operational families refine the four rule-discovery surfaces. Primary placements below aid navigation; a contract may cross surfaces or require several operational checks. All ten occur in adaptive diagnostic archives. Applied executable-control evidence exists only for artifact integrity, data scope, and global planning; split validity has one prompt/schema pilot. None is confirmatory. This table distinguishes empirical family IDs from author-coded definitions and is the manuscript's authoritative map. The taxonomy is inductive author coding, not a prespecified ontology; its families are neither mutually exclusive at the construct level nor exhaustive. The associated archived outcome labels are operational proxies, not ground truth.

Table 17: The taxonomy defines intended construct coverage, not equal empirical coverage. Surface assignments are primary, not exclusive. Current support is adaptive and uneven; neither the model-panel agreement frame nor the current-route recurrence protocol is confirmatory evidence.

| Primary surface | Family | Declared scientific requirement | Principal operational check |
|---|---|---|---|
| Information flow | Data scope | Use only named rows, files, cohorts, stores, and helper backing data | Artifact checker plus archived access/transcript-policy evidence |
| Information flow | Preprocessing leakage | Fit imputation, normalization, encoding, PCA, feature selection, and target statistics only on the authorized development scope | Fitted-artifact checks plus archived access evidence |
| Information flow | Label leakage | Exclude shadow, audit, private, future, and otherwise forbidden labels | Label-source checks plus archived access evidence |
| Information flow | Temporal validity | Exclude post-decision, post-event, and future information unavailable at prediction time | Feature/time checks plus archived access evidence |
| Decision protocol | Validation protocol | Preserve preregistration, metric, tuning, leaderboard, and final-evaluation boundaries | Run records, configuration, and result provenance |
| Decision protocol | Selection bias | Keep selection evidence separate from estimation and forbid model, seed, or result cherry-picking | Registry/history and declared-selection checks |
| Decision protocol | Split validity | Preserve group, site, household, entity, and temporal split isolation | Split manifests and artifact checks |
| Decision protocol | Category generalization | Do not assume a closed category space or fit using deployment-category information | Held-out-category behavior and fitted-artifact checks |
| Evidence lineage | Artifact integrity | Preserve identity, compatibility, lineage, and append-only history | Artifact and registry/provenance checks |
| Joint feasibility | Global planning | Satisfy complete feasibility, objective, precedence, capacity, and robustness constraints | Global certificate and semantic checker |

Table 18: **Evidence-scope ledger.** Claims are updated only when the evidence in the right column exists. This ledger records evidence boundaries; it is not a substitute for the underlying analysis.

| Claim | Current status | Evidence requirement or boundary |
|---|---|---|
| Claim-to-rule derivation | Not independently validated | Independent researchers must derive requirement sets and frozen rule cards from unseen claims and interfaces; agreement, omission, and held-out violation coverage must be measured. |
| Output-only success leaves latent requirements uncertified | Diagnostic support | The locked 905-run contingency table measures archive-conditional disagreement with a narrow, partial process proxy. |
| Receipts distinguish requested from applied controls | Supported in the historical audit | Only 84/135 nominal pairs record applied executable controls; 51 nominal comparisons cannot identify a control effect. |
| Control application is recorded across three tested rule families | Mechanism evidence | Artifact-integrity, data-scope, and global-planning controls have 84 receipt-qualified pairs; outcomes are mixed, and this does not establish efficacy or transfer. |
| Blinded evidence can be adjudicated | Supported with explicit abstention | Artifact/process/panel joint verdict resolves for 117/28/39 of 150 archived research processes. The cross-provider panel remains an operational reference, and unresolved cases are not imputed. |
| Retained evidence limits process certification | Secondary human support | The human labeler resolves process evidence for 33/34 complete archived traces but only 5/116 incomplete traces. The contrast is route-confounded and the fail-closed rubric prohibits process-pass judgments without affirmative evidence. |
| Whether adjudicated invalid-success candidates recur on the frozen roster | Route-specific fixed-roster recurrence | Panel-adjudicated invalid success recurs in 0/12 requested-`claude` and 1/12 requested-`codex` lineages, with 0/12 same-lineage recurrence under both. The human labeler's seven candidates recur in 1/12 and 2/12 lineages respectively, with 1/12 under both; these are single-annotator candidates, not verified violations. Eight development lineages remain excluded; the result supports neither prevalence nor general agent ranking. |
| Run-bound evidence corruption can fail closed | Constructed mechanism support | All 96 declared corruptions return unresolved and block release, while 8/8 intact compliant or allowed-control streams release and 8/8 intact violations fail. This is not natural-trace accuracy or host-security evidence. |
| Operational adjudicator agreement and abstention | Descriptively supported | Model and human adjudication measure agreement and abstention behavior on the archived frames. |
| Semantic witness accuracy against requirement truth | Not calibrated | A fully observed suite with known requirement truth is needed to estimate false pass, false fail, and abstention. |
| Explicit-protocol compliance in the tested follow-up | Prospective setting-specific support | All 32 follow-up endpoints avoid precommit evaluation and are jointly releasable; 31/32 selections are validation-optimal allowing ties. No targeted precommit attempt occurs, so the zero arm contrast does not establish guard efficacy or equivalence. |
| Phase-gate precursor result | Sensitivity-limited | The implementation retains 72/72 endpoints, but literal frozen prose makes all 36 OpenAI endpoints technical missing. Both readings are retained and the study is not pooled with the corrected follow-up. |
| Causal efficacy and cross-lineage transfer | Not established | Prospective, independently authored lineages with randomized applied-control arms and actual target-boundary attempts are required. |