RESEARCH / ENGINEERING / EVALUATION

Before the Next Agent Acts

Receiver Reliance

A research and engineering account

From a question about justified reliance to reference software, failed assumptions, completed experiments and a more precise contribution.

James Dimachkie

Independent, solo-led research with substantial AI assistance

Integrated working paper | Evidence through September 5, 2026 Pacific time

Completed full operational initial stage; historical study remains separate

Contents

The account follows the changing question, then the evidence that changed its answer.

Abstract

Receiver Reliance (RR) began with a question about delegation: when one AI agent hands work to another, what justifies the receiver's reliance on it? The project developed an explicit vocabulary for support and authority, a deterministic reference classifier, receiver adapters, formal analyses, independently authored implementations, and an ambitious prospective evaluation. Testing then exposed several different problems: evidence could be bound to the wrong subject; correct reference echoes could accompany lost task scope; test boards could agree while missing a stated requirement; and an experiment could observe the wrong bytes while reporting an apparently coherent result. The historical public/adversarial full-treatment study did not reach registration or execution. A separately qualified local successor later completed an amended 1,805-episode initial stage with the full operational classifier.

This integrated account follows those developments and the smaller completed investigations that followed. A retrospective 408-record replay does not establish comparative efficacy: its single additional detected defect was seeded and its comparator omitted an ordinary cross-record join. Two local operational development versions improved from 3/12, 2/12 and 2/12 correct tasks to 11/12, 12/12 and 12/12 for unchecked, conventional and narrow-RR packages respectively, on the same twelve constructed blocks. The conventional implementation matched RR in both versions. In a separate 160-state control model, all ten final learned policies underperformed a simple heuristic. A frozen prompt pilot on eighteen evaluation families found no demonstrated benefit from adding an evidence-use review procedure; its primary paired estimate was -16.7 percentage points with a wide conservative interval of [-50.9, +25.4]. These are different experiments and cannot be pooled.

A four-arm local study then compared full R-all with transparent U and the frozen static S4 and S1 comparators on synthetic arithmetic and item-selection workflows. All 1,805 selected episodes had valid measurements. Equal-stratum seed estimates for R minus U, S4 and S1 were -1.00, -1.00 and -0.75 currently improper presentations per episode, with all three conditional simultaneous intervals below zero. Clean-completion intervals crossed zero. Seeded-task completion was 80/162 for R, versus 150/151 for U, 154/155 for S4 and 139/157 for S1; persistent administrative defects often caused holds. Every arm completed all six semantic-control trajectories but obtained zero correct final answers. The local result demonstrates a structural-release effect under constructed conditions, not general task benefit or superiority to an equally expressive conventional policy. The sample was explicitly amended after collection began and before endpoint analysis; 619 original assignments remain deferred.

The surviving contribution is an inspectable implementation and evaluation case study: explicit contracts for evidence and intended use, conditional mathematical results, preserved local executions, and a record of claims revised when their support failed. Broad security efficacy, real-world benefit, and a novel security primitive remain unestablished. The account explains both what was built and why some of its original ambitions did not survive testing, so that another researcher can reuse the machinery, challenge a result, or avoid repeating an identified mistake.

1. A handoff is a decision

An upstream agent has finished its part of a task. It sends a polished record to the next agent: a recommendation, a revision identifier, perhaps a declaration that the work has been checked. The receiver must now decide what it can do with that record. It can accept the recommendation, verify the calculation, recover the original sources, ask for missing information, or decline to proceed.

Several different failures can look identical from the outside. The record may be an exact copy of an obsolete recommendation. It may name the current rule but omit one of the objects the user asked about. Its evidence may concern an earlier work product, before another agent transformed it. Or every structural relationship may be intact while its answer is simply wrong. A fluent handoff does not tell the receiver which of these situations it faces.

RR was an attempt to make part of that decision explicit. Its initial motivation was the possibility that errors acquire apparent authority as they move through delegated work. A downstream agent may treat an earlier output as a premise instead of reopening the question it answers. The project therefore placed responsibility at the receiving edge: inspect the relevant evidence before the record becomes a basis for further work.

That motivation is plausible. It is also broader than any one implementation or experiment in the record. The account that follows keeps three questions separate. Can the proposed relationship be specified precisely? Does the implemented system compute that relationship from the observations it actually receives? Does using that computation improve an outcome people care about, compared with a competent alternative? RR made progress on all three questions, but reached different answers and different levels of completion for each.

1.1 A running example: recommendation versus recomputation

Consider the parcel-allocation workflow later used in the operational experiments. Agent A prepares parcel facts and recommendations. Agent B turns that work into a grouped manifest. Agent C produces the final allocation under a current rule. A declaration about A's record does not automatically apply to B's new grouping. Even a declaration that does apply to B's grouping cannot make an incorrect grouping correct.

There is another distinction. C might be asked to execute B's plan, or it might be asked to read the parcel facts and recompute an allocation. In the second case, C can sometimes produce a correct result from an incorrectly grouped intermediate record. Calling the whole record unusable would erase a legitimate operation. Calling it valid without naming the intended operation would conceal the difference.

This example became more than an illustration. The local executions retained both kinds of failure: faithfully copied but incomplete inputs, and a correct final recomputation from an incorrectly grouped intermediate record. It forced the project to move from a general judgment about a record toward a narrower question about a particular use of a particular part of that record.

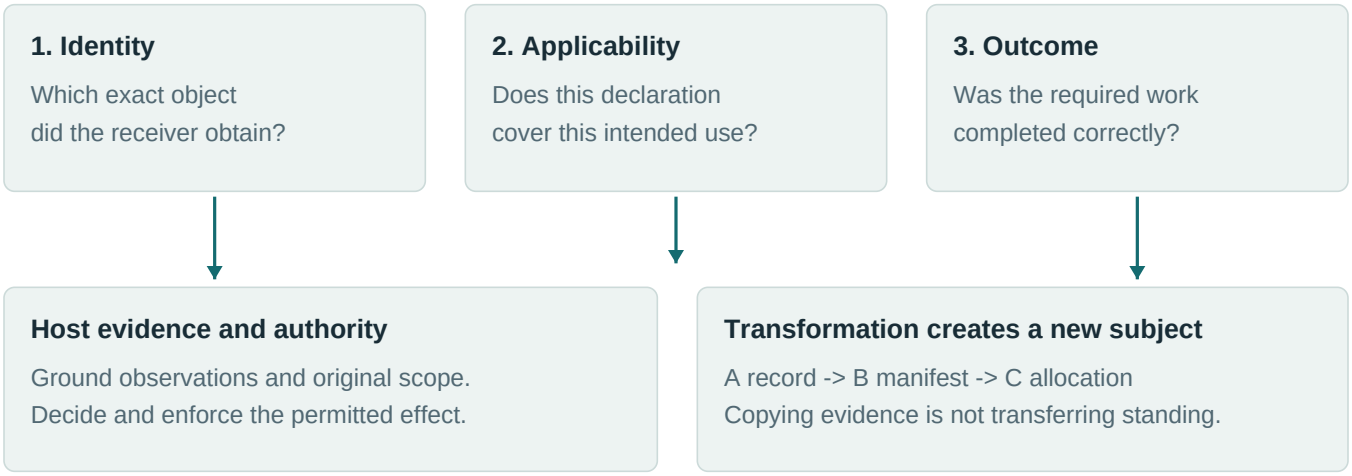


Figure 1. The receiving decision separates identity, applicability and task outcome. The arrows mark dependencies to examine, not proof that the host supplies them.

1.2 How to read this account

Chapters 2-4 reconstruct the early concept, reference software, integrations and replay. Chapters 5-7 explain the proposed full study and the work that prevented premature execution. Chapters 8-12 cover the subsequent binding, mathematical, operational, learning and prompt experiments. Chapter 13 reports the separately admitted full operational local study. Chapters 14-16 position the results in the field and state what is now worth reusing. The appendices provide an evidence inventory, corrections, reproduction routes and source locators.

This is a research account rather than a chronological export of every conversation. It integrates the technical v1 paper, the earlier 26-page illustrated reader, and the later study packages. It does not replace their exact authorities or rewrite their recorded outcomes. A local source map accompanies the paper so that the account remains navigable beyond its narrative.

The work was directed by James Dimachkie, with substantial AI assistance in specification, implementation, mathematical analysis, experiment execution, review and writing. Separated AI reviewers sometimes used different providers or contexts; all remained within one operator's project. That arrangement enabled extensive local work. It did not create independent human replication or eliminate shared assumptions.

Sources: earlier technical paper, earlier reader, operational record.

2. From epistemic vocabulary to executable questions

The earliest recovered August 6 material was concerned with more than stale files. It distinguished evidential support, permission, interpretation, independent supporting roots, and alternatives. These distinctions addressed a real conceptual problem: a statement can be well supported without authorizing an action, and an authorized action can still rest on a mistaken interpretation. Combining every favorable property into one label such as trusted would make the label hard to audit.

The early design also had a weakness. A broad profile or ontology can ask for more structure than a system can ground in observable evidence. Requiring a field does not establish how its value is known. A processor that compares supplied interpretations is not thereby a system that discovers the meaning of arbitrary language. The interpretation-integrity prototype made typed interpretations and resolutions executable, but its inputs still carried the semantic judgments it compared.

The foundations therefore moved toward identifiable predicates and receiver-derived applicability. Unknown information became an explicit state. Lifecycle events and non-transitivity mattered: a relation supported for one receiver, use or stage need not transfer automatically to another. Later addenda modified named parts of the foundation rather than replacing every earlier proposition. This detail would become decisive when the project's own retrospective corpus treated older filenames as evidence of invalid reliance.

2.1 Several version histories, not one ladder

RR's archive contains foundation versions, engine generations, software releases and study-authority versions. They are related, but their numbers are not interchangeable. Foundation 0.4 is not software v0.4; public v1.3 is not the later study's treatment-v6. Treating the largest number as the most complete version would combine incompatible authorities and overstate integration.

Family	What existed	What the account carries forward
Epistemic handoff and interpretation integrity	Distinct support, authorization and interpretation axes; a typed processor	Useful distinctions, with supplied semantics made explicit
Foundations 0.3, 0.4 and scoped addenda	Receiver-derived predicates, lifecycle, unknowns and non-transitivity	A use-sensitive question and respect for unchanged base clauses
Frozen engine 0.2 and composed 0.3	A 28-operation baseline and two supplemental operations	An executable reference law and conformance machinery
Grounded 0.4 / public software v1.3	Audited thirty-operation surface and field-authority inventory	A clearer statement of what the engine actually computes
Generation 0.5	Proposed derivation of compatibility from intended-use facts	A design input; not an adopted or measured engine
Local v1.4 proposals	Substantive field and contract changes; incomplete closure	Preserved proposals, not a silent replacement inside older studies
Scientific release and later narrow profiles	Treatment, instrument and statistical work; smaller completed experiments	Distinct evidence packages with their own scope and status

Two early proposal labels also require care. H-BIND had no eligible submission and was dropped by forfeit. It did not lose an empirical comparison. The remaining G-BIND route concerned generic representation and rule value. Its history already entertained an ordinary policy translation. The later decision to compare RR with conventional joins was therefore consistent with an early scientific question, not an attempt to move the goalposts after an unfavorable result.

2.2 What simplification meant

The useful simplification was to ask what each field or relation changes. If a field cannot be derived, its presence should not improve a verdict. If a new layer makes exactly the same decision as a conventional implementation on the same observations, the remaining question concerns usability, integration cost or explanation quality. If those benefits are not measured either, the extra layer has not earned a claim of practical advantage.

This principle did not make all earlier work unnecessary. The distinctions among identity, support, standing and use helped locate later failures. The archive's value is partly that it shows which abstractions became executable, which remained supplied judgments, and which were unnecessary for the smaller task eventually chosen.

Early foundations
Support, permission, interpretation and unknowns

Reference software
Frozen predicates, audited fields and host obligations

Integrations and replay
Real routes, missing evidence and false holds

Full study apparatus
Comparator, instrument, assignment and qualification

Narrow completed studies
Binding, use, scope, operational work and learning

Current contribution
Operational comparison, earlier results and corrections

Figure 2. The argument evolved across related families. This is a conceptual sequence; foundation, engine, software and authority version numbers are separate namespaces.

Sources: historical decisions and coverage limits, early-forms archaeology, engineering-forms archaeology. The chronology is a reconstruction from retained files, commits and handoffs; missing early probes and conversation records are identified in Appendix A.

3. The reference mechanism and the system around it

The reference engine consumes a host-assembled fact profile about a handed record. It evaluates a frozen predicate table in a fixed precedence order. The first matching defect class determines the result; if none fires, the residual class is VALID. That determinism is useful because another implementation can be checked against an explicit decision law. It does not establish the truth of the observations supplied to the engine.

Engine class	Meaning within the declared profile
MALFORMED_OR_BOUNDARY	A representation or structural boundary condition fails
BINDING_OR_CONFLICT	A required binding or consistency relation fails
OMISSION_OR_INCOMPLETE	Required evidence or structure is incomplete
VALID	No earlier defect predicate fires

Native preflight is a separate surface. It distinguishes READY, REJECTED_INVALID and INSUFFICIENT_EVIDENCE before the engine's classification is interpreted. Incomplete audits and protocol

errors also have their own representations. Combining these states into a single allow bit would destroy information the host needs. In particular, insufficient evidence cannot count as a correct answer or a successful task merely because it avoids a false hold.

3.1 What a seal does, and what it does not do

The audited decision surface binds a result to the input and witness from which it was derived, together with relevant decision identities. This enables integrity and identity checks: a later verifier can determine whether the presented pieces agree. Nothing in that relationship makes an untrusted input true. Nor is the project's seal a digital signature authenticating an independent authority. The historical errata explicitly distinguish self-verification from authentication.

The binding itself needed correction. Early receipts did not bind the actual decision input. A later review found that policy and engine identities did not by themselves identify the decision-law contracts. These are different omissions: one concerns what was evaluated, the other concerns the law under which it was evaluated. The additive audited surface addressed those gaps. This history illustrates why a digest should be treated as an answer to a specific identity question, not as a general badge of trustworthiness.

A further review found that some PASS records were declarations emitted by the machinery rather than conclusions recomputed by an external check. The correction was to make the verification path perform the relevant checks. A report that says verification succeeded and a verifier that independently recomputes the result are different artifacts. The project learned this distinction while building its own evidence machinery.

3.2 Seven host obligations

RR does not gather observations, decide organizational policy or execute a downstream effect. Its host supplies those capabilities. The host contract makes the dependency visible rather than hiding it inside the classifier's success rate.

Obligation	Why it matters
H1: truthful state	A perfect calculation over invented or stale observations can still support the wrong action
H2: atomicity and replay	The check and effect must not become unrelated through partial execution or reuse
H3: derive facts	Supplying the desired conclusion as an input would make the check circular
H4: applicability calibration	A rule must be applied only where the record supports its semantics
H5: input binding and transcripts	The decision must concern the record actually received, with inspectable evidence
H6: effects	Classification does not establish that the intended effect occurred safely or at all
H7: request admission	Unsupported requests need an explicit boundary disposition

The field-authority audit sharpened this contract. It identified fifty-eight required fields without value-based classification authority. Such fields could be part of a schema without their values governing the classification in the way a reader might assume. This was a disclosure about the implementation's actual authority, not evidence that all fifty-eight values were harmless or that the engine understood their intended meaning.

3.3 Integration work and its distinct limits

RR was more than a paper interface. Retained work includes live per-record and batched MCP calls, exact-byte middleware, an AG2 seam, native protocol capture, runtime measurements, and host transaction designs. These establish different things. A real harness call establishes that a route can invoke the mechanism. A scripted-model AG2 test establishes behavior at that seam. Neither measures the quality of an ordinary model's reasoning or the benefit of a deployed guard.

The retired rr-gate sibling recorded five workspace replays, including a dispatch/commit line-ending discrepancy. Its hook was staged but never installed. That is useful implementation experience, but not a history of continuous prevention. Native MCP traffic over fictional data also often lacked the identity or digest evidence the intended judgment required. Missing evidence was an integration finding; inventing it would have manufactured a more favorable demonstration.

The later transactional host work distinguished a non-releasing candidate from an effect, and explored durability and recovery obligations. Its unfinished status matters because a transaction design can be substantial without supplying an operationally qualified host. The current small reporting prototype does not need every element of that historical host program. The broader study would need the capabilities its claim depends on.

Sources: worked example, host obligations, root errata, historical integration assessment. Detailed original implementation claims remain governed by those files and the specific frozen release.

4. Dogfooding exposed the cost of the wrong question

The native-record replay initially appeared to offer an encouraging result. The retained comparison records eighteen detected defects for RR and seventeen for a smaller schema/policy comparator. A later calibration removed a large number of false holds. Read without its provenance, this looks like both a gain in detection and a cleaner decision boundary.

The later audit asked how those numbers were made. That changed their interpretation.

4.1 Reconstructing the denominator

The retained truth file contains 390 real-labelled clean records, five real-labelled defective records and thirteen seeded defective records: 408 in total. Thus 395 are real-labelled and thirteen seeded. Earlier prose reported 398 and ten. The illustrated reader further called the corpus held-out and described all eighteen defects as seeded. The retained record supports neither description.

The labels themselves are not independent authentication of origin. The first untracked extraction epoch was not recovered. Recomputing the retained file establishes its current arithmetic, not an unseen collection process. This is why the account uses real-labelled rather than turning a field in a local file into independent evidence of real-world provenance.

Both implementations detected the five real-labelled defects. Their sole difference was seeded. Moreover, the comparator omitted the ordinary cross-record join that detected that case. RR therefore demonstrated additional coverage relative to that smaller comparator, but the result does not establish an advantage over equally capable conventional verification. The proper response is to strengthen the comparator and preserve the original result with its scope intact.

Both methods caught all five real-labelled defects; the sole difference was seeded.

Figure 3. Corrected composition of the retained native truth file. Labels are retained provenance fields, not independently authenticated origin. Bar lengths use the common 408-record denominator.

4.2 False holds were an adapter lesson

The strict adapter produced 133 false holds among 390 clean records, approximately 34.1%. It imposed an ordering obligation on a family whose one-event tasks did not support that interpretation. The historical engineering assessment also records 768 manufactured schema values across 384 decisions. Supplying the fields needed to call an engine did not make their assumed semantics appropriate to those records.

A calibrated replay records zero false holds. A later three-state fallback records 208 insufficient-evidence outcomes, zero new false holds and 18/18 labelled-defect detection. These numbers describe different implementations and counters. The reduction in false holds is useful classification behavior on retained data, but replacing a forced decision with abstention does not tell us how much useful work the surrounding system completes.

Calibration was therefore a lesson about asking an answerable question. A missing observation should remain missing. A rule that presumes multiple ordered events should not be imposed on a record family that supplies only one. The receiving system needs a defensible path for incomplete evidence, rather than a schema adapter that quietly creates the facts necessary to obtain a verdict.

4.3 A stale filename is not a false proposition

The replay also exposed a semantic problem beyond arithmetic. Its extraction logic treated adjacent foundation versions as invalidations and inferred reliance from filename mentions. Yet the later addenda retained unchanged parts of the base. A citation to an older foundation can be legitimate when explaining history or invoking an unchanged clause.

To establish invalid reuse, one needs the proposition being used, the clause that changed, and the operation the reader is performing. Newness of a filename is an imperfect proxy for all three. Reproducing an old label exactly

would not repair that construct-validity problem. This recognition led directly to the later evidence-use work, where retaining legitimate reuse became part of the success definition.

4.4 The August 20 precursor

The precursor compared per-record RR, ordinary source recovery and a later batched RR condition. Its retained arithmetic is shown below. These are agent runs, not human participants.

Condition	Fully correct runs	Correct answers	Correct defect answers
A: per-record RR	7/8	63/64	32/32
B: ordinary recovery	5/8	61/64	32/32
C: later batched RR	4/4	32/32	16/16

The C cohort also answered all sixteen nondefect items correctly. An earlier 20/20 defect denominator was wrong. Serialized C reports contained leg B labels, while the aggregate recovered C from run identities; both the serialization problem and the recovered aggregate remain visible. Across conditions there were twenty runs and 160 answers.

The ordinary condition's perfect assigned-defect score is important counterevidence. Agents could often recover the relevant authoritative sources without an RR-specific representation. The design also had high defect density, adapted cases, exposure history, a later smaller cohort and a missing clean-control approval record. It cannot support a causal comparison. Its timing medians of 305.5, 264 and 360.5 seconds derive from agent-supplied timestamps, not independent measurement of human effort or wall-clock savings.

The original answer key, twenty reports and serialized scoring are preserved in the portable release package, alongside context documents. That makes the arithmetic inspectable. It does not recover every original interaction. The informative result is that conventional recovery already reached defect-answer parity and that apparent gains needed stronger provenance, a fairer comparison and an outcome beyond classification.

Sources: retained replay results, corrections, audit data, precursor source inventory, release errata.

5. Designing an experiment that could disagree with RR

Counting the defects a guard blocks is a weak test of usefulness. A system that withholds everything can prevent every measured presentation while preventing all useful work. A constructed corpus can also favor a mechanism simply by containing many examples of the exact conditions it was designed to recognize. RR's proposed scientific study attempted to address both problems.

Its unit was one complete delegation-chain episode. The assignment, intention-to-treat accounting and inference would happen at that level. Hops inside the same chain are dependent observations. Counting each receiving edge as a new independent trial would inflate the apparent sample size.

5.1 Three arms and a common host

Arm	Decision package	Intended contrast
U	Transparent capture and telemetry without classification or withholding	What the complete intervention changes relative to transparency
S-all	Independently implemented static comparator at every receiver	What RR adds beyond this active comparator
R-all	Prospectively frozen RR package at every receiver	Package-level receiving-side intervention

All arms were to share an arm-neutral capture, policy-enforcement, remediation, persistence, resource and observation shell. Only the prospective decision-module difference could vary. This constraint matters because a comparison is uninterpretable if one arm sees different evidence, gets better repair instructions or follows a different release route.

The original S-all comparator was interface-matched but not identical to the later conventional implementation of the narrow direct-mandate relation. These are distinct comparisons. A result against a particular static bit threshold would not establish superiority over every competent conventional verifier. Later work made that distinction more explicit.

The study proposed a direct handoff, a three-receiver chain and a four-receiver fork/join topology, across two task families. A frozen smoke set contained one clean and six single-defect classes. A development extension added 51 classes. Neither collection supplied a justified natural population. Final class weights would need to come from the stated threat model rather than the outcomes observed during development.

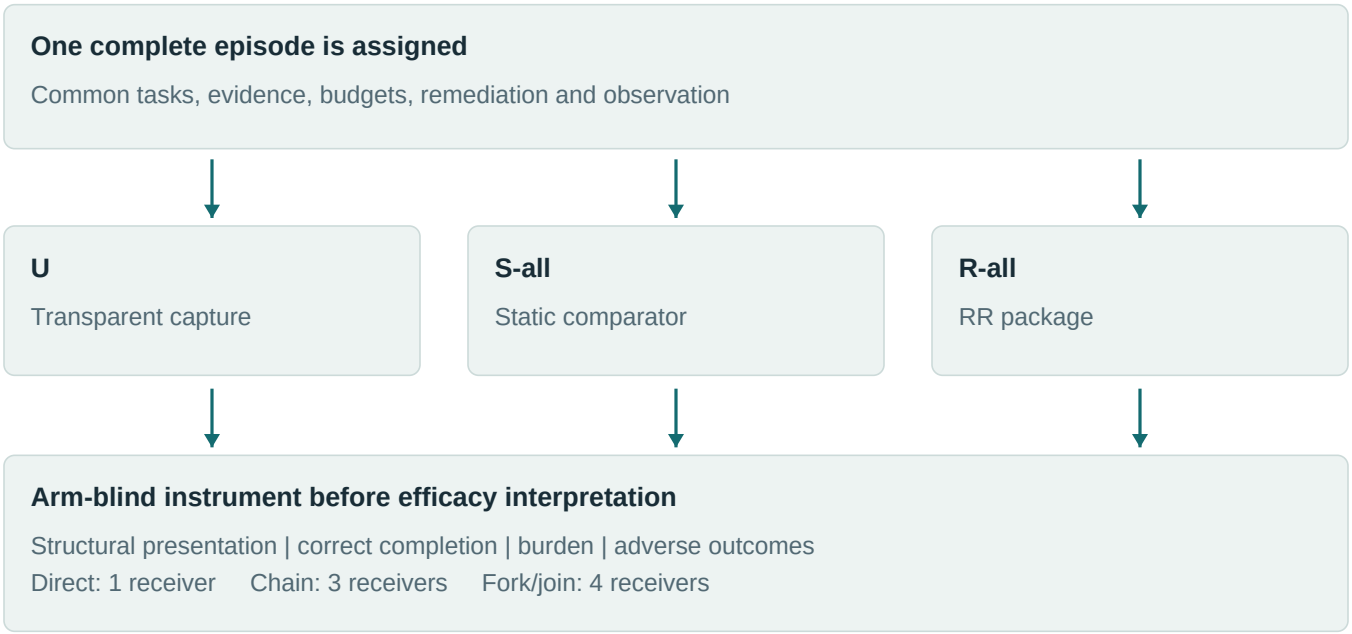


Figure 4. The proposed full-treatment experiment. The common shell and independent observation are design requirements, not demonstrated treatment effects. This historical public/adversarial study remains unrun; Chapter 13 reports a separately admitted local successor.

5.2 Count what arrives, and separately count what gets done

The structural endpoint concerned how many registered destinations received a constructed seed phenotype during an episode. Its proposed predicate, G_s, was to be truth-free and derived from captured witness bytes. Whether the final answer was correct belonged to a separate task oracle and usefulness path. A coherent false answer should not become a structural defect merely because the experimenter knows it is false.

The positive conclusion was conjunctive. A favorable structural contrast had to meet separate usefulness, burden, defect, no-harm and instrument-validity requirements. If containment improved but usefulness failed, the result would be a tradeoff. If RR improved over U but not S-all, the conclusion would be no demonstrated incremental efficacy beyond that comparator. An instrument-invalid run would support no efficacy interpretation, however favorable its numbers.

This design also distinguished a separate adaptive diagnostic population from the ordinary randomized population. Development probes and their selected survivors could not be pooled into the primary sample to estimate prevalence. The prospective challenge law, including selection and accounting, remained unfinished. No operational attack procedure is needed to understand that statistical separation.

5.3 Independence was an instrument property

The measuring instrument was intended to run separately, remain blind to the arm, receive exact witness material through its designated endpoint, and preserve append-only witness and attempt records. It needed to establish which routes were observed, what counted as presentation or nonrelease, and whether an episode had reached a final state. These are prerequisites for interpreting the outcome, not evidence that the treatment works.

The custody program initially pursued stronger native operating-system isolation. Qualification attempts failed, and the portable replan retained residual administrator trust. An experimenter who controls the machine cannot be made unable to fabricate all relevant evidence merely by adding more local hashes. Prospective commitments, exact archives and independent reproduction offer a more defensible credibility model. The change narrowed the claim to what the infrastructure could support.

The distinction between implementation and qualification persisted throughout the study. A script can exist and run without having demonstrated complete observation at the required scale. An authority can be reviewed without every host obligation being implemented. The full study remained unrun because those dependencies had not all been discharged.

Sources: technical protocol, Sections 4-5, reader, Sections 4 and 6.

6. When the apparatus challenged itself

The earlier paper's most important material was not a list of successful checks. It was a sequence in which an apparently satisfactory system failed a more discriminating question. Preserving those episodes is necessary to understand both the effort invested and the limits of its result.

6.1 Independent implementations found an untested requirement

Treatment, route and statistics authorities each had two implementations written without shared code. Agreement on their enumerated boards was useful evidence. It showed that independently constructed programs could implement the pinned examples consistently. It did not show that the examples exercised every requirement in the text.

A later expanded treatment case exposed a duplicate-identity requirement that one implementation did not enforce. A passed the record and B deferred it; the authority supported B's disposition. Both programs had passed their own registered boards. The disagreement therefore identified a gap in the tests as well as an implementation defect.

The scientific consequence was to block the treatment freeze. The appropriate sequence was to clarify the successor authority, add the missing coverage, repair the implementation and repeat confirmation. Merely retaining the original green board would have allowed a known discrepancy into the experiment. Conversely, the finding did not prove that RR improves agent outcomes. It established that the proposed treatment was not yet uniquely and correctly implemented on a known part of its declared surface.

6.2 Agreement over the wrong evidence

Several demonstration defects concerned the path from an event to its measurement. One scorer inferred the phenotype from a construction label instead of the bytes the receiver obtained. A consumer read a driver's dictionary rather than the released record. A remediation message disclosed arm-specific information. Host staging also differed across profiles.

Each defect could make the experiment appear internally coherent while answering a different question. If the scorer reads the generator's intended class, it may score a defect that never reached the receiver. If the consumer reads a convenient driver object, the release boundary may have no effect on what the model actually consumes. If repair instructions reveal the arm, a later difference may arise from the message rather than the decision module.

These findings explain why instrument validity came before efficacy. They also explain why a single successful end-to-end trace was insufficient. The route had to connect the actual emitted bytes, the actual received bytes, the allowed intervention and the recorded outcome. Parts of the rehearsal were repaired, but whole-host equality and the full instrument remained unqualified.

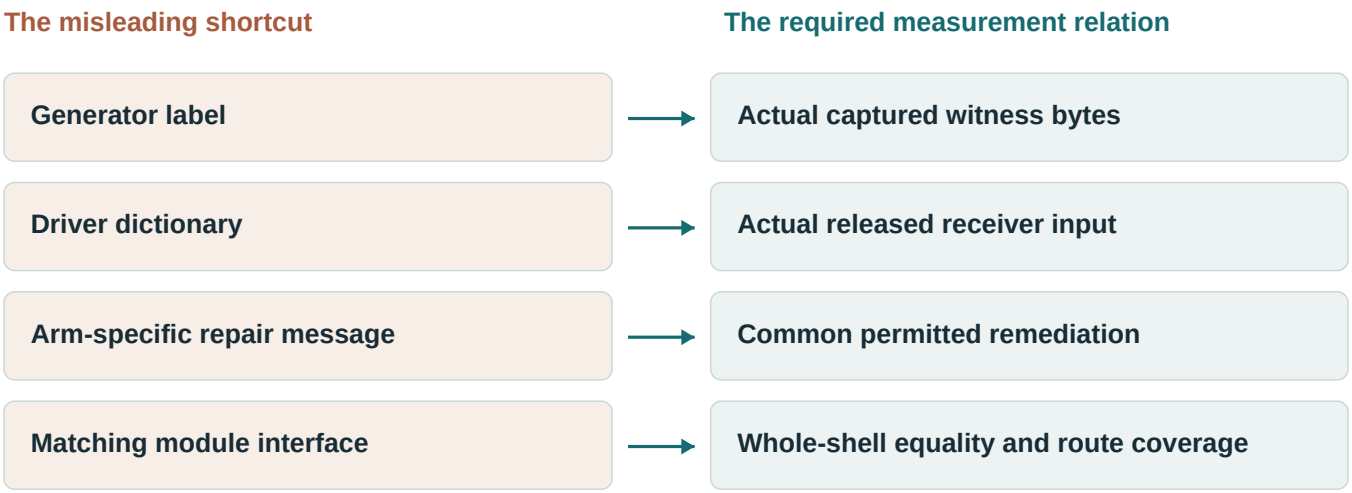


Figure 5. Four development findings changed what the experiment needed to establish. These repairs concern the measurement construct and treatment comparison; they do not demonstrate efficacy.

6.3 What the constructed worlds revealed

The original seven-class set included a structurally coherent false answer. Every arm passed it. This was consistent with the mechanism's intended boundary and important evidence against interpreting structural validity as truth detection.

The 51-class extension broadened acceptance, grant-use, lineage, delegation, size and negative-control coverage. Twenty-five classes were visible only to the intended RR treatment, eight only to the static comparator, nine to both and nine to neither. These counts describe constructed coverage, not the frequency of those conditions in real systems. Changing the mixture would change the apparent comparative advantage.

A specific expired-grant construction illustrates the importance of comparator semantics. In that case the static score was zero, so both thresholds 1 and 4 passed while RR blocked. The score had sixteen component bits under the authority. This example does not mean that every time or expiry-related mismatch was invisible to S-all; a separate validity-window condition was among its bits. A diagram or headline that generalized from the single construction would overstate the evidence.

6.4 Model outputs exposed a scoring choice

Twenty local model-authored source records were parseable and structurally consistent with the supplied host facts. Seven failed the required exact answer form: five contained the correct value inside a sentence, and two were numerically wrong. A more permissive canonical scorer changed the count substantially.

That observation supports neither scorer by itself. It shows that correctness depends on the declared task and output contract, and that a scoring convention can materially change the reported result. The rule needs to be fixed before an outcome is used as evidence. Since these source records had no structural defect, all three decision packages behaved identically by construction.

The demonstration also exercised a bounded remediation path, including eighteen withheld seeded records resubmitted once. It showed that parts of the route could classify, withhold, resubmit, release and observe a record. It did not qualify the arm-neutral remediation protocol for the full study.

6.5 Proofs at one layer did not certify another

The formal record reports 872 properties: 766 proved, 105 bounded and one refuted, with no checker errors in that report. A bounded check is not an unbounded proof. The refuted universal E6 obligation remains part of the record. A failed universal claim is not repaired by aggregating it with many successful checks.

The A8 atomic-predicate comparison reports 40,907,363 evaluations and 3,960 sealed pipeline calls. Its scope excludes important surrounding behavior, including wire parsing, effect receipts, transcripts and wrapper semantic rederivation. Elsewhere, the second complete implementation had 592 full-surface conformance differences across five mechanisms, leaving A3 open. These observations are compatible because they examine different layers.

The lesson is methodological and specific: agreement on the decision predicates does not certify the interface that admits a request or the host that applies its result. The counts establish the extent of the retained checks; they are not a composite security score. The formal and runtime records should be read together without allowing either to erase the other.

Sources: technical development and validation, Sections 6-8, contribution validation, engineering history, bounded report and source links.

7. The study that was not run

The full study's mathematical work created a different kind of resistance. Applying the first exact row law to the proposed collection of claims produced a confirmation requirement of 5,126,400 episodes. A precise number can still be scientifically or operationally unusable. Here it revealed that the proposed claim, bound and resources did not fit together.

The row law used exact-rational, Hoeffding-style calculations and dyadic error budgets. It controlled bounded outcomes without relying on their empirical variance. That conservatism could be expensive when a chain's structural outcome had a large possible range. More fundamentally, an admitted row calculator did not yet

amount to an admitted evaluator for the entire design: assignment, endpoint families, margins, missingness and claim disposition all had to agree.

7.1 Three numbers with three different meanings

Proposal	Confirmation episodes	Why the number was not a valid study decision
Original powered packet	5,126,400	Infeasible under the proposed resources; whole-design law incomplete
D8 reduced powered grid	10,448	Unsupported quarter margins, omitted required gates and pooled-topology assumptions
P fixed-precision proposal	3,072	Resource-selected allocation without a prospectively justified useful width

P's nominal interval half-width under the packet's arithmetic was 510,437/786,432 presentations per episode, about 0.65. Its coverage interpretation was conditional on resolving the relationship between assignment and the row law. It was an interval-estimation proposal, not a powered efficacy design. Choosing P because it was affordable would not establish that its interval could answer a useful question.

D8 reduced the computational burden in part by weakening or omitting conditions the intended positive claim still required. That made it a different claim, not merely a more efficient implementation of the same study. The unresolved design-selection decision, H5, remained open. This H5 is a planning label and is unrelated to host obligation H5 concerning input binding.

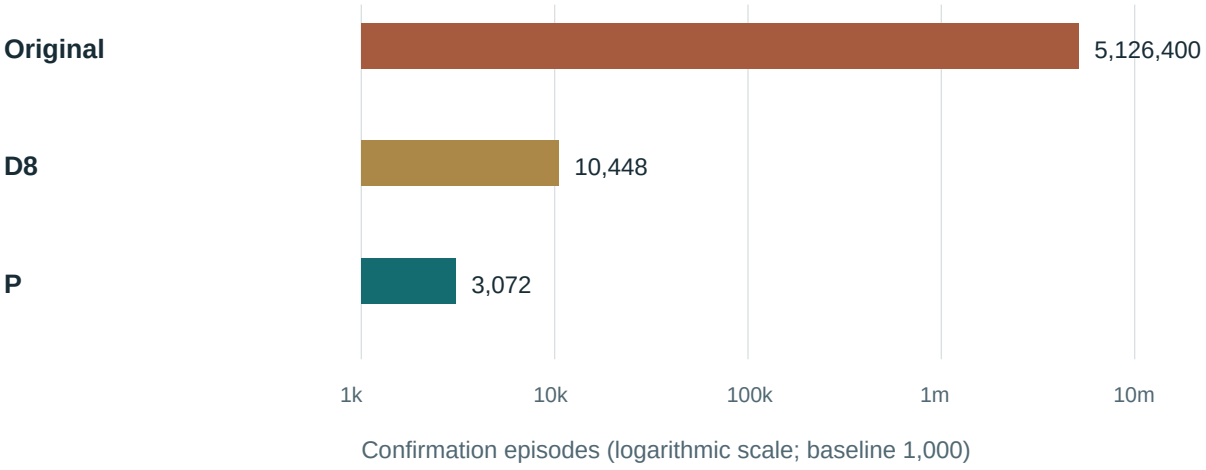


Figure 6. Resource pressure changed the scale of the proposals, but did not make D8 or P registrable. Different claim requirements prevent interpreting shorter bars as equivalent, more efficient studies.

7.2 Review also corrected the reviewers

An internal re-derivation reproduced the original floor, the eleven published powered-candidate boundaries and P's widths, then recommended neither current proposal. A separate exhaustive falsifier corrected two exact-binomial sensitivity values in that internal analysis, from 2,496 to 2,448 and from 1,840 to 1,744. The rounded predicate was non-monotone. The corrections did not change the recommendation, but they mattered to the reliability of the reasoning supporting it.

This episode is a useful example of rigorous iteration because the revised quantities and the unchanged decision were both preserved. A reviewer's conclusion does not make every intermediate calculation correct. The source record also preserves a disagreement about the provider identity of an earlier statistics constructor. A plausible resolution was not substituted for conflicting provenance.

7.3 A deterministic setting did not qualify a deterministic closure

The 8B development model produced identical bytes in a warm smoke check, then differed on one of three prompts after cold loads. Later observations were stable within particular placements but differed between CPU and GPU. A 14B candidate produced 200 matching pairs, representing 400 calls over 190 distinct prompt byte strings. That was evidence for the tested workload, not qualification at either proposed study scale.

The exact closure needed to name the model, runtime, host, accelerator, driver and operational schedule. Zero temperature alone did not guarantee identical executions. These failures were found during development, when they could still prevent an invalid registration. They did not become evidence that no reproducible model study is possible.

7.4 What remained unresolved

The full-treatment study still needed a treatment successor, statistics closure, fixed comparator policy, implementable scorer and task-oracle semantics, qualified observation instrument, whole-host equality, qualified executable closure, complete design and assignment law, justified population and margins, and a prospectively specified adaptive diagnostic. Its chronology would have been qualification, complete prepilot freeze, a blinded feasibility disposition, public preregistration of unchanged scientific bytes, and confirmation.

None of the smaller experiments later in this account completes that sequence. Leaving it unrun is a substantive outcome of the development program: the proposed study did not yet satisfy the conditions needed to interpret it. The apparatus and rejected design candidates remain reusable inputs, but their unfinished state must be visible to anyone considering the investment of completing them.

Sources: technical design law and feasibility, Section 5, collaboration runway, design-law candidate.

8. Smaller questions that could be completed locally

The next phase did not treat the unfinished full study as a reason to stop learning. It asked which useful obligations could be specified and checked without pretending to complete the larger program. Three contributions followed: exact subject binding, a finite audit of information loss, and an executable design-calculation candidate. Each supplied a narrower result with explicit assumptions.

8.1 Bind evidence to the work product it actually concerns

The development materializer carried extensive evidence fields, but an acceptance template could concern a different subject from the newly delivered task record. Capturing the record and the evidence together did not prove that the acceptance applied to that record. The new subject-binding profile made that missing relation explicit.

It derives an identity from the exact canonical delivered envelope and checks the claim and use references against it. A transformation creates a new envelope and therefore a new subject. The constructor creates no grant or declaration. Its positive example deliberately contains no acceptance basis: the identity check succeeds while acceptance remains unevaluated.

An independent review then identified an ambiguity in the specification's JSON escaping. Independent implementations could produce different bytes, and therefore different identifiers, while apparently following the same text. The repair pinned permitted escaping, number representation, and the absence of a byte-order mark or trailing newline. A literal vector was independently reconstructed. The issue was not that hashing failed; it was that the object being hashed had not been specified precisely enough.

8.2 What cannot be recovered from an information-poor view

The finite evidence audit asks whether two cases with different supplied labels become indistinguishable after projection to the same observed bytes. Partition the cases by their identical visible representation. Within each group, a deterministic classifier must emit the same answer for every member. Its best possible exact-label error count is the group size minus the largest label count. Summing over groups gives a tight lower bound on error for that finite, equally weighted set.

In notation, if n(v,y) is the number of cases with visible bytes v and label y, the minimum number of errors is:

sum over v of [sum over y n(v,y) - max over y n(v,y)].

The lower bound is attained by choosing a most frequent label in each group. This is a simple information argument, not a claim that the supplied labels are true or that the set represents deployment. It separates a representational limitation from an implementation bug: no different deterministic algorithm can recover a distinction that the observed view has erased.

Applied to fourteen existing development worlds, task-content bytes alone merged two pairs with different recorded revision relations, forcing at least two exact-label errors. The full COMMON representation retained those distinctions. The reduced projection was not the current candidate scorer input, so the finding was not a counterexample to that scorer. Conversely, no conflict in a finite full-input set does not establish universal sufficiency or authenticity.

8.3 Close the arithmetic around a candidate design

The exact design evaluator combined directional rows with assignment structure, budgets, missingness, arm mappings, capacity and resource limits. It did not select a real population, margin, model or sample size. Its synthetic example derived N=896, with the preceding lattice point failing two precision rows. That number is a demonstration value, not a replacement for P or D8.

A general counting argument explains why a uniform fixed-size segment of a random permutation is a simple random sample without replacement. Different arm segments can be dependent while each marginal segment still has that property. The argument depends on the stated permutation and potential-outcome assumptions; it does not qualify a real entropy source or establish independence from an operator's choices.

Finite checks covered 4,680 finite-word transcripts, thirty segment cases and 13,084 binary-population inequalities. More revealing than their size was a fresh review after the initial tests passed. It found that primary contrast labels were not joined to actual arm bindings, a single package claim could be split without corresponding multiplicity control, and observed intention-to-treat reasons could also be entered as unresolved. The revised candidate closed all three gaps and passed 59 tests. The failed candidate remains archived.

These results make the larger program easier to inspect and extend. They do not admit a study by arithmetic alone. An internally consistent design input can still omit an important endpoint or choose an unjustified margin. Scientific completeness and value choices cannot be recovered from the calculator's output if they were absent from its inputs.

Sources: contribution overview, subject-binding profile, finite evidence audit, design evaluator, validation and preserved failed candidate.

9. A direct applicability law, and its ordinary translation

The next operational question was deliberately smaller than the original lifecycle and delegation system: does one exact declaration apply to one exact recorded use under its named direct mandate and the observed withdrawal history? Naming the question allowed a short mathematical relation and a conventional implementation to be compared without importing every historical RR operation.

The declaration and use must agree on subject, receiver, purpose, scope and rule revision. The selected mandate must match the declaration's issuer and relevant tuple. The use must fall within the declaration and mandate intervals, and no operative withdrawal of that exact mandate may be known. A renewal under G2 does not silently substitute for a declaration that names G1; a new declaration must establish its own relation.

9.1 Why UNKNOWN is a logical result

Let K denote the conjunction of the admitted tuple and time predicates. If K is false, the selected declaration is inapplicable. If an operative withdrawal is observed, it is also inapplicable. If K holds and no withdrawal is known, applicability still depends on whether the receiver's relevant history is complete.

With closed history, the recorded history is the sole completion. With open history, one possible completion may add no operative withdrawal and another may add one. The first supports applicability and the second defeats it. The proper result is therefore UNKNOWN, rather than a guess that the missing event probably does not exist.

Admitted observation state	Verdict
A required tuple/time fact is false, or an operative withdrawal is known	INAPPLICABLE
All required facts hold, no operative withdrawal is known, and history is closed	APPLICABLE
All recorded required facts hold, no operative withdrawal is known, and history remains open	UNKNOWN

Malformed or ambiguously resolved references fail earlier prerequisites and return UNKNOWN with an explicit reason. Even a closed-history flag cannot turn a dangling named reference into a clean negative. The truth of history closure is supplied by the host; the evaluator does not derive it from a list or a digest.

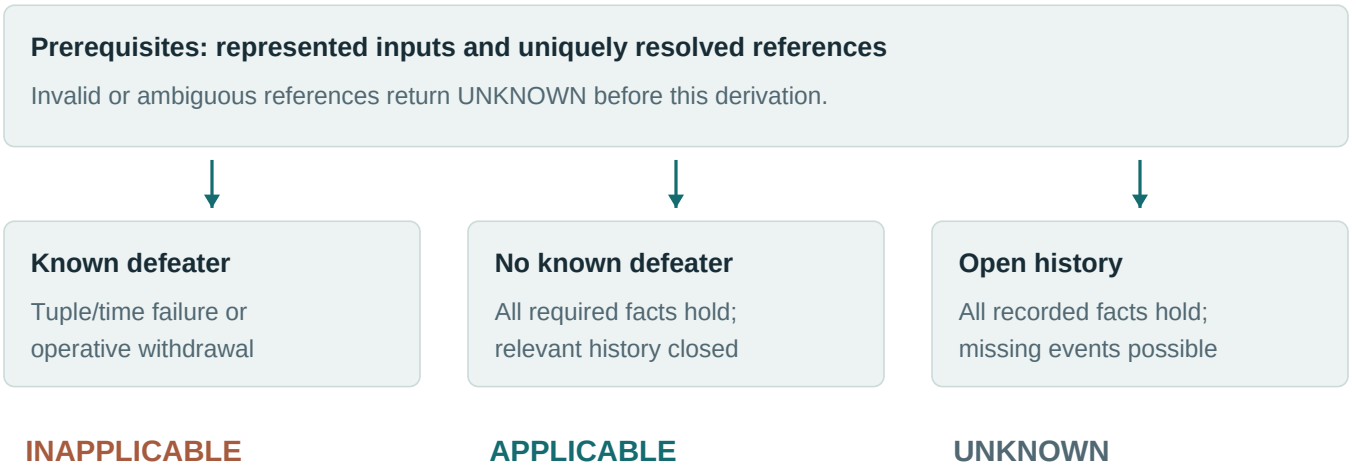


Figure 7. The direct-mandate relation under admitted observations. History closure is a host assertion whose truth remains outside the evaluator. The result is selected-declaration applicability, not authorization.

9.2 A conditional refinement argument

Under truthful immutable observations, uniquely resolved references and a history model that only adds valid withdrawal observations, further information narrows the set of possible completions. A conclusive verdict cannot flip; UNKNOWN may become conclusive. Changing the subject, mandate, revision or use time changes the question and is not such a refinement. Removing a known withdrawal also violates the premise.

The relation has an ordinary database interpretation. An exact join selects the mandate; tuple equality and interval filters compute K; an anti-join checks for operative withdrawals; and the closure rule chooses the three-state output. These operations compute the same predicates, so they produce the same verdict and reasons under the stated domain. This conditional algebraic equivalence is the reason a competent conventional comparator belongs in the experiment.

Two methods agreed across 110,592 finite combinations, with a separate completion check and 128 refinements. They shared validation and response formatting and had the same contributor. This is method agreement plus a conditional argument, not an independently authored complete RR implementation or a mechanical proof of the full engine.

9.3 Scope completeness is another relation

A separate coverage profile first requires distinct original object identifiers, then checks equality between their multiset and the multiset actually present. Under those prerequisites, equality means every required identifier occurs exactly once with no extras. Duplicate requirements are not admitted as a valid task scope. The finite test set contains 2,912 checks.

The task's original required set must come from an independently grounded source. Deriving it from the latest intermediate record would certify omissions already made upstream. Identifier coverage also does not prove the associated facts correct. The operational experiments that follow made both limits concrete.

The profile intentionally excludes recursive delegation, global conflict adjudication and broader supersession semantics. The archive preserves a real disagreement between a foundation addendum and a later candidate about unresolved replacement references and precedence. A small direct-mandate relation cannot silently settle that broader authority choice. Narrowing the schema kept the completed result interpretable.

Sources: direct-applicability specification and proof, coverage specification, first-principles redesign.

10. The operational experiment: identity survived, scope did not

Two local development versions executed the parcel workflow introduced in Chapter 1. A prepared recommendations, B formed a grouped manifest, and C recomputed a final allocation under the current rule. The measured guard assessed B at C's receiving boundary. It did not independently enforce every edge of the chain.

Each version used twelve fixed blocks: three parcel structures crossed with stable standing, same-rule renewal, changed-rule renewal and withdrawal. Every block ran unchecked U, conventional guard C and narrow-RR guard R. All six arm orders occurred twice. Each version therefore contained 36 chains and 120 actual model calls, with no missing episodes or transport errors. Across both versions there were 72 executions and 240 calls, but still only twelve constructed blocks repeated across conditions and development versions.

10.1 Prospective freezes, followed by disclosed refinement

The first source was committed before generation. Its result was poor: U completed three tasks correctly and each guarded package completed two. All first-version reference echoes were correct, yet task scope disappeared in upstream work. B could preserve A's facts faithfully while A had already omitted two parcels from the original task.

The second version was designed after inspecting all first-version outcomes. It replaced answer-shaped examples with neutral output-type descriptions, supplied the original required parcel identifiers to every role, checked exact coverage, and added a separate terminal notice path for withholding. Known withdrawal no longer triggered a futile regeneration. The direct applicability relation and underlying task oracle remained unchanged; the refined productive endpoint explicitly required a released manifest.

The refined source was separately frozen before its run. That chronology protects the identity of each version, but it does not turn the second into a held-out confirmation. Several changes were made together on reused cases. The improvement cannot be assigned causally to a single component.

10.2 Outcomes and their denominators

Version and package	Correct / 12 assigned chains	Productive-correct / 9 active cases	Correct explicit abstentions / 3 withdrawals	Calls
First U	3	3	0	36
First conventional	2	2	0	42
First RR	2	2	0	42
Refined U	11	8	3	36
Refined conventional	12	9	3	42
Refined RR	12	9	3	42

Whole-task correctness includes explicit withdrawal abstentions. Scale: 0-12.

Figure 8. Development improved the packages on reused cases. Conventional and RR outcomes match in each version. These are paired case counts, not estimates from independent new populations.

Each refined guarded package completed nine manifest-consuming C calls and three notice-consuming C calls. A notice contained no B payload and had its own identity. A correct guess in response to a notice would not count as productive use of a released manifest. All six observed notice responses were explicit correct abstentions.

The refined guarded packages improved on U by one correct allocation, in the three-parcel changed-rule case. They matched each other on every block. This is positive development evidence for the refined package on this fixed population and zero observed incremental outcome advantage for its RR formulation over the conventional implementation. Neither statement implies equivalence outside the tested cases.

10.3 Three transcript findings change the interpretation

First, exact identity was compatible with lost scope. In the original version A preserved all original parcel facts in 24/36 outputs. B preserved its actual A facts in 54/54 outputs, yet covered the original required identifiers in only 36/54. The local relation between A and B could be correct while the global task was already incomplete. The original task description had to remain an independent reference.

Second, an applicable-looking route could still lead to a wrong answer. In first-version B02, renewal changed the mandate identifier while the parcel facts and correct STANDARD grouping remained identical. Guarded C returned EXPRESS; unchecked C returned STANDARD. Binding and applicability did not ensure correct semantic use by the final model.

Third, an incorrect intermediate plan did not always prevent a correct final answer. The refined run retained two incorrectly grouped B records, but C recomputed their destinations correctly. This is direct evidence that consuming facts to recompute an answer differs from executing a plan. It limits any proposal that attaches one universal usability judgment to the entire record.

10.4 Cost, environment and independent checks

The guarded packages each used six more model calls than U, a 16.7% increase. Recorded refined model time was 125.513 seconds for U, 148.389 for conventional and 148.285 for RR. Loading and caching affected this single sequence. These are execution observations, not an estimate of intrinsic guard overhead or human time saved.

The local runtime was Ollama 0.33.3 with a pinned Qwen3:8B model digest, fixed seed, temperature zero, context 4096, output cap 640, JSON mode and thinking disabled. The hardware record names an RTX 4070 Laptop GPU with 8 GB VRAM. Capturing those identities enables a more exact attempted reproduction, but does not qualify a universally deterministic closure.

Separate outcome audits recomputed each version from raw responses and literal task answers, reconciling all 120 calls and 84 parent or notice links per version. These were author/context-separated local AI audits under the same operator. The raw prompts, responses, freezes, failed first version and post hoc transcript diagnosis remain available.

The completed result makes a practical design choice possible: retain original-scope coverage, a clear withdrawal path and a competent conventional implementation. It provides no basis for marketing the RR representation as uniquely effective, and no reason to count designed inapplicability cases as naturally discovered security incidents.

Sources: full operational results and frozen-run links, operational collaboration route.

11. Learning had to earn its place

The project also considered whether machine learning or reinforcement learning could improve recovery decisions. Once a record is incomplete or inapplicable, a system might refetch evidence, regenerate work, seek a new declaration, recompute or abstain. A policy could in principle allocate limited effort among these choices. But a learned action value cannot establish whether a source is authoritative or whether an answer is true.

The completed learning experiment therefore separated admissibility from optimization. It used a small, fully observed synthetic control model with 160 states, a perfect correctness diagnostic and stipulated dynamics. Ten seeds each trained for 10,000 episodes, totaling 100,000 episodes and 131,346 decisions. No policy or training setting was retuned after the outcome was observed.

11.1 Exact planning and a simple heuristic were strong references

Policy	Exact correct-completion probability	Abstention probability	Expected resource units
Exact planner	679/2560 = 0.265234	0.734766	0.533984
Dependency heuristic	661/2560 = 0.258203	0.741797	0.598047
Learned policies, mean of ten seeds	5041/25600 = 0.196914	0.803086	0.376172

These are exact policy evaluations under a uniform distribution over the stipulated states, not observed real-world success rates. The exact planner knows the transition model; the learner acquires values from sampled transitions. All ten final learned policies underperformed the heuristic. Lower resource use accompanied more abstention, rather than the same quality at lower cost.

Figure 9. Uniform exact evaluation over 160 synthetic states. Every final learned seed underperformed the heuristic. Lower learned resource use accompanies more abstention; these are not deployed-agent success rates.

A retained seed provides a concrete diagnosis. In a state where evidence and candidate were correct, the declaration current, standing active and execution free of remaining resource cost, the learned policy abstained. It had not sampled the terminal execution there; all action values remained zero and the tie rule selected abstention. This demonstrates finite exploration and value-propagation failure in that policy. It is not a safety rejection by the applicability mask.

11.2 A safe-looking final policy did not prove learned safety

Training summaries record 11,184 correct, 358 wrong and 88,458 abstaining terminals. The frozen greedy policies had no wrong completions under the toy evaluation, but a perfect correctness observation, reward structure, initialization and tie-breaking all favored that result. It would be misleading to credit the absence of wrong greedy outputs to a learned understanding of truth or safety.

Only the first two episodes per seed were retained as detailed traces, and all twenty are abstentions. The full transition stream has a checksum rather than a stored transcript. Thus individual productive training trajectories cannot all be independently inspected. A separate audit checked the retained evaluation vectors, final policy choices, sensitivities and accounting without importing the author's planner or rerunning the registered training experiment.

All 27 sensitivity settings retained a positive learned-policy gap. Because gaps already existed in the default model, this does not isolate degradation caused by distribution shift. A post hoc positive-budget-only diagnostic also left all ten policies below the heuristic; the result was not solely due to zero-budget states.

11.3 Why Deep RL was not the next step

The model had 160 known states and no demonstrated need for a neural representation. The operational chains did not supply a generalization dataset. Adding Deep RL would have increased complexity before establishing a problem the simpler policy could not handle. The supported local decision was to retain the heuristic or exact planning where the transition model is trustworthy.

Learning may still be relevant when diagnosis is partial, observation costs vary or dynamics must be estimated. That would require new calibrated observations, a real objective and a comparison against strong fixed policies. The completed experiment supplies evidence against this tested configuration, not against reinforcement learning as a discipline.

Source: frozen learning findings, source identities and independent outcome audit.

12. Testing the reporting lesson instead of assuming it helped

RR's own reporting errors suggested another possible application: helping a writer preserve the relationship between evidence and a changed claim. A recorded experiment can be cited historically without becoming evidence of current efficacy. A correction to one clause does not invalidate every unchanged clause. A source may support a narrow numerical statement but not a stronger recommendation built from it.

The project first made those distinctions executable in a structured reporting prototype, then tested whether an added generic evidence-use procedure improved model answers. These are separate steps. A deterministic prototype that correctly applies a supplied policy does not establish that asking a model to follow a similar procedure will improve its work.

12.1 A finite reporting contract, with conventional parity

Twelve paired cases produced twenty-four decisions per implementation. Candidate and conventional methods matched on all twenty-four: twelve legitimate uses retained, three wrong values contradicted, and nine changed uses judged insufficiently supported. Three generated evidence cards contained 38 statements through a shared renderer.

The semantic assignments and templates were human-authored within the project. The system did not infer whether arbitrary prose was justified. An ablation that omitted use checks was intentionally incomplete and could show that the checks mattered to that contract, but not that RR outperformed a competent alternative. Ordinary lookups and joins were sufficient for the measured workflow.

12.2 A frozen prompt pilot

The next study asked a narrower empirical question: given explicit task requirements and the same source documents, does an added generic evidence-use review improve final answers beyond a similarly sized general review checklist?

An independently tasked agent authored 24 constructed task families based on public Python, USGS and NIST documentation. Six were development families and eighteen evaluation families. The evaluation set covered transformation, correction and changed purpose across the three domains, with nine legitimate-use tasks and nine requiring qualification or revision. Scenarios and numbers were synthetic.

O was an ordinary source-grounded prompt. G added six general review steps, totaling 101 words. E added six evidence-use steps, totaling 102 words. All conditions already contained the same explicit task-specific requirements. The contrast was the additional generic procedure, not the presence versus absence of careful requirements, and not the whole RR system.

The prompts were frozen before the root inspected evaluation content. All conditions shared the final-answer schema, no tools, an 8,192-token context and a 1,024-token output cap. Gold answers and study strata were excluded from inference. Two installed quantized local 8B models were used, with Qwen primary and Hermes secondary. There was one output per model, family and condition, with fixed seed and randomized family/arm order within model blocks.

All 144 planned requests completed with nonempty, strictly shaped JSON and normal stopping: 36 development and 108 evaluation. There were no uncertain requests, retries, development repeats or prompt revisions. Successful transport and schema compliance therefore did not explain away the semantic failures.

12.3 The result and its uncertainty

Model and grading reading	Ordinary O / 18	General review G / 18	Evidence-use E / 18
Qwen: all three reported readings	8	9	6
Hermes: root	3	4	5
Hermes: independent agent and reference adjudication	3	3	3

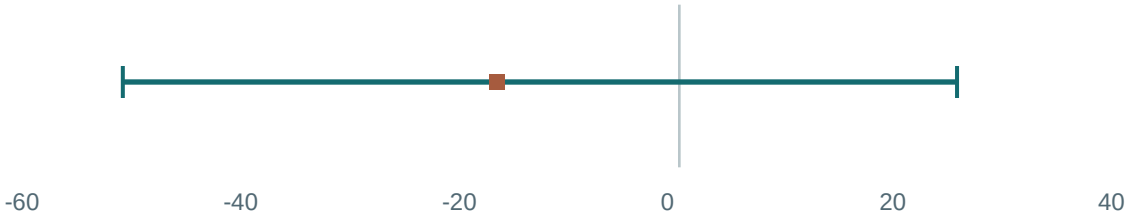
For Qwen's primary E/G comparison, five families passed both, one passed only E, four passed only G and eight failed both. The paired difference was -3/18, or -16.7 percentage points. The prospectively selected exact two-sided paired test gave p=.375. The conservative 95% interval was [-50.9, +25.4] percentage points. It remains wide enough to include useful benefit as well as substantial harm; the result does not establish equivalence or a general absence of benefit.

Paired Qwen families

Both pass: 5 E only: 1
G only: 4 Both fail: 8

Primary comparison

E: 6/18 G: 9/18
Difference: -16.7 percentage points



Conservative 95% interval [-50.9, +25.4] points; exact paired p = .375.

Figure 10. The point estimate favors general review, but the interval is wide. Constructed and potentially dependent families limit its sampling interpretation. This pilot tests a prompt addition, not the full RR treatment.

The interval subtracts simultaneous 97.5% Clopper-Pearson bounds on the E-only and G-only probabilities. Its sampling interpretation still assumes binomial behavior across families. These were purposively constructed cases, sharing sources and templates, rather than a random sample of real work. A three-domain block sensitivity gave p=.50. More model calls and multiple graders do not increase the number of independent task families.

The selected exploratory-benefit rule was not met. E also trailed O descriptively. The proper conclusion was no demonstrated benefit for this prompt addition under the tested conditions. Neither a secondary model nor a favorable scoring choice was promoted to replace the primary comparison.

12.4 All five primary disagreements, not selected anecdotes

Family	Observed E/G difference	What distinguished the answers
Timestamp correction	E passed; G failed	E retained the requested old-naive/new-aware distinction
Reflectance correction	E failed; G passed	E stated the right scale and offset but calculated values incorrectly
Footprint coverage	E failed; G passed	E invented extra quality filtering in its explanation
Seed reuse	E failed; G passed	E omitted the requested narrow generator compatibility guarantee
Uncertainty types	E failed; G passed	E assigned the statistically evaluated fit to the wrong uncertainty type

This complete discordance list gives the aggregate a concrete meaning. The failures include arithmetic, unsupported explanation and incomplete requested work. They suggest that stating a procedure and executing its demands are distinct. They do not identify why the extra procedure failed or prove that a particular new checker would fix it.

12.5 Grading was itself uncertain

Root and a fresh-context agent graded individual outputs with model and condition labels hidden before unblinding. The root had authored the prompts, and style could reveal conditions. Both graders were AI systems within one operator's project.

The graders differed on at least one criterion in 59/108 evaluation records, although only three whole-task outcomes differed. Reporting only whole-task agreement would conceal that uncertainty. Reference adjudication retained both original readings and documented changes. It did not become a uniquely correct human gold standard.

One Qwen answer failed for different reasons in the two original readings. Combining their favorable criterion judgments could make it pass. The resulting criterion envelope was E=6-7, G=9 and O=8 out of eighteen. Even within that favorable recorded envelope, the observed primary direction remained negative. This is a descriptive robustness check over known disagreements, not protection against shared grading mistakes. Hermes' envelope allowed either direction and supported no efficacy claim.

The frozen sources, exact requests and responses, both original grading files, adjudication and paired table remain available. The study answered a modest empirical question without tuning until it produced a favorable result. That is useful evidence for deciding what not to promote as an already validated intervention.

Sources: structured use-checking prototype, prompt pilot results and source archive, machine-readable results.

13. Full operational treatment in a qualified local host

The next investigation returned to the complete operational R-all classifier. The narrow operational law, learning exercise and prompt pilot described earlier were different interventions. Their results could motivate a new question, but they could not answer whether the full classifier changed what downstream models actually received. This successor therefore restored the full staged treatment and constructed a common executable host in which all arms produced their own model trajectories.

The implementation and observation boundary were deliberately local. An honest controller created and captured administrative records, invoked a fresh decision process, retained its engine-call and witness-construction observations, and released the captured input to a model only after PASS. Two independently authored treatment implementations were reconciled and checked on development inputs; package A was fixed for collection. Development resolutions in the treatment specification remain disclosed,

including namespace and parsing interpretations. The study does not claim that the original archival bytes ran without repair.

This was an admission of a bounded local study. It did not satisfy or inherit the historical public/adversarial program's different qualification, registration, adaptive diagnostic or administrative-resistance requirements. Complete operational treatment means the entire decision procedure is applied, not that four constructed conditions exhaust every failure it may encounter in deployment.

13.1 Development was allowed to change the candidate

The apparatus did not move directly from specification to measurement. An independent host review found that identical parent payloads could collapse two different parent identities when identity was only a content digest. The repair registered a node-specific canonical envelope and kept its identity distinct from the payload hash. The review also exposed arm-contaminated episode identity and an ambiguous sum instruction: a model could reasonably interpret the wording as adding a parent subtotal once per element rather than once to the aggregate. The task prompt was corrected and the earlier calibration was stopped. Its raw requests and completed episodes were retained as invalid development evidence for capability selection, not reassigned to the later study.

A new development screen used the corrected workload and a declared first-eligible selection order. The 8B nonthinking profile produced 3/24 correct episodes and 23/24 fully parsed episodes; the 8B reasoning profile produced 15/24 correct and 15/24 fully parsed episodes. The 14B reasoning profile produced 24/24 of each and was selected. These observations are calibration results on constructed worlds, not comparisons of RR arms or a general model ranking. Local execution was too slow for the fixed study, so the same model digest and decoding profile were qualified on the authorized Colab GPU. A separate 24-world development run there was fully correct and parsed, with 88 actual model calls. The qualification's median episode duration was about 19 seconds, compared with about 152 seconds for the selected local profile.

Qualification then checked the actual treatment/host boundary. Sixty-two normative R cases and three comparator checks on both implementations produced 130 expected outputs; plain and observed executions agreed. Twelve bounded wire fixtures checked a disclosed parsing-limit resolution. A task-message interpreter ran 264 synthetic host episodes with actual fresh decision workers and 714 A/B decision comparisons. Because that interpreter was a qualification stub, those episodes were not LLM efficacy evidence. A further 16 live development episodes on Colab exercised held, repaired, clean and semantic-control trajectories. Including four repeated warm requests, it retained 52 actual requests and responses and 44 A/B decision comparisons. The four repeats agreed with the original response contents; that small check is limited evidence about the runtime, not a proof of determinism.

The independently authored scorer reconstructed original task truth and current predicates without importing the treatment, host or workload code. It was checked on 282 retained development/qualification episodes and eight sensitivity families. Changing labels or a claimed summary did not change the score; missing or spliced witness evidence failed measurement. A recurrent downstream predicate was counted as a current occurrence rather than being replaced by an ancestry label. Separate deadline fixtures ensured that even a final output arriving after the episode cap could not count as timely completion.

The first final collection review required repairs to isolated observation-failure handling, interrupted active-time accounting and the placement of invocation finish evidence in the exported archive. Those repairs were

exercised with bounded lifecycle fixtures. The next exact-state review admitted the source and protocol with explicit root obligations for local custody and lost-VM accounting. This development chronology matters: candidate defects were repaired before the study freeze, and failed development attempts were not silently promoted into successful qualification or evaluation outcomes.

13.2 Tasks and populations

Every episode had a source model and either one receiver, a chain of three receivers, or a fork/join of four receivers. The two tasks were summing integer groups and selecting item identifiers by grade and count. The source operated on group A. Receivers combined their actual parent outputs with fresh task groups; the fork/join preserved overlap correctly in the task instructions. A standalone scorer computed ground truth directly from the original groups. Neither a reference answer nor another arm's model response was supplied to a model.

The finite evaluation reservoir contained 256 worlds in each task/graph cell. It was separate from the development reservoir. A fixed arm-neutral function assigned a decoding seed to each world. Each arm drew its own worlds uniformly with replacement, and a separately drawn full-list shuffle determined execution order. Repeated worlds remained repeated isolated intervention executions; they were not represented as new naturally occurring incidents.

Four equally weighted seed conditions operated at the first receiver: an absent adoption declaration, an adoption naming the wrong use scope, expiration of a local ingress grant before its benign file-copy operation, and delivery of a parent envelope at a different revision while preserving its payload. The host created the administrative event before capturing the resulting facts. It did not edit a completed observation to make a label appear true. Later receivers received their own current-use records, so a valid child declaration could clear an earlier administrative defect.

Two equally weighted clean conditions used direct standing and one-hop delegated standing. A separate semantic control changed the source answer after preserving the genuine source request and response: add one to a sum, or toggle one item identifier. This was a declared constructed input, not a model output attributed falsely to the source. It tested whether structural acceptance could coexist with an actual task error. A control's final correctness still had to be measured; the construction alone did not determine the final answer.

The four arms were transparent U, static S4, secondary static S1, and R-all. U applied shared admission and then passed transparently. S4 and S1 used the original independently implemented static comparator's three component and thirteen extended bits, with thresholds four and one respectively. These were fixed comparator packages; they were not a newly optimized conventional implementation of the complete R-all relation. The same common host governed capture, release, correction opportunities and resource limits. Equality meant the same rules for the same captured input and facts. It did not require identical later inputs after different earlier decisions or different actual model outputs.

13.3 Correction, observation and endpoints

A receiver's first non-PASS permitted one administrative refetch with unchanged parent payloads and intended use. Even-indexed reservoir worlds had a correcting service; odd-indexed worlds retained the defect. The reservoir mixture was exactly half each, while the actual sampled mixture could differ. Refetch did not consult gold, regenerate a parent or change an answer. A second non-PASS held the receiver. Dependents lacking a required output were cancelled; an available independent sibling could continue. Each receiver was presented at most once.

The primary seed endpoint counted actual presentations satisfying at least one currently evidenced registered structural predicate. A presentation counted once even when several predicates held. The count's support was zero to one, three or four by graph. All current presentations were checked, including a newly created downstream occurrence. An ancestor's label or defect did not itself make a later presentation erroneous. Descendants of an observed defective presentation were retained as a separate descriptive quantity.

The primary clean endpoint required every receiver output to parse under the fixed strict task grammar, the final answer to match independently computed truth, and the episode to remain within its deadline. A fully observed hold, malformed task output or known deadline remained an assigned outcome. Missing model or witness evidence was a measurement failure, not a successful zero. Completion in the seed population, semantic input and output errors, intermediate errors that later recovered, controller holds, retries, tokens, model calls and elapsed time were retained separately. Structural conformance did not establish answer truth.

The host used Ollama 0.33.3 and a pinned qwen3:14b model on a qualified Colab A100-SXM4-40GB image. Requests used temperature zero, context length 8,192, JSON mode, thinking enabled and a 1,536-token generation cap. There was one HTTP attempt per model operation, a 180-second transport timeout and a 1,200-second episode deadline checked between operations and after the final one. An in-flight call's overrun was retained and could not qualify as timely completion.

13.4 An original freeze and an explicit initial-stage amendment

The original schedule assigned nine draws per seed condition/task/graph cell per arm: 216 seed episodes for each arm. It assigned 32 per clean cell per arm: 384 clean episodes for each arm. The four arms therefore had 2,400 primary episodes, plus 24 semantic controls. The exact 2,424-slot schedule, reservoir, random draw tape, source identities, decoding profile and review admission were saved before evaluation. This was a local precommitment, not a public registry submission.

After collection began, James requested a smaller credible initial-stage study. An outcome-blind assessment inspected only the frozen order and design labels. It proposed the first 1,800 scheduled entries plus the five semantic controls outside that prefix, retaining every acquired observation and all 24 controls. The selected sample contained 1,805 episodes; the other 619 original primary assignments were deferred. They were not zeros, successful outcomes or missing required observations. The original schedule and its full 2,424-row accounting were preserved. This completed an amended initial stage, not the originally planned full sample or the historical public/adversarial study.

The assessment and implementation each received bounded independent internal review. James adopted the change while 1,400 episodes were locally verified and batch 15 was running. Operational counts, timing and

collection status had been visible, but evaluation endpoints had not been scored or inspected to choose the sample or stopping point. The dated record is a post-start, pre-analysis amendment; calling it an original preregistration would misrepresent its timing. The first 1,500 episodes used the original collector. The additive collector then ran the next three unchanged 100-slot slices and the five remaining controls, with exact identity checks and the same model, treatment and host. Original frozen source and scorer bytes remained unchanged.

Initial-stage population	R-all	U	S4	S1
Seed	162	151	155	157
Clean	292	299	291	274
Semantic controls	6	6	6	6

The episode remained the inferential unit. Receiver events did not increase the sample size. Within each primary population, every condition/task/graph cell retained equal prospective weight. Unequal realized cell counts entered the stratified estimator; pooled arm means would estimate a different mixture. Six fixed contrasts compared R minus U, S4 and S1: current structural presentation count in the seed population, and correct completion in the clean population.

The amendment allowed at most two fixed inferential datasets: this 1,805-member initial stage and the original full 2,424-member sample if separately authorized. No full extension was executed. The exact rational variance-bound term remained S = sum over cells of w squared times support-range squared times (1/n_R + 1/n_comparator). The radius rule changed explicitly from h squared = 3S to h squared = (25/8)S. An exact finite exponential certificate bounds each directional tail below 1/512; the union bound over 24 directional tails across the two possible six-interval reports gives joint coverage of at least 61/64, or 95.3125%, conditional on the original sampling and runtime assumptions. The two reports need not be independent. Roots were rounded outward to six decimals and intervals intersected with their possible contrast supports.

The selected resolution was at most about 0.599 current improper presentations per seed episode and 0.1494 clean-completion probability. These tolerances are coarse initial-study choices, not a universal minimum for credibility, a power calculation, a guarantee of significance or a stakeholder-justified harm margin. A missing required observation invalidated each contrast depending on that slot; a global source/runtime identity failure invalidated estimates globally. Known holds, malformed outputs and observed deadline failures remained assigned outcomes. Intentional deferral was outside the selected stage. This local invalidation law differs explicitly from the historical contract's worst-case completion envelopes.

Collection retained its 14-hour active-time cap and three-interruption-resumption ceiling. A durable local lease identified every assigned batch before execution. Each checkpoint was downloaded and checked against its emitted hash, complete file manifest, assignment identities, closed invocation and byte-identical prior overlap before the next lease. Interrupted slots were never rerun; issued but unverified slots would have become missing after confirmed VM loss. The final initial-stage evidence consists of nineteen locally verified checkpoints, including one five-control checkpoint. The complete original schedule remains in the final ledger, with the 619 deferred entries explicitly distinguished.

Sources: post-start amendment, fixed membership, amendment admission, independent amendment review, original local admission, outcome-blind sample assessment.

13.5 What a result can mean

The causal interpretation is conditional on stable seeded potential outcomes, or independent stochastic executions with the same conditional distribution. Fresh messages and processes, fixed decoding seeds and a small successful cold/warm repetition check did not prove stationarity. Numerical variability, caching, timing and interruptions remained limitations. Administrative facts and custody observations were auditable within an honest-controller experiment; they were not authenticated external facts or administrator-resistant attestation.

A result can support a package comparison over the frozen constructed population and execution boundary. It cannot identify individual components' effects, estimate natural error prevalence, establish semantic truth from provenance, demonstrate general security efficacy or field utility, or establish broad novelty. Whether structurally stricter release also improved task completion is an empirical question answered by separate observations in this study.

13.6 Results: structural release changed; clean utility remains uncertain

All 1,805 selected episodes were scored VALID. This status means the required measurement evidence was present and coherent, not that every task succeeded. The report had no global identity errors and all six contrasts were VALID_CONDITIONAL. Nineteen locally verified checkpoint archives accounted for 88,865 unique evidence files. No interrupted episode was replaced and no VM-loss resumption occurred. Independent internal checking recomputed all six estimates and exact radius calculations from the episode scores without importing the production analyzer or estimator. A fixed trace sample also checked assignment, observation and actual model-input correspondence; this was a bounded audit, not a second complete trace census.

Population and endpoint	Contrast	Estimate	Conditional simultaneous interval
Seed current improper presentations	R minus U	-1.000	[-1.599, -0.401]
Seed current improper presentations	R minus S4	-1.000	[-1.596, -0.404]
Seed current improper presentations	R minus S1	-0.750	[-1.334, -0.166]
Clean completion (percentage points)	R minus U	-0.375	[-14.973, +14.224]
Clean completion (percentage points)	R minus S4	+0.888	[-13.831, +15.607]
Clean completion (percentage points)	R minus S1	-0.429	[-15.362, +14.505]

The intervals use the amended two-report certificate: joint coverage at least 95.3125% conditional on the declared sampling/runtime assumptions. Table values are display-rounded; the report retains exact rational estimates, squared radii and outward bounds. The structural endpoint is a count, not a percentage or a rate of naturally occurring incidents. Lower seed values favor R; higher clean-completion values favor R.

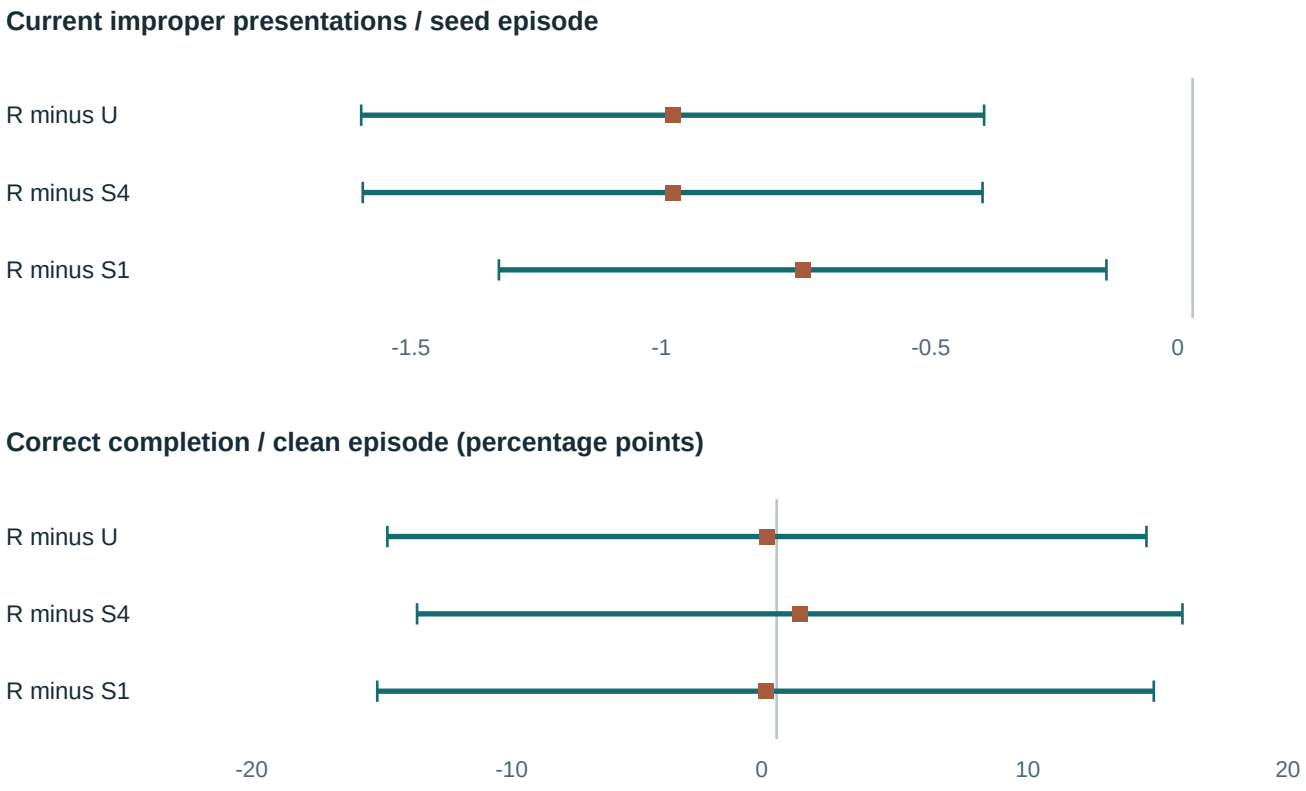


Figure 11. Six fixed initial-stage contrasts. Seed intervals are below zero; clean intervals cross zero. The two-report certificate gives joint conditional coverage of at least 95.3125%. Panels use different units. These intervals do not certify seed completion or semantic benefit.

The equal-stratum seed means were zero for R, one for U and S4, and three quarters for S1. No R seed episode presented a receiver while a registered current structural predicate held. U and S4 each retained one such presentation per episode. S1 removed that occurrence in one of the four equally weighted seed conditions. Because all three upper interval bounds were below zero, the fixed conditional comparisons support reduced structural exposure for the full package within this constructed population.

Clean completion does not support a parallel headline. Its estimated R differences were -0.375, +0.888 and -0.429 percentage points against U, S4 and S1. The intervals extended roughly 14-15 points in each direction. They establish neither equivalence, noninferiority, absence of harm nor a demonstrated utility gain. No stakeholder-derived acceptable-loss margin was tested.

13.7 Completion, semantic errors and cost must travel with the result

Population	Arm	Episodes	Correct completions	Model calls	Mean episode seconds
Seed	U	151	150/151	565	18.41
Seed	S4	155	154/155	561	18.07
Seed	S1	157	139/157	534	16.74
Seed	R	162	80/162	395	12.38
Clean	U	299	298/299	1109	18.73
Clean	S4	291	286/291	1070	18.37
Clean	S1	274	273/274	1009	18.50
Clean	R	292	290/292	1064	18.42
Control	U	6	0/6	22	20.25
Control	S4	6	0/6	22	18.17
Control	S1	6	0/6	22	18.88
Control	R	6	0/6	22	19.53

These are pooled descriptive counts and means over the realized draws, not the equal-stratum primary estimators. They must not be substituted into the six fixed contrasts. They also show why lower exposure is not automatically better completed work. R's 80/162 seed completions accompanied 81 terminal holds and 162 administrative retries. A correcting service could repair a missing or stale administrative relationship without changing the answer; a persistent defect stopped required work. The administrative intervention often concerned a mathematically usable payload. Holding it could reduce the structural endpoint while preventing an otherwise answerable task.

S1 made 36 retries and completed 139/157 seed episodes. U and S4 made no retries there and completed 150/151 and 154/155. These descriptive utility differences are substantial, but no additional inferential interval for seed completion was in the six-contrast family. The study does not silently promote them into a separately certified effect. The report retains all observed failures, including token-limited malformed model outputs, rather than deleting them as inconvenient trials.

All 24 semantic controls completed their trajectories. Each arm nevertheless had zero correct final answers out of six, with sixteen semantically erroneous receiver inputs and sixteen erroneous receiver outputs across its six graph/task cases. All had zero registered structural-error presentations. In these controls the handoff could be structurally acceptable and consistently wrong. R-all supplied no semantic repair or truth oracle.

Mean clean episode times ranged from 18.37 to 18.73 seconds across arms; model calls, tokens and wall times remain descriptive execution costs on one qualified GPU runtime. R's lower seed time and 395 seed model calls

accompanied fewer executed receivers, so they are not evidence of equal-quality efficiency. Across all populations the study retained 6,395 actual model calls. No monetary-cost advantage or general latency distribution was estimated.

13.8 The contribution after this comparison

This study adds a positive, bounded operational result to a record containing earlier negative and null results. Restored R-all, supplied with auditable local facts and applied to actual model releases, reduced the specific current structural exposures relative to the declared static packages. It did not show that the RR vocabulary or engine is uniquely necessary. The static comparators did not implement the complete dynamic relation; an ordinary policy that did so was not a fifth arm. The earlier conventional parity findings therefore remain intact in their narrow settings, while a universal claim of parity across all measured packages would now be inaccurate.

The study also makes the cost of the intervention visible: persistent structural defects blocked task completion, clean-utility uncertainty remained broad, and coherent semantic errors passed every arm. These observations support a separation between structural admissibility, semantic correctness and practical utility. Any stronger claim would require a separately designed population, comparator and outcome. The present study is complete at its amended initial stage. Neither the deferred 619 episodes nor the historical public/adversarial program were executed as part of this result.

Sources: full treatment and disclosed resolutions, original statistical specification, initial-stage report, all episode scores, original schedule accounting.

14. What the field comparison changed

A project can make a useful distinction without originating it. RR's language of receiver-owned evidence, scoped authority and justified use overlaps established work in security, provenance and evidence-based revision. The field review therefore had to ask a more demanding question than whether another paper used the name Receiver Reliance.

Proof-carrying authentication already develops authorization from formal proofs. RATS separates evidence appraisal by a verifier from a relying party's decision. Policy-carrying data connects data and use restrictions. CaMeL separates trusted control from untrusted data and applies out-of-band enforcement. These are substantive precedents for the broader architectural ideas, even though they have different mechanisms and targets. RR's classifier is not an authenticated proof system, and a receiver-boundary diagram alone does not establish novelty. Appel and Felten, 1999; RATS, RFC 9334; Padget and Vasconcelos, 2015; CaMeL, 2025.

The retained field assessment also considered newer agent-memory and execution work, including AuthMem, MemTX and context-to-execution integrity. Their problem statements overlap authority loss during transformation or consolidation, evidence validity and action boundaries. Those overlaps constrain positioning; their published performance is not reproduced by RR's local tests. The dated review records what was inspected and what remains uncertain. Retained field and direction assessment.

14.1 A proposed pivot encountered closer prior art

The reporting failures suggested maintaining evidence relationships as documents change. An initial assessment considered this a promising direction but did not identify the closest article-updating work. Subsequent checking found FRUIT, which explicitly studies updating existing text from new evidence while preserving unchanged material, and WiNELL, which addresses ongoing fine-grained article updates. That discovery narrowed the proposed contribution again. Complete-document updating was not an unoccupied research category. FRUIT, NAACL 2022; WiNELL, 2025.

This correction belongs in the research record because the AI-assisted assessment itself had missed relevant work. A confident statement that a direction looks distinctive is not a literature review. The useful next question would be whether a specified maintenance method improves complete updates or total effort against strong existing methods, with unchanged justified content counted as something to preserve. The existing RR cases can help develop such a test, but they are already inspected and unsuitable as an unseen evaluation population.

James also proposed a research skill for ordinary people: help them reach durable conclusions sooner and avoid AI sycophancy, misunderstanding and unsupported claims. The review found direct overlap with research companions, tutors and evidence-linked autonomous research. Its source-access limits are explicit: some full papers were unavailable, so publisher abstracts or author accounts supported only narrower comparisons. The broad idea was not established as novel, and no skill or human study was completed in this lane. Novelty corrections and access limits.

14.2 A fair comparison is part of the contribution

Across the project, the strongest revisions came from improving the comparison. A conventional verifier must be allowed ordinary joins, source recovery, caching, explicit requirements and appropriate repair. A reporting baseline must preserve correct content rather than rewrite indiscriminately. A learned policy must compete against a sensible heuristic and, in a small known model, exact planning. A prompt addition must compete against a similarly sized generic review, not only against an intentionally careless prompt.

These are established principles of evaluation applied to a concrete record. Their value here lies in the cases showing why they mattered. The archive makes it possible to see how a weaker comparison could have supported an impressive but misleading story, and how a stronger one changed the surviving claim.

15. What is now supported

The completed record supports several distinct statements. Some are software results, some mathematical, and some empirical. None should borrow the scale or certainty of the others.

Result	Supported interpretation	Boundary
Reference classifier and audited contracts	Specified structural decisions can be implemented and checked	Host observations and effects remain separate obligations
Conditional applicability and coverage relations	Exact use relations and task coverage can be made explicit	Assumptions, declared scope and conventional equivalence remain
Retrospective replay audit	Earlier provenance, denominator and comparator interpretations needed correction	No natural comparative advantage established
Operational development	Refined packages completed useful toy work; conventional policy matched RR	Reused small population, multiple joint refinements
Finite learning	This tested learner underperformed the heuristic in all ten final seeds	Known toy dynamics, no general result against learning
Frozen prompt pilot	The added evidence-use procedure did not demonstrate benefit	Wide uncertainty, constructed tasks and AI grading
Historical full-study apparatus	Substantial methods and development work exposed admission blockers	Original public/adversarial study unrun; broad efficacy unestablished
Local full operational initial stage	R reduced registered current seed exposures versus U, S4 and S1	1,805 selected episodes; amended sample; seed holds, broad clean-utility uncertainty and semantic failures

15.1 The most progressed form of the idea

The most defensible form of RR is an explicit contract between a received object, the evidence available about it, and the particular use being proposed. Identity answers which object. Applicability answers whether a selected declaration covers this use. Coverage asks whether the original task survived transformation. Correctness asks whether the resulting answer satisfies the task. Authorization and enforcement remain separate system responsibilities.

For the completed narrow workflows, ordinary deterministic composition is sufficient to express these relations. The measured record does not justify adding a distinct RR engine or a learned layer merely to obtain the same decisions. A useful implementation can retain the distinctions and use the simplest mechanism that reliably computes them.

This conclusion is narrower than the original aspiration for broad security efficacy, but it is not a claim that the receiving boundary is irrelevant. The results locate places where a system can go wrong and show that the proposed check alone is insufficient. They also identify the evidential step missing from several attractive claims: a comparative gain must persist when the conventional system receives the same facts and capabilities, and it must improve completed work rather than only a proxy.

15.2 What a collaborator can contribute without inheriting the whole project

Entry point	Concrete work	A useful result could be
Reproduce the completed evidence	Verify archived inputs, arithmetic and the paired outcome tables	A confirmed result or a precise discrepancy with source bytes
Reimplement the narrow relation	Work from the applicability specification before reading its methods	Independent agreement, an ambiguity or a counterexample to the stated contract
Reassess semantic grading	Read source packets and blinded pilot answers with an explicit rubric	A revised estimate and transparent uncertainty, not a preferred majority vote
Study a real receiving workflow	Select an authentic operation, competent comparator and complete task outcome	An incremental benefit, a cost-quality tradeoff or a credible negative result
Complete a mathematical dependency	Critique assignment, margins or the whole-design evaluator against its claim	A valid reduction in cost or a reason the desired claim is infeasible

The first three routes reuse a completed artifact and require no commitment to the unfinished full study. The last two are new research and need their own prospective worksets. There is no obligation to revive every historical subsystem. A collaborator should be able to choose one falsifiable question and contribute an answer whose scope can be stated clearly.

The reason to engage is therefore concrete: the package contains real specifications, retained failures, conventional translations, raw local outputs and disputed grading decisions. A critic can inspect the exact point of disagreement. The contribution is not contingent on accepting RR's original framing or producing a favorable next result.

16. What the journey demonstrates

RR began by asking how one agent could justify relying on another's work. During development, the same question applied to the research itself. A test result needed a correct denominator. A seal needed a specified subject and law. A review needed to recompute rather than repeat PASS. A claim of benefit needed a capable comparator. A statement about novelty needed closer prior art. Each was a receiving decision about whether a supplied artifact supported its next use.

The project's rigor is best judged through those decisions. The record retains the first operational failure alongside its refinement, the refuted formal obligation alongside the proved ones, the comparison that conventional policy matched, and the prompt result that did not support the proposed intervention. It also records where the evidence is incomplete: missing early extraction records, unretained training trajectories, shared operator and grader limitations, and an unrun historical public/adversarial study. The new local operational comparison is retained with its positive structural result, its completion tradeoff and its explicit post-start sample amendment.

There are limits to what this demonstrates about the research process itself. The project did not run a controlled study showing that its documentation, AI reviews or workflow made research faster or more reliable. Nor does the quantity of material establish expertise by itself. What a reader can assess is the observable work: problems specified, programs built, assumptions exposed, tests executed, interpretations corrected and source evidence preserved.

For James, that is the durable professional account this artifact is intended to provide. It presents a solo-led project with substantial AI assistance and makes the contribution available for technical scrutiny. It neither attributes every machine-produced analysis to unaided human work nor treats assistance as a substitute for accountability for the final claims.

The original broad claim remains open. The completed investigation nevertheless provides an answer worth keeping: exact structural evidence is useful only when it concerns the actual object and operation, and demonstrating that relationship is still different from demonstrating better outcomes. In the earlier narrow settings, conventional methods matched the proposed checks; the tested learner did not improve the policy, and the added review prompt did not establish benefit. The later full operational study reduced specified structural exposures against its fixed static comparators, while persistent defects often prevented completion and semantic errors remained unchecked. That result earns a bounded structural claim without settling broad practical benefit. The archive makes those findings available as results to build from, rather than reasons to conceal how the project developed.

Appendix A. Evidence inventory and recovery limits

This inventory keeps unlike units separate. It is not a sum of independent evidence or a claim that every historical artifact was reproduced during this editorial synthesis.

Package	Retained scale or result	Evidence status
Native replay	408 records: 390 clean real-labelled, five defective real-labelled, thirteen seeded	Retrospective classification; origin and semantic-label limits
August precursor	Twenty agent runs, 160 answers	Exploratory, uneven cohorts; arithmetic inputs recovered
Public reference software	Thirty operations; historical 907-case suite and twenty live gates	Historical conformance record for specified release
Formal report	872 properties: 766 proved, 105 bounded, one refuted	Conditional claims in declared formal domains
Second implementation	Atomic agreement and 592 full-surface differences	Different tested layers; complete-interface obligation open
Original demonstration	Seven smoke classes plus 51 extension classes; twenty model source records	Instrument and treatment development, not efficacy
Astra reference contribution	Subject binding, finite information audit, candidate design evaluator; 77 component tests	Narrow local contributions, not study admission
Applicability and coverage	110,592 comparisons; 128 refinements; 2,912 coverage checks	Finite checks supplement conditional arguments
Operational versions	Twelve blocks, three arms, two versions; 240 calls	Reused development population with raw local outputs
Learning	Ten seeds, 100,000 episodes, 131,346 decisions	Exact evaluation in 160 synthetic states
Structured reporting	Twelve paired cases, twenty-four decisions per implementation	Supplied semantic policy; conventional parity
Prompt pilot	Six development and eighteen evaluation families; 144 calls	Frozen local exploratory evaluation
Historical public/adversarial full-treatment study	No pilot or confirmatory execution	Original qualification obligations remain open

Package	Retained scale or result	Evidence status
Local full operational initial stage	1,805 valid episodes, 24 included controls, 19 checkpoints	Six conditional contrasts; 619 original primary assignments deferred

The historical assessment covered material families through retained files, commits, worktrees and handoffs. It did not recover every Markdown line, private conversation or retired prototype. Original Claude work was reached through retained artifacts rather than a complete private conversation export. The first unretained interpretation probe, exact initial untracked extraction epoch, full old event/token streams and every provisioning intermediate were not recovered. Claims requiring those records remain unsupported.

The twelve adapted incident shapes, their 27 records and native integration observations are historical illustrations. Ten incident shapes were outside the calibrated preflight because their native evidence was insufficient. This does not establish incident prevention. This successor adds only the benign synthetic local operational model study described in Chapter 13. It performed no exploit reproduction, human-subject study or registered training rerun.

Appendix B. Corrections that change what a reader may infer

Earlier or tempting interpretation	Corrected interpretation	Consequence
398 real and ten seeded records	395 real-labelled and thirteen seeded in retained truth data	Provenance arithmetic changes; origin remains unverified
Held-out replay; eighteen seeded defects	Retrospective corpus; five real-labelled and thirteen seeded defects	No unseen evaluation claim
One extra detection demonstrates RR advantage	Sole difference seeded; conventional cross-record join omitted	No advantage over equally capable ordinary verification
Fewer false holds means more successful work	Calibration and later abstention use different counters	Downstream utility must be measured separately
An older filename identifies invalid authority	Unchanged clauses and historical use may remain legitimate	Proposition and intended use must be inspected
Batch cohort caught twenty defects	Sixteen defect and sixteen nondefect answers	Preserve corrected denominator and serialization issue
Sealed means authenticated	Self-verifiable identity/integrity, not a digital signature	Source authenticity remains a host obligation

Earlier or tempting interpretation	Corrected interpretation	Consequence
Atomic agreement means full conformance	Parser, wrapper and other interface behavior lie outside that check	Preserve the full-surface discrepancies
Refined 12/12 establishes efficacy	Same cases, joint changes, conventional parity	Development improvement only
Cheap learned behavior is efficient	Lower cost accompanies greater abstention	Compare quality and cost together
Nonsignificant pilot means no effect	Wide interval includes benefit and harm	No demonstrated benefit; no equivalence claim
A structural-release improvement proves useful task benefit	Seed holds reduce completion; clean intervals cross zero; semantic controls fail	Report structural, semantic and completion outcomes together
The initial stage completed the original full sample	Explicit post-start amendment selected 1,805; 619 remain deferred	Preserve timing, denominator and conditional two-report law
A new research companion would be novel	Close public frameworks and updating methods exist	A distinct contribution needs its own evidence

Original bytes remain unchanged. This integrated account uses the corrected interpretations directly, so a new reader need not first absorb an erroneous headline and later discover its erratum.

Appendix C. Reproduction and artifact custody

The companion directory contains this Markdown source, its self-contained HTML and PDF, a deterministic builder, a source map with byte identities, and a verification record. The earlier technical paper, illustrated reader, seven-page summary and original study packages remain separate, unchanged historical artifacts in the same repository snapshot. The integrated paper is the main reading route for this revision; the earlier papers remain sources, not competing current summaries.

From the repository root, the retained portable evidence check is:

```
python -B study/2026-1/release/verify.py
```

That check verifies retained release identities and predecessor preservation, recomputes selected arithmetic, and runs bounded component tests. It does not call a model, rerun the original full study or establish external replication. The operational unit tests include two twenty-episode deterministic learning replays, forty small algorithm-check episodes in total; these are separate from the registered 100,000-episode experiment and do not replace its stored policies. The command does not rerun every historical conformance or formal checker.

The new initial-stage analysis can be reproduced without model calls from the locally retained evidence:

```
python -B study/2026-1/full-treatment/stage_one/analyze.py study/2026-1/full-treatment/
collection-local/evidence study/2026-1/full-treatment/analysis/reproduction
```

The analyzer verifies the admitted source identities and all selected observations before producing contrasts. The checkpoint archives and union evidence are local custody artifacts; a paper or source map alone is insufficient to reproduce their observations. The unamended collector must not be used to infer that the deferred assignments were completed.

The integrated rendering can be rebuilt with:

```
python -B study/2026-1/full-treatment/paper/build.py
python -B study/2026-1/full-treatment/paper/check.py
```

Rendering requires ReportLab and pypdf in addition to Python; the exact tested versions are recorded with the output. Visual inspection is a separate check. Text extraction and a successful build cannot establish that every table or figure is legible. Source-map hashes bind the material actually read by the document's links; they do not authenticate external claims or make unavailable local history portable.

The earlier integrated account used local commit 08f7fa7d29a35a125ba047983a5ae9df456d0c96. This successor preserves that 41-page account and uses worktree baseline c22018a2d66e673b765313a63517a389bb34a5d9. The unchanged prior distribution contains 2,847 files and preserves all 2,800 predecessor blobs from its earlier baseline. Its archive SHA-256 is 3e98dca323d86ee1da04393269ef88c7b1fa2862b77f53e61536537f2813310d. That identity applies to the prior archive, not to this integrated successor. The new paper's identities and validation are recorded separately.

The public software repository and concept DOI identify the existing RR project. They do not imply that this local integrated working paper or the new study packages have been published or assigned a new DOI. Receiver Reliance repository; software concept DOI.

Appendix D. Source guide

The chapter-end links are the shortest routes from the narrative to retained evidence. The machine-readable SOURCE_MAP.json records their exact local bytes and the sections that cite them. Linked historical documents may themselves name unavailable private inputs; this source map does not certify those transitive dependencies. The following routes identify the main authoritative packages without requiring a reader to understand the original worktree layout.

Topic	Primary entry point within study/2026-1
Original mechanism, study design, demonstration and validation	paper/RR_PROTOCOL_PREPRINT_DRAFT_v1.md
Earlier illustrated narrative and nine-figure delivery provenance	paper/RR_PROTOCOL_PREPRINT_DRAFT_v2_READER_EDITION.md; astra-contribution/reader/README.md

Topic	Primary entry point within study/2026-1
Historical families, recovered sources and missing records	astra-benefit/HISTORY_DECISIONS.md; astra-benefit/SOURCE_INDEX.json
Replay and precursor corrections	astra-benefit/CORRECTIONS.md; audit-results.json; release/originals/
Exact binding, finite audit and design candidate	astra-contribution/README.md; VALIDATION.md
Direct applicability, coverage and operational executions	astra-operational/APPLICABILITY.md; COVERAGE.md; RESULTS.md
Learning freeze, policies and audit	astra-operational/learning/FINDINGS.md
Structured reporting contract	astra-use/README.md; CONTRACT.md
Full operational treatment, amendment and results	full-treatment/TREATMENT_SPEC.md; stage_one/AMENDMENT.md; analysis/initial-stage-20260906/report.json
Prompt methods, inputs, responses, grading and analysis	evidence-use-empirical/RESULTS.md
Field comparison and subsequent novelty corrections	evidence-maintenance/RESEARCH_DIRECTION_ASSESSMENT.md; NOVELTY_UPDATE.md
Portable completed-evidence verification	release/README.md; verify.py; VERIFICATION.md

The original technical paper retains its broader bibliography. This account cites the primary field sources material to its synthesis rather than presenting that earlier bibliography as a fresh exhaustive literature search. Newer preprints and unavailable full texts retain their reported inspection limits. No source is credited with independently validating RR's results.