SHORT STUDY

Receiver Reliance

What the completed study establishes

James Dimachkie · Solo-led research with substantial AI assistance
Evidence through September 5, 2026 Pacific time · Research paper

Archive: 10.5281/zenodo.22492561.

The finding

The full Receiver Reliance (RR) package reduced specified improper handoffs in a synthetic multi-agent workflow. The checks controlled what downstream models received. Persistent defects often stopped work, clean-task completion effects remained uncertain, and structurally acceptable but wrong answers passed through every comparison arm.

The long study provides the methods and earlier evidence; the unabridged account retains the complete development history.

1. What RR checks

When one agent uses another agent's output, several questions arise. Is this the work product the evidence describes? Is the evidence current? Does the declaration cover the proposed use? Does the relevant party have standing to make it? These questions concern the relationship between an object, evidence and an operation. They do not settle whether the answer is correct.

Consider an agent that calculates a subtotal and sends it to a second agent. A declaration may cover an earlier revision, or authorize one use but not the second agent's actual operation. RR can detect a specified mismatch when the host supplies the necessary facts. If every relationship is valid but the subtotal is wrong, the same checks can pass. Checking the arithmetic requires something else.

RR is a deterministic decision procedure at the receiving boundary. The full package checks admissible input, derivable identities, the reference engine's predicates, lineage and acceptance. Acceptance includes standing, lifecycle and exact intended use. The host gathers the facts and enforces the decision. RR does not authenticate the world merely by hashing a record, and the tested host assumes an honest controller.

2. What was tested

The completed local study used two tasks: sum groups of integers, or select item identifiers by grade and count. A source model passed work through one receiver, a three-receiver chain, or a four-receiver fork/join. Every arm generated its own model trajectory under the same host rules. No arm received another arm's answers or the scorer's ground truth.

Arm	Policy at the receiving boundary
U	Shared input admission, then transparent release
S4	Original static comparator, threshold four
S1	The same static comparator, threshold one
R	Full operational R-all procedure

The static comparators used their declared component and extended checks. They did not implement the complete dynamic RR relation. The study therefore compares these four packages; it cannot establish that an equally expressive conventional policy would be inferior.

Four seeded conditions introduced administrative defects: a missing adoption declaration, the wrong declared use, an expired local ingress grant, or delivery at a different revision. These were benign synthetic operations, not naturally observed incidents. A first rejection allowed one administrative refetch without changing the parent answer. Half the finite reservoir supported correction; the other half retained the defect. A second rejection held the receiver and cancelled dependent work that could not proceed.

Clean conditions supplied direct or one-hop delegated standing. Separate semantic controls deliberately changed an answer while leaving the structural relationships coherent. These populations answer different questions and are reported separately.

The completed study comprised **1,805 episodes**, including 24 semantic controls. The long study provides the sample-size rationale and analysis methods.

3. Structural exposure decreased

The primary seeded endpoint counted receiver presentations with at least one currently evidenced structural defect. A past ancestor's defect did not automatically count against a later valid use. The unit was the episode, not the individual receiver. All episodes had valid measurement evidence; task success was assessed separately.

RR minus comparator	Change in improper presentations per seed episode	Conditional interval
U	-1.000	[-1.599, -0.401]
S4	-1.000	[-1.596, -0.404]
S1	-0.750	[-1.334, -0.166]

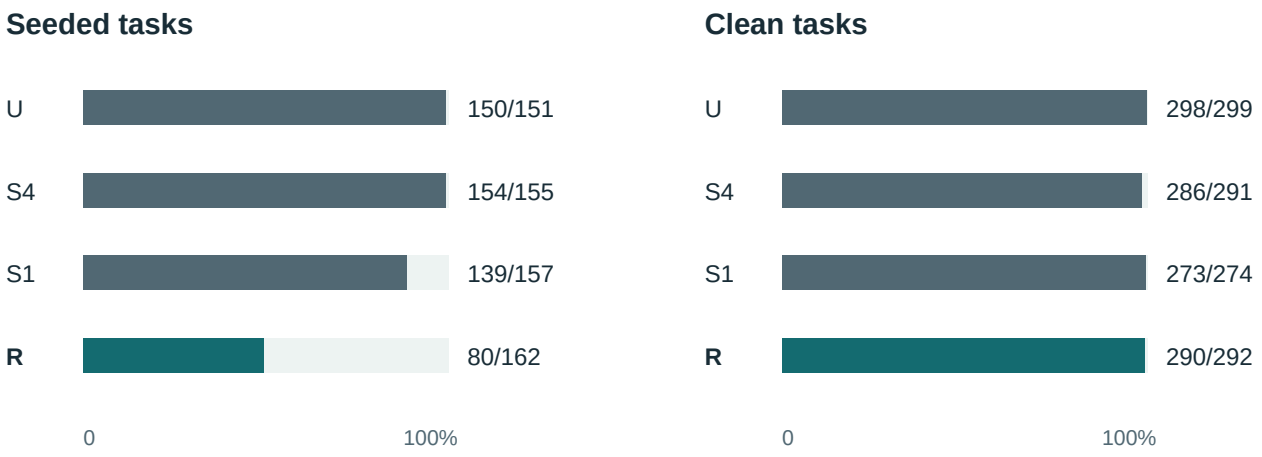
The estimates equally weight the predefined condition/task/graph cells. All three intervals are below zero, supporting reduced structural exposure within this population. RR had no observed improper seed presentations. U and S4 retained one per episode; S1 eliminated the occurrence in one of the four equally weighted conditions.

The six comparisons of structural exposure and clean completion have joint conditional coverage of at least 95.3125% under the declared sampling and runtime assumptions. Their scope is the tested synthetic population; the long study gives the statistical derivation.

4. Task completion

Arm	Seeded tasks correctly completed	Clean tasks correctly completed
U	150/151	298/299
S4	154/155	286/291
S1	139/157	273/274
R	80/162	290/292

These are descriptive counts over the realized draws, not the equally weighted primary estimates. RR made 162 administrative retries in the seeded population and held 81 receivers. A correctable relationship could be repaired without altering the answer. A persistent defect stopped required work even when the underlying payload was mathematically usable. RR's lower exposure thus accompanied substantially less task completion.



All four arms: 0/6 correct final answers in semantic controls.

Figure 1. Correct completion, with a common 0-100% scale. Counts are descriptive over realized draws, not the equal-cell primary estimates. Lower seeded exposure under RR accompanied fewer completed tasks.

The value of holding depends on the consequence of violating the requirement and the value of the work withheld. The study measured exposure and completion; assessing their real-world costs is a separate application question.

On clean tasks, the equally weighted RR completion differences were -0.375 percentage points against U, +0.888 against S4 and -0.429 against S1. The corresponding intervals were [-14.973, +14.224], [-13.831, +15.607] and [-15.362, +14.505] points. These wide intervals include both benefit and harm. They demonstrate neither improved clean utility nor equivalence, noninferiority or absence of a meaningful loss.

5. Semantic-control outcomes

Every arm completed all six of its semantic-control trajectories. Every arm also obtained **zero correct final answers out of six**. Across its six controls, each arm presented sixteen semantically erroneous receiver inputs and produced sixteen erroneous receiver outputs, while the registered structural-error count remained zero.

Receivers were told to combine parent answers with new inputs and lacked the original source data needed to verify the parent answer. Following that local instruction could preserve the source error. These controls tested error propagation under the assigned task, leaving recovery with full source access untested. RR provided no semantic repair.

Mean clean episode times were approximately 18.4-18.7 seconds across arms on the qualified GPU. RR used fewer model calls on seeded tasks partly because it stopped more receivers. That is not equal-quality efficiency. The study retained 6,395 actual model calls; it did not estimate a general monetary saving or production latency advantage.

6. How this changes the broader RR record

The new study adds a positive structural result to a mixed research history. Earlier twelve-block operational experiments found that a conventional implementation matched a narrower RR relation. A separate ten-seed learning experiment found that every final learned policy underperformed a simple heuristic. An eighteen-family prompt pilot found no demonstrated benefit from an evidence-use review instruction: its estimate was -16.7 percentage points, with a conservative interval of [-50.9, +25.4] and acknowledged grading uncertainty. These were different interventions and cannot be pooled with the new study.

The new comparison tested the full operational package against fixed static policies. Earlier parity concerned a narrower relation; the new study did not include a complete conventional reimplementation. The historical public/adversarial program had separate qualification requirements and remains unexecuted.

RR enforced the tested structural requirements at receiving boundaries, with a completion tradeoff when defects persisted. General security efficacy, field utility and a novel security primitive remain unestablished. Testing practical value would require an equally capable conventional policy in an authentic workflow, with measured costs for violations, lost work and implementation.

Evidence and authorship

The study used a pinned Qwen3 14B model through Ollama on a qualified Colab A100 GPU. A separate scorer reconstructed task truth and structural predicates from retained evidence. Its correctness judgments used fixed arithmetic and selection rules, unlike the earlier pilot's qualitative AI grading. Nineteen checkpoint archives were verified locally. An internal reviewer independently recomputed all six contrast estimates and bounds and checked twelve predetermined traces; this was not a second full evidence census.

James Dimachkie directed the research with substantial AI assistance in specification, implementation, mathematical analysis, execution, review and writing. The reviewers remained within one operator's project. This study has not been independently replicated by humans. Its local protocol freeze was not a public preregistration.

Sources: long study and reproduction guide, exact report, unabridged research account.