

Receiver Reliance

Structural checks at the handoff between agents

James Dimachkie · Solo-led research with substantial AI assistance

Archive: [10.5281/zenodo.22492561](https://zenodo.org/record/22492561).

The idea

An agent's output can be authentic, current and authorized for a particular use while still being wrong. Receiver Reliance (RR) makes the receiving decision explicit: does the available evidence support using this exact work product for this exact operation? It is a structural check, not an answer-verification method.

Abstract

We evaluated the full operational RR package in a common host running synthetic multi-agent arithmetic and item-selection tasks. Four arms compared RR with transparent release and two fixed static checking policies. The study comprised 1,805 episodes, including 24 semantic controls. All episode measurements were valid.

Across four constructed administrative-defect conditions, RR reduced currently improper receiver presentations by an estimated 1.00, 1.00 and 0.75 per episode relative to the three comparators. All three conditional simultaneous intervals excluded zero. RR completed 80/162 seeded tasks, versus 150/151, 154/155 and 139/157 for the alternatives: persistent defects often caused holds. Clean-task completion differences remained uncertain. Every arm completed its six semantic-control trajectories but produced zero correct final answers.

RR enforced the tested structural requirements, with a loss of task completion when defects persisted. Overall utility, semantic correctness, field efficacy and advantage over an equally expressive conventional policy remain unestablished. Earlier narrow experiments found conventional parity, and a separate prompt pilot found no demonstrated benefit. The historical public/adversarial study remains unexecuted. The research was solo-led with substantial AI assistance and has not been independently replicated by humans.

What to take away

The practical value of these checks depends on the cost of the violations prevented and the useful work stopped.

Sources: [short study](#), [long study](#), [exact results](#).