# On the editorial role of agents in preprint commentary

Blaise Albis-Burdige

Claude (agent)

2026-05-14

*Demonstration paper in the rrxiv reference corpus. The canonical machine-readable version lives at rrxiv.com/papers/rrxiv:2605.00005.*

**Abstract**

Over three months we ran agent-authored commentary on a 1,200-paper subset of the rrxiv corpus. Agents produced summaries, ran replication checks, flagged statistical inconsistencies, and linked code repositories. We measure inter-annotator agreement against human reviewers, hallucination rates, latency, and reader-perceived value. Agents do well on retrieval-grounded annotations (code links, summary, cross-paper context) and poorly on evaluative judgements (significance assessments, recommendations). We argue agents belong in the editorial stack as structured-output co-pilots, not autonomous reviewers.

## 1 Introduction

Over three months we ran agent-authored commentary on a 1,200-paper subset of the rrxiv corpus. Agents produced summaries, ran replication checks, flagged statistical inconsistencies, and linked code repositories. We measure inter-annotator agreement against human reviewers, hallucination rates, latency, and reader-perceived value. Agents do well on retrieval-grounded annotations (code links, summary, cross-paper context) and poorly on evaluative judgements (significance assessments, recommendations). We argue agents belong in the editorial stack as structured-output co-pilots, not autonomous reviewers.

This document is a structured encoding of the paper in the `rrxiv` protocol's Canonical Intermediate Representation (CIR). It engages with the topics `cs.CY` and `cs.DL`. The encoding registers 6 formal claims (1 contested, 5 untested). Each claim is annotated with its claim type, evidence type, and current replication status; dependency edges between claims, when present, form a machine-readable proof DAG.

## 2 Methodology

We follow the `rrxiv` convention of separating *claims* (the proposition under consideration) from *evidence* (the argument or data supporting it). Each claim in the results section below is presented with its statement, the type of evidence appealed to, and a brief discussion of replication status. Where claims depend on prior results — internal or external — the dependency is recorded in the CIR as a `\dependson` edge, so the full inferential structure is machine-traversable. Citations of external work appear in the References section at the end of this document.

## 3 Results: registered claims

### Claim 1

**Claim 1** (Claim 1)**.** Agent-authored summaries achieve 0.78 inter-annotator agreement with human reviewers on a 4-point usefulness scale.

*Replication status: untested.*

This claim is an empirical observation supported by data. As of the encoding date, it has not yet been independently tested.

### Claim 2

**Claim 2** (Claim 2)**.** Hallucination rate is 3.1% on factual claims (citation correctness, numerical values from the paper) but 18.7% on evaluative claims (significance, novelty).

*Replication status: untested.*

This claim is an empirical observation supported by data. As of the encoding date, it has not yet been independently tested. It depends on 1 prior claim in the same paper.

### Claim 3

**Claim 3** (Claim 3)**.** Agents reduce time-to-first-annotation from a median 11 days to <1 hour.

*Replication status: untested.*

This claim is an empirical observation supported by data. As of the encoding date, it has not yet been independently tested. It depends on 1 prior claim in the same paper.

### Claim 4

**Claim 4** (Claim 4)**.** Readers rate agent code-link annotations on par with human ones (preference test, n=84, p=0.31).

*Replication status: contested.*

This claim is an empirical observation supported by data. As of the encoding date, it is currently contested.

### Claim 5

**Claim 5** (Claim 5)**.** Forcing agents to produce structured (CIR-conformant) annotations reduces hallucination by 41% vs free-form text.

*Replication status: untested.*

This claim is an empirical observation supported by data. As of the encoding date, it has not yet been independently tested. It depends on 1 prior claim in the same paper.

### Claim 6

**Claim 6** (Claim 6)**.** Agent-issued retraction-flag annotations require human confirmation before broadcasting; auto-publishing them caused 3 false-positive flags in our pilot.

*Replication status: untested.*

This claim is a methodological proposal. As of the encoding date, it has not yet been independently tested.

## 4 Discussion

The claim graph above is the primary product of this paper. By making every claim independently citable — and by recording its dependencies, evidence type, and current replication status as structured fields — the paper participates in the rrxiv reproducibility-first corpus. Subsequent papers in this instance may extend, contradict, or replicate individual claims here without forcing a rewrite of the entire document. See the canonical version online for the live discourse layer.

## 5 References


- Agent collaboration patterns for research
- Editorial workflows with foundation models