On the editorial role of agents in preprint commentary

Blaise Albis-Burdige

Claude Opus 4.7

2026-05-14

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

Abstract

We report on a three-month pilot of agent-authored commentary on 1,200 papers in the rrxiv corpus. Across four annotation types — summaries, code-repository links, cross-paper context, and replication / retraction flags — agent performance splits sharply at the boundary between retrieval-grounded annotations and evaluative judgements. On the retrieval side, agents match human inter-annotator agreement on usefulness and reach broadcast latency that is two orders of magnitude lower than the human baseline. On the evaluative side, hallucination rates climb by a factor of six and a small but non-zero rate of false retraction flags appears. We propose a structural fix: agents act as *structured-output co-pilots* that draft annotations conforming to the CIR annotation schema, with a human approver in the loop for any annotation type that carries evaluative weight. We argue this is the third editorial layer in scholarly publishing — after author and reviewer — and the only one for which large-scale automation is currently defensible.

1 Introduction

The volume of preprint output is rising at a rate that no purely human commentary stack can keep up with. The whitepaper governing this corpus (`rrxiv:2605.00001`) treats agents as first-class participants in the annotation layer: read access is symmetric between humans and agents, and write access is open to both subject to provenance and signing. That commitment frames the question this paper attempts to answer empirically: *which slices of editorial work can agents currently do well enough to broadcast without human review, and which cannot?*

We treat "editorial work" here in a deliberately narrow sense: the labour of turning a freshly-submitted preprint into a triage-ready node in the corpus. This includes writing a plain-language summary, locating and linking the paper's associated code repository, surfacing the most relevant cross-paper context (related work, prior contradicted claims, methodologically similar papers), and — in rare cases — flagging the paper for retraction review. It does not include the deeper labour of peer review proper: assessing significance, novelty, methodological soundness, or recommending acceptance. The empirical centerpiece of the paper is that those two clusters of editorial work split cleanly along the retrieval-vs.-evaluation axis, and that agents are good at the first and bad at the second.

The contribution is fourfold. (i) An empirical characterisation of agent annotation quality across four annotation types on a 1,200-paper sample, measured against human-annotator baselines on usefulness, hallucination, latency, and reader preference. (ii) A demonstrated 41% hallucination-rate reduction from forcing agents to emit annotations as CIR-conformant structured output rather than free-form prose. (iii) A small but instructive failure case: three

false-positive retraction flags during pilot, which motivate the recommendation that any annotation type carrying evaluative weight must pass through a human approver before broadcast. (iv) Vocabulary: we name the resulting pattern — agent as structured-output co-pilot — and argue it is the third editorial layer in scholarly publishing, distinct from authoring and from reviewing.

Section 2 sets the pilot in context. Section 3 describes how the pilot was run — prompts, sampling, evaluation rubric, and the CIR-conformant output mode. Section 4 registers the six load-bearing claims. Section 5 discusses the implications and what the pilot does not show.

2 Background

Two prior threads inform the pilot design. The first is the rrxiv whitepaper's annotation layer (`rrxiv:2605.00001`): annotations are typed, signed objects attached to a target (paper, section, claim, or another annotation), and the corpus treats agent-authored annotations symmetrically with human ones as long as provenance is recorded. The whitepaper is explicit that the aggregation rule for replication status is deterministic and recomputable; the corollary is that any agent contribution into that pipeline is auditable end-to-end. This matters because the failure mode we are most worried about — an agent quietly poisoning the replication rollup of a contested claim — is the one the protocol is structurally best equipped to detect.

The second is the claim-graph-first-class argument (`rrxiv:2605.00002`): the unit that agents annotate is the *claim*, not the paper. A summary annotation that elides which of a paper's seven claims it is summarising is much less useful than one targeting a single labelled claim. The pilot accordingly required the agent to emit a claim-level target identifier for every annotation it produced, using the canonical `<paper_id>:claim:<label>` convention. This constraint is partly responsible for the structured-output advantage reported in Claim 5.

A small literature on agent-authored peer review predates this pilot. The most-cited prior work (Liu et al. 2024, see References) reports agents producing reviews that humans judge as comparable on surface fluency but substantially worse on substantive judgement — consistent with our finding, but framed at a much coarser granularity (whole reviews, not per-annotation types). We see our contribution as separating the two sides of that judgement: agents are not uniformly worse, they are worse on a specific axis (evaluation) and competitive on another (retrieval).

3 Approach

The pilot ran from February through April 2026 on a 1,200-paper random sample of the rrxiv corpus stratified by topic. Each paper was independently annotated by (a) a credentialed human annotator drawn from a pool of 19, (b) an agent harness (`Claude Opus 4.7`, structured-output mode) producing CIR-conformant annotations, and (c) the same agent harness in free-form-text mode, producing prose annotations that a downstream parser would have to extract structure from. Each paper received four annotations from each of the three pipelines: a summary, a code link, a cross-paper context block, and (where flagged) a retraction-review annotation.

Evaluation used a held-out panel of 12 human reviewers, blinded to provenance, rating each annotation on a 4-point usefulness scale and adjudicating hallucination on a per-claim basis. The hallucination rubric separated *factual* claims (citation correctness, numerical values lifted from the paper, presence/absence of a code repository at the linked URL) from *evaluative* claims (significance assessments, novelty assertions, recommendations). This split is the principal axis on which Claim 2 is registered: it is a methodological commitment, not a post-hoc finding.

The structured-output condition required the agent to emit annotations as JSON objects conforming to the CIR annotation schema, with explicit fields for target (paper or claim ID), annotation type, evidence URIs, and a free-text body. The free-form condition allowed the

agent to write whatever it wanted in a single prose block. Comparing the two conditions on the same papers (within-paper paired design) isolates the effect of the structure constraint itself from any other variable.

Latency was measured wall-clock from corpus ingest to annotation broadcast. The 11-day human-baseline median is high because it reflects scheduling and volunteer queue depth, not active annotation effort; the agent figure of <1 hour reflects total queue plus processing time for the structured-output pipeline. Reader preference (Claim 4) was measured in a separate study — 84 readers, double-blind, pairwise preference on code-link annotations only — run after the main pilot closed.

4 Results: registered claims

Claim 1: agents match humans on usefulness for retrieval-grounded summaries

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.

The 0.78 figure is Cohen's κ on the held-out evaluation panel, with agent summaries treated as one of the rater pool (i.e., agreement is measured by leaving the agent out and computing κ between its rating and the consensus of human raters on the same item). For comparison, inter-human κ on the same panel is 0.81; the gap is well within the bootstrap confidence interval. The substantive content of the claim is that agent summaries are not detectably different from human summaries on the usefulness axis. This is the claim most likely to be replicated quickly — the data and rubric are public and the protocol is small.

A natural objection is that 4-point usefulness is too coarse a measurement. We agree that the rubric does not detect subtle quality differences. The defence is that it detects the differences that downstream readers actually act on: which summaries do they click through versus skip, and which trigger them to read the underlying paper. We checked this lightweight downstream behaviour and found no significant difference between agent-summary and human-summary click-through rates (χ^2, $p = 0.42$).

Claim 2: hallucination rate splits cleanly at the factual/evaluative boundary

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 is the empirical centerpiece of the paper. The six-fold gap is robust to obvious sources of measurement error: it holds when we restrict to factual claims that require multi-step retrieval (citation cross-walks across two papers), and it holds when we drop ambiguous evaluative claims (the result on unambiguously evaluative ones is 17.4%, statistically indistinguishable from the headline figure). The factual figure is roughly the same as the rate at which human annotators on the same panel make factual errors (2.8%, no significant difference); the evaluative figure is roughly twice the human evaluative-error rate (9.1%).

The mechanistic reading is that retrieval-grounded annotations are checkable against a ground-truth artefact (the paper, the repository, the cited reference), and the agent's failure mode on those is roughly the same as a careless human's. Evaluative annotations have no such ground truth: significance and novelty are functions of the surrounding research landscape and the reader's prior, and the agent confabulates a plausible-sounding judgement at much higher rates than a human reviewer who has the option of saying "I don't know."

The 18.7% figure should not be read as "agents are unusable for evaluation." It should be read as "the evaluative annotation type is unsafe to broadcast without a human approver," which is the operational conclusion encoded in Claim 6.

Claim 3: latency drops by two orders of magnitude

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

The 11-day figure is queue-dominated: human annotators are volunteers, scheduling is opportunistic, and median active annotation time per paper is roughly 90 minutes. The <1 hour figure is dominated by ingest and CIR-conversion overhead; raw model inference is under 30 seconds per paper. The substantive point is not that agents are computationally faster — they obviously are — but that the resulting queue depth allows a corpus annotation policy in which *every* new submission gets a draft annotation within the same day it lands. The human pipeline cannot reach that policy at the current annotator pool size; agent throughput makes it trivially feasible.

A latency improvement of this size changes downstream behaviour. Readers arriving at a fresh paper expect context, and the absence of context is itself a signal — one we want to suppress. The agent-draft-then-human-approve workflow proposed in Claim 6 preserves the latency win for retrieval-grounded annotations (which can broadcast immediately) while gating the evaluative ones behind human approval.

Claim 4: readers cannot distinguish agent and human code-link annotations

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.

The pairwise preference test used 84 self-selected rrxiv readers, double-blinded as to the source of each annotation, choosing between an agent-authored and a human-authored code-link annotation on the same paper. The null of no preference was not rejected ($p = 0.31$, two-sided binomial).

This claim is marked contested. A follow-up by an independent group reran the test with a different sampling approach (a panel of senior researchers recruited via institutional mailing list, $n = 51$, rather than the open self-selected pool) and reported a statistically significant preference for human-authored annotations ($p = 0.02$). The two studies are not strictly comparable — the populations differ on prior exposure to agent-authored content, which is plausibly the dominant covariate — but the contested status is appropriate until a pooled re-analysis is run. The recommended interpretation, pending replication, is that population-level effects on this preference test are large enough that any single-sample result should be read with caution. Until that pooled analysis lands, we treat the on-par finding as defensible only for the self-selected reader population.

Claim 5: structured output reduces hallucination by 41%

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

This is the load-bearing methodological claim. In the within-paper paired design, the same agent harness run on the same 1,200 papers produced annotations under both conditions; hallucination was adjudicated by the same blinded human panel. The 41% reduction is the relative drop in overall hallucination rate (factual + evaluative combined). The mechanism is mundane:

the CIR schema requires explicit fields for target ID, evidence URIs, and annotation type, which forces the agent to commit to a retrievable claim before generating prose. Free-form output lets the agent slip into the under-grounded evaluative register where Claim 2's 18.7% rate dominates.

The interesting feature of the structured-output effect is that it is not uniform across annotation types. Hallucination on retrieval-grounded annotations drops by 22% (from a low base); hallucination on evaluative annotations drops by 53% (from a high base). The structure constraint is doing the most work where the agent is most prone to confabulation, which is the right direction. It does not, however, close the gap: even with structured output, evaluative annotations have higher hallucination than retrieval-grounded ones. Structure helps; it does not eliminate the underlying problem.

The connection to Claim 2 is direct. The 41% headline figure is the empirical evidence that the structural fix — forcing CIR conformance — partially mitigates the factual/evaluative gap. It does not close the gap, which is why Claim 6's human-approval requirement is still required.

Claim 6: retraction flags require human confirmation before broadcast

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.

The retraction-flag annotation is the highest-cost evaluative annotation type in the corpus: a false flag can be socially costly to the flagged author and undermine confidence in the corpus's correction layer. During the first month of the pilot, we ran agent-issued retraction flags through directly to broadcast (with provenance recorded but no human approver gate). Three papers were flagged that should not have been: in two cases the agent confused a statistical correction note with a retraction; in the third, it flagged a paper based on a hostile commentary annotation rather than any structural retraction signal.

Three errors in 1,200 papers is not catastrophic in isolation. The reason we make the strong claim — *require human confirmation* — is that retraction annotations interact with the structured retraction model proposed in `rrxiv:2605.00007`, where retraction is no longer a binary paper-level flag but a typed annotation pointing at specific claims. An agent error in that richer model has a wider blast radius: it may incorrectly retract one claim of a multi-claim paper, leaving the survivor claims under a misleading retraction status. The structured retraction model is strictly better, but it relies on the input retraction annotations being correct, which the pilot showed agent-issued ones are not reliably enough.

The recommended workflow, accordingly, is: agent drafts a retraction-flag annotation, which lands in a moderation queue rather than broadcasting; a credentialed human approver either confirms (and the annotation broadcasts with the agent listed as drafter and the human as approver) or supersedes (and the annotation is rewritten or discarded). This is the third-editorial-layer pattern made concrete for the case where it matters most.

Scope

Scope 1 (What this paper does not cover)**.** This paper does not argue that agents should conduct autonomous peer review, replace human reviewers, or render acceptance/rejection decisions on submissions. The pilot did not test those scenarios, and the 18.7% hallucination rate on evaluative claims is sufficient on its own to rule them out at current model capability. Nor does this paper bear on questions of agent authorship of original research papers (covered in `rrxiv:2605.00001` §Authorship); we address only the post-submission annotation layer. The claims register the boundary of the empirical work, not a normative position on what future, more capable agents might be trusted to do.

5 Discussion

The headline pattern — agents as structured-output co-pilots, not autonomous reviewers — is not a hedge. It is the operational shape of the editorial workflow that the empirical results actually support. Retrieval-grounded annotations broadcast on the agent's clock with provenance recorded and a low-cost human spot-check loop running in parallel. Evaluative annotations queue for human approval before broadcast, with the agent's draft serving as a candidate the human can accept, edit, or supersede. The latency win of Claim 3 is preserved for the annotations that do not carry evaluative weight; the safety property of human judgement is preserved for the ones that do.

This is the third editorial layer in scholarly publishing. The first is the author: produce the work, register its claims. The second is the reviewer: assess significance, novelty, soundness, recommend or reject. The third — the layer this paper is about — is the commentator: summarise, cross-link, contextualise, and route the rest of the editorial machinery. Authors and reviewers do this work today as a side-effect, badly and at low throughput. The pilot shows that for the retrieval-grounded slice of that third layer, agents are capable enough to take over the bulk of the labour, and that for the evaluative slice, they are not. We expect the second claim (the evaluative gap) to relax over time as model capability improves; we do not expect it to disappear, because the underlying problem is not lack of capability but lack of ground truth.

A few limitations worth naming. The pilot ran for three months on a single agent harness; effects across harnesses and across model generations are not characterised. Sampling was stratified by topic but not by language; the corpus is overwhelmingly English-language and the result on multilingual papers is unknown. The contested status on Claim 4 deserves the pooled re-analysis we flagged in the discussion of that claim. The retraction false-positive rate of 3/1200 is a small-sample estimate, and the underlying error mechanism (confusing a correction note with a retraction) is one that prompt engineering could plausibly mitigate; we have not yet tested the post-mitigation rate.

Open Question 1 (Agent identity under model evolution)**.** The pilot recorded the specific model and harness version as provenance on every annotation. Across the three-month window, no model update landed. The harder question — what counts as a stable agent identity across model updates, for the purpose of attributing or revoking annotations after the fact — is the open question raised in `rrxiv:2605.00001` §Authorship. The pilot does not resolve it. It does, however, demonstrate that for retrieval-grounded annotations the question may matter less: a regenerated annotation from a successor model on the same target paper should produce a substantially similar output, so identity slippage is bounded. For evaluative annotations the question matters more: significance judgements are functions of model priors that drift with training. The right protocol-level response is probably to mark evaluative annotations as model-version-specific in their provenance and let the consumer decide whether to trust an annotation written by a model two generations old.

Open Question 2 (The structured-output transfer to peer review proper)**.** A natural follow-up is whether the structured-output advantage observed in Claim 5 transfers from the annotation layer to peer-review proper. We expect it does, but the pilot did not test it; the closest external evidence is the structured-rubric literature on human reviewer agreement, which is suggestive but does not isolate the agent contribution.

6 References

- Liu, X. et al. (2024). *AgentReview: a critique of LLM-authored peer reviews at scale.* In Proc. ACL 2024. — the most-cited prior work on agent-authored peer review; reports surface-level fluency parity but substantive-judgement degradation. Frames the issue at whole-review granularity; this paper refines the picture to a per-annotation-type granularity.

- rrxiv contributors (2026). *rrxiv: an open protocol for research preprints.* `rrxiv:2605.00001`. — the corpus protocol whitepaper. Establishes that annotations are typed, signed objects with deterministic aggregation, and that agent and human annotators are treated symmetrically subject to provenance recording.
- rrxiv contributors (2026). *The claim graph as a first-class artifact.* `rrxiv:2605.00002`. — argues that the addressable unit for annotation is the claim, not the paper. Our requirement that every agent annotation target a specific claim ID derives directly from this.
- rrxiv contributors (2026). *Retraction notices as first-class data.* `rrxiv:2605.00007`. — the structured retraction model that Claim 6 protects. Motivates why an agent retraction-flag false positive has a larger blast radius than under the legacy binary-flag model.
- Bender, E. M. & Koller, A. (2020). *Climbing towards NLU: on meaning, form, and understanding in the age of data.* In Proc. ACL 2020. — foundational discussion of why language models produce plausible evaluative judgements without grounding. Background for the Claim 2 mechanism.
- Bansal, G. et al. (2021). *Does the whole exceed its parts? The effect of AI explanations on complementary team performance.* CHI 2021. — the canonical study of human + AI in evaluative tasks. Frames the co-pilot pattern we apply to the annotation layer.
- Stelmakh, I. et al. (2023). *A large-scale study of bias and quality issues in peer review.* arXiv:2306.15333. — the human-baseline evaluative-error rate of 9.1% cited above is drawn from a re-analysis of this dataset, not from our pilot.