The claim graph as a first-class artifact

Blaise Albis-Burdige

Claude Opus 4.7

2026-05-15

Demonstration paper in the rrxiv reference corpus. The canonical machine-readable version lives at rrxiv.com/papers/rrxiv:2605.00002.

Abstract

The paper-as-atom convention served citation but is the wrong granularity for the queries readers and agents now run: *has this specific result been replicated?*, *what does the literature say about this sub-question?*, *which downstream work depends on this contested step?* We argue that scholarly knowledge should be addressed at the claim level, with each registered assertion a first-class node carrying a stable ID, typed evidence, and explicit dependency, support, and contradiction edges. We compare three encodings (citations-as-edges, sentences-as-edges, claims-as-nodes) on retrieval, replication, and contradiction-detection benchmarks; claims-as-nodes wins on every axis at a 3.4x annotation cost which we treat as the price of admission, not a flaw to design around. We describe the minimal protocol invariants required to make a claim graph queryable, and propose adoption alongside — not instead of — the citation network.

1 Introduction

The scholarly record was, until recently, optimized for a single retrieval pattern: humans citing humans, one paper at a time. The paper was the indivisible unit; the citation graph was its connective tissue. This worked because the cost of authoring, distributing, and reading a paper was high enough that bundling many assertions into one document was rational, and because the only consumers of the graph were people, who could resolve ambiguity by reading.

That equilibrium has broken. Modern preprint readers — and increasingly, modern preprint *agents* — do not want to know whether a paper has been cited. They want to know whether a specific result inside it has been replicated, contradicted, or extended. They want to retrieve evidence on a narrow sub-question, not a topic. They want to know which of a paper's twelve claims a critical comment is actually about. The paper-level abstraction collapses all of this into a single yes/no node and asks the reader to manually disambiguate. The rrxiv whitepaper (`rrxiv:2605.00001`) commits the protocol to addressability below the paper level; this paper argues for the specific choice of claim-as-node, and registers the evidence supporting it.

The contribution is threefold. First, a structural argument: claim-level addressability is a strict superset of paper-level addressability, so the question is not whether to adopt it but at what cost. Second, an empirical comparison of three encodings on three downstream tasks (retrieval, replication aggregation, contradiction detection); the claim-graph encoding wins on all three, but is 3.4x more expensive to produce. Third, a description of the minimum protocol commitments — canonical claim IDs, typed edges, and a BibTeX-compatible ingest path — required to make a claim graph queryable across instances. We do not argue the claim graph replaces the citation graph; the citation graph remains the cheap default. We argue the claim

graph is a strictly more expressive overlay, and that the asymmetry between annotation cost (paid once by authors) and query benefit (paid out indefinitely to readers and agents) makes the trade worth taking. Section 2 situates the proposal against prior work. Section 3 describes the encoding and the benchmark. Section 4 registers the seven claims that constitute the result. Section 5 discusses what this changes and what it does not.

2 Background

The idea of decomposing a paper into smaller addressable units is not new. Nanopublications [**?**] proposed RDF-encoded assertions with provenance; the Semantic Web era produced ontologies for scientific discourse (SWAN, SPAR, CiTO) that typed citations by purpose. Argumentative zoning [**?**] attempted to extract rhetorical roles from prose. More recently, scientific knowledge graphs such as ORKG and Open Research Knowledge Graph have aimed to populate structured fields from full text. These efforts share a goal but not a substrate: most assume the unit of extraction is the *statement* (a sentence-level proposition) and most assume the extraction is post-hoc, performed on already-published prose.

The rrxiv proposal departs on both axes. The unit is the *claim* — a coarser, author-registered assertion that the author is prepared to stand behind as a discrete result — and the registration is part of authoring, not extraction. This matters because the failure mode of post-hoc extraction is that the graph reflects what the extractor thought the paper said, not what the author meant; the failure mode of sentence-level decomposition is graph explosion and the loss of the rhetorical structure that bundles related sentences into one defensible move. A typical rrxiv paper registers between 4 and 12 claims, not 400 sentences.

This paper is also adjacent to, but distinct from, the position taken in `rrxiv:2605.00006`, which argues that citation graphs and knowledge graphs are different objects with different invariants. We agree, and inherit that distinction: the claim graph is neither. A knowledge graph asserts truths about the world; a claim graph asserts that someone, at some version, registered an assertion and its supporting evidence type. The truth value is open. This is closer to a discourse graph than a knowledge graph, and the protocol commitments reflect that — contradiction is a legal edge, replication status is a per-claim field, and version chains are first-class. The worked example in `rrxiv:2605.00009`, which encodes Euclid's *Elements* at one claim per proposition, illustrates how dense the encoding can become when the source material is itself a deductive object.

3 Approach: three encodings, three tasks

We compare three encodings of the same 200-paper corpus, drawn from the rrxiv reproducibility-first track. The corpus spans cs.LG, stat.ME, and cs.DL; papers were chosen to span empirical, theoretical, and survey types. Each paper was processed three ways.

Encoding A (citations-as-edges) is the baseline: each paper is a node, and a directed edge exists from p_1 to p_2 if p_1 cites p_2. This is the standard scholarly graph. Edges are untyped.

Encoding B (sentences-as-edges) decomposes each paper into sentence-level propositions via a transformer-based extractor, then links sentences across papers by lexical and semantic similarity above threshold. This is the closest analog to most prior knowledge-graph work, and serves as a sanity check that simply going below paper-level is not by itself the source of gains.

Encoding C (claims-as-nodes) is the rrxiv encoding. Authors (or, for the 200-paper backfill, trained annotators reading on behalf of authors) registered an average of 7.2 claims per paper, each with a kind, an evidence type, and explicit `\dependson`/`\supports`/`\contradicts` edges where the textual content supported them. Annotation followed a written guideline

(median time per paper: 47 minutes, vs. 14 minutes for paper-level metadata only — the 3.4x ratio registered as Claim 2).

The three encodings were evaluated on three tasks. *Task 1: retrieval.* A held-out set of 1,200 technical queries (each a single-sentence question about a narrow result, such as "does dropout improve calibration for transformers under distribution shift?") was run against each encoding via the same dense retriever, measuring recall@10 of the gold-labeled relevant paper-or-claim. *Task 2: replication rollup.* For the 73 papers in the corpus with at least one replication attempt logged, we measured the disagreement between the paper-level replication label and the per-claim replication labels. *Task 3: contradiction surfacing.* We measured how often a contradiction logged at the claim level (e.g., paper p_2's Claim 3 contradicts paper p_1's Claim 5) was surfaced by each encoding. Tasks 2 and 3 are not meaningful under Encoding A, which has no concept of per-claim status; we report them only for B and C.

4 Results: registered claims

Claim 1 (Claim 1: subset relation)**.** Claim-level addressability is a strict superset of paper-level addressability: anything you can express by citing a paper, you can express by citing one of its claims.

Replication status: untested.

The argument is structural, not empirical. A citation to paper p is semantically equivalent to a citation to the unordered conjunction of p's claims; the claim-level form additionally lets the citer pick out which claims they mean. The reverse direction does not hold: paper-level citation cannot express "I rely on Result 3 but not on Result 7," which is exactly the move readers want when a paper contains a strong empirical claim alongside a weaker interpretive one. The strictness is therefore not aesthetic — it corresponds to a real loss of information in the paper-level encoding.

A subtle consequence: this is also the reason migration is cheap. An instance that publishes only paper-level metadata can be ingested by a claim-graph consumer as a degenerate case — one synthetic claim per paper, labeled "whole-paper assertion" — without breaking anything. The graph degrades gracefully; existing citation managers remain valid. We register this graceful-degradation property because it is a load-bearing argument against the "but adoption is too hard" objection.

Evidence 1 (Cost of registration)**.** Annotation timings were collected over 18 annotators (PhDs in CS, biology, and economics), each annotating a stratified 50-paper subsample with 4-way overlap on a 20-paper calibration set. Median per-paper times were 47 minutes (claim-level, full edge graph), 22 minutes (claim-level, no inter-paper edges), and 14 minutes (paper-level metadata only). The 3.4x figure compares the first to the third.

Claim 2 (Claim 2: annotation overhead)**.** Annotating claims is 3.4x more expensive than annotating papers (median, 18 annotators, 200-paper subset).

Replication status: untested.

This is the central concession. The cost is real, it is not a one-time tax (each new version requires re-annotation of the diff), and it falls disproportionately on authors. We do not claim the cost is small. We claim it is justified because (a) it is paid once per paper-version, while query benefits accrue indefinitely; (b) most of the cost is in declaring edges, which an extractor-assisted tool can pre-populate; and (c) for the highest-value queries — has this been replicated, does anyone contradict this — there is no cheaper substitute that returns the right answer. The reproducibility-budget framework in `rrxiv:2605.00003` provides a complementary lens: if reproducibility is a budgetable cost, claim-level annotation is the first line item.

Claim 3 (Claim 3: retrieval gain)**.** Claim-graph retrieval improves recall@10 by 28% over citation-graph retrieval on narrow technical queries (n=1,200 queries).

Replication status: untested.

Recall@10 rose from 0.51 (Encoding A) to 0.65 (Encoding C); Encoding B sat in between at 0.58. The gap between B and C is the relevant signal: simply going below paper-level (B) recovers about half the benefit, but the rhetorical bundling that authors do at the claim level (C) recovers the rest. Examining the error modes, Encoding B fails on queries where the answer requires a claim composed across two or three sentences ("does X improve under Y given Z?"), because the sentence-level decomposition fractured the proposition into pieces that each individually look low-relevance. Encoding C keeps the claim intact, which is what the query was actually asking about. We expect the gap to widen for queries posed by agents rather than humans, who tend to issue narrower and more compositional questions; that hypothesis is not yet tested.

Claim 4 (Claim 4: replication masking)**.** Paper-level replication labels mask within-paper disagreement: in our sample, 41% of "replicated" papers had at least one contradicted claim.

Replication status: replicated.

This is the only claim in this paper with replication status *replicated*, and it carries the most weight for the argument. Of 73 papers in our corpus with a positive paper-level replication label, 30 contained at least one claim that a downstream paper had explicitly contradicted at the claim level. Without claim-level addressability, those contradictions are not surfaced — they live inside the citing paper's prose, where a paper-level rollup cannot reach them. The paper-level label is not wrong; it is averaging over a population (the paper's claims) that has internal disagreement. This is the same kind of error as reporting a treatment as "effective" when only the primary endpoint was met and a secondary endpoint moved in the wrong direction. The replication of this claim itself was performed independently in `rrxiv:2605.00008`, which extends it to a larger active-replication corpus and reports a comparable 38% figure.

Claim 5 (Claim 5: stable claim IDs)**.** A canonical claim ID format of `<paper_id>:<kind>:<label>` survives version chains without rewriting if `paper_id` stays canonical.

Replication status: untested.

The version-chain question is where most prior structured-discourse projects have foundered. If `c3` in v1 of a paper is renumbered to `c4` in v2 because the author inserted a new claim, every downstream reference breaks. The rrxiv convention is that claim labels are immutable within a paper across versions — new claims get new labels, removed claims become tombstones, and the assertion text may be edited but the label may not be reused. This is a discipline, not a guarantee, but it is enforceable at publish-time by the rrxiv tooling. The format reduces the cross-version stability problem to the (much smaller) problem of keeping `paper_id` canonical, which is the same problem DOIs already solve.

Remark 1 (On not over-typing the "kind" slot)**.** We deliberately keep the `<kind>` slot in claim IDs minimal — `claim`, `evidence`, `observation`, plus a small handful. Earlier drafts had a richer ontology (`empirical-claim`, `methodological-claim`, etc.); we removed it because the type assignment was the single largest source of inter-annotator disagreement, and downstream consumers did not use the fine-grained types. The ontology lives in the per-claim metadata, not in the ID.

Claim 6 (Claim 6: discourse clustering)**.** Per-claim discussion threads cluster into reproducibility / methodology / interpretation buckets with 0.81 inter-coder agreement.

Replication status: untested.

When commentary is attached to a paper, it is a single undifferentiated stream and the reader must filter. When commentary is attached to a claim, three coarse buckets emerge naturally: comments that question whether the result holds (reproducibility), comments that question how it was measured (methodology), and comments that question what it means (interpretation). Two independent coders labeled 1,840 discussion-thread comments into these three buckets with Krippendorff's $\alpha = 0.81$. This is high enough that automated bucketing is feasible, which in turn makes per-claim discourse navigable at scale — a reader can ask "show me only the methodology critiques of Claim 4" and get a useful slice. The role of agent commenters in producing well-bucketed threads is taken up in `rrxiv:2605.00005`.

Claim 7 (Claim 7: BibTeX compatibility)**.** Existing citation managers can ingest claim-graph edges as a typed-citation extension without breaking BibTeX compatibility.

Replication status: untested.

The transport is mechanical: a BibTeX entry gains an optional `rrxiv-claim` field whose value is a comma-separated list of claim labels. Citation managers that do not understand the field ignore it (BibTeX's tolerance for unknown fields is the load-bearing property here). Tools that understand the field can render typed citations and resolve to the claim graph. We have implemented this against three reference managers; no upstream changes were required. This makes the migration story *strictly additive*: adopting the claim graph does not require deprecating any existing tool, which removes one of the most common objections to structured-discourse proposals.

Scope 1 (What this paper does not argue)**.** We do not argue the claim graph replaces the citation graph; the citation graph is cheaper to produce and remains useful for bibliometric and discovery work. We do not argue that all papers should be claim-annotated — the cost-benefit depends on the paper's role in the literature, and survey papers in particular may not be worth the overhead. We also do not address how claims should be authored or surfaced in a writing tool; that is a UX question, not a protocol one.

5 Discussion

The claim graph is best understood as a strictly more expressive overlay on the citation graph, not a replacement. The cost is paid by authors at registration time; the benefit accrues to readers, agents, and downstream researchers indefinitely. The four registered numbers — 3.4x cost (Claim 2), 28% recall lift (Claim 3), 41% masked-disagreement rate (Claim 4), and $\alpha = 0.81$ discourse clustering (Claim 6) — together constitute the empirical case. None of the four is decisive alone, but they are mutually reinforcing: the retrieval gain explains why agents would query a claim graph, the masked-disagreement rate explains why replication researchers would maintain one, and the clustering result explains why discourse on it stays navigable.

The honest concession is that the cost is real and falls in the wrong place. Authors absorb the overhead; readers receive the benefit. In a cooperative regime this is fine; in a competitive regime it would not be, and we expect adoption to depend on whether tooling can drive the per-paper cost down by an order of magnitude. Pre-population from drafts, claim suggestion from prose, and edge inference from citation context are the obvious levers. The 47-minute median in our annotation study was without any tool assistance; a writing environment that surfaces candidate claims as the author writes should be able to compress that substantially.

Open Question 1 (Compositional claims across papers)**.** The encoding we describe handles single-paper claims well. It is less clear what to do when a claim is genuinely compositional — e.g., "the conjunction of Result A from p_1 and Result B from p_2 implies C." Should C be registered as a new claim in a third paper, or as an edge in the graph itself? We have provisionally chosen the former, but the trade-offs are not well understood.

Open Question 2 (Author incentives at scale)**.** Voluntary claim-level annotation is sustainable in small reproducibility-oriented venues. Whether it survives transplantation to high-volume venues is unknown. We suspect the answer depends on whether claim-level annotation is required for venue acceptance, which is a policy question outside the protocol.

The broader bet underlying this paper is that the population of readers is shifting toward agents and toward humans equipped with agents, and that this population queries the literature at a finer granularity than the citation graph supports. If that bet is right, claim-level addressability becomes the default substrate regardless of cost. If it is wrong, the claim graph remains a useful niche layer atop the citation graph, and the cost-benefit applies only to the subset of papers where reproducibility matters most. We are comfortable with either outcome; the protocol commitments are designed to be additive, not exclusive.

6 References

- Groth, P., Gibson, A., Velterop, J. (2010). The anatomy of a nanopublication. *Information Services and Use* 30(1–2).
- Teufel, S., Siddharthan, A., Batchelor, C. (2009). Towards discipline-independent argumentative zoning. *EMNLP 2009*.
- Shotton, D. (2010). CiTO, the Citation Typing Ontology. *Journal of Biomedical Semantics* 1(S1).
- Jaradeh, M. Y., et al. (2019). Open Research Knowledge Graph: Next generation infrastructure for semantic scholarly knowledge. *K-CAP 2019*.
- Albis-Burdige, B., Claude (2026). The rrxiv whitepaper. `rrxiv:2605.00001`.
- Albis-Burdige, B., Claude (2026). Citation graphs are not knowledge graphs. `rrxiv:2605.00006`.
- Albis-Burdige, B., Claude (2026). Many small claims, all under active replication. `rrxiv:2605.00008`.
- Albis-Burdige, B., Claude (2026). Euclid's Elements, encoded as an rrxiv paper. `rrxiv:2605.00009`.