From a Web of Documents to a Web of Claims: A Claim Layer for Human and AI Science

Jingbo He and Michael Färber

TU Dresden & ScaDS.AI, Germany
{jingbo.he,michael.faerber}@tu-dresden.de

Abstract. We propose the *Web of Claims* as the next stage of the Web for science: after documents and data, claims. Scoped to the scholarly record, every scientific assertion becomes a first-class object with a persistent identifier, a signature, stated conditions, and an equivalence class over its surface forms. Two halves of this system exist separately today: the Semantic Web community solved the *representation* of such assertions, but the infrastructure was never populated at scale; generative AI has made population cheap, but its pipelines discard the claims they extract after every run. Conditioned claims form a graph that neither earlier Web contains, and over it a third kind of inference becomes definable, *acceptability*: which claims can be jointly held, under which conditions, on what evidence. A running example develops the claim and this inference step by step. Both rest on two open problems, claim identity and contextual truth; we show against four criteria that no existing line of work solves all four, and outline a twenty-five-year research agenda.

Keywords: Scientific Claims · Knowledge Graphs · Contextual Truth · Scholarly Communication.

1 Introduction

The scholarly record is under strain from both sides of the transition to generative AI. Retractions reached a record ten thousand papers in 2023 [24], yet only 5.4% of more than thirteen thousand post-retraction citation contexts in biomedicine acknowledge the retraction [14]. Machine-generated papers have meanwhile begun to enter the record: in 2025 a fully AI-generated manuscript passed a first round of peer review, and the system behind it was published in *Nature* a year later [18]. Scientific question-answering systems now add per-claim citation and evidence filtering on top of retrieval [3]: support and contradiction hold between claims, not between passages. The narrative document has been science's atomic unit since 1665 [8]; it can neither absorb the current publication volume nor hold its authors, human or machine, accountable for individual assertions.

The Semantic Web community largely solved *representation*. Nanopublications [11], micropublications [8], and scholarly knowledge graphs [16] give individual assertions identity, attribution, and citability. What was never solved is *population*. Field studies attribute the stall to extraction accuracy that stayed poor for

anything beyond the most basic tasks [7], and to manual curation that was costly and brought little credit [22]. Large language models have since changed the economics, decomposing text into atomic claims at negligible cost [21]; what they have not built is representation. The extracted claims are disposable, re-derived on every run with no identity, provenance, or reuse, and the decomposition itself is unstable unless made aware of conditions [15]. Representation without population failed; population without representation is failing.

We think joining the halves could occupy this community for the next twenty-five years. The Web began as a web of documents; the Semantic Web re-imagined it as a web of data [2]. We propose a third stage, a *Web of Claims*, in which every scientific assertion is a first-class object whose persistent identity and equivalence class span papers, corpora, and authors, human or machine. Truth is contextual: a claim holds under stated conditions, and disagreements are preserved rather than averaged away. Each stage of the Web brought its own inference: relevance over documents, entailment over data. Conditioned claims form a graph neither predecessor contains, and over it a third becomes definable, *acceptability* [10]: which claims are jointly held, under which conditions, on what evidence.

Does this vision not already exist as nanopublications plus provenance? The adoptable machinery does (Sect. 5); what no deployed system provides is identity across surface forms (two differently worded nanopublications are simply two nanopublications) and a truth type indexed by conditions, whose absence shows as a hard 0.7 ceiling that no better model removes (Sect. 4.2).

2 The Web of Claims

In a Web of Claims the units identified, linked, and held to account are claims rather than documents; documents do not disappear, they stop being what the infrastructure addresses. We call the claim-level infrastructure the *claim layer*.

A claim on this layer has six components. *Identity* is a persistent identifier together with an equivalence class over surface forms. *Content* pairs a structured core with natural-language forms, since demanding full formalization would suppress population. *Context* states the conditions under which the claim is asserted to hold, and *Provenance* records, under signature, who or what asserted, extracted, or endorsed it. Typed *evidence links* run to other claims and to data. *Standing*, finally, is a profile of verdicts indexed by condition. The structured core lives in RDF and builds on the triple terms of RDF 1.2 [26], but a claim is not a triple. A false triple is an error to repair; a refuted claim remains a first-class object whose standing records the refutation.

A boundary question follows: is not every statement on every web page a claim? In principle, yes; on the layer, no. An assertion is admitted only if it carries the six components, above all a signing agent, stated conditions, and typed evidence; scientific assertions are the ones for which these are already social expectations. We therefore scope the Web of Claims to the scholarly record.

LLM pipelines mint candidate claims [21] and candidate equivalences [17], and hybrid verification refines them; verification of this kind already runs at

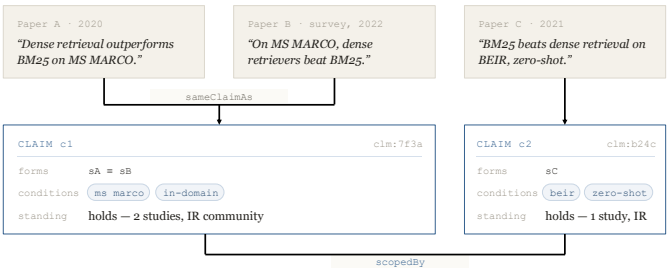


Fig. 1. Two forms are one claim; the third differs in its conditions — scope, not conflict.

scale over hundreds of thousands of machine-extracted assertions, with domain-expert feedback in the loop [20]. Deposit by authors comes second, inverting the authors-formalize-first assumption that stalled earlier semantic publishing. Consumption mirrors population: what gets cited and grounded on is the claim rather than the document, and every write is signed, so an AI scientist accrues or loses standing exactly as a human author does.

Consider the three sentences of the running example, Fig. 1. One paper reports that dense retrieval beats BM25 on MS MARCO; a survey restates the finding in different words; a third paper reports the opposite on BEIR, in a zero-shot setting. The first two express the same claim. Differing in surface form and in nothing else, they join a single equivalence class, and their evidence aggregates. The third expresses a different claim, because the corpus and the transfer setting are part of what it asserts. The layer records `scopedBy` rather than `conflictsWith`, and a reader obtains both claims with conditions attached. A deduplicator on surface text would merge all three and record a contradiction that exists only because the conditions were discarded. MedScore documents its precondition, the loss of conditions during decomposition, for biomedicine [15]. Nothing here requires machinery beyond making conditions first-class.

If the layer lowers the cost of writing, why is it not itself a target for pollution? Pollution becomes visible: a bad-faith write either mints an isolated claim that nobody endorses, or collides with an existing class's evidence profile as an anomaly. And verification persists and is reused, so defence compounds where attack does not. Standing affects the weight a claim carries, never whether it can be written.[1]

3 Acceptability over Evidence

The claim graph supports a computation with no analogue at either predecessor layer. Abstract argumentation supplies the template: given a relation of attack between arguments, an acceptability semantics returns the sets of arguments that can be jointly held [10]. The scientific setting breaks that template in three

[1] Repo: `https://github.com/faerber-lab/web-of-claims`

ways. Attacks are relative to conditions: the two claims of Fig. 1 conflict on their surface forms and are compatible in fact. Verdicts are indexed by the community that evaluates them. And support is graded by signed provenance rather than simply present or absent. To the best of our knowledge, no acceptability semantics indexes arguments by the conditions under which they hold. Value-based argumentation comes closest, making acceptability relative to an audience [1]; but conditions are not audiences, and no framework we have found carries evidence or a trust model. The running example makes the first break concrete: read without conditions, s_C attacks s_A and a Dung semantics rejects one; indexed by conditions, c_1 and c_2 are jointly acceptable. A genuine attack needs a fourth report of BM25 winning on MS MARCO in-domain, colliding with c_1 under the *same* conditions, weighed by graded evidence and community standing.

Two families of computation follow. The first is revision. Truth maintenance systems propagate a withdrawal through a justification structure [9], but they have never been run over a literature, which exposes no such structure; the unread retraction notices of Sect. 1 are what withdrawal looks like when the unit is the paper. Suppose the paper behind s_A is retracted: the retraction is a signed write against one provenance edge into c_1, the standing of c_1 recomputes to a single study, and a traversal flags every answer and downstream claim whose support chain passes through it. The second family is generation: literature-based discovery becomes the prediction of missing edges between conditioned claims, making the discovered hypotheses sound with respect to their conditions. Both become well defined once identity fixes nodes and contextual truth types edges.

4 Two Constitutive Problems

Much of the machinery a claim layer needs is *adoptable*: signatures, provenance, and trust propagation operate on whatever unit they are given [20]; only the type they compute over must change. A problem is *load-bearing*, by contrast, when it determines what the unit is, and only two meet that criterion: deciding when two assertions are the same claim, and saying what it is for a claim to hold. Both presuppose conditions as first-class structure: identity tells claims apart under conditions; standing judges them under the same.

4.1 Global Claim Identity and Equivalence

When are two differently worded assertions the *same claim*? Almost everything else presupposes an answer: deduplication, aggregation, verdicts, accountability. The relation is graded: two assertions can be paraphrases, one can entail or specialize the other, or both can state identical content under different conditions. Which of these count as "the same" depends on the task, since a systematic review aggregates what a replication audit must keep distinct. Identity is therefore a family of typed equivalence and subsumption relations, not a single clustering. This community has met the problem before in another guise: `owl:sameAs` is a single identity predicate routinely pressed into service for the weaker relations

it cannot express [13]. SEPIO [4] ties an assertion to the agent and provenance of its making but lives inside curated databases; claim verification and matching [17] yield labels and clusters valid for a single run: equivalence without identity. Context, on our account, is part of identity rather than an annotation upon it. The milestone is convergence: identity works when independent extraction runs arrive at substantially the same identifiers. Studies of manual formalization already measure the analogous convergence among four experts who were themselves the method's designers: full or majority agreement in 56% of cases before discussion and 84% after it [6]. No automated pipeline has been evaluated this way. Standing attaches to equivalence classes, making equivalence assertions signed, contestable claims that no single party adjudicates globally.

4.2 Contextual Truth

What is a claim's truth status across conditions, communities, and time? A scalar, whether binary or probabilistic, is the wrong type for it: a claim may be supported in one population and contradicted in another, and one number discards exactly what a scientific record exists to preserve. The closest built system makes the point empirically: three LLMs assigned scalar truth values to 45,649 assertions, and the aggregate agent built from them cannot exceed 0.7 for any assertion, because combining scalars weighted by each model's accuracy caps the result there [20]. A fact beyond dispute scores 0.7: the ceiling is the type failing, not the models. Standing should instead be a profile: verdicts indexed by conditions and by evaluating community, with disagreement retained. One index already exists: PROV-K records a trusted truth value per agent, timestamped, with supporting and conflicting evidence [20]; what no deployed model indexes is the condition. What averaging costs is visible in the same system: two models return 0.10 and 0.05 on an assertion where a third returns 0.90, and the aggregate reports 0.26 [20]. Three agents in flat contradiction become one unremarkable number. This is conditionalization, not relativism: a bare claim is incomplete, its conditions complete its content, and the layer records a verdict issued *by* a community rather than truth *for* it, itself an objective, signed, contestable fact. GRADE already does this in clinical practice, rating certainty outcome by outcome rather than assigning one score to a finding [12]. Nor does preserving disagreement serve those who manufacture it: once endorsements carry the structure of their evidence, volume without depth shows up as thinness.

5 The Research Landscape and the Gap

Technical preconditions keep arriving faster than the unit that would organize them. Table 1 maps related lines of work onto four criteria that together define the claim: (a) persistent global identifiers, (b) equivalence across surface forms, (c) conditions as part of identity, (d) typed disagreement. Each realizes a strict subset; none realizes all four. We state this as a deliberately falsifiable claim, since a system satisfying all four would make this engineering rather than research.

Table 1. Related lines of work against criteria (a)–(d) from the text.

approach	(a) identity	(b) equivalence	(c) conditions	(d) disagreement
nanopublications	per artifact	no	yes, context slot	supersede, retract
micropublications, SEPIO	per publication	no	partial	support, challenge
ScholOnto, discourse graphs	per project	no	no	support, oppose
argumentation frameworks	per framework	no	no	typed attack
claim verification	per run	clusters	rare	three-way label
Wikidata, ORKG	per item, paper	no	annotation	ranks

Nanopublications come closest as infrastructure: they provide persistent identifiers, and the super-pattern's context slot was used in 76% of the claims of the original study [6]. But the identifier names the artifact, not the claim's meaning, so two differently worded nanopublications are simply two nanopublications. Wikidata approximates context, provenance, and disagreement through qualifiers, references, and ranks [25], yet its identifiers name entities rather than assertions, and its ranks resolve disagreement editorially. The ambition is a quarter-century old: ScholOnto set out to build a semantic network of research claims in 2000 [5]. Claim-native proposals multiply again, yet neither of the newest closes Table 1: Blueprints [19] scope identity to the artifact, so one claim in two blueprints is two components, and ClaimFlow [23] canonicalizes surface forms within a paper but not across them, so a claim advanced in two papers remains two nodes. The unit is reached for from several directions, and fixed by none.

6 A Twenty-Five-Year Research Agenda

Foundations (2026–2031). The first task is infrastructure for claim identity: identifier schemes, typed equivalence vocabularies, open registries, and benchmarks that evaluate extraction as an identity task, alongside a first corpus-scale claim graph over millions of open-access papers. Retrieval is the natural first consumer; living systematic reviews the natural first domain, since review teams already perform claim identity, conditioned verdicts, and evidence linking by hand, and none of that labour is reusable on the document layer. The economics run consumers-first: every assistant today re-derives claims and discards them; the layer converts that repeated private cost into work done once and reused. Success by 2031 would mean two independent pipelines converging on substantially the same identifiers over a shared open corpus, and one living review running on the layer instead of beside it. What distinguishes 2026–2031 from the preceding twenty-five years is not the machinery but the evaluation target: no automated pipeline has ever been assessed on identifier convergence.

Institutions and infrastructure (2031–2051). Peer review then becomes claim-native, with reviewers examining the change a submission makes to the layer, as piloted by nanopublication-based reviewing [7]; machine authors acquire accountability protocols, and citation and credit follow the claim. By 2040, at least one venue should read those changes and citation indices should resolve claim identifiers; by 2051 the layer should be the default grounding target for scientific

AI, with documents persisting as narrative interfaces, and the measure of success is disappearance: a layer as unremarkable as DOIs are today.

Three failure modes concern us. Extraction error at scale, pollution with better packaging, is why the benchmark comes first and verification stays hybrid. Cold start is why the order runs consumers-first: assistants already extract the claims they would deposit. Capture argues for federated minting and contestable equivalence writes, since whoever controls minting or equivalence controls standing. No mitigation is complete; they mark where the design is tested first.

7 Conclusion

Twenty-five years ago this community proposed that the Web's knowledge could be given identity, structure, and meaning, and built the machinery it required; that machinery is one half of the system the generative-AI era needs, and the other half, cheap population, has arrived from outside. Claim identity and contextual truth are open problems squarely in this community's competence, and acceptability over evidence is an inference still awaiting its semantics. We offer the Web of Claims as a shared research programme rather than a finished design, and claim registries with identity-aware retrieval are a first step available now.

Use of Generative AI

Claude was utilized to generate sections of this Work including text.

Acknowledgments. The authors acknowledge the financial support by the Federal Ministry of Research, Technology and Space of Germany (BMFTR) and by the Saxon State Ministry for Science, Culture and Tourism in the programme Center of Excellence for AI Research "Center for Scalable Data Analytics and Artificial Intelligence Dresden/Leipzig", project identification number: ScaDS.AI.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Bench-Capon, T.J.M.: Persuasion in practical argument using value-based argumentation frameworks. Journal of Logic and Computation **13**, 429–448 (2003)
2. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific American **284**, 34–43 (2001)
3. Besrour, I., He, J., Schreieder, T., Färber, M.: SQuAI: Scientific question-answering with multi-agent retrieval-augmented generation. In: CIKM. pp. 6603–6608 (2025)
4. Brush, M.H., Shefchek, K., Haendel, M.: SEPIO: A semantic model for the integration and analysis of scientific evidence. In: ICBO/BioCreative, CEUR 1747 (2016)
5. Buckingham Shum, S., Motta, E., Domingue, J.: ScholOnto: An ontology-based digital library server for research documents and discourse. International Journal on Digital Libraries **3**, 237–248 (2000)

6. Bucur, C.I., Kuhn, T., Ceolin, D., et al.: Expressing high-level scientific claims with formal semantics. In: K-CAP. pp. 233–240 (2021)
7. Bucur, C.I., Kuhn, T., Ceolin, D., et al.: Nanopublication-based semantic publishing and reviewing: A field study with formalization papers. PeerJ Computer Science **9**, e1159 (2023)
8. Clark, T., Ciccarese, P.N., Goble, C.A.: Micropublications: A semantic model for claims, evidence, arguments and annotations in biomedical communications. Journal of Biomedical Semantics **5**, 28 (2014)
9. Doyle, J.: A truth maintenance system. Artificial Intelligence **12**, 231–272 (1979)
10. Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artificial Intelligence **77**, 321–357 (1995)
11. Groth, P., Gibson, A., Velterop, J.: The anatomy of a nanopublication. Information Services and Use **30**, 51–56 (2010)
12. Guyatt, G., Oxman, A., Vist, G., et al.: GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. BMJ **336**, 924–926 (2008)
13. Halpin, H., Hayes, P.J., McCusker, J.P., et al.: When `owl:sameAs` isn't the same: An analysis of identity in linked data. In: ISWC. pp. 305–320 (2010)
14. Hsiao, T.K., Schneider, J.: Continued use of retracted papers: Temporal trends in citations and (lack of) awareness of retractions shown in citation contexts in biomedicine. Quantitative Science Studies **2**, 1144–1169 (2021)
15. Huang, H., DeLucia, A., Tiyyala, V.M., Dredze, M.: MedScore: Generalizable factuality evaluation of open-ended long-form medical answers by domain-adapted claim decomposition and verification. In: Findings of ACL. pp. 14149–14180 (2026)
16. Jaradeh, M.Y., Oelen, A., Farfar, K.E., et al.: Open research knowledge graph: Next generation infrastructure for semantic scholarly knowledge. In: K-CAP. pp. 243–246 (2019)
17. Kazemi, A., Garimella, K., Gaffney, D., et al.: Claim matching beyond English to scale global fact-checking. In: ACL-IJCNLP. pp. 4504–4517 (2021)
18. Lu, C., Lu, C., Lange, R.T., et al.: Towards end-to-end automation of AI research. Nature **651**, 914–919 (2026)
19. Ma, J.W.: Toward an engineering of science: Rebalancing generation and verification in the age of AI. arXiv:2605.10425 (2026)
20. Menotti, L., Marchesin, S., Giachelle, F., et al.: Provenance-driven nanopublications: Representing source lineage and trust networks for multi-source assertions. International Journal on Digital Libraries **26**, 24 (2025)
21. Min, S., Krishna, K., Lyu, X., et al.: FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In: EMNLP. pp. 12076–12100 (2023)
22. Open Research Knowledge Graph: Call for applications: Curation grants 2025. `https://orkg.org/page/call-curation-grants-2025` (2025), accessed July 2026
23. Pramanick, A., Hou, Y., Mohammad, S.M., Gurevych, I.: ClaimFlow: Tracing the evolution of scientific claims in NLP. arXiv:2603.16073v2 (2026)
24. Van Noorden, R.: More than 10,000 research papers were retracted in 2023 — a new record. Nature **624**, 479–481 (2023)
25. Vrandečić, D., Krötzsch, M.: Wikidata: A free collaborative knowledgebase. Communications of the ACM **57**, 78–85 (2014)
26. W3C RDF and SPARQL Working Group: RDF 1.2 concepts and abstract data model. W3C Candidate Recommendation (2026), `https://www.w3.org/TR/rdf12-concepts/`