

A Rhetorical Citation Graph Model for Characterizing Scientific Findings: Resource Construction and Application to Plant Health

Anne-Sophie Foussat^{1,2,3}, Robert Bossy¹, Vincent Guigue², Nicolas Sauvion³,
Claire Nédellec¹

¹ Université Paris-Saclay, INRAE, MaIAGE, 78350 Jouy-en-Josas, France,

² AgroParisTech, MIA Paris-Saclay, Palaiseau, France,

³ PHIM, Univ Montpellier, INRAE, CIRAD, IRD, Institut Agro, Montpellier, France

Correspondence: anne-sophie.foussat@inrae.fr

Abstract

Understanding how scientific findings are discussed and evaluated over time provides valuable evidence for assessing their reliability. We propose the Rhetorical Citation Graph (RCG), a finding-centric graph model that represents citation events by explicitly linking scientific findings, citing articles, rhetorical functions, and citation contexts. We instantiate this model as the Plant Literature Reliability Network (PLIN), a graph of 262,807 citation events automatically constructed from the plant-health literature using Natural Language Processing methods for citation and finding extraction, and rhetorical stance classification. By integrating provenance data with the rhetorical interpretation expressed in citation contexts, the RCG provides structured evidence for contextual analyses of scientific findings and lays the foundation for longitudinal studies of their evolution, scientific consensus, and reliability. Finally, we illustrate the utility of PLIN through a case study on the curation and enrichment of the EPPO Global Database, a reference knowledge base for plant health.

1 Introduction

As scientific production accelerates, distinguishing well-established findings from contested or weakly supported ones has become increasingly difficult. Nissen et al. (2016) describe *canonization* as the process by which a scientific claim becomes widely accepted as fact. In empirical sciences, however, knowledge evolves through replication, agreement, and controversy: a consensus may emerge, be challenged, and later be reformulated (Shwed and Bearman, 2010). Scientific knowledge is therefore dynamic and provisional, continually improved or corrected in light of technical advances and new evidence (Miller, 2013). A temporal perspective is consequently essential for identifying the trajectories through which consensus develops (Shwed and Bearman, 2010).

Citation networks provide a means of tracing the evolution of scientific knowledge. Bibliometric studies examine temporal patterns across papers, authors, journals, institutions, and fields (Yan and Ding, 2012). However, understanding rhetorical function of citations is essential to reveal how peers interpret, support, or challenge a finding over time. Such interpretation requires analysing the context surrounding each citation.

Citation counts alone cannot capture these interpretive relations and may lead to misleading conclusions about scientific quality (Moravcsik and Murugesan, 1975). Understanding how scientific consensus emerges therefore requires representing citation events at the level of individual scientific findings rather than publications. In this paper, we introduce the Rhetorical Citation Graph (RCG), a finding-centric graph model for representing citation events at the level of individual findings. In RCG, a citation is modeled as an event linking a cited finding to a citing article, together with its rhetorical function and the surrounding citation context. Findings are represented as normalized relational assertions together with original textual realization, enabling equivalent findings to be compared across publications while preserving traceability to their original formulation. RCG is constructed from scientific documents using Natural Language Processing methods for identifying citation contexts, extracting findings, and classifying rhetorical functions.

By integrating normalized findings, with citation contexts, provenance information and rhetorical functions, RCG provides structured evidence for studying the dynamics of consensus formation and supporting downstream reliability assessment. We instantiate RCG as the Plant Literature Reliability Network (PLIN), a graph constructed from plant-health literature, and use it as a case study throughout the paper. The graph construction pipeline is organized into successive

processing stages, for which we adapted language models to perform finding extraction and rhetorical function classification. The code used to construct the dataset and the graph is publicly available on GitHub.¹ Finally, through a case study on the curation and enrichment of the EPPO Global Database, we illustrate how PLIN provides structured evidence to support knowledge base curation by facilitating the interpretation of scientific findings.

2 Previous work

2.1 Mapping scientific knowledge with graphs

Previous research has proposed graph-based representations to model different aspects of the scientific literature, including citation networks and structured scientific knowledge. For example, (Viswanathan et al., 2021) combine citation graph embeddings with citances, i.e. the sentences containing the citation marker, to improve information extraction by incorporating inter-articles relationships. Rather than extracting information from individual papers, other studies focus on analyzing corpora of articles to map the evolution of knowledge. For instance, Luo et al. (2018) combine an influence-relation graph, representing causal-like relations between concepts extracted from children’s health publications, with a citation graph to identify gaps in scientific knowledge. Similarly, Kargupta et al. (2025) propose a structured framework to explore the novelty of scientific claims.

2.2 Citation graph

Graph-based models of citation relationships differ in the types of entities and relations they emphasize. Berrebbi et al. (2022) proposes a graph representation in which nodes represent papers, authors, or venues, and edges are the citations between these entities. Quan et al. (2025) also propose a model that emphasizes multiple citation intents for the same citing-cited pair, since a single article may be cited in different sections and for different rhetorical purposes. However, in their representation, citations remain directly attached to the cited article, without distinguishing its different scientific contributions. Addressing this limitation, Liang et al. (2024) introduce a fine-grained representation by modeling both

papers and their individual findings as nodes, with edges representing citation links between a finding and the citing article. This approach accounts for the fact that different findings within the same paper can be cited independently. However, in Liang et al.’s representation, the citation type is not characterized, and the model does not capture the possibility of multiple citation relations between the same finding and a citing article. Others normalize findings into structured semantic assertions to facilitate semantic comparison and integration Pramanick et al. (2026). Existing citation graph models, however, do not jointly represent normalized findings, individual citation events, and their rhetorical functions. Because an article may cite the same finding multiple times for different purposes, methods are needed to extract and normalize findings and link them to individual citation events.

2.3 Scientific Finding Extraction and Normalization

Scientific finding extraction methods often represent findings as textual spans extracted from articles. Their contextual embeddings are then compared to match statements across publications (Cohan and Goharian, 2017; Tan et al., 2026), supporting tasks such as citation contextualization, claim alignment, and miscitation detection. Such textual representations do not enable interpretable comparisons across publication statements or between statements and structured knowledge bases. Recent work has proposed richer semantic decompositions of scientific claims, for example by distinguishing their topic, stance, and claim object (Gangi Reddy et al., 2022), but these representations remain centered on the source document rather than on how claims are subsequently cited and interpreted. In contrast, scientific knowledge graph construction represents findings as normalized semantic relations or events, obtained through entity recognition, relation extraction, entity linking, and ontology grounding enabling comparison and integration across publications (Viswanathan et al., 2021). Recent prompt-based large language models have further improved scientific information extraction, including in plant health through the EPOP corpus (Nédellec et al., 2026). Although these methods provide the building blocks for extracting and normalizing scientific findings, they are generally developed independently of citation analysis and

¹https://github.com/AnneSophie148/PhD/tree/main/new_dataset

therefore do not support modeling how normalized findings are interpreted through individual citation events.

2.4 The rhetorical citation classification task

Capturing citation events at a fine-grained level also requires characterizing their rhetorical function. Rhetorical citation classification, also referred to as citation intent classification (Zhang et al., 2025), aims to characterize the role of citations within the argumentative structure of scientific discourse. Early work relied on rule-based methods (Garzone and Mercer, 2000), which were later replaced by supervised machine learning approaches, using manually engineered linguistic features and document structure features (Teufel et al., 2006; Jurgens et al., 2018). More recently, pretrained language models such as SciBERT have become the dominant approach (Beltagy et al., 2019; Zhang et al., 2022; Jiang and Chen, 2023; Xu et al., 2025), while recent studies have investigated large language models under prompting and in-context learning settings (Kunnath et al., 2023; Koloveas et al., 2025). Other methods explore ways to incorporate relational information between citing and cited entities, using graph embeddings to model the citation intent (Berrebbi et al., 2022; Tsai et al., 2023; Quan et al., 2025). These approaches aim to capture structural properties of citation. While studies on rhetorical citation classification provide a rich characterization of citations, they rely on different annotation schemas with varying typologies and number of classes, thereby limiting cross-study comparison.

3 A Fine-Grained Rhetorical Citation graph

We propose the Rhetorical Citation graph (RCG) a fine-grained graph model that represents how individual scientific findings are cited and rhetorically interpreted over time. Its structure is shown in Figure 1. The central element of the graph is the *Citation Event*, which links a unique *In-context Citation* and *Finding* pair together with the corresponding citing and cited *Article* nodes. The explicit representation of *Citation Events* as nodes is a conceptual modeling choice that gives each unique (*In-context Citation*, *Finding*) association its own identity and makes it directly addressable as the unit of rhetorical analysis. A *Finding* node represents a scientific

finding from a cited article. A single *Finding* node may participate in multiple *Citation Events* when the same finding is cited in different citation contexts. The *In-context Citation* node stores the citation passage, including a sentence-level decomposition that enables contextual analysis, the citation marker, and the cited findings extracted from the passage. Findings are represented as normalized relational assertions in which the two entity arguments are mapped to reference classes defined in hierarchical semantic resources while the relation is directed and labeled. We distinguish between original findings extracted from the article that states the finding, and cited findings extracted from the context in which another article cites the original article. *Citation Event* nodes store the result of the alignment between the original finding and cited findings, indicating whether one of the cited findings matches the original finding. This design preserves every citation event, including cases where no reliable alignment can be established. Modeling citation events as first-class nodes allows each association between a citation context and a scientific finding to be represented and characterized individually. A single citation context may give rise to multiple *Citation Event* nodes when it refers to several findings, while a finding may participate in multiple citation event nodes across different citation contexts or citing articles. Conversely, a cited article typically contains multiple findings that can each be cited independently. This representation preserves fine-grained citation patterns that are collapsed in conventional article-level citation graphs. *Rhetorical Class*, *In-context Citation*, *Section* and *Self-Citation* nodes are connected to *Citation Event* nodes, thereby enriching it with rhetorical contextual information. *Article* nodes stores bibliographic metadata, including DOI, authors, title, and year of publication, and are linked to the findings they report.

The *Rhetorical Class* nodes represent the rhetorical function expressed in a *Citation Event*. The *Comparison or Contrast* class is further refined into four additional categories that indicate whether the rhetorical function concerns methods or results, and distinguish *Comparison*, *Contrast* and *Contradictory results* in order to identify citation events expressing contrasting results. The *Self-Citation* nodes indicate whether the citation is a self-citation, with possible values “yes”, “no”, or “unknown”. The *Section* nodes specify the

structural section of the citation within the citing article.

4 RCG Automated construction

4.1 Implementation in plant health

To illustrate the application of the RCG graph model, we implemented it on a corpus of plant-health literature, resulting in the Plant Literature Reliability Network (PLIN). We focus on scientific publications describing pathogen transmission by insect vectors. This domain provides a coherent body of literature with relatively homogeneous citation practices while involving complex biological interactions that have been the subject of both scientific consensus and ongoing debate. It therefore offers an appropriate testbed for structuring citation events over time. The construction of the PLIN graph comprised four main stages: (i) document acquisition and preprocessing; (ii) finding extraction, and alignment of original finding and cited finding; and (iii) rhetorical function classification.

4.2 Document acquisition and preprocessing

To construct a plant-health citation corpus, we first identified articles on the transmission of plant pathogens by insect vectors in the *Web of Science* (WoS) database. The query targeted English-language research articles and combined plant-related terms, insect-related terms derived from CABI arthropod datasheets, and vector-transmission terms, while excluding terms related to human health. Articles without a DOI were discarded to facilitate full-text retrieval. The resulting reference collection is referred to as the *Query Article Set*. The metadata of the citing-articles were then retrieved primarily from *OpenAlex*, selected for its broad coverage, and enriched with self-citation information through *COCI* API and duplicates identified. During this process, we applied a set of filters to reduce errors caused by bibliographic metadata inconsistencies. The citing-article links were then consolidated from full-texts.

Full-text documents were then collected from multiple sources, including *PMC*, *ISTEX*, *Crossref*, *Unpaywall*, *ScienceDirect*, *Wiley Online Library*, and the *PLOS* Open Dataset. PDF full text documents were processed with *GROBID* (v0.8.2) ([gro](#)) to extract document structure, bibliographic references, and in-text citation markers. We also

used *ScienceDirect*, and *PLOS* citation information. Finally, sentence segmentation was performed to enable consistent extraction of citation contexts.

To associate extracted citation context with the correct cited articles, we matched bibliographic references in the full text to the citing-cited DOI pairs recorded in the citation metadata. This matching relied on article metadata, including DOI, author names, publication year, and title, and was complemented by *CrossRef* queries when exact matching was not possible. Only articles from the *Query Article Set* with at least one citation were indexed in the graph. *Article*, *Section*, and *In-context Citation* nodes were instantiated. *Article* nodes store bibliographic metadata, while each *In-context Citation* node stores the sentence containing the in-text citation marker and a citation window consisting of the three preceding and three following sentences. At this stage, the graph captures the document structure and the bibliographic citation network. Appendix B displays the distribution of the citations.

4.3 Relation extraction and finding identification

Findings are automatically extracted from cited article abstracts and citation contexts. We used the hard prompting approach of ([Nédellec et al., 2026](#)) with the same prompt as it achieved the highest F₁-score (72%) among the methods evaluated by the authors on the EPOP plant health corpus. Following [Nédellec et al. \(2026\)](#), we selected Qwen3-32B for the LLM-based experiments because it was among the best-performing models, alongside DeepSeek-V3, while being substantially less computationally demanding.

To distinguish the original findings from background knowledge in the cited article abstract, candidate findings are classified by a Qwen3-32B LLM. These findings are instantiated as *Finding* nodes storing their normalized relational representations. For each citation context–finding pair, a *Citation Event* node is then instantiated and linked to the corresponding *Finding* and *In-Context Citation* nodes. The findings extracted from the citation context are stored in the corresponding *In-context Citation* node.

The cited findings are compared with the associated original finding to determine whether the citation context refers to that finding. A positive match is assigned when the relation labels are identical and the normalized arguments are semantically

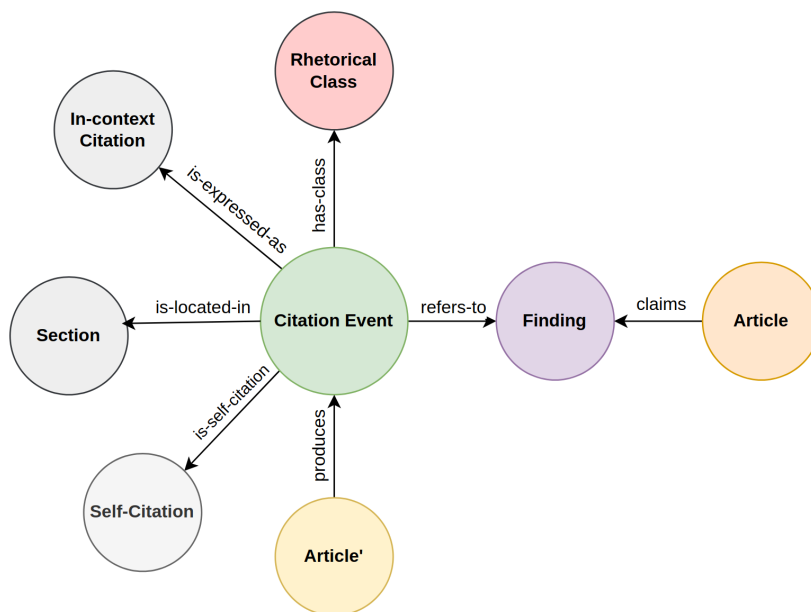


Figure 1: The Rhetorical Citation Graph schema

compatible according to the hierarchical resources used for normalization. Depending on the outcome, the alignment attribute of the *Citation Event* node is set to "yes" or "no". For the plant health application we adopted the same relational schema as the EPOP corpus. Relation labels are *transmits* (between insect and pathogen), and *hosts* (between insect and plant). Relation arguments are species normalized to NCBI taxonomy classes that considered compatible if one belongs to the taxonomic lineage of the other. For example, (Bin et al., 2026) cites (Zhang et al., 2019). Abstract of (Zhang et al., 2019): *This is the first report of the ability of D. citri to transmit CYVCV from infected to healthy citrus under laboratory conditions.*

Citation by (Bin et al., 2026): *Additionally, laboratory studies have shown that CYVCV can also be transmitted by insect vectors, including Aphis spiraecola, A. craccivora, and Dialeurodes citri (Onelge et al. 2010; Zhang et al. 2019a). However, the mechanisms underlying insect-mediated transmission of CYVCV under natural field conditions remain unclear.* The abstract of Zhang et al. (2019) yields the finding *D. citri (Dialeurodes citri) - transmits - CYVCV (Citrus yellow vein clearing virus)*, where the reference names of the vector insect and the transmitted virus according to the NCBI taxonomy are given in parentheses. This finding is represented by a single *Finding* node. The citation context in Bin et al. (2026) yields three cited findings involving *Aphis spiraecola*, *A. craccivora*,

and *Dialeurodes citri*. They are stored in the *In-context Citation* node. During alignment, the relation involving *Dialeurodes citri* is identified as matching the original finding. The alignment attribute of the Citation Event Node is then set to *yes*. The two other cited findings are aligned with different findings extracted from the cited article through their respective Citation Event nodes.

4.4 Rhetorical citation function classification

To assign rhetorical functions to citations and populate the corresponding *Rhetorical Class* nodes, we used successively two language models: a coarse-grain classifier followed by a refinement model designed to distinguish comparison from contrast, and method from result.

Six-class rhetorical citation function classification

In the absence of an annotated corpus in the plant health domain, we adopted a transfer learning strategy. We first fine-tuned language models on rhetorical function datasets from the NLP domain, where such resources are well established, and then applied them to plant health. To evaluate the effectiveness of this transfer, we manually annotated a new corpus, PD100Cit, comprising 100 citation contexts centered on Pear Decline, a disease affecting pear trees. For training, we selected the citation function corpus of Jurgens et al. (2018) because of the high quality of annotation and the relevance of the six-class rhetorical

annotation scheme: *Background*, *Motivation*, *Uses*, *Extends*, *Comparison/Contrast*, and *Future*.

We compared five BERT-based transformer models. The input consists of the citance together with its surrounding context sentences, truncated to a maximum of 512 tokens. Citation markers are replaced by a special token, CITSEG, whose contextual representation is extracted by the pretrained language model. The contextual representation of the CITSEG token is used as a compact representation of the citation, under the hypothesis that it captures the argumentative cues expressed in its local discourse context. The CITSEG embedding is then passed to a multi-layer perceptron consisting of a single hidden layer of size 768 with a non-linear activation, followed by a softmax output layer for rhetorical function classification.

We experimented under different experimental settings to evaluate the impact of the size of citation context on rhetorical function classification. Citation context was evaluated using three context-window sizes: 1–1, 2–2, and 3–3, where an N–N window includes the citance together with the N preceding and N following sentences.

Because the rhetorical classes are highly imbalanced with *Background* as the dominant class, we also evaluated two data augmentation strategies. First, we incorporated citation instances from Teufel et al. (2006), using the class mapping proposed by Jurgens et al. (2018). Second, to increase the representation of the *Comparison or Contrast* class, we added citation instances from Cohan et al. (2019) labeled as *Comparing Results*. Training hyperparameters and complete evaluation results are detailed in Appendix A. On the Jurgens et al. (2018) dataset the best configuration used a 3–3 context window with BioBERT, reaching a macro-F1 score of 76.14%. Data augmentation further improved performance yielding a best macro-F1 score of 77.81% with PubMedBERT.

The resulting performance is comparable to the best transformer-based models reported in the literature, although it remains slightly below the graph-enhanced HGfN model of Quan et al. (2025).

To assess the transferability of the BERT-based classifiers, we evaluated them on the PD100Cit dataset. The best performances were obtained with BioBERT using a 3–3 context window, which achieved a macro-F1 score of 58.89% and a weighted-F1 score of 83.66%. Increasing

the context window consistently improved performance across all models. The performance gap between the source and target datasets suggests that the task remains challenging in the plant health domain. Performance was dominated by the majority *Background* class ($F_1 = 89.20\%$), whereas minority classes were classified less accurately. The *Uses* class achieved the second-highest F_1 score (66.20%), while the *Comparison/Contrast* class exhibited high recall (87.50%) but low precision (36.20%), indicating a tendency to overpredict this class. The *Motivation* class proved the most difficult to identify ($F_1 = 29.20\%$). The *Extends* and *Future* classes were absent from the PD100Cit test set and were therefore excluded from the class-level evaluation.

Prompting-based refinement of comparative and contrastive citations

The *Comparison/Contrast* class, as defined by Jurgens et al. (2018), encompasses citation events expressing both agreement and disagreement, as well as comparisons of methods and results. To better support the analysis of consensus and disagreement, we further refined the citation events assigned to this class using prompt-based classification with Qwen3-32B. This targeted approach enables finer-grained classification from only a few examples without requiring reannotation of the original training corpus.

Two prompt-based classification tasks were defined: the first distinguishes method- from result-based comparisons, and the second, applied only to result-based citations, distinguishes agreement (*Compare*) from disagreement, divergence or contradiction (*Contrast*). Each prompt included a task definition, a description of the expected input and output format, and two examples per class.

The first prompt was evaluated using the original Teufel et al. (2006) corpus. Since this distinction is not explicitly annotated, we derived a binary evaluation set from the original annotation by mapping *Contrast/Comparison in Goals or Methods* (CoCoGM) to *Method* and *Contrast/Comparison in Results* neutral (CoCoR0), and negative CoRes-, CoCo- to *Result*. This yielded an evaluation set of 104 method-based and 42 result-based citation events. The model achieved a macro-F1 score of 87.19% and a weighted-F1 score of 89.61%, as reported in Table 1. Performance was higher for the *Method*

class (F1 = 92.89%) than for the *Result* class (F1 = 81.48%).

The second prompt, which distinguishes result-based comparisons from contrasts, was evaluated using the Cohan et al. (2019) corpus. Citations labeled as *comparing results* were assigned to *Compare* or *Contrast* according to their secondary *supportive/not supportive* annotation. To increase the number of contrast examples, we additionally incorporated Teufel et al. (2006) citations labeled *CoRes-*, corresponding to cases in which the citing article reports lower results than the cited work. The resulting evaluation set comprised 289 *Compare* and 89 *Contrast* citation events.

Qwen3-32B achieved a macro-F1 score of 81.58% and a weighted-F1 score of 86.41% on this task, as shown in Table 2. The *Compare* class was identified more accurately (F1 = 90.71%) than the *Contrast* class (F1 = 72.45%), reflecting both the smaller number of contrast examples and the greater difficulty of identifying disagreement in citation contexts.

Table 1: Performance of Qwen3-32B on *Method/Result* classification.

Class	Precision	Recall	F1-score	Support
Method	0.916	0.942	0.929	104
Result	0.846	0.786	0.815	42
Macro average	0.881	0.864	0.872	146
Weighted average	0.896	0.897	0.896	146

Table 2: Qwen3-32B performance on *Compare/Contrast* classification.

Class	Precision	Recall	F1-score	Support
Compare	0.958	0.862	0.907	289
Contrast	0.664	0.798	0.725	89
Macro average	0.811	0.830	0.816	378
Weighted average	0.888	0.847	0.864	378

These results indicate that prompt-based classification provides an effective means of refining the *Comparison/Contrast* class into more informative rhetorical categories. The best configuration was applied to the citation contexts of the PLIN graph and the predicted rhetorical classes were used to populate the corresponding *Rhetorical Class* nodes.

Table 3: Corpus statistics.

Collection	Count
WoS primary references	8,181
Query Article Set (with DOI)	7,318
Citing references	94,042
Union	97,800
Full texts retrieved	62,878

5 Plant Literature Reliability Network

Together, these components define a generic and modular pipeline for constructing graphs conforming to the Rhetorical Citation Graph (RCG) model. Each component was selected based on reported or experimentally validated performance and can be replaced as improved methods become available.

5.1 PLIN Graph Description

Table 3 shows the corpus statistics for the article collection retrieved by 5 August 2025. The dataset exhibited an average of 31 citations per article, with a standard deviation of 48.59. The maximum number of citations for a single article was 1,208, while the minimum was 1. A summary of the counts per node type is presented in Table 4.

Table 4: Numbers of nodes and edges in the graph.

Graph element	Instances
Articles (primary and citing)	87,974
Original findings	18,854
Citation events (CEs)	262,807
Characterized CEs	169,925

Note: Characterized CEs are citation events linked to a rhetorical class, an in-context citation, a section, and a self-citation node.

Among all citation events, 169,925 were characterized with rhetorical class and contextual information. 92,882 citing-cited article pairs could not be characterized, either because the citing article was not successfully retrieved or could not be converted to a structured format due to missing OCR, accounting for 78.52% of characterization failures, or because the association between a cited DOI and its corresponding in-text citation failed, accounting for 21.48% of characterization failures. To establish this association, metadata for the cited article recorded in the graph (first-author name, publication year, and title) were compared with the bibliography entries in the citing article. When

this comparison produced no match, Crossref was queried using the complete bibliographic reference. The association was accepted only when the DOI returned by Crossref matched the DOI recorded in the graph.

The rhetorical citation function classification was applied on the PLIN citations using BioBERT with the 3–3 context window. The distribution after classification and refinement of the *Comparison or Contrast* class is reported in Table 5. *Contrast – contradictory results* denotes cases in which the results of the citing study explicitly contradict those reported in the cited study for the same study object. It is distinguished from *Contrast results*, which refers to broader differences or discrepancies between the results of the citing and cited studies that do not constitute a direct contradiction. For example, "*The transmission efficiency of *M. hiroglyphicus* (55%) is higher than that of *Y. flavovittatus* (45%) (CITSEG)*" is classified as *Contrast results* while "*Our observations contradict the conclusions of (CITSEG).*" is classified *Contrast – contradictory results*.

Table 5: Distribution of predicted rhetorical classes across citation events (CE)

Rhetorical Class	CE number
Background	124,401
Compare results	12,043
Usage	10,009
CoCoGM	8,865
Contrast results	8,651
Extends	2,623
Motivation	1,428
Contrast – contradictory results	968
Future	937
Total	169,925

6 Supporting EPPO Base Curation

The EPPO Global Database (EPPO, 2026) contains curated structured information about pathogen transmission to plants. Curating these interactions requires extensive examination of the scientific literature. We investigated whether the PLIN graph could facilitate this process by providing rapid access to the evidence supporting or questioning a reported interaction.

As a case study, we considered interactions that EPPO had annotated as uncertain, namely insect–pathogen relations labeled "Potential vector" and insect–plant relations labeled "Doubtful host", representing 129 interactions. These interactions correspond to cases where the

available literature was considered insufficient to establish the vector role with certainty. For these interactions, we identified 181 original articles in the PLIN graph whose abstracts mentioned both the source and target entities of the corresponding interaction. To keep the manual analysis tractable, we retained only articles associated with more than five citation events, yielding 134 articles with citation contexts. For each, we calculate the average distribution of rhetorical functions across all citation events in order to assess whether citation behaviour could provide evidence of scientific uncertainty or disagreement. As shown in Table 6, the debate-related rhetorical classes (Contrast – contradictory results, Contrast results, Compare results, and CoCoGM) account for only a small proportion of citation events on average (0.34%, 5.30%, 7.24%, and 5.28%, respectively), indicating that the EPPO uncertainty labels generally do not arise from conflicting claims in the literature.

Table 6: Mean percentage of citation events assigned to each rhetorical class across the 134 selected articles.

Rhetorical class	Mean (%)
Background	73.87
Motivation	0.68
Uses	4.97
Extends	1.45
Future	0.87
Contrast – contradictory results	0.34
CoCoGM	5.28
Compare results	7.24
Contrast results	5.30

Manual examination of the original articles showed that the EPPO annotations reflected a lack of conclusive evidence for disease transmission rather than conflicting claims in the literature. The low average prevalence of debate-related rhetorical functions was consistent with this interpretation, although citation behaviour varied substantially among articles.

This case study constitutes a proof of concept, illustrating the potential of the PLIN graph to support knowledge base curation by providing rapid access to normalized findings together with the citation contexts in which they are discussed. This structured representation enables to trace the evidence underlying reported interactions and to identify the specific biological relationships addressed in the literature. Furthermore,

debate-related rhetorical functions could provide a useful means of prioritising findings for manual review in cases where the literature contains explicit scientific disagreement.

7 Conclusion

In this work, we introduced the Rhetorical Citation Graph (RCG), a graph model that represents scholarly citations as fine-grained citation events linking citation contexts to individual scientific findings together with their rhetorical function. We instantiated this model by constructing PLIN, a graph of the plant health literature that integrates articles, normalized findings, citation contexts, citation events, and rhetorical classifications.

By moving beyond document-level citation networks, the proposed framework enables the analysis of how individual findings are cited, interpreted, and discussed across the scientific literature. Although illustrated in the plant health domain, the model is generic and can be applied to other scientific fields. The current PLIN graph corresponds to a partial instantiation of the RCG model centered on the Query Article Set. As additional collections are incorporated, previously distinct subgraphs can be merged through shared article nodes, progressively yielding a more comprehensive representation of the citation network. The proposed framework could also support the development of scientific fact-checking systems in experimental fields (Labbé et al., 2019), as well as scientific claim verification tasks (Wadden et al., 2020) by enabling the structured analysis of how findings are cited, interpreted, and evaluated through citation context and rhetorical functions.

Future work will focus on improving the automatic identification of findings and how they are framed as scientific claims within texts, for example using recent finding identification and generation methods (Liang et al., 2024). This would strengthen the alignment of findings across publications and further enhance applications such as scientific claim verification and expert knowledge curation. More broadly, the proposed framework provides a computational foundation for studying the epistemic dynamics of scientific knowledge. Beyond characterizing the rhetorical profile of individual findings, it opens the way to longitudinal analyses of how findings are interpreted, debated, corroborated, and

progressively established through scientific communication.

8 Limitations

PLIN graph was constructed from English-language documents. Extending PLIN to multilingual corpora would enable a more comprehensive and inclusive analysis of citation dynamics across scientific communities.

Our approach does not explicitly associate extracted relations with individual references within a citation context. Consequently, residual attribution ambiguities may arise when multiple cited works report semantically equivalent findings. Since the objective of the pipeline is to characterize citation behaviour around normalized findings rather than to reconstruct exact citation attributions, this ambiguity has only a limited impact on the analyses presented here. Finally, the rhetorical classification assigns a single label to each citation passage whereas prior work suggests that citation intent can be multi-faceted, with multiple rhetorical functions potentially co-occurring within the same citation context (Zhang et al., 2025; Lauscher et al., 2022). The RCE schema is extensible to multi-label annotation: a *Citation Event* could be linked to multiple *Rhetorical Class* nodes. Implementing this extension would require adapting the classification, and evaluation but would not require redesign of the graph structure.

Acknowledgments

We are grateful to the INRAE MIGALE bioinformatics facility (MIGALE, INRAE, 2020. Migale bioinformatics Facility, doi: 10.15454/1.5572390655343293E12) for providing help and/or computing and/or storage resources. This work was initiated through collaborations established within the BEYOND project (ANR grant no. 20-PCPA-0002). Anne-Sophie Foussat is currently undertaking a PhD funded by the INRAE Directorate General.

References

- GROBID. <https://github.com/grobidOrg/grobid>.
- Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural*

- Language Processing (EMNLP-IJCNLP 2019)*, pages 3615–3620. Association for Computational Linguistics.
- Dan Berrebbi, Nicolas Huynh, and Oana Balalau. 2022. [GraphCite: Citation Intent Classification in Scientific Publications via Graph Embeddings](#). In *Companion Proceedings of the Web Conference 2022 (WWW '22 Companion)*, page 779–783.
- Yu Bin, Liqing Jiang, Aobo Zhuang, Muzi Li, Junjun Li, Zhen Song, Changyong Zhou, Taiyun Wei, and You Li. 2026. [Semipersistent transmission of citrus yellow vein clearing virus by asian citrus psyllid: Insights from field surveys and infectious clone-based studies](#). *Phytopathology Research*, 8(1):32.
- Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. [Structural scaffolds for citation intent classification in scientific publications](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*, pages 3586–3596, Minneapolis, Minnesota. Association for Computational Linguistics.
- Arman Cohan and Nazli Goharian. 2017. [Contextualizing citations for scientific summarization using word embeddings and domain knowledge](#). In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2017)*, page 1133–1136. Association for Computing Machinery.
- EPPO. 2026. [EPPO Global Database](#).
- Revanth Gangi Reddy, Sai Chetan Chinthakindi, Zhenhailong Wang, Yi Fung, Kathryn Conger, Ahmed ELsayed, Martha Palmer, Preslav Nakov, Eduard Hovy, Kevin Small, and Heng Ji. 2022. [NewsClaims: A new benchmark for claim detection from news with attribute knowledge](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022)*, pages 6002–6018, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Mark Garzone and Robert E. Mercer. 2000. Towards an Automated Citation Classifier. In *Advances in Artificial Intelligence: Canadian AI 2000*, pages 337–346. Springer Berlin Heidelberg.
- Xiaorui Jiang and Jingqiang Chen. 2023. [Contextualised segment-wise citation function classification](#). *Scientometrics*, 128:5117–5158.
- David Jurgens, Srijan Kumar, Raine Hoover, Daniel A. McFarland, and Dan Jurafsky. 2018. [Measuring the Evolution of a Scientific Field through Citation Frames](#). *Transactions of the Association for Computational Linguistics*, 6:391–406.
- Priyanka Kargupta, Ishika Agarwal, Tal August, and Jiawei Han. 2025. [Tree-of-Debate: Multi-Persona Debate Trees Elicit Critical Thinking for Scientific Comparative Analysis](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025)*, pages 29378–29403. Association for Computational Linguistics.
- Paris Koloveas, Serafeim Chatzopoulos, Thanasis Vergoulis, and Christos Tryfonopoulos. 2025. [Can LLMs Predict Citation Intent? An Experimental Analysis of In-Context Learning and Fine-Tuning on Open LLMs](#). In *Linking Theory and Practice of Digital Libraries: 29th International Conference on Theory and Practice of Digital Libraries (TPDL 2025)*, page 207–224.
- Suchetha N. Kunnath, David Pride, and Petr Knuth. 2023. [Prompting Strategies for Citation Classification](#). In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23)*, pages 1127–1137. Association for Computing Machinery.
- Cyril Labbé, Natalie Grima, Thierry Gautier, Bertrand Favier, and Jennifer A. Byrne. 2019. [Semi-automated fact-checking of nucleotide sequence reagents in biomedical research publications: The Seek & Blastn tool](#). *PLoS ONE*.
- Anne Lauscher, Brandon Ko, Bailey Kuehl, Sophie Johnson, Arman Cohan, David Jurgens, and Kyle Lo. 2022. [MultiCite: Modeling realistic citations requires moving beyond the single-sentence single-label setting](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2022)*, pages 1875–1889. Association for Computational Linguistics.
- Yuan Liang, Massimo Poesio, and Roonak Rezvani. 2024. [A Fine-grained citation graph for biomedical academic papers: the finding-citation graph](#). In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing (BioNLP 2024)*, pages 416–426. Association for Computational Linguistics.
- Fan Luo, Marco A. Valenzuela-Escárcega, Gus Hahn-Powell, and Mihai Surdeanu. 2018. [Scientific Discovery as Link Prediction in Influence and Citation Graphs](#). In *Proceedings of the Twelfth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs@NAACL-HLT-2018)*, pages 1–6.
- Boaz Miller. 2013. [When is consensus knowledge based? Distinguishing shared knowledge from mere agreement](#). *Synthese*, 190:1293–1316.
- Michael J. Moravcsik and Poovanalingam Murugesan. 1975. [Some Results on the Function and Quality of Citations](#). *Social Studies of Science*, 5(1):86–92.
- Silas Boye Nissen, Tali Magidson, Kevin Gross, and Carl T. Bergstrom. 2016. [Research: Publication bias and the canonization of false facts](#). *eLife*, 5.

- Claire Nédellec, Marine Courtin, Xinzhi Yao, Marie Grosdidier, Isabelle Pieretti, Sandy Duperier, and Robert Bossy. 2026. [Epop: A benchmark corpus for assessing nlp models on structured information extraction in plant health](#). In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 1331–1340, Palma, Mallorca, Spain. European Language Resources Association (ELRA).
- Aniket Pramanick, Yufang Hou, Saif M. Mohammad, and Iryna Gurevych. 2026. [Claimflow: Tracing the evolution of scientific claims in nlp](#). *Preprint*, arXiv:2603.16073.
- Zhibang Quan, Jin Mao, and Gang Li. 2025. [Multi-intent prediction of scientific literature based on heterogeneous graph fusion network](#). *Scientometrics*.
- Uri Shwed and Peter S. Bearman. 2010. [The Temporal Structure of Scientific Consensus Formation](#). *American Sociological Review*, 75(6):817–840.
- Neşet Tan, Niket Tandon, Oyvind Taffjord, Michael Witbrock, Peter Clark, and Mark Gahegan. 2026. [Echoes of citations: Automated extraction of claims from full scientific papers](#). *Proceedings of the AAAI Symposium Series*, 8:568–576.
- Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. [Automatic Classification of Citation Function](#). In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP '06)*, pages 103–110. Association for Computational Linguistics.
- Hong-Jin Tsai, An-Zi Yen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2023. [Citation Intent Classification and Its Supporting Evidence Extraction for Citation Graph Construction](#). In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23)*, page 2472–2481. Association for Computing Machinery.
- Vijay Viswanathan, Graham Neubig, and Pengfei Liu. 2021. [CitationIE: Leveraging the citation graph for scientific information extraction](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*, pages 719–731. Association for Computational Linguistics.
- David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. [Fact or fiction: Verifying scientific claims](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, pages 7534–7550. Association for Computational Linguistics.
- Xiujuan Xu, Yueyue Xie, Xiaowei Zhao, and Yu Liu. 2025. [MF-Cite: Citation intent classification in scientific papers based on multi-feature fusion](#). *Scientometrics*, 130(8):4465–4493.
- Erjia Yan and Ying Ding. 2012. [Scholarly network similarities: How bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks relate to each other](#). *Journal of the American Society for Information Science and Technology*, 63(7):1313–1326.
- Yan-Hua Zhang, Chun-Hua Liu, Qian Wang, Yan-Ling Wang, Chang-Yong Zhou, and Yan Zhou. 2019. [Identification of Dialeurodes citri as a vector of Citrus yellow vein clearing virus in china](#). *Plant Disease*, 103(1):65–68.
- Yang Zhang, Yufei Wang, Quan Z. Sheng, Lina Yao, Haihua Chen, Kai Wang, Adnan Mahmood, Wei Emma Zhang, Munazza Zaib, Subhash Sagar, and Rongying Zhao. 2025. [Deep learning meets bibliometrics: A survey of citation function classification](#). *Journal of Informetrics*, 19(1):101608.
- Yang Zhang, Rongying Zhao, Yufei Wang, Haihua Chen, Adnan Mahmood, Munazza Zaib, Wei Emma Zhang, and Quan Z. Sheng. 2022. [Towards Employing Native Information in Citation Function Classification](#). *Scientometrics*, 127(11):6557–6577.

Appendix A. Rhetorical classification training and scores

Models were implemented in PyTorch 1.10.2+cu102 and trained using the following hyperparameters: batch size of 32, 20 epochs, AdamW optimiser, learning rate of $2e-5$, weight decay of 0.002, dropout of 0.5, and warmup ratio of 0.2. Experiments were repeated using four random seeds: 42, 5171, 798, and 1965. The dataset was split into 65% training, 15% validation, and 20% test data.

The average macro-F1 and weighted-F1 scores are reported in Table 7. Adding section information as a 32-dimensional embedding concatenated to the CITSEG embedding did not consistently improve performance. Larger section embeddings of 256 dimensions, and positional section embeddings were evaluated only on the ACL-ARC dataset, but they produced limited variation across models, as shown in Table 7, which reports experiments conducted using ACL-ARC data only.

The shortest context window, 1-1, yielded the lowest scores. Therefore, in the augmented-data experiments using additional examples from Teufel et al. (2006) and Cohan et al. (2019), we retained only the 3-3 and 2-2 context windows, with either no section embedding or a 32-dimensional section embedding. Results for the five evaluated models are reported in Table 8.

Table 7: Macro-F1 scores for six-class rhetorical citation function classification on the Jurgens dataset. Results are shown for different context-window sizes and section representations without additional training data. Bold values indicate the best score for each model, and the best mean score across models.

Context	Section representation	SciBERT	BioBERT	BioLinkBERT	PubMedBERT	RoBERTa	Mean
3-3	256	0.7597	0.7398	0.7293	0.7213	0.7458	0.7392
3-3	32	0.7424	0.7351	0.7453	0.7370	0.7178	0.7355
3-3	Position	0.7163	0.7614	0.7333	0.7466	0.7403	0.7396
3-3	None	0.7481	0.7501	0.7605	0.7411	0.7543	0.7508
2-2	256	0.7513	0.7505	0.7402	0.7431	0.7592	0.7488
2-2	32	0.7356	0.7442	0.7440	0.7304	0.7558	0.7420
2-2	Position	0.7374	0.7400	0.7379	0.7267	0.7373	0.7359
2-2	None	0.7532	0.7221	0.7509	0.7420	0.7394	0.7415
1-1	256	0.7519	0.7064	0.7179	0.7276	0.7543	0.7316
1-1	32	0.7369	0.7333	0.7314	0.7139	0.7320	0.7295
1-1	Position	0.7448	0.6991	0.7203	0.7557	0.7392	0.7318
1-1	None	0.7474	0.7237	0.7388	0.7411	0.7274	0.7357

Table 8: Macro- F_1 scores for six-class rhetorical citation function classification using additional Teufel and Cohan training data. Results are shown according to citation-context size, section representation, and language model.

Context	Section representation	Model	Macro-F1
3-3	None	SciBERT	0.7518
3-3	None	BioBERT	0.7474
3-3	None	BioLinkBERT	0.7365
3-3	None	PubMedBERT	0.7733
3-3	None	RoBERTa	0.747440
3-3	32	SciBERT	0.7471
3-3	32	BioBERT	0.7537
3-3	32	BioLinkBERT	0.7436
3-3	32	PubMedBERT	0.7782
3-3	32	RoBERTa	0.753724
2-2	None	SciBERT	0.7454
2-2	None	BioBERT	0.7529
2-2	None	BioLinkBERT	0.7356
2-2	None	PubMedBERT	0.7321
2-2	None	RoBERTa	0.7682
2-2	32	SciBERT	0.7544
2-2	32	BioBERT	0.7464
2-2	32	BioLinkBERT	0.7467
2-2	32	PubMedBERT	0.7419
2-2	32	RoBERTa	0.7709

Table 9: Average macro- F_1 and weighted- F_1 scores across seeds for each model and context window on PD100Cit.

Model	2-2			3-3		
	N	Macro	Weighted	N	Macro	Weighted
BioBERT	4	0.562	0.794	4	0.589	0.837
BioLinkBERT	4	0.519	0.812	4	0.560	0.823
PubMedBERT	4	0.492	0.735	4	0.512	0.786
RoBERTa	4	0.536	0.780	4	0.560	0.804
SciBERT	3	0.470	0.787	4	0.494	0.796

Per-class performance on ACL-ARC augmented are reported in Table 11. Scores on PD100cit are reported Tables 9 and 10.

Table 10: Average precision, recall, and F_1 scores across four seeds for each class using BioBERT with a 3-3 context window on PD100Cit.

Class	Support	P	R	F1
Background	83	0.940	0.849	0.892
Compare/contrast	6	0.362	0.875	0.509
Motivation	2	0.375	0.250	0.292
Uses	9	0.742	0.611	0.662

Table 11: Per-class precision (P), recall (R), and F_1 scores obtained using a 3–3 context window without section information on ACL-ARC after data augmentation. Results are reported as mean $\pm$ standard deviation across four seeds.

Class	BioBERT			BioLinkBERT			PubMedBERT			RoBERTa			SciBERT		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Background	0.833 $\pm$	0.898 $\pm$	0.864 $\pm$	0.843 $\pm$	0.884 $\pm$	0.863 $\pm$	0.856 $\pm$	0.905 $\pm$	0.879 $\pm$	0.847 $\pm$	0.881 $\pm$	0.863 $\pm$	0.847 $\pm$	0.895 $\pm$	0.870 $\pm$
	0.037	0.023	0.019	0.008	0.032	0.015	0.036	0.024	0.018	0.025	0.021	0.015	0.034	0.020	0.018
Compare/contrast	0.817 $\pm$	0.829 $\pm$	0.822 $\pm$	0.814 $\pm$	0.823 $\pm$	0.818 $\pm$	0.832 $\pm$	0.824 $\pm$	0.826 $\pm$	0.812 $\pm$	0.820 $\pm$	0.815 $\pm$	0.805 $\pm$	0.829 $\pm$	0.814 $\pm$
	0.034	0.031	0.005	0.043	0.025	0.023	0.064	0.037	0.016	0.058	0.049	0.035	0.071	0.054	0.027
Extends	0.644 $\pm$	0.532 $\pm$	0.579 $\pm$	0.540 $\pm$	0.590 $\pm$	0.557 $\pm$	0.618 $\pm$	0.586 $\pm$	0.593 $\pm$	0.547 $\pm$	0.625 $\pm$	0.581 $\pm$	0.630 $\pm$	0.554 $\pm$	0.580 $\pm$
	0.053	0.055	0.019	0.084	0.072	0.034	0.095	0.090	0.052	0.070	0.088	0.059	0.073	0.103	0.042
Future	0.796 $\pm$	0.575 $\pm$	0.660 $\pm$	0.729 $\pm$	0.536 $\pm$	0.599 $\pm$	0.834 $\pm$	0.725 $\pm$	0.774 $\pm$	0.765 $\pm$	0.612 $\pm$	0.676 $\pm$	0.719 $\pm$	0.712 $\pm$	0.708 $\pm$
	0.151	0.066	0.053	0.146	0.119	0.050	0.050	0.034	0.018	0.131	0.105	0.096	0.150	0.112	0.096
Motivation	0.859 $\pm$	0.694 $\pm$	0.756 $\pm$	0.854 $\pm$	0.680 $\pm$	0.754 $\pm$	0.786 $\pm$	0.683 $\pm$	0.727 $\pm$	0.820 $\pm$	0.654 $\pm$	0.724 $\pm$	0.821 $\pm$	0.652 $\pm$	0.722 $\pm$
	0.076	0.133	0.061	0.065	0.075	0.043	0.104	0.036	0.041	0.089	0.091	0.075	0.078	0.078	0.038
Uses	0.840 $\pm$	0.823 $\pm$	0.829 $\pm$	0.844 $\pm$	0.815 $\pm$	0.828 $\pm$	0.849 $\pm$	0.833 $\pm$	0.840 $\pm$	0.839 $\pm$	0.814 $\pm$	0.826 $\pm$	0.845 $\pm$	0.813 $\pm$	0.827 $\pm$
	0.051	0.041	0.013	0.032	0.020	0.005	0.025	0.028	0.008	0.017	0.016	0.011	0.039	0.052	0.023

Appendix B. Citations per article

Figure 2 displays the distribution of citations per article in the Plant Health Query Article Set.

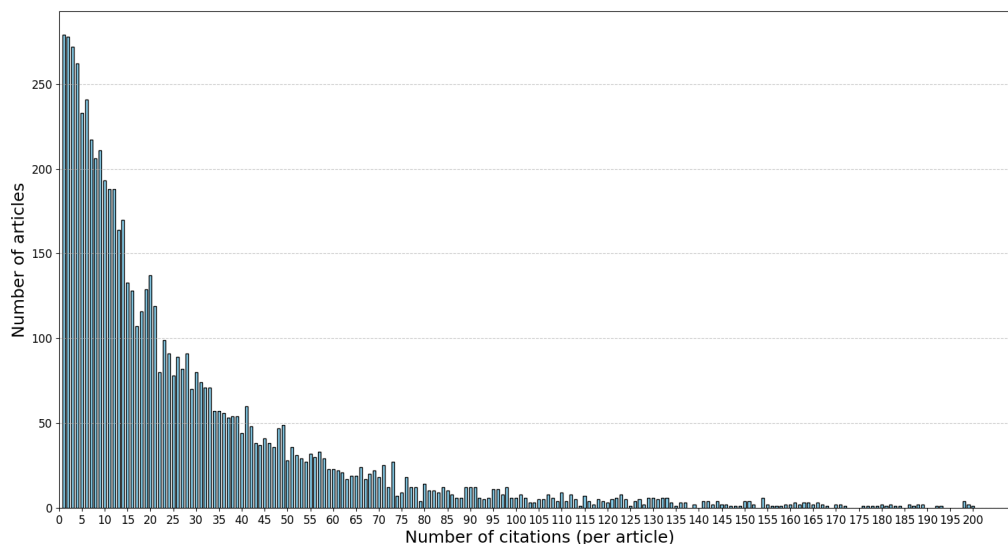


Figure 2: Distribution of citations per article in the Query Article Set (up to 200 citations per article).

C Prompting Qwen3-32B for six-class citation-function classification

Tables 12 and 13 report the performance of Qwen3-32B with thinking disabled and enabled, respectively.

Table 12: Per-class performance of Qwen3-32B on the PD100Cit dataset with thinking disabled.

Class	Precision	Recall	F1-score	Support
Background	0.9000	0.6506	0.7552	83
Motivation	0.2500	1.0000	0.4000	2
Uses	0.4167	0.5556	0.4762	9
Compare or contrast	0.1000	0.3333	0.1538	6
Macro average	0.4167	0.6349	0.4463	100
Weighted average	0.7955	0.6300	0.6869	100

Table 13: Per-class performance of Qwen3-32B on the PD100Cit dataset with thinking enabled.

Class	Precision	Recall	F1-score	Support
Background	0.9355	0.6988	0.8000	83
Motivation	0.2222	1.0000	0.3636	2
Uses	0.4615	0.6667	0.5455	9
Compare or contrast	0.2143	0.5000	0.3000	6
Macro average	0.4584	0.7164	0.5023	100
Weighted average	0.8353	0.6900	0.7384	100