UniCite: A Dataset and Unified Hierarchical Taxonomy for Multi-Dimensional Citation Analysis

Amina Mourky, Elena Leitner, Julian Moreno Schneider,
Raia Abu Ahmad, Ekaterina Borisova, Georg Rehm
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI)
Salzufer 15/16, 10587 Berlin, Germany
{firstname.lastname}@dfki.de

Abstract

Research in Citation Context Analysis (CCA) has produced numerous taxonomic schemes that vary from three to 12+ categories, with different granularities and no mappings between frameworks, severely limiting systematic comparison and progress. Despite decades of study, CCA methods have largely relied on fragmented frameworks that treat citation tasks independently, ignoring systematic relationships between function classification, sentiment analysis, and importance assessment. To address these research gaps, we present three integrated contributions. First, we develop UniCite, a two-level taxonomy (six primary functions, 12 subcategories, two orthogonal dimensions) that systematically integrates three existing schemes. Second, we develop a comprehensive dataset of 4,017 citations combining established resources with 1,547 newly extracted citations from 2018-2024 publications, all manually annotated under our unified framework. Third, we demonstrate systematic task relationships through multi-task learning, achieving 21.1% relative improvement in subfunction classification over single-task approaches.

Keywords: Citation Context Analysis, Scholarly Document Processing, Citation Classification

1. Introduction

Citations encode complex rhetorical relationships that extend far beyond simple bibliographic acknowledgment. These relationships reflect how scholar communities build knowledge, evaluate, and position new work within existing discourse (Garfield et al., 1964; Swales, 1986). Understanding citation functions has implications for scientific evaluation, literature analysis, and computational approaches to scholarly communication.

Citation Context Analysis (CCA) studies how and why authors cite previous work. This can involve different aspects including the classification of citation functions, sentiment toward cited work, and evaluation of citation importance to the citing paper's contributions. As scholarly publications accelerate and evolve, accurate citation analysis becomes increasingly important for bibliometric evaluation and understanding knowledge flow patterns.

However, CCA research faces two fundamental challenges that limit systematic progress. First, different groups have developed taxonomic schemes that vary in granularity and focus (e. g., Jurgens et al., 2018; Cohan et al., 2019). This variation in category boundaries, terminology, and annotation guidelines prevents cumulative research and systematic comparison across studies. Second, many established datasets focus on pre-2019 publications, failing to capture contemporary scholarly practices (Jurgens et al., 2018; Cohan et al., 2019; Lauscher et al., 2022).

We address these challenges through UniCite, a unified approach that integrates existing taxonomies while extending temporal coverage. Our hierarchical framework preserves semantic distinctions from existing schemes by integrating the argumentative categories from Teufel et al. (2006), the citation frames from Jurgens et al. (2018), and the cross-domain applicability from Cohan et al. (2019), while enabling systematic mapping and comparison as shown in Figure 1. The taxonomy operates on two-levels: Level 1 captures six broad functions (Background, Methodology, Extension, Comparison, Motivation, Future Work), while Level 2 provides twelve targeted categories that distinguish important cases within major functions, such as separating General Background from Previous Work citations, or differentiating Direct Use from Tool Usage in methodology citations. We supplement this with the comprehensive manual annotation of contemporary citations and validate our design through multi-task learning experiments that reveal systematic relationships between citation analysis tasks. Our data and code are publicly available at `https://github.com/aminamourky/UniCite.git`.

This paper makes three primary contributions:

1. **Unified Hierarchical Taxonomy:** We develop UniCite, a framework that systematically integrates existing classification schemes while preserving their semantic distinctions through a two-level hierarchy with orthogonal sentiment and importance dimensions.

2. **Comprehensive Annotated Dataset:** We develop a dataset of 4,017 citations, annotated by experts, that addresses temporal gaps by

Proceedings of the 3rd Int. Workshop on Natural Scientific Language Processing (NSLP 2026) @ LREC 2026, pages 277–288
12 May 2026.

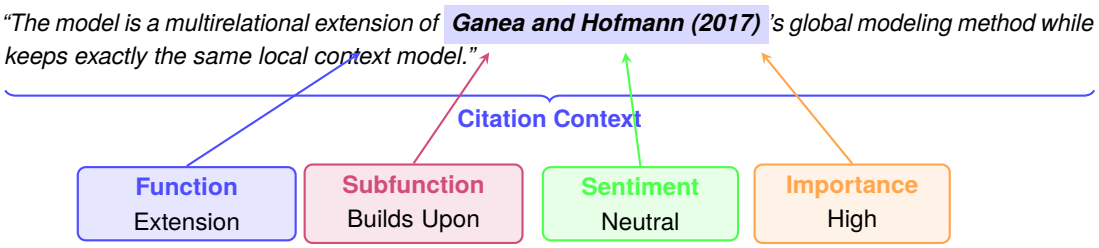

Figure 1: Example of an annotated citation from the UniCite dataset.

integrating established resources with 1,547 recent citations from 2018-2024 publications.

3. **Multi-Task Learning Validation:** We explore systematic relationships between citation analysis tasks through multi-task learning experiments, achieving substantial performance improvements that validate our taxonomic design and reveal beneficial task interactions.

2. Related Work

The computational analysis of citations has evolved along several research threads, each attempting to capture different aspects of how scholars use references in their writing. The foundations of CCA can be traced back to the early bibliometric work of Garfield et al. (1964), which established that citations fulfill distinct rhetorical roles, followed by early structured approaches to citation categorization (Moravcsik and Murugesan, 1979).

Early work established that citations could be automatically classified based on their rhetorical purpose. Teufel et al. (2006) developed a twelve-category system that provided detailed argumentative distinctions, proving that computational models could learn citation intent from textual features. Different groups subsequently developed frameworks reflecting their particular priorities. Jurgens et al. (2018) created a six-category framework targeting main citation functions, demonstrating how citation functions vary across paper sections. Cohan et al. (2019) introduced the SciCite dataset with a simplified three-category scheme arguing for a broader applicability across different scientific domains. Pride and Knoth (2020) developed the Argumentative Citation Types (ACT) framework by presenting a multi-dimensional approach that separates citation function from citation importance, establishing an orthogonal dimension that captures how central a citation is to the work's contributions.

The diversity of these approaches shows the genuine disagreements about how citation behavior should be modeled. Some researchers have argued for even more complex representations. Jochim and Schütze (2012) developed four orthogonal dimensions to capture complex citation relationships that one-dimensional frameworks fail to represent. Hernández-Álvarez et al. (2017) created the CONCIT scheme, which provided a rich annotation of citation contexts including detailed markup of citation aspects and evaluative indicators, emphasising the critical importance of fine-grained contextual information in citation analysis applications. Lauscher et al. (2022) challenged the single-label assumption, allowing citations to be assigned multiple functions simultaneously and thus revealing that compound citation purposes are common in scholarly writing.

These parallel developments have created a field where different communities work with diverse data and evaluation frameworks that restrict direct comparison. The lack of standardization makes it difficult to assess which approaches actually work better for real applications. The OpenCitations Data Model (Daquino et al., 2020) provides an ontological framework for representing bibliographic entities and citation relationships, enabling structured representation that complements taxonomic approaches. Additional challenges arise from metadata limitations in some existing datasets. For example, MultiCite (Lauscher et al., 2022) lacks paper titles. While it provides paper IDs, these are no longer traceable to the original publications, limiting researchers' ability to access broader publication context when needed. Such gaps can constrain analysis capabilities and cross-domain studies.

An additional limitation involved the artificial separation of related analytical tasks. Function, classification, sentiment analysis, and importance assessment are typically studied separately, even though they all require nuanced contextual interpretation of citation text. Research in related areas has shown that jointly modeling such tasks can improve performance through shared representations (Ruder, 2017). Recent work in citation analysis has begun exploring these approaches: Su et al. (2019) demonstrate performance improvements through neural multi-task learning for citation function and provenance, while Baig et al. (2021) achieve competitive results using multi-module architectures for citation context classification. Recent work also

emphasizes the importance of comprehensive context extraction, with Jantsch et al. (2025) showing that fine-grained citation contexts capturing semantic dimensions beyond single sentences improve classification performance by up to 25%. However, citation analysis has yet to systematically explore whether function, sentiment and importance annotations benefit from shared modeling or exhibit overlapping representational structure across comprehensive taxonomic frameworks.

Recent work has explored complementary approaches to citation classification. Jiang (2025) demonstrates that ensemble methods combining diverse base classifiers substantially improve performance, addressing the observation that no single model excels universally. Similarly, Duan and Tan (2026) propose a semantically orthogonal framework separating citation intent from content type, achieving superior cross-domain generalization. While LLM-based approaches have emerged (Koloveas et al., 2025), Kunnath et al. (2023) systematically evaluate prompting strategies for citation classification, finding that dynamic context-based prompting outperforms standard fine-tuning, while zero-shot GPT-3.5 performs well on multidisciplinary data but poorly on domain-specific datasets. Fogelson et al. (2025) further identify significant reproducibility challenges with LLM-based approaches, validating fine-tuned domain-specific models for methodological transparency. Beyond citation classification, Bolaños et al. (2025) propose a complementary annotation schema for classifying rhetorical roles in literature review sections, demonstrating that fine-tuned LLMs achieve strong performance on scholarly sentence classification tasks.

3. Multidimensional Taxonomy

To enable systematic comparisons in CCA, we develop a unified taxonomy (Figure 2) that integrates existing approaches[1] such as SciCite (Cohan et al., 2019), MultiCite (Lauscher et al., 2022), and Jurgens-CFC (Jurgens et al., 2018) into a hierarchical framework considering citation function, sentiment, and importance.

The citation function consists of two levels of granularity – six primary categories in Level 1 and twelve subcategories in Level 2.

Level 1 Functions cover the main ways researchers use citations. *Background* establishes context or foundations. *Methodology* involves applying cited methods, tools, or datasets. *Extension* shows how the current work builds upon or modifies the cited work. *Comparison* evaluates cited work against current research or other studies. *Motivation* justifies the need for the current research by pointing to gaps in existing work. *Future Work* discusses potential future research directions.

Level 2 Subcategories distinguish important cases within major functions: **Background** separates domain knowledge (*General Background*) from specific research findings (*Previous Work*). **Methodology** differentiates between applying methods (*Direct Use*), using software tools (*Tool Usage*), or utilising datasets (*Dataset Usage*). **Extension** separates conceptual foundations (*Builds Upon*) from adapted applications (*Adaptation*). **Comparison** distinguishes conceptual evaluation (*Conceptual Comparison*), performance assessment (*Performance Comparison*), and negative assessment (*Critique*). **Motivation** and **Future Work** remain without subcategories, as they represent relatively focused functions. We adopt a single-label design to maintain annotation feasibility and enable systematic comparison with existing benchmarks, with multi-label extension left for future work.

Orthogonal properties capture evaluative and structural aspects independent of function. **Sentiment** reflects the author's stance: *Positive* (favorable), *Neutral* (objective), or *Negative* (critical). **Importance** assesses centrality of the cited work to the current paper: *High* (essential to the paper's contribution), *Low* (peripheral), *Cannot Determine* (unclear from context).

4. Dataset

We develop a comprehensive dataset through a two-stage process: (1) systematic conversion of all existing citations to our unified Level 1 taxonomy, and (2) manual annotation of a representative subset for complete multi-dimensional labeling.

4.1. Source Data Collection and Mapping

Three established datasets form our foundation, selected for their complementary coverage and annotation quality. **SciCite** (Cohan et al., 2019) provides 11,020 citation contexts spanning computer science and biomedical domains with three-category function labels. **MultiCite** (Lauscher et al., 2022) covers 12,653 citation contexts from computational linguistics with seven function categories and detailed context modeling. For our dataset, we extracted only the single-label citations to maintain consistency with our single-label classification approach. **Jurgens-CFC** (Jurgens et al., 2018) offers 1,969 expertly annotated citations with six function categories that established a benchmark for citation function classification.

We supplement these resources with 1,547 new citations from publications spanning

[1] For more details on the compatibility with existing systems, see Appendix A, Section A.1.

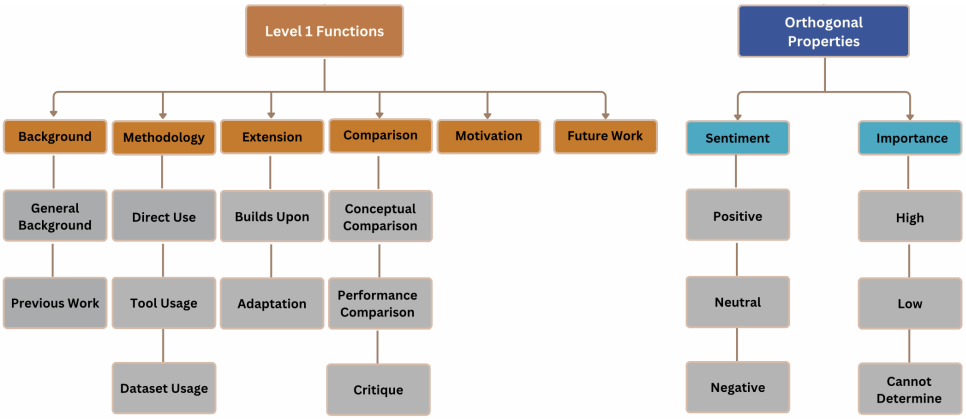

Figure 2: UniCite taxonomy with its hierarchical organization of citation functions and orthogonal properties.

2018-2024, a period marked by the rise of preprint culture, accelerated publication cycles, and growing interdisciplinary collaboration, which may have introduced new citation patterns not captured in earlier datasets. Computer science venues (ACL/EMNLP/NAACL, NeurIPS/ICML/ICLR, CVPR/ICCV/ECCV, ICSE/FSE/ASE) provided 1,200 citations. Interdisciplinary venues contributed 347 citations from robotics (RSS/ICRA/IROS) and human-computer interaction (FAccT/CHI). Paper selection was stratified by impact metrics and temporal distribution. We used PyMuPDF[2] with citation recognition to process PDF files. We extracted citation context by keeping 2–3 sentences surrounding citation markers with validation for author name, publication year, and format consistency, following established practices showing that multi-sentence contexts significantly improve classification performance over single sentences (Jantsch et al., 2025; Lauscher et al., 2022).

Our approach first converts all 27,189 citations from source datasets to our Level 1 taxonomy through systematic mapping rules with full automation. Direct mapping rules handle 89% of conversions, with the remaining 11% requiring merging operations where multiple source categories map to single target categories. However, complete multi-dimensional annotation (including Level 2 subcategories, sentiment, and importance) requires manual annotation that is expensive for the full dataset. We therefore selected a representative 4,017-citation subset for complete manual annotation, balanced across source datasets and supplemented with 1,547 newly extracted recent citations. This sampling strategy ensures coverage of existing taxonomic diversity while maintaining annotation feasibility and adding crucial temporal coverage.

[2]https://pymupdf.readthedocs.io/en/latest/

4.2. Annotation Process and Agreement

We developed comprehensive annotation guidelines[3] through an iterative process involving a pilot annotation phase and edge case analysis. Two graduate students with backgrounds in computational linguistics, both with prior experience in citation analysis, performed the annotation using INCEpTION (Klie et al., 2018). For the 2,470 citations converted from existing datasets, Level 1 labels were obtained through automated mapping, while Level 2 subcategories and orthogonal properties were annotated manually. For the 1,547 newly extracted citations, annotators provided complete annotations across all taxonomic dimensions. The workflow proceeded hierarchically: the annotators first assigned the Level 1 function, then selected the Level 2 subfunction, followed by sentiment and importance. For importance assessment, annotators consulted the full paper when available, which was the case for all but 31 citations (0.8% of the dataset).

We followed established protocols for quality assurance from CCA literature (Teufel et al., 2006; Cohan et al., 2019; Jurgens et al., 2018). The two annotators underwent 6-hour guideline training, 2-hour pilot annotation, and 3-hour calibration sessions addressing edge cases. We monitored the annotation quality weekly throughout the 8-week annotation period. Inter-annotator agreement assessment covered 1,110 citations (28% of total) with independent dual annotation. Level 1 function classification achieved Cohen's κ = 0.88 (Artstein and Poesio, 2008), demonstrating excellent agreement on primary categories. Level 2 subcategory classification achieved κ = 0.63, reflecting the inherent difficulty of fine-grained distinctions while remaining within acceptable ranges for complex annotation tasks. Sentiment classification achieved κ

[3]The annotation guidelines are available at https://github.com/aminamourky/UniCite/tree/main/annotation.

Dataset	No.	F L1	F L2	Sen	Imp
New citations	1,547	1,547	1,547	1,547	1,547
Jurgens-CFC	892	892	892	892	892
SciCite	828	828	828	828	828
MultiCite	750	750	750	750	750
Total	**4,017**	**4,017**	**4,017**	**4,017**	**4,017**
Manual	–	**1,547**	**4,017**	**4,017**	**4,017**

Table 1: Annotation coverage in UniCite across source datasets and taxonomic levels. Manually annotated labels are in blue.

= 0.92, and importance assessment achieved κ = 0.66. Our agreement scores are strong compared to prior work in citation analysis, considering Cohan et al. (2019) reported 86% agreement on 100 samples, while Lauscher et al. (2022) achieved 0.76 accuracy for function classification. We resolved all disagreements through careful examination of citation contexts and consistent application of our annotation guidelines.

4.3. Dataset Statistics and Composition

Table 1 presents the annotation coverage across source datasets and taxonomic levels. The UniCite dataset balances established annotation quality with contemporary coverage: new paper extractions are the largest component with 1,547 instances (38.5%), followed by Jurgens-CFC (892 instances, 22.2%), SciCite (828 instances, 20.6%), and MultiCite (750 instances, 18.7%).

Temporal distribution addresses existing dataset limitations: 1,750 citations (43.6%) are from 2018-2024 publications, with peak years including 2019 (367 citations), 2023 (313 citations), and 2021 (300 citations), while 2,267 citations (56.4%) maintain historical representation for comparative analysis.

Function distribution reflects realistic scholarly practices (see Figure 3): Background dominates with 1,600 instances (39.8%), followed by Methodology (1,028 instances, 25.6%) and Comparison (790 instances, 19.7%). Subcategory distribution shows Previous Work constituting 1,230 instances (30.6%) of background citations, while General Background comprises 370 instances (9.2%). Methodology subdivides into Direct Use (640 instances), Tool Usage (208 instances), and Dataset Usage (180 instances). Comparison categories include Conceptual Comparison (512 instances), Performance Comparison (230 instances), and Critique (48 instances). Orthogonal properties reflect academic discourse characteristics: Sentiment remains predominantly Neutral (3,753 instances, 93.4%), with Positive (164 instances, 4.1%) and Negative (100 instances, 2.5%) representing evaluative minorities. Importance assessment shows High (2,088 instances, 52.0%) slightly exceeding Low (1,712 instances, 42.6%), with Cannot Determine cases comprising 217 instances (5.4%).

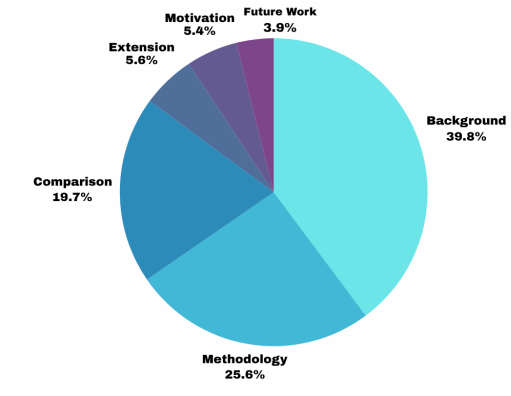

Figure 3: Distribution of function citations.

Domain representation spans Natural Language Processing (1,439 citations, 36.3%), Machine Learning (485 citations, 12.2%), Computer Science and Medicine (476 citations, 12.0%), Computer Vision (405 citations, 10.2%), Software Engineering (403 citations, 10.2%), Robotics (394 citations, 9.9%), and Human-Computer Interaction (360 citations, 9.1%), enabling both domain-specific and cross-domain evaluation.

5. Experiments

We evaluate the effectiveness of the framework through systematic experiments across multiple modeling paradigms. The experimental design addresses three primary research questions: (1) How do different citation analysis tasks interact in multi-task learning scenarios? (2) What benefits does hierarchical conditioning provide for subfunction classification? (3) How do multi-task approaches compare to specialized single-task models?

We use stratified sampling to maintain consistent class distributions across data partitions: Training Set (2,814 citations, 70%), Validation Set (602 citations, 15%), and Test Set (601 citations, 15%). Model performance is measured using accuracy and macro F_1, which addresses class imbalance by providing equal weight to all taxonomic categories. We supplement these with a hierarchical consistency metric that measures the agreement between function and subfunction predictions with respect to our taxonomic structure (Sun and Lim, 2001; Plaud et al., 2024). Further details can be found in Appendix A, Section A.2.

5.1. Reference Models

Large Language Models (LLMs): We evaluate GPT-3.5 Turbo[4] and GPT-4o mini[5] in few-shot configurations to assess zero-training performance on citation analysis tasks. These models were selected to represent two capability tiers of widely accessible commercial LLMs, enabling assessment of how model scale affects few-shot citation classification. GPT-3.5 evaluation uses 1,105 citations with taxonomy definitions plus 2 examples. GPT-4o mini evaluation uses 1,093 citations with expanded prompts containing 12 examples covering diverse label combinations. The difference in citation counts (12 less for GPT-4o mini) results from excluding the example citations used in the expanded prompts to prevent data leakage. Each prompt contains taxonomic guidelines, representative examples, and explicit formatting requirements. The full prompts used for both LLM evaluations are available in the project repository.

Single-Task Fine-Tuned Models: We train individual SciBERT models (Beltagy et al., 2019) for each task using task-specific classification heads. Function and subfunction models use weighted cross-entropy loss with a learning rate of 2e-5. Sentiment classification employs focal loss (Lin et al., 2017) (γ=2.0) for the 93% neutral class imbalance. Importance assessment uses weighted cross-entropy after filtering "Cannot Determine" labels.

5.2. Multi-Task Learning Approaches

Our multi-task learning (MTL) approach builds on established work in CCA. Su et al. (2019) demonstrate that neural architectures can effectively share representations between citation function and provenance tasks, while Baig et al. (2021) show that multi-module designs combining diverse feature types achieve competitive performance. We extend these foundations by investigating systematic relationships across function, sentiment, and importance dimensions.

Hierarchical MTL: Our hierarchical approach follows the hard parameter sharing paradigm established by Ruder (2017), where tasks share hidden layers while maintaining task-specific output layers. The design leverages principles from hierarchical classification (Silla and Freitas, 2011), where taxonomic structure provides natural learning progression from general to specific categories. Task selection follows insights from Standley et al. (2019), who showed that beneficial task relationships depend on semantic similarity rather than superficial domain overlap. The shared SciBERT encoder (Beltagy et al., 2019) generates common representations while task-specific heads process specialized features. A learned embedding of the predicted function category (192-dimensional, i.e., hidden_dim/4) is concatenated with the encoder features (768-dimensional), forming a 960-dimensional combined representation before subfunction classification. Training proceeds through three phases: individual task optimization, hierarchical conditioning introduction, and end-to-end joint learning across 8 epochs.

Progressive MTL: We train tasks sequentially to analyze interaction patterns. Phase 1 trains function and subfunction jointly (10 epochs). Phase 2 adds importance classification (6 epochs). Phase 3 incorporates sentiment analysis (6 epochs). Phase 4 performs final joint optimization (4 epochs). Each phase trains an independent model from scratch with its respective task combination, enabling direct comparison of how different task sets affect performance without confounding transfer effects.

All experiments run across three random seeds (42, 123, 456) with identical data splits. Epoch counts were determined empirically with early stopping (patience of 3–4 epochs) to prevent overfitting. We report means and standard deviations for all metrics, with 95% confidence intervals and t-tests for statistical significance assessment.

6. Results

The results of the few-shot approach, fine-tuning, and MTL on Function (F L1), Subfunction (F L2), Sentiment (Sen) and Importance (Imp) are shown in Table 2. The results of the fine-tuned models represent the mean $\pm$ standard deviation in three random seeds.

Approach	F L1	F L2	Sen	Imp
LLM Few-Shot				
GPT-3.5 Turbo	29.8	24.7	**47.4**	29.8
GPT-4o mini	50.8	39.3	46.3	37.6
Fine-Tuned Models				
SciBERT	68.7±1.7	53.9±0.7	40.9±1.4	71.3±1.2
Multi-Task Learning				
Hierarchical MTL	**72.5**±0.8	**65.3**±0.6	–	–
Progressive MTL	70.1±0.9	60.4±3.2	32.5±0.0	**74.2**±0.9

Table 2: Classification results across all approaches. Best values per task shown in bold.

LLM Performance. The LLM experiments establish task difficulty hierarchies while revealing limitations of general-purpose models for specialized citation analysis. GPT-3.5 Turbo with two examples achieves moderate performance across tasks. GPT-4o mini with 12 examples shows substantial

[4] https://platform.openai.com/docs/models/gpt-3.5-turbo

[5] https://platform.openai.com/docs/models/gpt-4o-mini

Ph	Tasks	F L1	F L2	Imp	Sen
1	Fun+Sub	71.2±0.3	56.7±3.4	N/A	N/A
2	Fun+Sub+Imp	70.1±1.0	58.3±1.6	74.2±0.8	N/A
3	Fun+Sub+Imp+Sen	70.1±0.9	60.4±3.2	74.2±0.9	32.5±0.0

Table 3: Progressive MTL results.

improvements on Function (+21%), Subfunction (+14.6%), and Importance (+7.8%), with a minimal decrease in Sentiment performance (-1.1%) compared to GPT-3.5 Turbo.

Single-Task Performance. Function classification achieves 68.7% ± 1.7% macro F_1. Subfunction classification shows increased difficulty at 53.9% ± 0.7% macro F_1. Sentiment classification exhibits high accuracy (93.0% ± 0.4%) but low macro F_1 (40.9% ± 1.4%) due to class imbalance. Importance assessment demonstrates stable performance (71.3% ± 1.2% macro F_1) with the lowest coefficient of variation (0.018).

Hierarchical MTL. Hierarchical conditioning between function and subfunction tasks produces significant improvements. Subfunction classification gains 11.4 percentage points in macro F_1 (53.9% → 65.3%, $p < 0.01$), representing a 21.1% relative improvement. This improvement substantially exceeds the 2% gains reported by Su et al. (2019) for function-provenance MTL, demonstrating the particular effectiveness of hierarchical conditioning for taxonomically structured tasks. Function performance also shows some improvement with an increase from 68.7% to 72.5% ($p = 0.013$). The conditioning architecture achieves 83.3% ± 0.5% hierarchical consistency between function and subfunction predictions (95% CI: [0.821, 0.845]).

Progressive MTL. Progressive training enables systematic analysis of task interaction patterns (see Table 3). Phase 1 (Function + Subfunction) establishes stable baseline joint performance with 0.712 $\pm$ 0.003 function F_1 and 0.567 $\pm$ 0.034 subfunction F_1. Phase 2 importance addition produces modest function degradation (-0.011 $\pm$ 0.010 F_1, $p > 0.05$) while providing modest subfunction improvement (+0.016 $\pm$ 0.016 F_1), with importance achieving robust performance (0.742 $\pm$ 0.008 F_1). Phase 3 sentiment integration shows negligible additional function impact (+0.001 $\pm$ 0.001 F_1) while providing modest subfunction improvements (+0.021 $\pm$ 0.022 F_1). Sentiment classification achieves consistent but poor performance (0.325 $\pm$ 0.000 F_1) across all seeds due to severe class imbalance.

To examine our architectural choices, we conducted ablation studies (see Appendix A, Section A.3, for more details).

7. Analysis and Discussion

The learning experiments provide evidence about the semantic relationships between citation analysis tasks. The hierarchical relationship between function and subfunction classification demonstrates significant synergistic effects, with joint learning yielding a substantial 11.4% improvement (21.1% relative improvement) for subfunction classification when conditioned on function predictions. This enhancement stems from the elimination of hierarchically inconsistent predictions, reducing the effective classification space and providing contextual guidance. The hierarchical consistency score of 83.3% indicates that the model successfully learns meaningful structural relationships between functional categories and their constituent subfunctions. This finding validates our design principles, demonstrating that the semantic hierarchy reflects genuine linguistic patterns in scholarly discourse rather than arbitrary categorization boundaries.

7.1. Task Interaction Patterns

The results show a clear distinction between tasks that benefit from joint learning and those that exhibit independence. Function predictions provide conditioning information that guides subfunction classification, while sentiment and importance exhibit orthogonal relationships to the hierarchical tasks, showing minimal beneficial interaction and, in some configurations, interference effects. This aligns with established work (Ruder, 2017), which emphasizes that successful joint learning depends on identifying tasks with genuinely beneficial semantic relationships. The conditioning mechanism reveals that function predictions serve as powerful contextual priors for subfunction classification, with implications for any hierarchically structured classification scheme in citation analysis.

The conditioning mechanism employed in hierarchical MTL reveals the significant impact of semantic structure on model performance. Function predictions serve as powerful contextual priors for subfunction classification, eliminating impossible category combinations and focusing the model's attention on semantically coherent options. This insight has implications beyond our specific taxonomy, suggesting that any hierarchically structured classification scheme in citation analysis could benefit from similar conditioning approaches.

7.2. Error Analysis and Boundary Ambiguities

Function Classification Confusion Patterns. The systematic analysis of function classification errors reveals consistent patterns of semantic ambiguity at taxonomic boundaries. Figure 4 presents

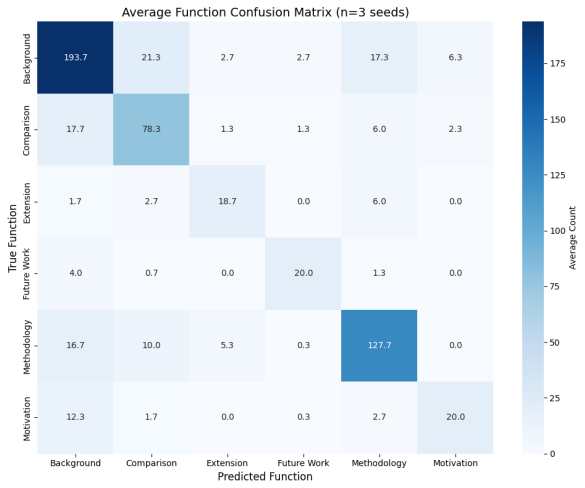

Figure 4: Average function confusion matrix across 3 random seeds.

the comprehensive confusion matrix for hierarchical MTL averaged across three random seeds, illustrating the distribution of classification errors across functional categories. The most prevalent confusion pairs demonstrate the inherent challenges in distinguishing between contextual establishment and comparative analysis in scholarly discourse. Background → Comparison confusion accounts for 21.3 ± 2.1 errors, while Comparison → Background confusion represents 17.7 ± 0.6 errors. Background → Methodology confusion shows 17.3 ± 1.2 errors. The confusion matrix visualization reveals the concentrated nature of these errors, with the majority of misclassifications occurring between semantically similar categories rather than exhibiting random distribution patterns. Citations serving dual purposes, such as "*Previous studies have shown...*" (labeled Background, predicted Comparison), illustrate the difficulty of imposing single-label constraints on multi-functional citations.

The error analysis reveals systematic bias toward majority classes: Background citations (39.8% of the dataset) show the strongest prediction bias, with 33.2% of true Motivation instances incorrectly predicted as Background.

Error Propagation in Hierarchical Systems. The most significant finding concerns error propagation in the hierarchical system. Function errors occur in 24.0% of instances, limiting subfunction performance: the conditioning mechanism leads subfunction predictions to categories within the predicted function, so function misclassification eliminates the possibility of selecting subfunctions from the correct function category. The subfunction confusion matrix shows 23.1% overall error rate across the hierarchical system. Given that function errors (24.0%) automatically propagate to subfunction errors, the conditioning architecture creates a bottleneck where upstream accuracy directly bounds downstream performance. When function predictions are correct, the model demonstrates strong subfunction classification capability, but this performance depends critically on accurate function-level decisions.

7.3. Class Imbalance Impact Assessment

Class imbalance effects demonstrate varying severity across different classification tasks, with the most dramatic impact observed in sentiment classification. The severe neutral class dominance (93.4% of instances) overwhelms the learning signal for positive and negative sentiment, resulting in systematic failure to learn minority class patterns. This imbalance creates a prediction distribution heavily skewed toward neutral classifications (95.5% of predictions), effectively rendering the sentiment classification task ineffective.

Sentiment Classification Failure: Sentiment classification demonstrates systematic failure across all configurations, achieving only 40.9% macro F_1 despite 93.0% accuracy. The model produces neutral predictions, with positive and negative sentiment achieving near-zero F_1 in multi-task configurations. Multi-task learning shows no improvement (F_1 = 0.325 across all phases), suggesting that academic writing's formal tone constrains evaluative expression to subtle contextual cues that current models cannot capture. Notably, LLMs outperform fine-tuned SciBERT on sentiment (GPT-3.5: 47.4% , GPT-4o mini: 46.3% vs. 40.9%), likely because their broader pretraining provides richer signal for nuanced evaluative language that domain-specific fine-tuning cannot recover under severe class imbalance.

Rare Category Performance: Rare subfunction categories exhibit variable performance depending on linguistic distinctiveness rather than sample frequency. *Future Work* achieves 76.9% accuracy with 26 test samples due to clear linguistic markers ("future work", "next steps"), while *Critique* shows 50.0% accuracy with only 4 samples despite lacking distinctive surface features. This indicates that category viability depends more on semantic coherence and linguistic distinctiveness than training data volume alone, though extremely small sample sizes (n<5) remain problematic regardless of feature clarity.

Mitigation Strategy Effectiveness: Loss function modifications show limited effectiveness for severe imbalance. Weighted cross-entropy provides modest improvements for moderate imbalance, while focal loss demonstrates meaningful but insufficient impact on subfunction classification. The hierarchical conditioning mechanism proves substantially

more effective, achieving 65.3% macro F_1 compared to 50.9% without conditioning (28.3% relative improvement). This architectural solution leverages semantic relationships rather than attempting optimization-based corrections alone.

8. Conclusions and Future Work

We established a unified taxonomic framework for multi-dimensional CCA and developed a comprehensive dataset of 4,017 citations, annotated by experts, integrating contemporary scholarly works with established schemes. Our MTL experiments demonstrate that citation analysis tasks exist within structured semantic hierarchies, with hierarchical conditioning providing substantial improvements for subfunction classification. The hierarchical multi-task learning approach achieved 65.3% $\pm$ 0.6% macro F_1 for subfunction classification, representing a 21.1% relative improvement over single-task approaches. The unified framework addresses critical temporal coverage limitations through contemporary annotation protocols while enabling systematic comparison across diverse taxonomic approaches. Future research should prioritize cross-dataset evaluation using external benchmarks and explore methodological extensions including soft parameter sharing and multi-label classification frameworks.

Limitations

Our evaluation focuses exclusively on the integrated dataset we developed, limiting claims about cross-dataset generalization and universal applicability across different academic fields. The single-label classification approach cannot capture citations that serve multiple functions simultaneously, with boundary ambiguities evident where citations establish context while making comparisons. Class imbalance presents persistent challenges despite mitigation efforts, with sentiment classification proving largely ineffective and rare categories showing poor performance with limited training examples. Additionally, the taxonomy does not include an 'Other' or catch-all category, which means every citation is assigned to one of the six defined functions. While this reflects our design goal of exhaustive functional coverage, it may force classification of ambiguous or multi-purpose citations into the closest available category rather than leaving them unclassified.

Acknowledgments

This work was supported by the consortium NFDI for Data Science and Artificial Intelligence (NFDI4DS)[6] as part of the non-profit association National Research Data Infrastructure (NFDI e. V.). The consortium is funded by the Federal Republic of Germany and its states through the German Research Foundation (DFG) project NFDI4DS (no. 460234259). This work was also co-funded by the EU Horizon Europe project SciLake (grant agreement 101058573).

9. Bibliographical References

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Yasa M. Baig, Alex X. Oesterling, Rui Xin, Haoyang Yu, Angikar Ghosal, Lesia Semenova, and Cynthia Rudin. 2021. Multitask learning for citation purpose classification. In *Proceedings of the Second Workshop on Scholarly Document Processing*, pages 134–139, Online. Association for Computational Linguistics.

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)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Francisco Bolaños, Angelo Salatino, Francesco Osborne, and Enrico Motta. 2025. Modelling and classifying the components of a literature review. *arXiv preprint arXiv:2508.04337*.

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, Volume 1 (Long and Short Papers)*, pages 3586–3596, Minneapolis, Minnesota. Association for Computational Linguistics.

Marilena Daquino, Silvio Peroni, and David Shotton. 2020. The opencitations data model. In *The Semantic Web – ISWC 2020*, volume 12507 of *Lecture Notes in Computer Science*, Cham. Springer.

[6] https://www.nfdi4datascience.de

Changxu Duan and Zhiyin Tan. 2026. Semantically orthogonal framework for citation classification: Disentangling intent and content. *arXiv preprint*.

Alex Fogelson, Ana Trišović, and Neil Thompson. 2025. Llms in citation intent classification: Progress, precision, and reproducibility challenges. In *Proceedings of the 3rd ACM Conference on Reproducibility and Replicability*, ACM REP '25, page 250–253, New York, NY, USA. Association for Computing Machinery.

Eugene Garfield, Mary Elizabeth Stevens, and Vincent E. Giuliano. 1964. Can citation indexing be automated? In *Statistical Association Methods for Mechanical Documentation, Symposium Proceedings*, volume 269, pages 189–196. National Bureau of Standards. Accessed August 4, 2025.

Myriam Hernández-Álvarez, José M. Gómez Soriano, and Patricio Martínez-Barco. 2017. Citation function, polarity and influence classification. *Natural Language Engineering*, 23(4):561–588.

Lasse M. Jantsch, Dong-Jae Koh, Seonghwan Yoon, Jisu Lee, Anne Lauscher, and Young-Kyoon Suh. 2025. FineCite: A novel approach for fine-grained citation context analysis. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24525–24542, Vienna, Austria. Association for Computational Linguistics.

X. Jiang. 2025. Ensembling approaches to citation function classification and important citation screening. *Scientometrics*, 130:1371–1419.

Charles Jochim and Hinrich Schütze. 2012. Towards a generic and flexible citation classifier based on a faceted classification scheme. In *Proceedings of COLING 2012*, pages 1343–1358, Mumbai, India. The COLING 2012 Organizing Committee.

David Jurgens, Srijan Kumar, Raine Hoover, Dan 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.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico. 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.

Suchetha N. Kunnath, David Pride, and Petr Knoth. 2023. Prompting strategies for citation classification. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 1127–1137.

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*, pages 1875–1889, Seattle, United States. Association for Computational Linguistics. Code and dataset available at `https://github.com/allenai/multicite`.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007.

Michael J. Moravcsik and P. Murugesan. 1979. Citation patterns in scientific revolutions. *Scientometrics*, 1(2):161–169. Accessed August 5, 2025.

Roman Plaud, Matthieu Labeau, Antoine Saillenfest, and Thomas Bonald. 2024. Revisiting hierarchical text classification: Inference and metrics. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 231–242, Miami, FL, USA. Association for Computational Linguistics.

David Pride and Petr Knoth. 2020. An authoritative approach to citation classification. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, JCDL '20, pages 337–340, New York, NY, USA. Association for Computing Machinery.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Carlos N. Silla and Alex A. Freitas. 2011. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, 22(1-2):31–72.

Trevor Standley, Amir R. Zamir, Dawn Chen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. 2019. Which tasks should be learned together in multi-task learning? *arXiv preprint arXiv:1905.07553*. Presented at ICML 2020.

Xuan Su, Animesh Prasad, Min-Yen Kan, and Kazunari Sugiyama. 2019. Neural multi-task learning for citation function and provenance. In *Proceedings of the 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 394–395. IEEE.

Aixin Sun and Ee-Peng Lim. 2001. Hierarchical text classification and evaluation. In *Proceedings of the 2001 IEEE International Conference on Data Mining (ICDM)*, pages 521–528. IEEE.

John Swales. 1986. Citation analysis and discourse analysis. *Applied Linguistics*, 7(1):39–56. Accessed August 4, 2025.

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*, pages 103–110, Sydney, Australia. Association for Computational Linguistics.

A. Appendix

A.1. Compatibility with Existing Schemes

Our framework integrates with existing classification schemes through structured mappings. The three established schemes we work with – SciCite (Cohan et al., 2019), MultiCite (Lauscher et al., 2022), and Jurgens-CFC (Jurgens et al., 2018) – align well with our Level 1 categories, allowing direct automated conversion. Citation functions appear across taxonomies with similar meanings, though terminology differs. Level 2 subcategories provide finer distinctions not always present in existing schemes, requiring manual annotation while maintaining broad Level 1 compatibility. This design allows researchers to use existing datasets while gaining access to more detailed analysis. Direct mappings handle 89% of conversions, with only MultiCite's "Similarities" and "Differences" requiring merging into our COMPARISON category.

This approach allows researchers to work with a unified framework while preserving insights from existing classification schemes. The high proportion of direct mappings shows that our taxonomy captures core distinctions that the field has developed, while the hierarchical structure provides additional analytical dimensions not available in individual existing approaches.

A.2. Experimental Configuration

All experiments were conducted on Google Colab with NVIDIA Tesla T4 GPU (16GB VRAM). All three random seeds (42, 123, 456) completed successfully. Single-task experiments required 5.2–8.8 minutes per task with consistent GPU utilization (75–77% mean) and peak memory usage of 3.2–3.6 GB. Hierarchical MTL trained in 24.8 minutes using only 2.0 GB peak memory (45% reduction), though with lower mean GPU utilization (15%) due to early stopping dynamics. Progressive MTL required the longest training time at 108.7 minutes total across three sequential phases, with 3.4 GB peak memory and 66% mean GPU utilization. Table 4 presents the complete hyperparameter configuration used across all experiments.

Category	Parameter	Value
Model Architecture		
Base Model	SciBERT	allenai/scibert_scivocab_uncased
Max Sequence Length	tokens	256
Hidden Size	dimensions	768
Dropout	Hierarchical MTL	0.1
	Progressive MTL	0.3
Training Configuration		
Batch Size	Function	16
	Subfunction	12
	Sentiment	16
	Importance	16
	Hierarchical MTL	8
	Progressive MTL	16 (GPU)
Learning Rate	Function	1.5e-5
	Subfunction	1e-5
	Sentiment	1e-5
	Importance	1e-5
	Hierarchical MTL	2e-5
	Progressive MTL	2e-5
Optimizer		AdamW
Weight Decay		0.01
Warmup Ratio	Function	0.05
	Other tasks	0.1
Max Epochs	Function	6
	Subfunction	10
	Sentiment	8
	Importance	8
	Hierarchical MTL	10
	Progressive MTL	Variable by phase
Early Stopping Patience	Function	4
	Subfunction	4
	Sentiment	3
	Importance	4
	Hierarchical/Progressive	3
Gradient Clipping	max norm	1.0
Loss Functions		
Function Classification		Weighted Cross-Entropy
Subfunction (Single-task)		Weighted Cross-Entropy
Subfunction (MTL)		Focal Loss ($\gamma = 2.0$)
Sentiment Classification		Focal Loss ($\gamma = 2.0 - 3.0$)
Importance Classification		Weighted Cross-Entropy
Multi-Task Learning Weights		
Progressive Phase 1		Function: 0.7, Subfunction: 0.3
Progressive Phase 2		Func: 0.5, Sub: 0.3, Imp: 0.2
Progressive Phase 3		Func: 0.4, Sub: 0.3, Imp: 0.2, Sen: 0.1
Hierarchical MTL		Function: 0.7, Subfunction: 0.3
Data Configuration		
Training Set		2,814 citations (70%)
Validation Set		602 citations (15%)
Test Set		601 citations (15%)
Random Seeds		[42, 123, 456]
Stratified Sampling		Yes (by citation_id)
Hierarchical MTL Architecture		
Conditioning Mechanism		Concatenation
Function Logits Dimension		6
Function Embedding Dim		192 (hidden_dim/4)
Combined Feature Dim		960 (768 + 192)
Use Function Conditioning		True

Table 4: Hyperparameters for all experimental configurations.

A.3. Ablation Study

Removing function conditioning reduces subfunction performance to 50.9% macro F_1, demonstrating the critical importance of hierarchical structure. Full hierarchical MTL with conditioning achieves 65.3%, representing a 14.4% absolute improvement. Replacing focal loss with standard cross-

entropy yields 56.6% macro F_1, showing an 8.7% benefit from focal loss adaptation for class imbalance. The conditioning mechanism proves far more effective than loss function modifications, emphasising the importance of leveraging known semantic relationships rather than purely optimization-based approaches.