Towards Reliable Paper Contributions Annotation in the ACL Rolling Review

Anonymous ACL submission

Abstract

With the rapid growth of scientific publications, researchers struggle to efficiently assess the relevance of numerous papers. Identifying the types of contributions an article makes can help readers quickly grasp its significance. The ACL Rolling Review (ARR) introduced a typology requiring authors to specify their contributions to improve review quality and fairness. However, the current typology lacks clear definitions and guidance, leading to inconsistent labeling and raising concerns about its reliability. Our extensive re-annotation campaign reveals substantial disagreement between authors and domain experts. Evaluation of LLMs on paper contribution identification shows competitive performance relative to authors against annotator consensus, highlighting a potential path toward more reliable annotation.

1 Introduction

As the volume of research articles continues to grow, researchers face increasing difficulty in keeping up with the literature given limited time and resources. Staying up to date is essential for advancing research, yet evaluating the relevance and significance of numerous evolving works is time-consuming and demands substantial expertise. Rapidly assessing an article's content and identifying key trends is crucial for keeping pace with developments in the field, and can be aided by identifying the paper contributions. By categorizing the nature of the research, researchers can more efficiently assess its relevance and impact.

Identifying the type of contributions an article makes can offer several useful applications such as *deciding what to read* by curating reading lists for learners based on pedagogical value and concept structure (Gordon et al., 2017), *setting expectations before reading* by anticipating the kind of contributions presented by a research paper (Wobbrock and Kientz, 2016; Rogers et al., 2023), *summarizing articles* using disentangled contributions from paper content (Hayashi et al., 2023; Liu et al., 2023), *automatically identifying and structuring knowledge* across different paper contributions types within a collection of articles (Auer et al., 2018; D'Souza et al., 2021), *grasping emerging trends* by classifying research contributions of a collection of articles to conduct analysis over the field (Pramanick et al., 2025; Kaltenhauser et al., 2025).

Several typologies have been proposed to categorize research contributions, but no clear consensus exists due to differences in scope and granularity (see Appendix A). The most widely adopted typology in NLP comes from the ACL Rolling Review (ARR), which requires authors to explicitly state their contributions to support clearer and fairer evaluation (Bawden, 2019; Rogers et al., 2023). In this paper, we critically examine the ARR typology, focusing on the lack of clear annotation guidelines and its consequences for annotation reliability. Through a dedicated annotation campaign, we assess the reliability of author-assigned contribution labels and explore the potential of LLMs to perform this annotation task.

In summary, our contributions are:

- We conduct a controlled re-annotation of a subset of ARR-accepted papers by domain experts, enabling direct comparison with author-assigned contribution labels.
- We quantitatively assess the reliability of ARR contribution labels through inter-annotator agreement analysis and link the main sources of disagreement to inconsistencies in the typology and the lack of clear guidelines.
- We investigate the ability of LLMs to reproduce contribution annotations and show that they perform competitively with authors, highlighting their potential to assist in annotating contributions within the ARR process.

Our code, data and model are available on GitHub.

2 Data Collection

We focus on articles submitted to the ARR via OpenReview, specifically on accepted papers that include a *contribution type* following the introduction of the new guidelines in 2023[1]. The ARR typology defines 11 contribution types (Rogers et al., 2023), from which authors are asked to select one or more labels that best characterize their submission: 1) *NLP engineering experiment* (most papers proposing methods to improve state-of-the-art), 2) *approaches for low-compute settings, efficiency*, 3) *approaches for low-resource settings*, 4) *data resources*, 5) *data analysis*, 6) *model analysis and interpretability*, 7) *reproduction studies*, 8) *position papers*, 9) *surveys*, 10) *theory*, and 11) *publicly available software and pre-trained models*.

For each paper, we retain only the most recent version and consider only conferences or tracks with more than ten accepted ARR articles in order to limit sampling bias. The resulting dataset contains 2,050 articles annotated with author-assigned contribution labels. We split the data into training, validation, and test sets (80–10–10) using multi-label stratification (Sechidis et al., 2011). On average, each paper is associated with 2.11 labels ($\sigma = 1.06$), highlighting the multi-faceted nature of many submissions. Figure 1 shows the label distribution across splits, revealing substantial class imbalance, largely driven by the predominance of method- and data-oriented papers. Additional distribution (§C.1) and correlation analysis (§C.2) on the dataset are provided within Appendix.

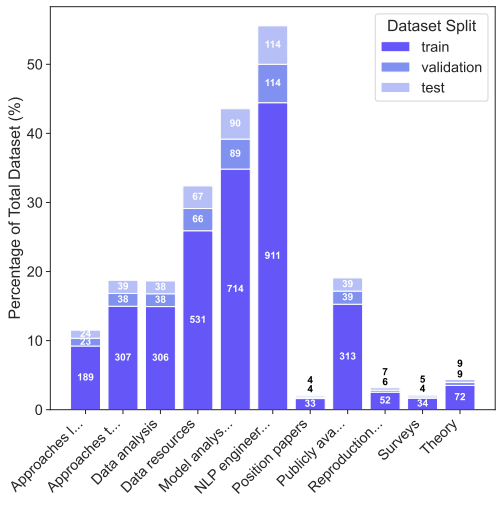


Figure 1: Contribution distribution across data splits.

[1] https://aclrollingreview.org/cfp

3 How Reliable Are Author-Assigned Paper Contributions?

3.1 Motivations & Re-annotation Task

A major limitation of the ARR typology is the lack of precise definitions for contribution categories, which may lead to inconsistent author-assigned annotations. To assess this issue, we introduce refined definitions to support more consistent annotation, organized into a systematic framework with four complementary components:

1. A general definition of the paper contribution, intended to guide readers unfamiliar with this category of work.
2. Methodologies or techniques typically implemented in such papers.
3. Clarifying criteria that distinguish the contribution from other types in cases of ambiguity.
4. Excerpts or language patterns frequently associated with this contribution.

An example of a paper contribution with our DICE-enhanced definition is provided below. See §D.3 in appendix for extensive definitions.

Label *Approaches to low-resource settings*
Description Papers investigating scenarios where labeled data, computational resources, or linguistic tools are limited.
Implementation These works typically employ techniques such as transfer learning, unsupervised or semi-supervised learning, or data augmentation to address these limitations.
Clarification Submissions focusing on well-resourced settings or small improvements that don't address major resource limitations are excluded.
Examples "Given constraints on computing resources in our deployment environment, we fine-tuned a distilled model to perform efficient and accurate intent classification.", ...

3.2 Annotation Framework

Using the refined contribution definitions, four domain experts (1 senior and 3 junior researchers) annotated the 207-document test split. Given the time-consuming nature of the task (experts required an average of 8 minutes per document), expert annotations are directly compared against author labels. Since papers are associated with few labels on average (2.11, $\sigma = 1.06$), we consider a Jaccard index below 2/3 as indicative of meaningful disagreement, triggering the assignment of an additional expert annotator. As a result, 72% of the samples were double-annotated. Additional details

and the full annotation guidelines are provided in Appendix D.

3.3 Authors Reliability Analysis

We assess the reliability of author-provided annotations by comparing them to expert annotations, treated as the reference, using Krippendorff's α (Krippendorff, 2004; Castro, 2017). On the test set, experts achieve an agreement of $\alpha = 0.55$, with substantial variability in pairwise agreement, reflecting the inherent complexity of the task. Agreement between authors and experts is lower but remains moderate ($\alpha = 0.47$; see Figure 2).

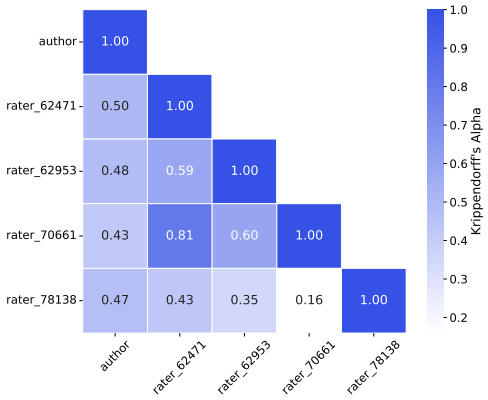


Figure 2: Pairwise Krippendorff's α between authors and domain experts on the test set.

As shown in Figure 3, author-assigned labels that are not selected by any expert occur frequently across categories, particularly for meta-level contributions. These discrepancies are also reflected in overall label proportions (Appendix E), with experts and authors assigning certain labels at very different rates. Experts tend to prefer generic labels and use meta-paper labels more conservatively, indicating differing views on the type or role of the annotated articles. Notably, the *Reproduction Study* label is never selected by experts, highlighting a clear mismatch between authors and experts understanding of this category.

These findings suggest that this aspect of the ARR process leads to inconsistent and unreliable labels, reducing their usefulness for reviewers.

4 Can Language Models Serve as Reliable Paper Contributions Annotators?

4.1 Identification of Paper Contributions

We formalize paper contribution identification as a multi-label classification task. Given a document

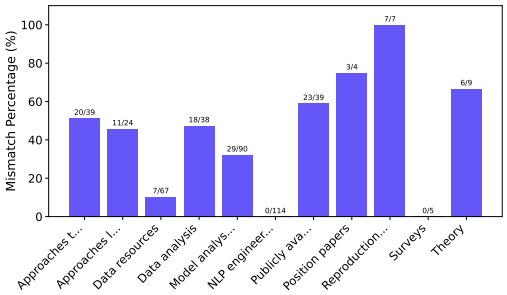


Figure 3: Percentage of author labels not confirmed by experts on test split.

D and a set of paper contributions $C = c1, ..., cn$, the goal is to learn a function $f : D \rightarrow L$, that maps D to one or more applicable labels $L \subseteq C$.

4.1.1 Experimental Setting

We evaluate the ability of generative models to infer paper contributions across a collection of articles. Specifically, we evaluate LLMs such as LLaMA (3.2–3B), Mistral (7B), and GPT-4.1, as well as fine-tuned general-domain Pre-trained Language Models (PLMs) (BERT, RoBERTa), fine-tuned science-focused PLMs (SciBERT, SPECTER2), and TF-IDF as baseline. Detailed models configurations (§F.1), instructions (§F.2) and prompting procedures (§F.3) are provided in Appendix.

In real-world scenarios, articles are often behind paywalls, leaving only metadata accessible. We therefore represent each article using the concatenation of its title and abstract. LLMs are evaluated under both author and expert conditions, using either labels only or the same annotation guidelines and DICE-enhanced typology employed by expert annotators. We further vary the number of few-shot examples to evaluate the models' ability to generalize from prior annotations, simulating an annotator refining decisions based on previously seen documents. The ground truth for the test set consists of labels on which a majority of domain experts and the authors agree. Model performance is evaluated using micro-averaged F1, reflecting overall labeling accuracy, and macro-averaged F1, to account for class imbalance and assess performance on under-represented labels.

4.2 Experimental Results

Results are presented in Table 1. Overall, the best-performing model is GPT-4.1 using 3-shot prompting. When evaluated against annotators consensus, this model achieves performance comparable to

author annotations in terms of micro-averaged F1.

TF-IDF performance is promising, achieving notably high $F1_{micro}$ scores, which indicates that lexical cues in the documents strongly support contribution identification. However, relatively low $F1_{macro}$ scores suggest that certain label types cannot be reliably inferred from lexical information alone. Similar trends are observed for PLMs on $F1_{micro}$, although $F1_{macro}$ scores are generally higher with fine-tuned domain-specific models, suggesting that their contextualized embeddings improve detection of underrepresented labels while maintaining stability across models. LLMs show more variable performance across settings. The best configurations of each model generally outperform both baselines and encoder-based approaches, suggesting that effective generalization from contribution definitions is more beneficial than relying solely on labeled training data. Providing models with the refined definitions used by expert annotators (DICE) generally improves alignment with expert labels, particularly when combined with few-shot sampling. Interestingly, for GPT-4.1, few-shot prompting largely mitigates the need for precise definitions, an effect not observed for other models.

Model		Annotators Consensus $F1_{micro}$	Annotators Consensus $F1_{macro}$
Authors		0.76	0.69
TF-IDF		0.62	0.27
BERT Base $_{Uncased}$		0.64	0.30
RoBERTa $_{Base}$		0.64	0.39
SciBERT $_{SciVocab, Uncased}$		0.66	0.40
SPECTER2 $_{Base}$		0.65	0.36
LLaMA 3.2 $_{3B\ Instruct}$	0-Shot	0.48	0.33
	1-Shot	0.47	0.29
	3-Shot	0.50	0.30
	0-Shot & DICE	0.48	0.32
	1-Shot & DICE	0.65	0.37
	3-Shot & DICE	0.45	0.25
Mistral 7B Instruct $_{v0.3}$	0-Shot	0.55	0.36
	1-Shot	0.54	0.33
	3-Shot	0.61	0.36
	0-Shot & DICE	0.55	0.36
	1-Shot & DICE	0.64	0.42
	3-Shot & DICE	0.69	0.49
GPT-4.1 $_{2025-04-14}$	0-Shot	0.67	0.54
	1-Shot	0.70	0.54
	3-Shot	**0.74**	**0.56**
	0-Shot & DICE	0.70	0.55
	1-Shot & DICE	0.72	0.55
	3-Shot & DICE	0.72	**0.56**

Table 1: Paper contribution identification performance (averaged over five seeds ; full results in Appendix G)

GPT-4.1 is the highest overall performing model, yet its per-label F1 remains lower than authors for several contribution types (Figure 4). Conversely, models outperform authors on a few labels (e.g., *NLP engineering experiment*, *approaches for low-resource settings*), indicating that while LLMs can surpass authors on generic categories, some contribution labels remain challenging to automatically identify even with improved guidelines.

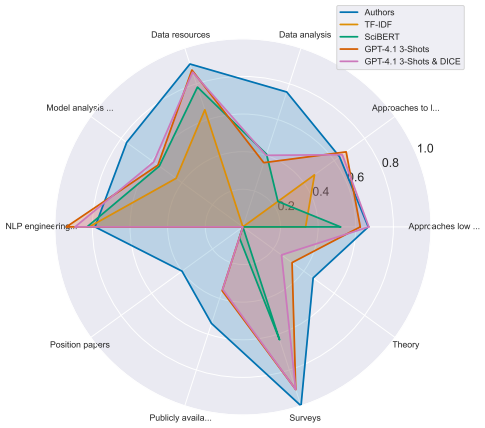


Figure 4: Per-label F1 performance of authors and best-performing AI systems on annotators consensus.

These results suggest that, while still below human-level performance, dedicated models could assist authors in annotating contributions within the ARR process. Moreover, the performance gains observed with DICE highlight the value of refined contribution definitions in guiding annotation.

5 Conclusion

Paper contributions, as currently defined and self-assigned, show limited reliability in the ARR process due to inconsistent author labeling. Re-annotation with improved guidelines shows discrepancies between authors and domain experts, highlighting unreliability of current ARR definitions. Our evaluation shows that LLMs are promising annotators: their predictions are relatively close to authors, though they do not yet reach human-level accuracy on all labels. Combining refined typologies with LLM assistance could make contribution annotation more consistent, transparent, and scalable, especially in rapidly evolving fields like NLP. To promote broader adoption of best practices and enhance authors and reviewers perception of paper contributions, we release refined definitions along with a pretrained SciBERT model intended to provide guidance for annotating paper contributions. Moreover, automatically extracting contributions can benefit not only ARR but also a wide range of downstream tasks, paving the way for more structured and accessible scientific knowledge.

Limitations

Enhancing ARR Typology Definitions. We proposed refinements to the ARR typology definitions using the DICE framework (Description, Implementation, Clarification, Examples), which is designed to support maintainability in evolving research fields. These refined definitions are informed by typology interpretations derived from author-annotated articles and feedback from expert annotators. While the present effort represents an initial step toward more structured and explicit definitions, it is possible that ARR editors emphasize aspects of the typology that differ from our interpretation of the papers contributions, which may in turn influence agreement with the proposed guidelines. The DICE-based definitions are therefore intended to remain adaptable and may be further refined through more extensive analyses and curated to better align with the needs of ARR organizers.

Annotation Campaign Scope. Re-annotation efforts were focused on the test set, reflecting practical constraints related to annotation time (approximately 8 minutes per document on average). To strengthen annotation quality, 72% of the data split was double-annotated by experts, and the original authors were incorporated as additional annotators when computing consensus. While the resulting typology analysis is conducted within this setting, extending annotation to the full dataset with a larger pool of annotators would be a natural direction for future work to further consolidate these findings.

Processed Documents. In these experiments, the identification of paper contributions is performed using article titles and abstracts. This choice reflects both the context length limitations of PLMs and the objective of evaluating models under comparable conditions. Nonetheless, restricting the input to partial document content may result in the omission of relevant contribution information, as annotators reported typically relying not only on the title and abstract but also prioritizing the introduction and conclusion sections, and more generally consulting the full article when assessing the scope of a paper's contributions. Although the dataset includes full-text content extracted using GROBID[2], this information was not exploited in the present study.

[2]https://github.com/kermitt2/grobid

Ethical Considerations

Annotation Campaign. In this study, we conducted internal annotation campaigns with a team of seven NLP experts, including three researchers and four NLP PhD students. Two of the four annotators selected for the test set annotations are also co-authors of this paper. Given the nature of the task, the annotation campaign was carried out over a two-week period, with annotators spending an average of approximately 8 minutes assessing the contributions of a single paper. Participants were compensated through non-financial means.

Use of generative AI in Scientific NLP. The use of generative AI models for assessing paper contributions entails inherent risks associated with this technology. In scholarly contexts, such systems must account for potential false positives and false negatives, overlooked contributions, and inaccurate representations of article content. We therefore advocate for a measured and responsible use of these tools and emphasize that, while they are intended to support authors and annotators during the annotation process, they are not designed to replace or fully automate this component of the ARR workflow.

Use of LLMs for the making of this study. In this study, LLMs were used only for writing style, code refactoring, and generating a few generic DICE excerpts. All conceptual contributions, analyses, and experiments were done without AI.

References

Liz Allen, Alison O'Connell, and Veronique Kiermer. 2019. How can we ensure visibility and diversity in research contributions? how the contributor role taxonomy (credit) is helping the shift from authorship to contributorship. *Learned Publishing*, 32(1):71–74.

Sören Auer, Viktor Kovtun, Manuel Prinz, Anna Kasprzik, Markus Stocker, and Maria Esther Vidal. 2018. Towards a knowledge graph for science. In *Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics*, WIMS '18, New York, NY, USA. Association for Computing Machinery.

Rachel Bawden. 2019. One Paper, Nine Reviews.

Emily M. Bender and Leon Derczynski. 2018. COLING 2018: Paper Types. Accessed: 2025-01-29.

Jordan Boyd-Graber, Naoaki Okazaki, and Anna Rogers. 2023. Paper-reviewer matching at ACL 2023: types of contributions and track sub-areas .

Amy Brand, Liz Allen, Micah Altman, Marjorie Hlava, and Jo Scott. 2015. Beyond authorship: attribution, contribution, collaboration, and credit. *Learned Publishing*, 28(2):151–155.

Santiago Castro. 2017. Fast Krippendorff: Fast computation of Krippendorff's alpha agreement measure. https://github.com/pln-fing-udelar/fast-krippendorff.

Haihua Chen, Huyen Nguyen, and Asmaa Alghamdi. 2022. Constructing a high-quality dataset for automated creation of summaries of fundamental contributions of research articles. *Scientometrics*, 127(12):7061–7075.

Liyue Chen, Jielan Ding, Donghuan Song, and Zihao Qu. 2025. Exploring scientific contributions through citation context and division of labor. *Scientometrics*, 130(5):2901–2921.

Jennifer D'Souza, Sören Auer, and Ted Pedersen. 2021. SemEval-2021 task 11: NLPContributionGraph - structuring scholarly NLP contributions for a research knowledge graph. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 364–376, Online. Association for Computational Linguistics.

Jennifer D'Souza and Sören Auer. 2021. Sentence, Phrase, and Triple Annotations to Build a Knowledge Graph of Natural Language Processing Contributions—A Trial Dataset. *Journal of Data and Information Science*, 6(3):6–34.

Alexander Fabbri, Irene Li, Prawat Trairatvorakul, Yijiao He, Weitai Ting, Robert Tung, Caitlin Westerfield, and Dragomir Radev. 2018. TutorialBank: A manually-collected corpus for prerequisite chains, survey extraction and resource recommendation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 611–620, Melbourne, Australia. Association for Computational Linguistics.

Jonathan Gordon, Stephen Aguilar, Emily Sheng, and Gully Burns. 2017. Structured generation of technical reading lists. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 261–270, Copenhagen, Denmark. Association for Computational Linguistics.

Komal Gupta, Ammaar Ahmad, Tirthankar Ghosal, and Asif Ekbal. 2021. *ContriSci: A BERT-Based Multitasking Deep Neural Architecture to Identify Contribution Statements from Research Papers*, page 436–452. Springer International Publishing.

Komal Gupta, Ammaar Ahmad, Tirthankar Ghosal, and Asif Ekbal. 2024. A bert-based sequential deep neural architecture to identify contribution statements and extract phrases for triplets from scientific publications. *International Journal on Digital Libraries*, 25(4):1–28.

Carolin Haeussler and Henry Sauermann. 2020. Division of labor in collaborative knowledge production: The role of team size and interdisciplinarity. *Research Policy*, 49(6):103987.

Hiroaki Hayashi, Wojciech Kryscinski, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2023. What's new? summarizing contributions in scientific literature. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1019–1031, Dubrovnik, Croatia. Association for Computational Linguistics.

Annika Kaltenhauser, Gian-Luca Savino, Nick von Felten, and Johannes Schöning. 2025. CHI's Greatest Hits: Analyzing the 100 Most-Cited Papers in 43 Years of Research at ACM CHI. *Interactions*, 32(1):28–33.

Klaus Krippendorff. 2004. Reliability in content analysis. *Human Communication Research*, 30(3):411–433.

Meng-Huan Liu, An-Zi Yen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2023. Contributionsum: Generating disentangled contributions for scientific papers. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 5351–5355, New York, NY, USA. Association for Computing Machinery.

Aniket Pramanick, Yufang Hou, Saif M. Mohammad, and Iryna Gurevych. 2025. The nature of NLP: Analyzing contributions in NLP papers. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25169–25191, Vienna, Austria. Association for Computational Linguistics.

Anna Rogers, Marzena Karpinska, Jordan Boyd-Graber, and Naoaki Okazaki. 2023. Program chairs' report on peer review at acl 2023. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages xl–lxxv, Toronto, Canada. Association for Computational Linguistics.

Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. On the stratification of multi-label data. In *Machine Learning and Knowledge Discovery in Databases*, pages 145–158, Berlin, Heidelberg. Springer Berlin Heidelberg.

Emily Sheng, Prem Natarajan, Jonathan Gordon, and Gully Burns. 2017. An investigation into the pedagogical features of documents. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 109–120, Copenhagen, Denmark. Association for Computational Linguistics.

Jacob O. Wobbrock and Julie A. Kientz. 2016. Research contributions in human-computer interaction. *Interactions*, 23(3):38–44.

A Existing Contribution Typologies

ARR Typology	
Rogers et al. (2023)	Approaches to low-resource settings , Approaches low compute settings-efficiency , Data resources , Data analysis , Model analysis & interpretability , NLP engineering experiment , Publicly available software and/or pre-trained models , Position papers , Reproduction study , Surveys , Theory
ARR Inspirations	
Bender and Derczynski (2018)	Computationally-aided linguistic analysis , NLP engineering experiment paper , Reproduction paper , Resource paper , Position paper , Survey Paper
Boyd-Graber et al. (2023)	Computationally-aided linguistic analysis , NLP engineering experiment , Approaches for data- and compute efficiency , Reproduction study , New data resources , Position papers , Surveys , Theory , Publicly available software and pre-trained models
Other Contributions Typologies	
D'Souza et al. (2021)	Research problem , Approach , Model , Code , Dataset , Experimental setup , Hyper parameters , Baselines , Results , Tasks , Experiments , Ablation analysis .
Chen et al. (2022)	Dataset/Resources creation , Theory proposal , Model construction or optimization , Algorithms/Methods construction or optimization , Performance evaluation , Applications
Liu et al. (2023)	Approach , Analysis , Result , Topic or Resource
Pramanick et al. (2025)	Knowledge (k-Dataset , k-Language , k-Method , k-People , k-Task), Artifact (a-Dataset , a-Method , a-Task)
Chen et al. (2025)	Theoretical , Methodological , Experimental , Data-based , Other
Other Non-Contributions Typologies	
Brand et al. (2015)	Conceptualization , Methodology , Software , Validation , Formal Analysis , Investigation , Resources , Data curation , Writing – Original Draft , Writing – Review & Editing , Visualization , Supervision , Project Administration , Funding acquisition
Sheng et al. (2017)	Survey , Tutorial , Resource , Reference Work , Empirical Results , Software Manual , Other

Table 2: Colors indicate generic concepts from the ARR typology, as well as similar labels found in other typologies based on the definition provided by their authors. The highlighted categories include: Optimization , Resources , Analysis , Experimental , Models/Softwares , Position paper , Reproduction study , Survey , and Theory . Please note that due to differences between typologies, the highlighted concepts may not exactly match the original labels in every case.

B Related Works

Contributions are commonly defined in the literature as novel scientific advancements attributed to the authors of a research paper (Pramanick et al., 2025). The identification of such contributions is typically performed at the document or statement level. However, the notion of what constitutes a "contribution" frequently extends beyond disciplinary boundaries, presenting similarity with other fields (Sheng et al., 2017; Brand et al., 2015).

Paper Contributions. Paper contributions identification aims at capturing an holistic view of scope of a paper. This task has been particularly explored in approaches designed to support reviewers during the submission process, helping them more accurately assess the nature of the work they evaluate (Bender and Derczynski, 2018; Bawden, 2019; Boyd-Graber et al., 2023; Rogers et al., 2023). Paper-level identification schemes have also been used to analyze submissions to top-tier conferences, revealing research trends and suggesting ways to adapt to evolving fields (Rogers et al., 2023; Kaltenhauser et al., 2025).

Contributions Statements. This line of research focuses on identifying individual contribution statements within scientific papers, providing fine-grained insights into the specific achievements of a scientific work. Several works have explored the extraction of contribution-specific data in order to build knowledge graphs, sometime directly from the statement (Gupta et al., 2021, 2024) or combined with contribution types (D'Souza and Auer, 2021; D'Souza et al., 2021). Other studies have leveraged contribution statements to summarize the key contributions of research articles, enabling readers to quickly grasp their main findings (Chen et al., 2022; Hayashi et al., 2023; Liu et al., 2023). More recently, Pramanick et al. (2025) identified and analyzed contribution statements across ACL Anthology papers to find research trends and characterize the nature of NLP research.

Pedagogical Roles of Papers. The identification of pedagogical roles is closely related to contribution analysis, as it seeks to characterize how useful a document is for individuals aiming to learn specific concepts (Sheng et al., 2017). Although pedagogical roles do not directly represent a paper's contributions, they are often closely aligned with them in existing typologies (see Table 2). These roles have been notably exploited by approaches that model relationships between papers to support resource recommendation, whether focusing on conceptual links (Gordon et al., 2017) or prerequisite chains (Fabbri et al., 2018).

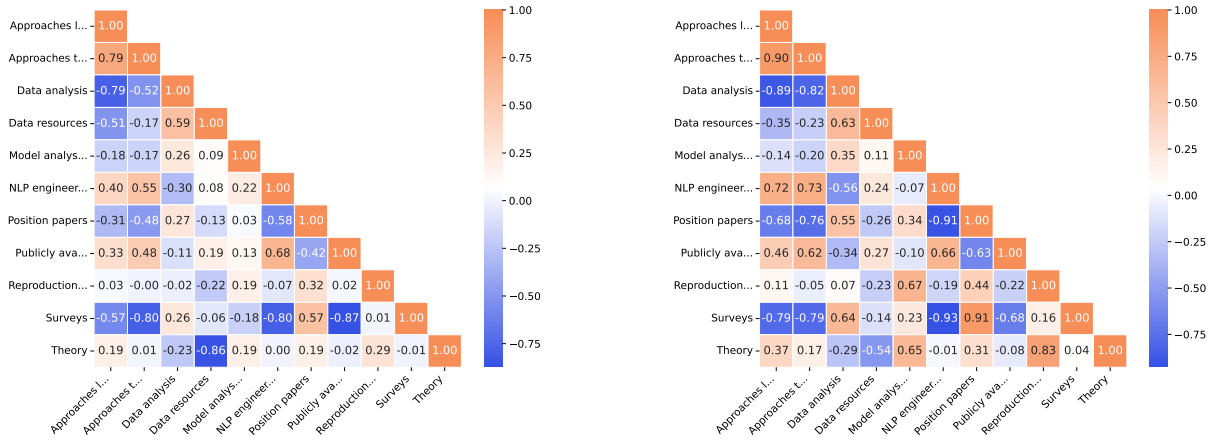
Contributor Roles. In Science Studies and related fields, researchers have explored the Division of Labor (DOL) in scientific collaboration, analyzing how responsibilities are distributed among contributors (Allen et al., 2019; Haeussler and Sauermann, 2020). Taxonomies such as the Contributor Role Taxonomy (CRediT) (Brand et al., 2015) allow to model the roles of scientists during production and publication of research articles. Interestingly, this taxonomy presented relationships between paper contributions and DOL, as shown by Chen et al. (2025). Their work leveraged contribution statements and citation contexts to examine whether the contributions reported in papers align with authors' self-perceived roles.

C Contributions Analysis

C.1 Distribution of Paper Contributions

Figure 1 shows the distribution of paper contributions labels across the collection, revealing a strong imbalance in the labels selected by authors. Some labels account for more than 30% of the dataset (e.g., NLP Engineering Experiment, Model Analysis and Interpretability, Data Resources), while others represent less than 5% of the instances (e.g., Theory, Reproduction papers, Surveys, Positions papers). These discrepancies between labels could be explained by multiple factors. First, the ARR typology is designed to make contributions more explicit to reviewers, helping them better recognize and fairly evaluate papers with less common types of contributions (Rogers et al., 2023). Second, certain contribution types may be incidental to others; for example, an experimental paper may include a model analysis, or a data resource paper may present an accompanying dataset analysis. Finally, the general lack of clear definitions or annotation guidelines within the typology may lead authors to favor broader, more generic categories, or to select all labels that appear even loosely relevant. This imbalance motivates the need for clearer contribution definitions and points to possible biases in models trained on this dataset.

C.2 Correlation Analysis of Contributions



(a) Pearson

(b) Spearman

Figure 5: Correlation analysis of paper contribution labels.

We compute correlation matrices over the co-occurrence of contribution labels to identify potential relationships between paper contribution types. Due to the imbalance in label distribution, we normalize the co-occurrence matrix using Pointwise Mutual Information (PMI). Both Pearson's r and Spearman's ρ correlation coefficients are computed, showing similar trends (see Figure 5).

The results reveal strong associations between certain contribution types: papers focusing on low-resource settings and efficiency are highly correlated ($\rho = 0.91$); surveys and position papers also co-occur frequently, as do theory papers with reproduction studies ($\rho > 0.80$). Conversely, and perhaps unsurprisingly, NLP engineering experiment papers exhibit strong negative correlations ($\rho < -0.90$) with both position and survey papers. Papers providing publicly available models show notable variation, being most positively correlated with experimental papers ($r = 0.68$) and most negatively correlated with survey papers ($r = -0.87$).

Overall, the observed correlations indicate that the current typology may benefit from refinement, as ideally each contribution type should capture a distinct facet of a paper with minimal overlap.

D Annotation Campaign

D.1 Pilot Annotation

Before moving to the main annotation campaign, we conduct a preliminary annotation campaign with seven annotators (three senior and four junior researchers, all specializing in NLP) aimed at refining guidelines and identifying the experts with the highest agreement. This was conducted on 65 full papers, deliberately oversampling underrepresented classes to ensure comprehensive coverage of the typology. The distribution of this pilot is as follows:

- Approaches low compute settings-efficiency: 15.38%
- Approaches to low-resource settings: 23.08%
- Data analysis: 24.62%
- Data resources: 26.15%
- Model analysis and interpretability: 44.62%
- NLP engineering experiment: 55.38%
- Position papers: 6.15%
- Publicly available software and/or pre-trained models: 13.85%
- Reproduction study: 9.23%
- Surveys: 6.15%
- Theory: 15.38%

Each paper is annotated by at least three annotators, with up to two additional annotators added when agreement, measured by the Jaccard index, fell below 30%. Feedback from this campaign is used to refine and consolidate the typology definitions through direct exchanges with annotators, addressing the difficulties they encounter throughout the annotation process. The *Clarification* and *Examples* components of the DICE framework notably emerged from this initial annotation phase. Agreement of this pilot annotations is presented in Table 6, showing Krippendorff's α among domain-experts. We selected annotators with higher agreement scores in this preliminary campaign for the final annotation campaign.

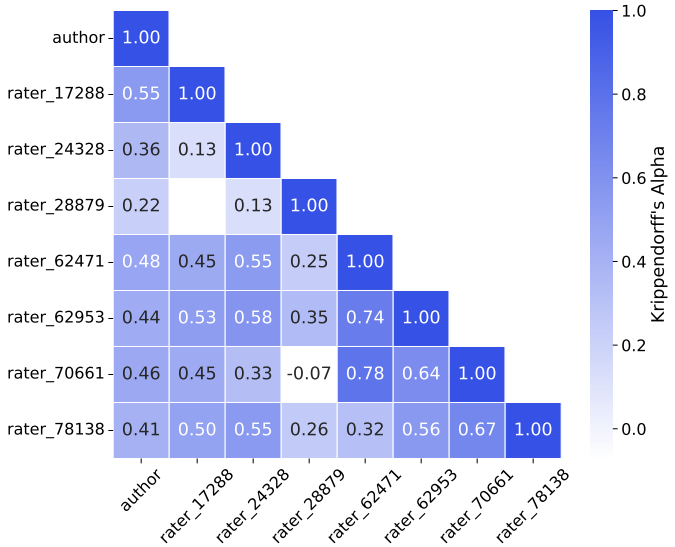


Figure 6: Pairwise Krippendorff's α between authors and domain experts on the test set within pilot campaign.

D.2 Annotation Interface

Figure 7: Annotation campaign interface

D.3 Annotation Guidelines

```
# Guidelines

You will be provided with a scientific paper and a typology of potential contribution types.
Analyze the paper and classify the type(s) of contributions it makes based on the provided typology.


Follow these rules when making your decisions:

1. Assign applicable labels. A paper may fall under multiple contribution types, assign those that are
     clearly present in the paper content.
2. Focus on notable contributions. Only assign label if the paper makes a substantial and deliberate
     contribution of that type, not just a minor mention or incidental inclusion.

# Typology

Approaches to low-resource settings
Papers investigating scenarios where labeled data, computational resources, or linguistic tools are limited.
      These works typically employ techniques such as transfer learning, unsupervised or semi-supervised
     learning, or data augmentation to address these limitations. Submissions focusing on well-resourced
     settings or small improvements that don't address major resource limitations are excluded.

Examples:
-"Given constraints on computing resources in our deployment environment, we fine-tuned a distilled model to
      perform efficient and accurate intent classification.
-"We propose an adversarial representation alignment model to mitigate performance degradation in low-
     resource parsing by selectively transferring relevant knowledge from high-resource languages.
-"To address limited annotation budgets, we trained a weakly supervised classifier using automatically
     generated noisy labels based on keyword heuristics."

Approaches low compute settings-efficiency
Papers focusing on computational efficiency in NLP environments. These works typically employ techniques
     such as reduced memory usage, optimized training or inference time, or ways to reduce energy
     consumption, making models more accessible or deployable in low-compute environments. Submissions
     focusing on minor efficiency gains in high-resource contexts or lacking practical impact on
     accessibility are excluded.

Examples:
-"To enable deployment on mobile devices, we reduced model size by pruning and quantization, significantly
     lowering memory usage without sacrificing accuracy.
-"We improve WER on long-form ASR and achieve up to 20x faster inference through batched parallel decoding.
-"We propose an optimized training schedule that cuts down GPU hours by 40%, making large-scale NLP model
     training more feasible for smaller labs."

Data resources
Papers providing new language-related resources such as datasets, annotated corpora, annotation standards or
      evaluation benchmarks. These works typically provide detailed documentation of the resource creation
     process, including methodology and quality assurance. Submissions lacking substantial novelty in
     resource creation or failing to share them publicly are excluded.

Examples:
-"We present a corpus of historical legal texts annotated for named entities and syntactic structures, along
      with an open-access tool for browsing and querying the data.
-"This work introduces a benchmark, including a standardized evaluation protocol and carefully curated test
     sets.
-"We propose a taxonomy and release a dataset of ~2,000 annotated NLP paper abstracts, capturing and
     categorizing their scientific contributions."

Data analysis
Papers conducting detailled analysis of data resources. These works generally focus on annotation quality,
     bias, linguistic patterns, or how models interact with data. Submissions should present novel insights
     that contribute to better data practices or enhanced model design.
```

```
Examples:
-"This work propose a comprehensive analysis of gender bias in a widely used sentiment analysis dataset,
     revealing systematic annotation inconsistencies that affect model predictions.
-"We analyze the existing data resources and identify areas for improvement in future iterations.
-"Model performances across subsets of a dependency parsing corpus show that annotation errors
     disproportionately impact low-frequency syntactic constructions."

Model analysis & interpretability
Papers investigating the internal mechanisms or external behavior of NLP models. These works typically
     employ techniques such as ablation studies, model probing, or interpretability visualization to make
     model decisions more transparent and understandable. Submissions should provide novel insights that
     deepen our understanding of model behavior rather than merely measuring performances.

Examples:
-"We probe the model representations across syntactic constructions and show that errors on low-frequency
     patterns stem from weak internal encoding, rather than data scarcity alone.
-"Our analysis identifies the amount of targeted privacy data and the extent of edited privacy neurons as
     the two key factors contributing to this model behaviour.
-"We visualize attention patterns across multiple layers to understand how models resolve coreference,
     uncovering systematic biases in entity tracking."

NLP engineering experiment
Papers describing the design, implementation, or deployment of NLP systems. These works generally propose
     new models, enhancements over state-of-the-art methods, or solutions to engineering challenges in NLP
     systems. Submissions should demonstrate clear technical contributions and real-world applicability.

Examples:
-"To address these challenges, we propose a novel fine-tuning method that employs sentence concatenation
     with augmented random facts to regularize generation.
-"Our model achieve state-of-the-art performance in machine translation by introducing efficient attention
     mechanisms, resulting in faster inference with comparable accuracy.
-"We deploy an end-to-end NLP pipeline for document classification, addressing engineering challenges
     related to scalability and latency."

Publicly available software and/or pre-trained models
Papers providing pretrained models or NLP-related softwares, APIs, or libraries intended for broad community
      use. These works typically provide public code repositories or access to models. Submissions should
     demonstrate significant utility or innovation and ensure open availability to the community.

Examples:
-"We release a pretrained multilingual language model optimized for low-resource languages, along with a
     public API for easy integration in downstream applications.
-"We provide a comprehensive toolkit for named entity recognition, featuring pretrained models and
     customizable annotation interfaces, all available via a public repository.
-"We will publicly release our code and pre-trained models on the following URL."

Position papers
Papers presenting a strong perspective or argument on existing research. These works challenge existing
     norms, give a new set of ground rule, or offer visions for the future of the field. Submissions should
     provide well-founded arguments and contribute meaningfully beyond opinion or commentary.

Examples:
-"We argue for a paradigm shift in NLP research, advocating for more human-centered evaluation methods that
     prioritize interpretability and fairness.
-"Recent debates about whether large language models understand text often stem from differing definitions
     of understanding and views on consciousness, illustrated here by a thought experiment with a high-
     performing but seemingly non-conscious chatbot.
-"We challenge the assumption that larger models inherently lead to better understanding, proposing
     alternative evaluation metrics that better capture semantic comprehension."

Reproduction study
Papers analyzing, replicating and validating prior work. These works often provide new insights, clarify
     ambiguities, or expose inconsistencies into existing works, allowing to improve confidence in research
     findings. Submissions should offer substantial contributions beyond mere repetition or summary of
     previous results.

Examples:
-"We replicate the experiments from the original paper to verify the reported results and assess their
     robustness across different datasets.
-"By reanalyzing a landmark sentiment analysis study, we identify ambiguous evaluation metrics and propose
     clearer standards for future work.
-"Our validation of recent models reveals inconsistencies in reported results."

Surveys
Papers synthesizing and structuring existing literature on a specific topic. These works are generally
     indicated as such and outline methods, categorize trends, highlight gaps, and suggest future directions
     , serving as a roadmap for researchers. Submissions should extends well beyond the typical related work
      section of a research paper.

Examples:
-"This survey provides a comprehensive overview of recent advances in transformer-based architectures for
     natural language processing.
-"In this article, we present a state-of-the-art of the main text generation approaches, including
     evaluation data, methods, and metrics.
-"We provide a comprehensive roadmap for explainable NLP, reviewing current techniques, evaluating their
     applicability, and proposing a unified framework for future research."
```

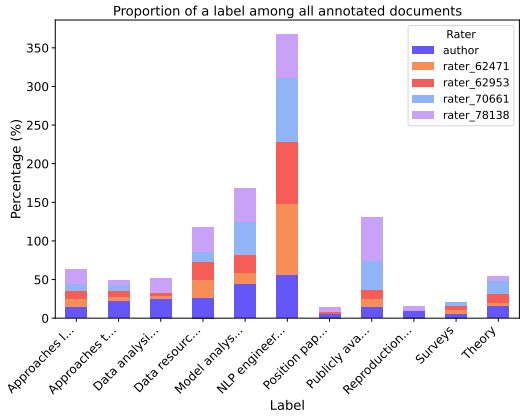
```
Theory
Papers contributing to the formal or mathematical foundations of NLP. These works may include new algorithms
     , formal grammar models, or computational theories. Submissions should emphasize rigorous theoretical
     development rather than empirical evaluation.

Examples:
-"This paper presents a new algorithm with proven computational guarantees for efficient parsing in natural
     language processing.
-"We combine theoretical proofs and experimental results to establish a foundation for improving Chain-of-
     Thought distillation within a multitask learning framework, guided by information-theoretic principles.
-"We propose a novel formal grammar model that rigorously characterizes syntactic structures without relying
      on annotated datasets."
```

D.4 Preliminary Annotation Guidelines

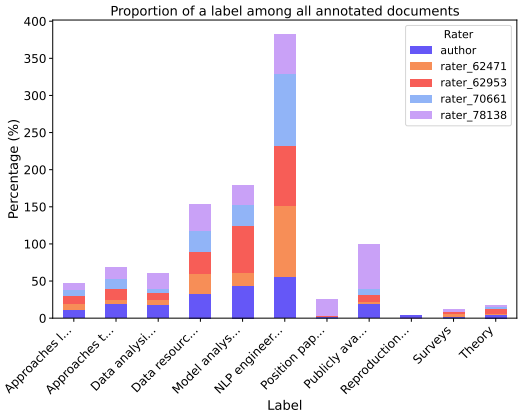
```
# Context

In scholarly articles, contributions refer to the scientific achievements or innovations attributed to the
     authors. These may include additions to existing knowledge, theoretical advancements, methodological
     innovations, or the development of new artifacts or tools.

Within the ACL Rolling Review process, authors are encouraged to specify one or more contribution types that
      their submission addresses. This typology is informed by prevailing research practices and thematic
     trends within the field of Natural Language Processing (NLP).

However, this typology is lackluster on ACL RR, no definitions are defined and it is not clear what should
     constitute a candidate contribution types.

# Objective

The aim of this annotation task is to assess whether independent annotators-that is, individuals other than
     the original authors-can reliably identify the same contribution types by providing more extensive
     guidelines.

These annotations will help evaluate the clarity and communicative effectiveness of scientific writing
     regarding contribution statements.

# Annotation Guidelines

1. Open the provided PDF file
2. Using the provided typology and following definitions, please assign labels that are relevant to the
     document.


Multiple contribution types can be assigned to the same document if appropriate.

Please keep in mind that even if some contributions seems to be present in the document, the goal is to
     asses if the document propose a notable enough contribution so it can be qualified by that type. (e.g.
     A data resource-oriented paper is not the same as merely making some data available, a paper taking a
     stance is not the same as a position paper, a paper analysis some results does not necessarily implies
     a model analysis, etc.)

# Typology definitions

Approaches to low-resource settings
Papers that propose methods or tasks designed for scenarios where labeled data, computational resources, or
     linguistic tools are limited. Examples of such works can include approaches leveraging generalizability
     , transfer learning, data augmentation or unsupervised/semi-supervised learning to address these
     limitations among other techniques.

Approaches low compute settings-efficiency
Papers that propose approaches or techniques to make NLP models more computationally efficient. Examples of
     such works can include optimizing memory usage, inference/training time or energy consumption to reduce
      computing costs among other techniques.

Data resources
Papers that introduces new datasets, annotated corpora, benchmarks, or other evaluation tools. Such works
     are often identifiable by detailed descriptions of the data creation process and by making the
     resources publicly accessible.

Data analysis
Papers focused on analyzing trends or patterns in data, such as linguistic phenomena, annotation quality,
     data biases or how models interact with datasets. Emphasis is on insight rather than system performance
     .

Model analysis & interpretability
Papers investigating how NLP models function internally. Examples of such works can include ablation studies
     , probing techniques, interpretable evaluations or tools designed to make model behavior more
     understandable and transparent to humans.

NLP engineering experiment
Papers focusing on the design, implementation, or deployment of NLP systems. Examples of such works can
     include experiments developping new NLP systems, proposing models that improve on the State-of-the-Art,
      or solving engineering challenges for NLP purpose.
```

```
Publicly available software and/or pre-trained models
Papers releasing tools, APIs, libraries, or pre-trained models that are intended for broad use by the
     research and development community. Such works are often identifiable by the inclusion of public
     repository URLs.

Position papers
Papers that articulate a clear stance or perspective on a significant issue in NLP. Such works may argue for
      specific changes in methodology, evaluation, ethical standards or future directions, often
     characterized by a more opinionated or speculative tone.

Reproduction study
Papers that attempts to replicate and validate the results of prior studies. Example of such works can
     include generalization to new settings or identification of gaps, ambiguities and errors in the
     original work.

Surveys
Papers that summarize and organize existing literature on a particular topic, method, or subfield. Unlike a
     standard related work section, surveys aim to synthesize trends, highlight gaps, and provide a roadmap
     for future research.

Theory
Papers that contribute to the formal or mathematical foundations of NLP. Example of such works can include
     new algorithms, computational models of grammar, or rules, often without direct empirical results.
```

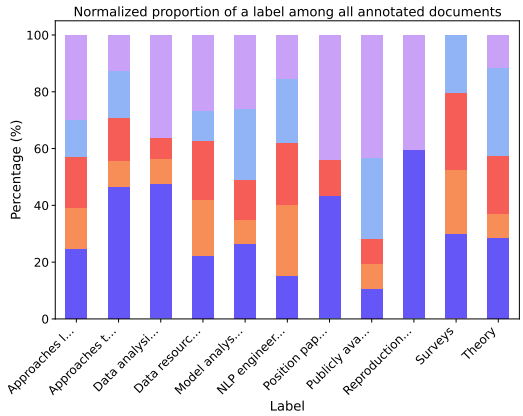
E Additional Annotations Analysis



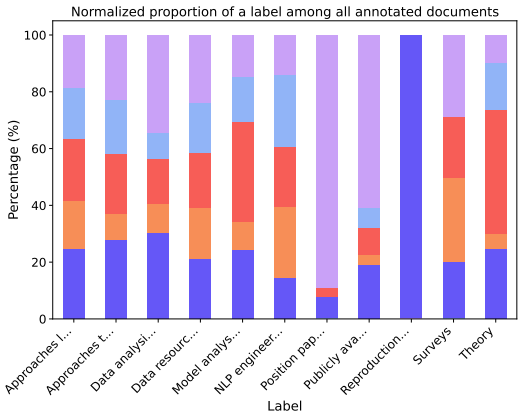
(a) Preliminary



(b) Final



(c) Normalized Preliminary



(d) Normalized Final

Figure 8: Proportion (top) and Normalized proportion (bottom) of labels selected across all annotations by the four best annotators, shown for the preliminary campaign (left) and the testing set campaign (right).

F Architecture Details

F.1 Configurations

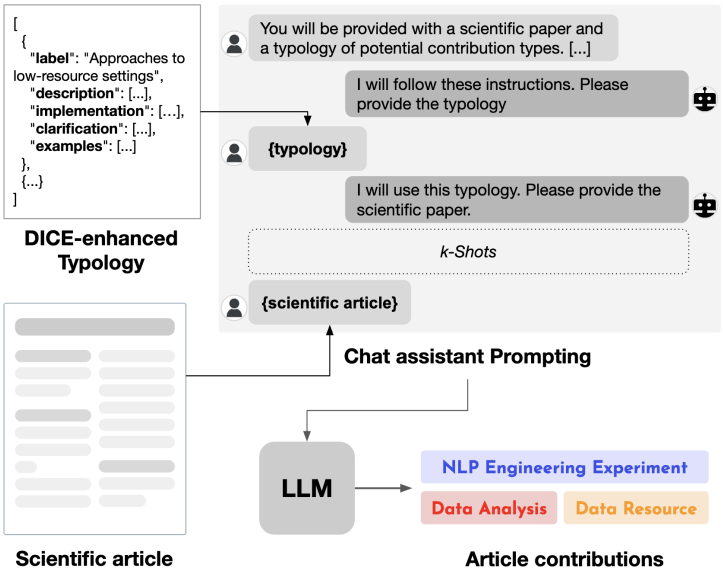


Figure 9: Workflow for identifying paper contributions via LLM prompting and DICE-enhanced definitions.

Large Language Models (LLMs). The LLM architectures included in our study are Llama (3.2–3B)[3] and Mistral (7B)[4]. We use the following generation parameters for reproducibility: sampling is disabled (`do_sample=False`), the repetition penalty is set to 1.0, the no repeat n-gram size is 0, and the maximum number of generated tokens is 128.

GPT-4.1. In addition to open-weight models, we also evaluate the closed-source model GPT-4.1[5] using the OpenAI Completion API. The temperature and presence penalty are both set to 0, and the maximum number of generated tokens is 128.

Pre-trained Language Models (PLMs). The PLMs included in our study includes both general-purpose models (BERT[6], RoBERTa[7]) and science-focused models (SciBERT[8], SPECTER2[9]). The models are fine-tuned using a learning rate of 1e-5, a batch size of 16, 20 training epochs with early stopping after 5, and a weight decay of 0.01.

Baselines. The baseline methods include (1) randomly label sampling from all available classes, and (2) a logistic regression classifier trained on TF-IDF features with 1 000 iterations using the liblinear solver.

① Experiments presented in this paper ran for under 50 GPU hours on two NVIDIA RTX 4500 GPUs (24GB each) and cost approximately $35 in OpenAI credits.

F.2 LLMs Instructions

The annotation guidelines provided to the LLMs are identical to those used in the human annotation campaign.

```
You will be provided with a scientific paper and a typology of potential
    contribution types.
Analyze the paper and classify the type(s) of contributions it makes based on the
    provided typology.

Follow these rules when making your decisions:
```

[3] https://huggingface.co/meta-llama/Llama-3.2-3B
[4] https://huggingface.co/mistralai/Mistral-7B-v0.3
[5] https://platform.openai.com/docs/models/gpt-4.1
[6] https://huggingface.co/google-bert/bert-base-uncased
[7] https://huggingface.co/FacebookAI/roberta-base
[8] https://huggingface.co/allenai/scibert_scivocab_uncased
[9] https://huggingface.co/allenai/specter2

```
1. Assign applicable labels. A paper may fall under multiple contribution types,
    assign those that are clearly present in the paper content.
2. Focus on notable contributions. Only assign label if the paper makes a
    substantial and deliberate contribution of that type, not just a minor mention
    or incidental inclusion.
```

F.3 Prompting Layout

```
messages = [
    {"role": "system", "content": "Follow the provided instructions. Reply ONLY the
        list of labels from the typology applicable to the scientific paper, no
        explanation"},
    {"role": "user", "content": instructions},
    {"role": "assistant", "content": "I will follow these instructions. Please
        provide the typology."},
    {"role": "user", "content": typology_instructions},
    {"role": "assistant", "content": "I will use this typology. Please provide the
        scientific paper."},
]

#Handle few-shots when requested
for i in range(nb_shots):
    messages.extend([
        {"role": "user", "content": dataset["train"][i]["document"]},
        {"role": "assistant", "content": f"Here is the list of contribution types
            present in the scientific paper: {dataset["train"][i]["
            contribution_types"]}"},
    ])

#Current document to process
messages.extend([
    {"role": "user", "content": doc},
    {"role": "assistant", "content": "Here is the list of contribution types present
         in the scientific paper:"},
])
```

G Additional Experimental Results

	Annotators Consensus			
	P	**R**	**F1 micro**	**F1 macro**
Authors	0.75 ±0.00	0.77 ±0.00	0.76 ±0.00	0.69 ±0.00
Baselines				
Random	0.20 ±0.00	0.55 ±0.02	0.29 ±0.01	0.23 ±0.01
TF-IDF	0.81 ±0.00	0.50 ±0.00	0.62 ±0.00	0.27 ±0.00
PLMs				
BERT Base (Uncased)	0.75 ±0.01	0.56 ±0.02	0.64 ±0.01	0.30 ±0.01
RoBERTa Base	0.74 ±0.01	0.57 ±0.01	0.64 ±0.01	0.39 ±0.02
SciBERT (SciVocab, Uncased)	0.75 ±0.00	0.59 ±0.01	0.66 ±0.01	0.40 ±0.01
SPECTER2 Base	0.74 ±0.02	0.58 ±0.01	0.65 ±0.01	0.36 ±0.03
LLMs				
LLaMA 3.2 (3B Instruct) 0-Shot	0.33 ±0.00	0.91 ±0.00	0.48 ±0.00	0.33 ±0.00
+ Definition	0.41 ±0.00	0.83 ±0.00	0.55 ±0.00	0.32 ±0.00
+ Implementation	0.44 ±0.00	0.82 ±0.00	0.57 ±0.00	0.34 ±0.00
+ Clarification	0.39 ±0.00	0.82 ±0.00	0.53 ±0.00	0.33 ±0.00
+ Examples (DICE)	0.34 ±0.00	0.83 ±0.00	0.48 ±0.00	0.32 ±0.00
LLaMA 3.2 (3B Instruct) 1-Shot	0.36 ±0.00	0.65 ±0.00	0.47 ±0.00	0.29 ±0.00
+ Definition	0.52 ±0.00	0.60 ±0.00	0.56 ±0.00	0.37 ±0.00
+ Implementation	0.61 ±0.00	0.68 ±0.00	0.64 ±0.00	0.39 ±0.00
+ Clarification	0.59 ±0.00	0.66 ±0.00	0.62 ±0.00	0.38 ±0.00
+ Examples (DICE)	0.63 ±0.00	0.67 ±0.00	0.65 ±0.00	0.37 ±0.00
LLaMA 3.2 (3B Instruct) 3-Shots	0.47 ±0.00	0.53 ±0.00	0.50 ±0.00	0.30 ±0.00
+ Definition	0.54 ±0.00	0.53 ±0.00	0.54 ±0.00	0.30 ±0.00
+ Implementation	0.59 ±0.00	0.57 ±0.00	0.58 ±0.00	0.33 ±0.00
+ Clarification	0.53 ±0.00	0.49 ±0.00	0.51 ±0.00	0.29 ±0.00
+ Examples (DICE)	0.44 ±0.00	0.45 ±0.00	0.45 ±0.00	0.25 ±0.00
Mistral 7B Instruct (v0.3) 0-Shot	0.44 ±0.00	0.74 ±0.00	0.55 ±0.00	0.36 ±0.00
+ Definition	0.52 ±0.00	0.76 ±0.00	0.62 ±0.00	0.42 ±0.00
+ Implementation	0.52 ±0.00	0.73 ±0.00	0.61 ±0.00	0.37 ±0.00
+ Clarification	0.52 ±0.00	0.71 ±0.00	0.60 ±0.00	0.43 ±0.00
+ Examples (DICE)	0.46 ±0.00	0.68 ±0.00	0.55 ±0.00	0.36 ±0.00
Mistral 7B Instruct (v0.3) 1-Shot	0.46 ±0.00	0.65 ±0.00	0.54 ±0.00	0.33 ±0.00
+ Definition	0.52 ±0.00	0.71 ±0.00	0.60 ±0.00	0.39 ±0.00
+ Implementation	0.57 ±0.00	0.73 ±0.00	0.64 ±0.00	0.43 ±0.00
+ Clarification	0.58 ±0.00	0.69 ±0.00	0.63 ±0.00	0.42 ±0.00
+ Examples (DICE)	0.60 ±0.00	0.69 ±0.00	0.64 ±0.00	0.42 ±0.00
Mistral 7B Instruct (v0.3) 3-Shots	0.53 ±0.00	0.72 ±0.00	0.61 ±0.00	0.36 ±0.00
+ Definition	0.62 ±0.00	0.75 ±0.00	0.68 ±0.00	0.44 ±0.00
+ Implementation	0.66 ±0.00	0.73 ±0.00	0.69 ±0.00	0.47 ±0.00
+ Clarification	0.67 ±0.00	0.74 ±0.00	0.70 ±0.00	0.51 ±0.00
+ Examples (DICE)	0.66 ±0.00	0.73 ±0.00	0.69 ±0.00	0.49 ±0.00
GPT-4.1 (2025-04-14) 0-Shot	0.55 ±0.00	0.85 ±0.01	0.67 ±0.01	0.54 ±0.00
+ Definition	0.60 ±0.01	0.80 ±0.01	0.69 ±0.01	0.55 ±0.00
+ Implementation	0.63 ±0.00	0.78 ±0.01	0.70 ±0.00	0.56 ±0.01
+ Clarification	0.64 ±0.00	0.76 ±0.00	0.69 ±0.00	0.56 ±0.01
+ Examples (DICE)	0.65 ±0.01	0.75 ±0.00	0.70 ±0.00	0.55 ±0.01
GPT-4.1 (2025-04-14) 1-Shot	0.62 ±0.01	0.80 ±0.00	0.70 ±0.00	0.54 ±0.00
+ Definition	0.68 ±0.00	0.79 ±0.01	0.73 ±0.00	0.57 ±0.02
+ Implementation	0.69 ±0.00	0.75 ±0.01	0.72 ±0.01	0.56 ±0.00
+ Clarification	0.69 ±0.00	0.73 ±0.01	0.71 ±0.01	0.55 ±0.01
+ Examples (DICE)	0.71 ±0.01	0.74 ±0.01	0.72 ±0.01	0.55 ±0.01
GPT-4.1 (2025-04-14) 3-Shots	0.68 ±0.00	0.80 ±0.00	0.74 ±0.00	0.56 ±0.01
+ Definition	0.69 ±0.00	0.78 ±0.00	0.73 ±0.00	0.56 ±0.00
+ Implementation	0.69 ±0.00	0.76 ±0.00	0.72 ±0.00	0.56 ±0.01
+ Clarification	0.70 ±0.00	0.74 ±0.00	0.72 ±0.00	0.56 ±0.00
+ Examples (DICE)	0.69 ±0.00	0.75 ±0.00	0.72 ±0.00	0.56 ±0.01

Table 3: Detailed performances of models with ablation comparison of the typology components. Reported scores are averaged across five random seeds.