DICE: Clarifying Paper Contributions Definitions through an Extensive Analysis of ACL Rolling Review Submissions

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, unclear definitions and guidance in the current ARR typology lead to inconsistent labeling and impact acceptance rates. To support this, we introduce DICE, a novel method aimed at improving the definition of paper contributions and helping authors clearly articulate their contributions when submitting to ARR. We also analyze the current typology through an extensive annotation campaign and explore whether large language models (LLMs) can identify contributions as effectively as domain experts. Our findings suggest that refining ARR definitions could lead to a more user-friendly typology and help authors better articulate their core contributions[1].

1 Introduction

As the volume of articles continues to grow, researchers increasingly face the challenge of effectively engaging with this overwhelming influx of information. With limited time and resources, it becomes increasingly difficult to read everything in-depth. To advance research, it is crucial to stay current with the latest progress in a field. However, assessing the significance of numerous works can be time-consuming and requires substantial experience, as research fields continuously evolve and individual knowledge may quickly become outdated. The ability to quickly assess an article's content and identify key trends can provide valuable insights, enabling researchers to stay up-to-date with the field's developments.

One way to achieve this is by identifying the type of contributions an article makes to the field. 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 (Hayashi et al., 2023), or from full papers Liu et al. (2023), *structuring articles in a collection* by automatically identifying, extracting, and structuring research contributions from scholarly papers into knowledge graphs (Auer et al., 2018; D'Souza et al., 2021; Gupta 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).

Over the years, several typologies have been proposed to categorize the contributions of research papers. Although some overlap exists among them, there is still no clear consensus, with existing frameworks differing in their scope and level of granularity. The most widely adopted typology in the Natural Language Processing (NLP) field originates from the ACL Rolling Review (ARR), which introduced mandatory guidelines requiring authors to explicitly describe their contributions at submission. This initiative aims to help reviewers better identify the nature of each paper's contribution, and promote a fairer and more consistent evaluation process (Rogers et al., 2023; Bawden, 2019).

In this paper, we take a closer look at the ARR typology, starting by analyzing its current structure and highlighting the issues it presents, includ-

[1] Our annotation guidelines, code, and data are available on https://anonymous.4open.science/r/ContributionTypes-8DBE

ing the lack of clear guidelines. Through an annotation campaign, we examine the reliability of author-provided contribution labels and finally investigate whether large language models (LLMs) can be leveraged to perform this annotation task.

In summary, our contributions are:

- To address the lack of guidance on using the typology, we propose a systematic approach to defining paper contributions through a Definition–Implementation–Clarification–Excerpts (DICE) schema.
- We collect papers accepted through the ARR process and annotate their contribution types, enabling a direct comparison between expert-assigned labels and those selected by the authors.
- Our analysis reveals a lack of agreement between authors and domain experts, pointing to inconsistencies in the application of the typology. This discrepancy highlights ambiguous areas within the typology and underscores the need for more objective and standardized annotation methods.
- We explore the use of large language models (LLMs) for annotation, showing that they tend to align more closely with domain experts than with original authors, suggesting their potential to serve as substitutes for authors in this scenario.

2 Related Work

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). A comprehensive comparison of existing contribution typologies is provided in Table 5 in the appendix.

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 scientific terms and their relations to contribution sentences in order to build knowledge graphs (D'Souza and Auer, 2021; D'Souza et al., 2021; Gupta et al., 2021, 2024). 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 Figure 5). 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.

3 Dataset

3.1 Dataset Construction

We focus on articles submitted to the ACL Rolling Review (ARR) via OpenReview, especially on accepted papers that include a *contribution type* following the introduction of the new guidelines in 2023[2]. The ARR typology proposes 11 contribution types (Rogers et al., 2023), from which authors are asked to select one or more that best describe the contributions of 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, 11) publicly available software and pre-trained models.

Our initial collection comprises a total of 7 122 articles. As some papers have multiple submitted versions with potentially different contribution labels, we retain only the most recent version of each paper. To avoid bias from under-represented conferences or tracks, we include only those with more than ten accepted articles in ARR. We then split our dataset into training, validation, and test sets using an 80-10-10 ratio, ensuring label balance through multi-label stratification strategy (Sechidis et al., 2011). The resulting dataset contains 2 050 articles, each associated with author-defined paper contributions.

3.2 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.

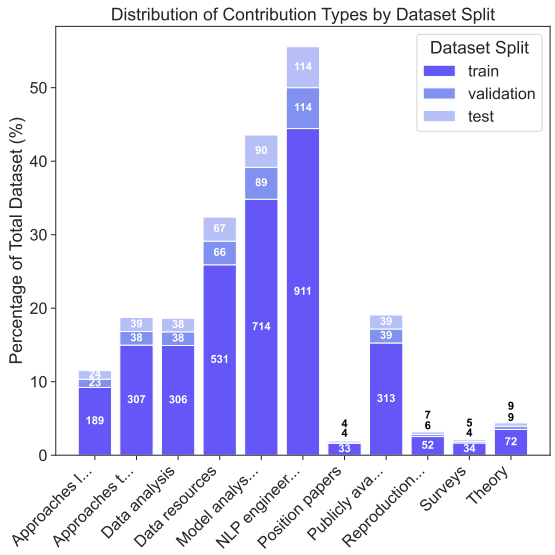


Figure 1: Distribution of paper contributions across training, validation, and test splits, highlighting the high imbalance in label representation.

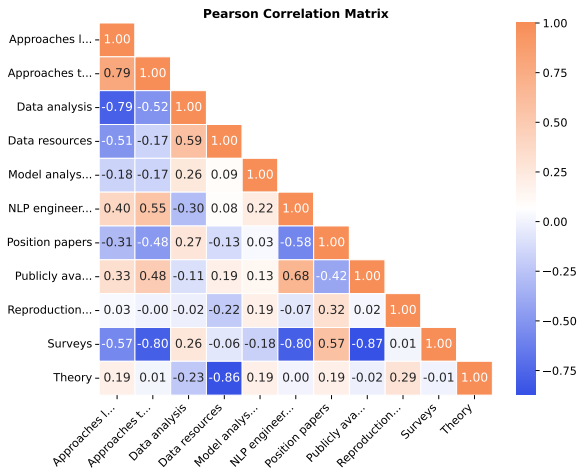


Figure 2: Correlation analysis of paper contribution labels using Pearson coefficients.

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

3.3 Correlation Analysis of Contributions

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 2). Spearman's results are presented in the appendix due to their close similarity to Pearson's (cf. Figure 10).

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.

4 Experiments

4.1 Re-annotation with Contribution Definitions

One major issue with the ARR typology is the lack of precise definitions for each paper's contributions. While some categories are explicit and can be intuitively assigned to well-defined papers (e.g., Surveys, Data Resources), others are more ambiguous, and the boundaries between specific paper contributions remain unclear. This lack of clarity poses challenges both within the ARR process, since papers with incorrectly assigned contribution types are less likely to be accepted (Rogers et al., 2023), and for automatic prediction using NLP methods such as LLM prompting, where insufficient guidance can reduce accuracy. To address this limitation, we conduct a preliminary annotation campaign with seven annotators (three researchers and four PhD students, all specializing in NLP) aimed at refining guidelines and identifying the experts with the highest agreement. We carry campaign using Label Studio (see Appendix B for guidelines).

4.1.1 Refining Definitions of ARR typology

We first clarify each paper contribution label in the ARR typology to reduce ambiguity and facilitate annotation. To do so, we review representative papers and created initial definitions based on their contribution statements. We conduct an annotation campaign on 65 full papers, deliberately oversampling underrepresented classes to ensure comprehensive coverage of the typology. 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 further refine and consolidate the typology definitions.

We structure these refined definitions using a framework we call DICE (Definition-Implementation-Clarification-Excerpts), which provides a systematic approach for constructing precise and interpretable descriptions consisting of 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. Clarification of the paper contribution, further distinguishing it from other types in cases where the label may be ambiguous.
4. Excerpts or language frequently associated with this paper contribution.

Table 1 illustrates our DICE framework for the *Approaches to Low-Resource Settings* label. Our objective is twofold: first, to refine the existing ARR typology with more precise definitions; and second, to support future iterations of the typology that reflect the evolving needs of the NLP community. The proposed definitions are generalized from observations made on the training set, with excerpts drawn from a mix of reformulated sentences and LLM-generated instances.

4.1.2 Annotation of the Test Set

After implementing the DICE definitions, we conduct an annotation campaign to label the test set of the dataset, consisting of 207 documents. Our preliminary annotation allows us to identify the four most reliable annotators, who are then selected for the main campaign. Given constrained resources and the need for rapid content assessment of entire articles, we establish an annotation procedure

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.", ...

Table 1: Example of a paper contribution type with a DICE-enhanced definition. Each row correspond to a specific component present in DICE. *Italicized* text indicates content visible to authors at the time of submission.

that considers the author as one annotator. Each document is annotated and compared against the author's original labels. When the Jaccard index between the domain expert and the author fell below 66%, an additional annotator is assigned to review the item.

4.1.3 Observations from the Annotation Campaign

We use Krippendorff's α (Krippendorff, 2004; Castro, 2017) to quantify agreement between expert annotators and original authors. An agreement score of $\alpha = 0.55$ is observed among experts on the testing set. This value indicates a complex task with probable disagreement among annotators even with enhanced guidelines. This value is below the suggested threshold of 0.67 for acceptability suggested in Krippendorff (2004). This suggests that the typology, even with enhanced definitions, is difficult to interpret. In this section, we summarize key observations from these results.

How reliable are ARR authors in their annotations? Here, we hypothesize that expert annotators represent the ground truth and evaluate the reliability of author-provided annotations in terms of agreement, where a higher value indicates greater reliability. Figure 4 presents the pairwise agreement results and offers a more nuanced picture: while author-expert agreement remains low at around 0.45, we observe that certain domain experts show substantially higher agreement with one another than with the authors. To better understand areas of disagreement, we examined instances with more than two domain experts and identified cases where authors assigned a label that was not selected by any expert. As shown in Figure 3, several categories exhibited substantial mismatch rates. Notably, some of the mismatched labels explicitly correspond to meta-papers, which is concerning since these articles represent a distinct type of literature characterized by specific style and phrasing. These results indicate that current annotation instructions and support for authors are insufficient, leading to inconsistent and unreliable labels, especially in complex categories.

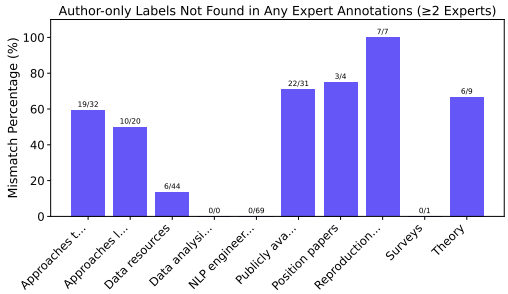


Figure 3: Percentage of author-assigned labels not identified by any expert, for instances with ≥ 2 expert raters.

To what extent does the typology exhibit ambiguity? We further explore annotator agreement through a per-label analysis, which reveals discrepancies across labels. Results are presented in Table 2. While some labels show relatively high agreement (e.g., *Surveys*, *Data resources*), others perform no better than chance (*NLP Engineering Experiment*, *Position Papers*, *Reproduction Study*). Examining the annotation ratios for individual labels (see Figure 12) further shows that certain la-

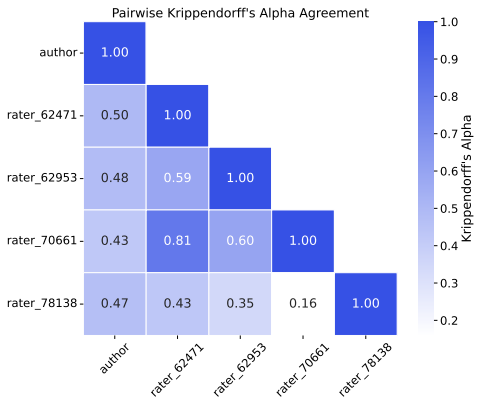


Figure 4: Pairwise Krippendorff's α between authors and domain experts on the test set.

bels are assigned at markedly different rates by experts and authors. Experts generally appear to be much more lenient to use generic labels such as *NLP Engineering experiment*, while being more sparing with meta-papers labels such as *Reproduction Study* on average. This analysis reveals that the typology exhibits significant ambiguity. Differences in labeling behavior between experts and authors suggest a need to refine label definitions and improve guidance to reduce ambiguity and enhance consistency.

Label	**Agreement** (α)
Approaches low compute settings-efficiency	0.58
Approaches to low-resource settings	0.41
Data analysis	0.40
Data resources	0.76
Model analysis and interpretability	0.24
NLP engineering experiment	0.06
Position papers	0.03
Publicly available software and/or pre-trained models	0.21
Reproduction study	-0.01
Surveys	0.78
Theory	0.32

Table 2: Krippendorff's α per contribution label.

4.2 Evaluating LLMs abilities to annotate Paper Contributions

A key advantage of LLMs is their ability to follow detailed annotation guidelines with minimal conditioning. Here, we investigate how effectively LLMs can identify paper contributions when prompted with our DICE-formatted guidelines.

4.2.1 Paper Contributions Identification

We define the identification of paper contributions as a multi-label classification problem, where, given a scientific document D (e.g., full text, abstract, specific sections) and a set of paper contributions $C = [c1, c2, ..., cn]$, the goal is to learn a function $f : D \rightarrow L$, which maps a document to the most appropriate contribution label L of the typology such as $L \subseteq C$. Each document must be associated with at least one paper contribution label, and may be associated with multiple if applicable.

4.2.2 Experimental setting

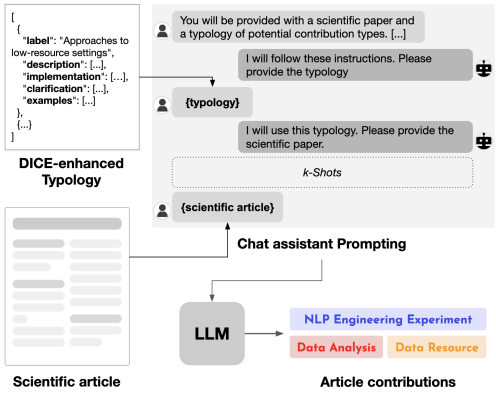


Figure 5: Workflow for identifying paper contributions via LLM prompting and DICE-enhanced definitions.

We evaluate how effectively LLMs can identify paper contributions under conditions comparable to those of human annotators. Each model is provided with the article title and abstract, the annotation guidelines, the DICE-enhanced typology, and a varying number of sample documents under few-shot conditions (see Figure 5). We also evaluate pre-trained language models (PLMs), which have shown good performance in sentence-level contribution identification (Pramanick et al., 2025), alongside several baseline approaches to gain further insights. Our experiments cover multiple architectures and model sizes, including both open-weight and API-based models. The experiments ran for under 50 GPU hours on two NVIDIA RTX 4500 GPUs (24GB each) and cost approximately $35 in OpenAI credits.

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.

[3]https://huggingface.co/meta-llama/Llama-3.2-3B
[4]https://huggingface.co/mistralai/Mistral-7B-v0.3

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.

4.2.3 Observations on LLMs Performance

We conduct evaluations on both the author-annotated test set and a consensus-based test set created through majority agreement between domain experts and authors. These experiments aim to assess whether LLMs can perform annotations on par with domain experts when provided with guidelines and detailed definitions, whether their annotations align more closely with experts or authors, and to what extent definitions help improve model performance.

Are LLMs suited for Paper Contributions identification? We begin by evaluating model performance on the author-annotated test set. Results are shown in Table 3. Interestingly, PLMs are competitive with LLMs in terms of micro-averaged F1 scores with only GPT-4.1 outperforming domain-specific PLMs. However, when considering macro-averaged F1, GPT-4.1 achieves superior performance, particularly in predicting less frequent labels (cf. Figure 6). Same phenomenon is observed to a lesser extent in other LLMs, suggesting that increased parameter size may lead to better generalization on underrepresented classes (Figure 11 in appendix). These results show promising potential for LLMs at predicting paper contributions.

To what extent do enhanced definitions improve LLMs ability to identify paper contributions?

[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

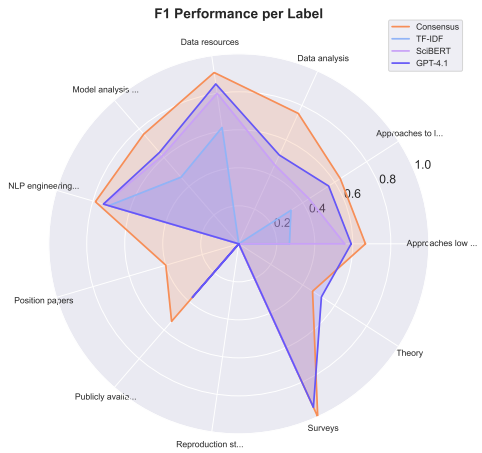


Figure 6: Performance of the best-performing AI systems on individual labels, compared to expert consensus on the Authors test set. Results are reported as F1 scores. GPT-4.1 uses the DICE-enhanced typology and a 3-shot prompting strategy.

We evaluate several models across different few-shot prompting strategies using different DICE configurations (Definition, Implementation, Clarification, Excerpts). Detailed results are provided in Table 6 (Appendix). Since the expert consensus did not include the label *Reproduction Study* in the test set annotations, model predictions for this label are not counted as errors in the reported results, which in turn lead to an increased macro-average score. We compute the performance gains (Δ) of each DICE strategy over the labels-only typology for LLMs. For comparison, results from other methods, including PLMs and baseline approaches, are also provided. In general, DICE-enhanced typologies consistently improve model performance, with only a notable exceptions observed for the best-performing model under 3-shot prompting. Importantly, the optimal DICE configuration varies with the model architecture and few-shot strategy. While some models perform best using the full DICE setup, most achieve better results with subsets such as Definition–Implementation or Definition–Implementation–Clarification. This suggests that, while excerpts illustrating how to detect contribution types may help expert annotators better define category boundaries, they do not necessarily improve LLM performance. This insight is interesting, as it suggests that the optimal strategy for leveraging LLMs as annotators is largely model-dependent and requires careful tuning.

Can LLMs serve as reliable alternatives to au-

	P	R	F1 $_{\text{micro}}$	F1 $_{\text{macro}}$
Experts Consensus	0.77	0.74	0.75	0.63
Baselines				
Random	0.19	0.56 ± 0.02	0.28 ± 0.01	0.25 ± 0.01
TF-IDF	0.67	0.39	0.49	0.22
PLMs				
BERT Base (Uncased)	0.57 ± 0.02	0.63 ± 0.03	0.60	0.35 ± 0.02
RoBERTa Base	0.61 ± 0.01	0.61 ± 0.02	0.61 ± 0.01	0.40 ± 0.03
SciBERT (SciVocab, Uncased)	0.60 ± 0.01	0.64 ± 0.02	0.62 ± 0.01	0.41 ± 0.01
SPECTER2 Base	0.61 ± 0.02	0.62 ± 0.02	0.61 ± 0.01	0.40 ± 0.02
Open-Weight LLMs				
LLaMA 3.2 (3B Instruct) $_{\text{0-Shot}}$	0.32	0.76	0.45	0.31
LLaMA 3.2 (3B Instruct) $_{\text{1-Shot}}$	0.52	0.53	0.53	0.31
LLaMA 3.2 (3B Instruct) $_{\text{3-Shot}}$	0.43	0.42	0.42	0.24
Mistral 7B Instruct (v0.3) $_{\text{0-Shot}}$	0.42	0.62	0.50	0.35
Mistral 7B Instruct (v0.3) $_{\text{1-Shot}}$	0.50	0.56	0.53	0.39
Mistral 7B Instruct (v0.3) $_{\text{3-Shot}}$	0.55	0.59	0.57	0.43
GPT-4.1				
GPT-4.1 (2025-04-14) $_{\text{0-Shot}}$	0.62	0.67	**0.64**	**0.52**
GPT-4.1 (2025-04-14) $_{\text{1-Shot}}$	0.64 ± 0.01	0.63 ± 0.01	**0.64** ± 0.01	0.50 ± 0.01
GPT-4.1 (2025-04-14) $_{\text{3-Shot}}$	0.63	0.65	**0.64**	0.51

Table 3: Performance of paper contributions identification on Authors test set. Results report Precision (P), Recall (R), and F1-Score (F1), with F1 calculated using both micro and macro averages. $\pm$ indicates standard deviation. LLMs use DICE-enhanced typology. PLMs and GPT-4.1 results average over 5 seeds; open-weight LLMs use one seed (Greedy Sampling) due to reproducibility. Underline marks the best within each model category; **bold** marks the best overall.

Annotation Sets	α
Experts $\cup$ Authors	0.54
Experts $\cup$ GPT-4.1$_{\text{Labels Only}}$	0.46
Experts $\cup$ GPT-4.1$_{\text{DICE}}$	0.49
Authors $\cup$ GPT-4.1$_{\text{Labels Only}}$	0.37
Authors $\cup$ GPT-4.1$_{\text{DICE}}$	0.42

Table 4: Krippendorff's α is computed to assess agreement across sets of annotations by human annotators, authors, and GPT-4.1 (0-Shot) on test set.

thors for annotation? To address this question, we employ GPT-4.1 as an LLM-based annotator to compute additional agreement scores. With this, we aim to evaluate how closely the model aligns with human annotators and its consistency as an alternative annotation source. The instructions and prompting procedure are detailed in Appendix C. We use GPT-4.1 with a 0-shot prompting strategy as it performs best on the Author test set. We conduct experiments using both the DICE-enhanced typology and a labels-only control. This setup reflects the expectation that, by following the same definitions as the experts, the model will align more closely with expert annotations than the authors do. However, even without explicit definitions, the model demonstrates stronger agreement with expert labels. These results suggest that LLMs could be more reliable annotators than relying solely on author-provided labels granted correct insctructions and clear definitions are provided (Table 4).

5 Conclusion and Discussion

This paper highlights that clearer, more precise definitions can improve the ARR contribution typology. We find through an extensive analysis that contribution labels are often applied inconsistently, reducing their value in review. The DICE method offers a systematic approach to defining paper contributions, improving expert agreement while revealing ongoing ambiguities. Our evaluation shows LLMs as promising annotators,though not yet matching human accuracy. These models align better with expert than author labels, indicating potential for annotating paper contributions in rapidly evolving fields like NLP. We recommend refining the contribution typology to reduce ambiguity, starting with clearer definitions and guidelines, and potentially reconsidering the categories themselves. Incorporating LLM-assisted contribution identification may further support the review process amid ARR's growing volume. We hope this work inspires continued efforts to improve typologies and strengthen the transparency and objectivity of scientific publishing.

Limitations

Enhancing ARR Typology Definitions. We proposed DICE as a method allowing for maintainability in evolving fields. However, specified definitions in the context of that study are largely dependent on unveiling typology definition based on author annotated articles. Despite a preliminary campaign aimed at structuring enhanced definition, ARR editors might have indented something else that our comprehension of the paper contributions and therefore bias agreement toward our guidelines. Proposed DICE definitions could be further modified granted more extensive analyses are conducted and more generally depending on the reviewers needs.

Annotation Campaign Exhaustibility. Re-annotation of the dataset is limited to the testing set due to limited resources and a time consuming task. Because of that low amount, analysis over the typology remains limited and more definitive conclusion would probably require annotating the full dataset with more annotators.

Processed Documents. The task of identification of paper contributions is conducted using articles title and abstract in these experiments. This is mainly due to PLMs context size limitations and the aim to evaluate models in similar scenarios. However, these partial documents could be missing contribution information, as annotators expressed that they are generally using title, abstract, introduction and conclusion sections for that purpose.

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. Due to its nature, the annotation campaign is conducted over the span of 2 weeks with annotators spending on average 5 minutes to assess the paper contributions of a single document. Participants were compensated through non-financial means.

Use of generative AI in Scientific NLP. Use of generative AI models for assessing paper contributions presents inherent risk related to the use of such technology. Generative AI systems used in scholarly contexts must consider the risks of false positives, false negatives, overlooked contributions, and inaccurate representation of articles.

Use of LLMs for the making of this study. In this study, LLMs were used for writing style reformulation, code refactoring and generating a few generic excerpts for DICE definition. All ideas, analyses and experiments were conducted without the use of 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`.

Francisco Charte, Antonio Rivera, María José del Jesus, and Francisco Herrera. 2013. A first approach to deal with imbalance in multi-label datasets. In *Hybrid Artificial Intelligent Systems*, pages 150–160, Berlin, Heidelberg. Springer Berlin Heidelberg.

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.

Grigorios Tsoumakas and Ioannis Katakis. 2007. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, 3(3):1–13.

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

A Supplementary Material

A.1 Measuring Imbalance of the Dataset

Label cardinality (LC) and *label density* (LD) are metrics allowing to characterize the multi-label nature of the dataset (Tsoumakas and Katakis, 2007). *Label cardinality* represents the average number of labels assigned to each instance where D is a dataset, Y_i denotes the set of labels assigned to instance i, and L is the set of unique labels:

$$\mathrm{LC}(D) = \frac{1}{|D|} \sum_{i=1}^{|D|} |Y_i|, \tag{1}$$

(2)

Label density is the fraction of the total label space assigned per instance, such as:

$$\mathrm{LD}(D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i|}{|L|} = \frac{\mathrm{LC}(D)}{|L|}. \quad (3)$$

Label imbalance in a multi-label dataset can be quantified using the *label-wise imbalance ratio* (IRperLabel) and its average, the *Mean Imbalance Ratio per label* (MeanIR) where $freq(l)$ is the frequency of a label l in the dataset and $\max_{k \in L} freq(k)$ the frequency of the most common label k (Charte et al., 2013):

$$freq(l) = \left|\{i \in D \mid l \in Y_i\}\right|, \quad l \in L, \quad (4)$$

$$\mathrm{IRperLabel}(l) = \frac{\max_{k \in L} freq(k)}{freq(l)}, \quad l \in L, \quad (5)$$

$$\mathrm{MeanIR} = \frac{1}{|L|} \sum_{l \in L} \mathrm{IRperLabel}(l). \quad (6)$$

To characterize the multi-label nature of the dataset, we compute its *label cardinality* (LC) and *label density* (LD) as well as the *imbalance ratio per labels* (IRperLabel & MeanIR). On average, each instance is associated with $\mathrm{LC} = 2.11$ labels ($\sigma = 1.06$), indicating a moderate degree of multi-labeling in the data. The dataset exhibits a low label density of $\mathrm{LD} = 0.192$, indicating that only a small fraction of the typology is typically assigned to each instance. Additionally, a mean imbalance ratio of $\mathrm{MeanIR} = 9.29$ indicates that some labels are much more frequent than others, reflecting a considerable imbalance in the label distribution (see Figure 7 in appendix for a more detailed view).

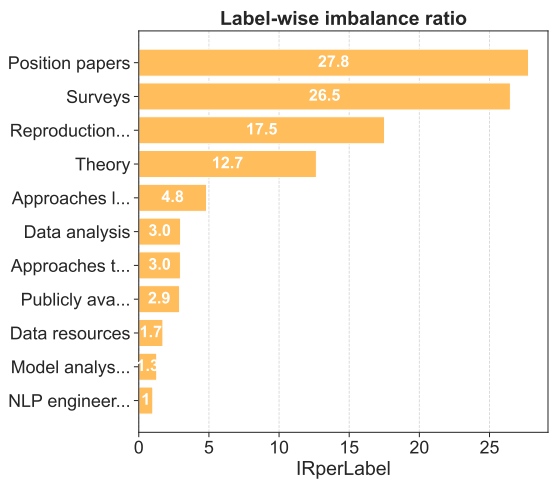


Figure 7: Imbalance ratios (IRperLabel) for paper contribution labels, where the most frequent label is 1 and others have higher ratios due to lower frequency.

B Annotation Campaign

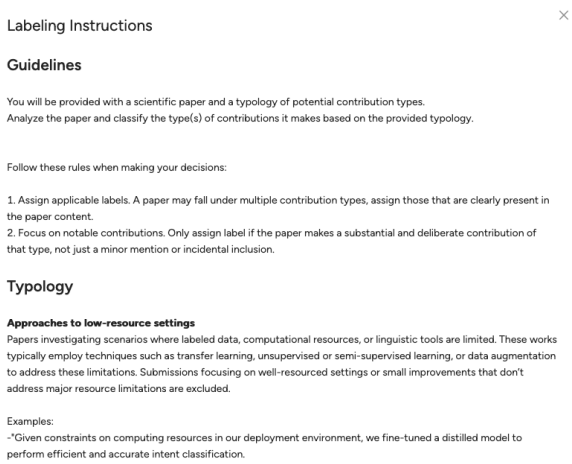


Figure 8: Guidelines provided to annotators.

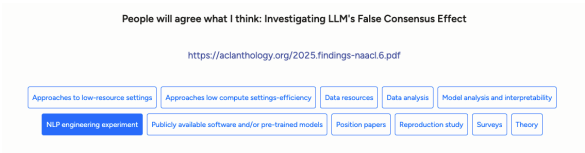


Figure 9: Annotation campaign interface

C LLM Prompts & Instructions

C.1 Guidelines 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:

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.
```

C.2 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:"},
])
```

C.3 Typology

Full DICE-enhanced typology is accessible at: https://anonymous.4open.science/r/ContributionTypes-8DBE/experiments/templates/typology.json. Refer to Table 1 for a partial illustration.

D Supplementary Tables & Figures

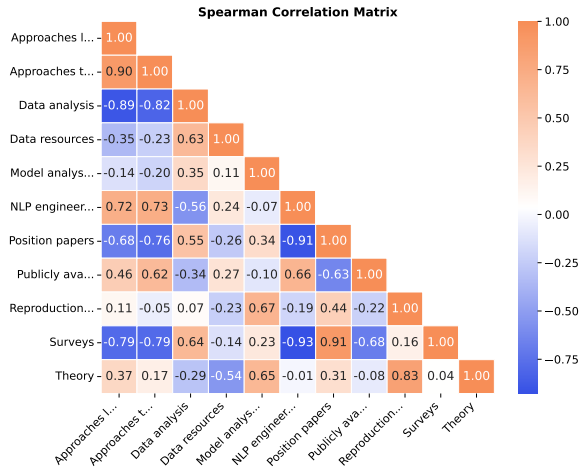


Figure 10: Correlation analysis of paper contributions labels using Spearman coefficient.

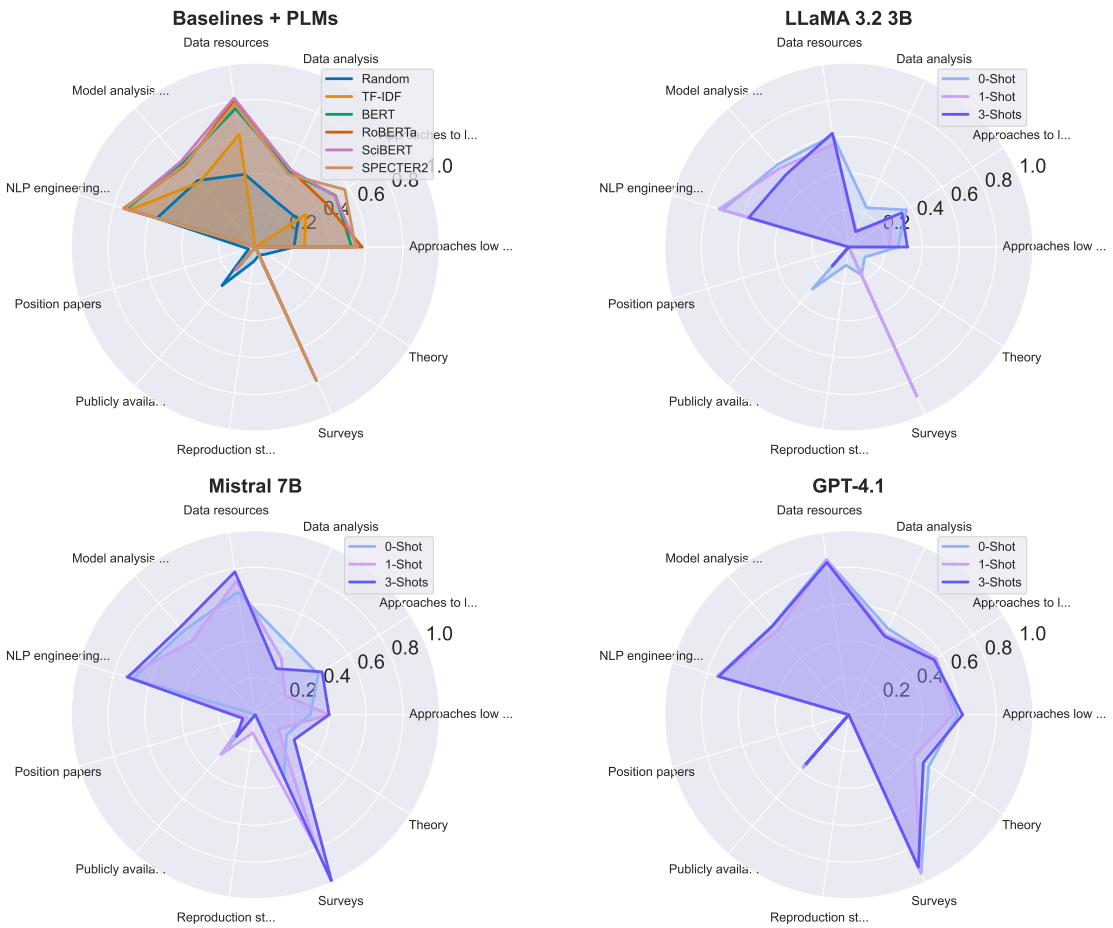
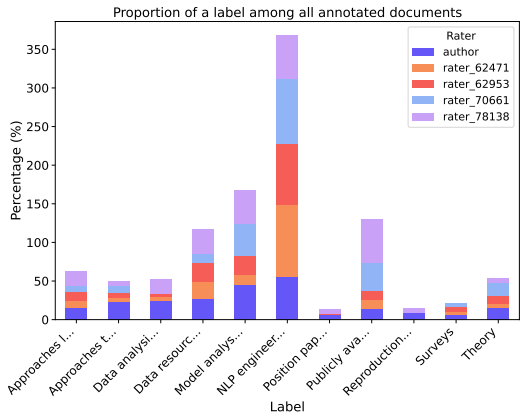
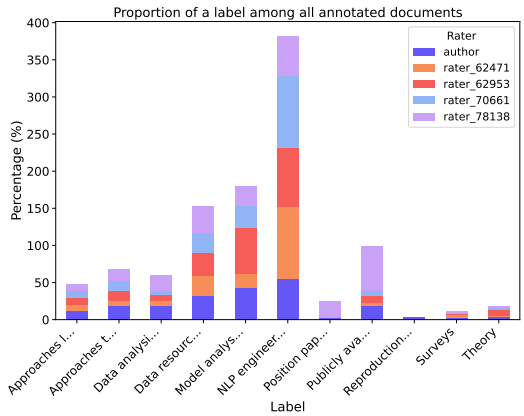


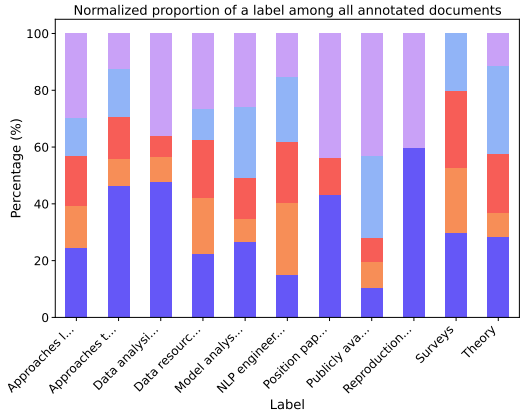
Figure 11: Detailed performance of AI systems on individual labels measured in F1 score. Result indicates LLMs are performing better than other approaches.



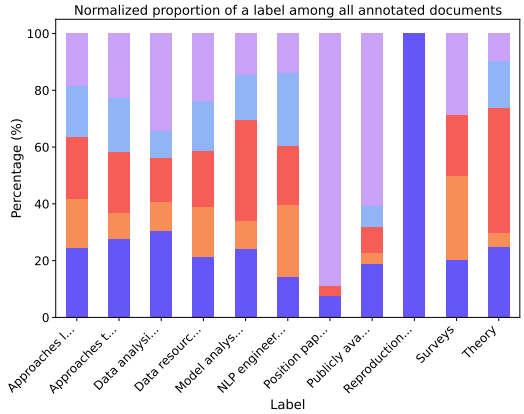
(a) Preliminary



(b) Full testing set



(c) Preliminary



(d) Full testing set

Figure 12: 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).

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 5: 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.

	Authors				Expert Consensus			
	F1 micro	**F1 macro**	**Δ F1micro**	**Δ F1macro**	**F1 micro**	**F1 macro**	**Δ F1micro**	**Δ F1macro**
Authors	1.00	1.00			0.76*	0.69*		
Experts Consensus	0.75*	0.63*			1.00	1.00		
Baselines								
Random	0.28	0.25			0.29	0.23		
TF-IDF	0.49	0.22			0.62	0.27		
PLMs								
BERT Base (Uncased)	0.60	0.35			0.67	0.39		
RoBERTa Base	0.61	0.40			0.67	0.44		
SciBERT (SciVocab, Uncased)	0.62	0.41			0.69	0.46		
SPECTER2 Base	0.61	0.40			0.69	0.45		
LLaMA 3.2 (3B Instruct)								
Labels + 0-Shot	0.45	0.34			0.48	0.33		
+ Definition	0.48	0.30	**+0.03**	-0.04	0.55	0.32	**+0.07**	-0.01
+ Implementation	0.49	0.34	**+0.04**	-	0.57	0.34	**+0.09**	**+0.01**
+ Clarification	0.47	0.32	**+0.02**	-0.02	0.53	0.33	**+0.05**	-
+ Examples	0.45	0.31	-	-0.03	0.48	0.32	-	-0.01
Labels + 1-Shot	0.43	0.28			0.47	0.29		
+ Definition	0.47	0.33	**+0.04**	**+0.05**	0.56	0.37	**+0.09**	**+0.08**
+ Implementation	0.55	0.36	**+0.12**	**+0.12**	0.64	0.39	**+0.17**	**+0.10**
+ Clarification	0.52	0.35	**+0.09**	**+0.07**	0.62	0.38	**+0.15**	**+0.09**
+ Examples	0.53	0.31	**+0.10**	**+0.03**	0.65	0.37	**+0.18**	**+0.08**
Labels + 3-Shot	0.44	0.27			0.50	0.30		
+ Definition	0.45	0.26	**+0.01**	-0.01	0.54	0.30	**+0.04**	-
+ Implementation	0.50	0.29	**+0.06**	**+0.02**	0.58	0.33	**+0.08**	**+0.03**
+ Clarification	0.44	0.25	-	-0.02	0.51	0.29	**+0.01**	-0.01
+ Examples	0.42	0.24	-0.02	-0.03	0.45	0.25	-0.05	-0.05
Mistral 7B Instruct (v0.3)								
Labels + 0-Shot	0.52	0.37			0.55	0.36		
+ Definition	0.53	0.40	**+0.01**	**+0.03**	0.62	0.42	**+0.03**	**+0.05**
+ Implementation	0.51	0.37	-0.01	-	0.61	0.37	**+0.06**	**+0.01**
+ Clarification	0.54	0.44	**+0.02**	**+0.07**	0.60	0.43	**+0.05**	**+0.07**
+ Examples	0.50	0.35	-0.02	-0.02	0.55	0.36	-	-
Labels + 1-Shot	0.47	0.30			0.55	0.34		
+ Definition	0.51	0.35	**+0.04**	**+0.05**	0.60	0.39	**+0.05**	**+0.05**
+ Implementation	0.53	0.37	**+0.06**	**+0.07**	0.64	0.42	**+0.09**	**+0.07**
+ Clarification	0.53	0.39	**+0.06**	**+0.09**	0.64	0.43	**+0.09**	**+0.08**
+ Examples	0.53	0.39	**+0.06**	**+0.09**	0.64	0.43	**+0.09**	**+0.08**
Labels + 3-Shot	0.54	0.36			0.61	0.36		
+ Definition	0.57	0.38	**+0.03**	**+0.02**	0.68	0.44	**+0.07**	**+0.08**
+ Implementation	0.57	0.40	**+0.03**	**+0.04**	0.69	0.47	**+0.08**	**+0.11**
+ Clarification	0.59	0.43	**+0.05**	**+0.07**	0.70	0.51	**+0.09**	**+0.15**
+ Examples	0.57	0.43	**+0.03**	**+0.07**	0.69	0.49	**+0.08**	**+0.13**
GPT-4.1 (2025-04-14)								
Labels + 0-Shot	0.61	0.50			0.67	0.54		
+ Definition	0.63	0.52	**+0.02**	**+0.02**	0.69	0.55	**+0.02**	**+0.01**
+ Implementation	0.64	0.52	**+0.03**	**+0.02**	0.70	0.56	**+0.03**	**+0.02**
+ Clarification	0.65	0.52	**+0.04**	**+0.02**	0.69	0.56	**+0.02**	**+0.02**
+ Examples	0.64	0.52	**+0.03**	**+0.02**	0.70	0.55	**+0.03**	**+0.01**
Labels + 1-Shot	0.61	0.49			0.70	0.54		
+ Definition	0.64	0.52	**+0.03**	**+0.03**	0.73	0.57	**+0.03**	**+0.03**
+ Implementation	0.63	0.50	**+0.02**	**+0.01**	0.72	0.56	**+0.02**	**+0.02**
+ Clarification	0.64	0.51	**+0.03**	**+0.02**	0.71	0.55	**+0.01**	**+0.01**
+ Examples	0.64	0.50	**+0.03**	**+0.01**	0.72	0.55	**+0.02**	**+0.01**
Labels + 3-Shot	0.63	0.50			0.74	0.56		
+ Definition	0.65	0.51	**+0.02**	**+0.01**	0.73	0.56	-0.01	-
+ Implementation	0.64	0.51	**+0.01**	**+0.01**	0.72	0.56	-0.02	-
+ Clarification	0.65	0.51	**+0.02**	**+0.01**	0.72	0.56	-0.02	-
+ Examples	0.64	0.51	**+0.01**	**+0.01**	0.72	0.56	-0.02	-

Table 6: Measurement of the impact of different DICE strategies with LLMs. Values are measured in F1 (micro and macro) as well as the difference (Δ) between Labels Only and DICE definitions. DICE strategies are additive in these experiments.Underline indicates best DICE strategy for each model configuration; **bold** marks positive Δ. *Discrepancies between the Authors and Consensus scores are partly due to the absence of the "Reproduction Study" label in expert annotations, slightly favoring evaluation on the Consensus set; this mainly affects the macro F1 score.