Dataset Cartography for Implicit Discourse Relation Recognition: Promises and Pitfalls

Daniil Ignatev[1], Denis Paperno[1], Massimo Poesio[1,2],

[1]Utrecht University, [2]Queen Mary University of London,

Correspondence: d.ignatev@uu.nl

Abstract

Crowdsourced data for implicit discourse relation recognition, IDRR, has been shown to contain both plausible interpretations and noisy annotations. We present a case study of dataset cartography (Swayamdipta et al., 2020) on IDRR-focused DiscoGeM corpus (Scholman et al., 2022). Our findings show that error identification via low confidence proves unreliable, as confidence is strongly affected by label rarity. However, high-confidence datapoints reveal a different use case: auditing the cue-rich regions of the dataset. Our lexical probe demonstrates an association between high confidence items and (mostly temporal) intra-argument cue words. Dataset cartography can thus serve a diagnostic of cue-driven *easy-to-learn* cases, which need to be balanced out to ensure the robustness of IDRR learning.

1 Introduction

The task of discourse annotation remains a challenge in data collection methodology. Scalable annotation solutions (crowdsourcing; silver data) cannot fully ensure the desired high data quality; on the other hand, expert annotation lacks scalability due to high costs (Das et al., 2017; Scholman and Demberg, 2017; Yung et al., 2019).

With crowdsourced discourse annotation, maintaining a balance between a descriptive and a prescriptive approach is particularly challenging. Although there is a need for error filtering, a prescriptive approach is hard to justify, as the annotation schemes tend to treat diverging interpretations permissively (Prasad et al., 2017), while errors are hard to tease apart from plausible readings:

(1) Arg1: *This time he did not go back to the boat.* Arg2: He sat down in the dark by Bilbo. A) SUBSTITUTION / B) CAUSE

Dataset cartography (Swayamdipta et al., 2020) would appear to be a useful analytic tool in these circumstances, as it has proven effective at detecting noisy labels in crowdsourced datasets (Klie et al., 2023). Dataset cartography hypothesizes that data instances that appear *hard-to-learn* from their training dynamics often coincide with annotation errors (see Anand et al., 2024 for a counter-argument). As such, they could be removed to improve model convergence or avoided in active learning (Zhang and Plank, 2021). Likewise, *easy-to-learn* data should not be prioritized in training.

For our case study, we use DiscoGeM (Scholman et al., 2022; Yung and Demberg, 2025), a large corpus annotated in the Penn Discourse Treebank standard (PDTB, Prasad et al., 2019). Notably, unlike PDTB or most other discourse resources, DiscoGeM retains a wide spectrum of crowd-collected labels for each item: 10+ labels. Thus, it offers a realistic image of label distribution in the PDTB standard; similarly, it offers insights into crowd workers' individual labeling behavior. At the same time, DiscoGeM incorporates noisy labels in some proportion, which stresses the need for automatic assessment of data quality (Yung et al., 2026).

DiscoGeM challenges dataset cartography in several ways. First, prior studies have shown that diverging annotations in DiscoGeM can at least partly be attributed to individual, subjective biases (Pyatkin et al., 2023). Anand et al. (2024) show that items with subjective variation, while not erroneous, can still fall into the *hard-to-learn* region of single-label classifiers; in response, we use both single-label and annotator-aware classifiers (Cabitza et al., 2023).

A second challenge is posed by label imbalance, as infrequent labels often constitute a small fraction of the DiscoGeM data (Yung and Demberg, 2025). E.g., at level-3, the most granular level of the PDTB sense hierarchy, 9 out of 29 classes are picked by the majority in less than 1% of cases. It can thus be expected that infrequent classes will display low confidence, which indeed occurs in practice. Due

Proceedings of the 2nd Joint Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences and Computational Models of Reference, Anaphora and Coreference (CODI-CRAC 2026), pages 53–64

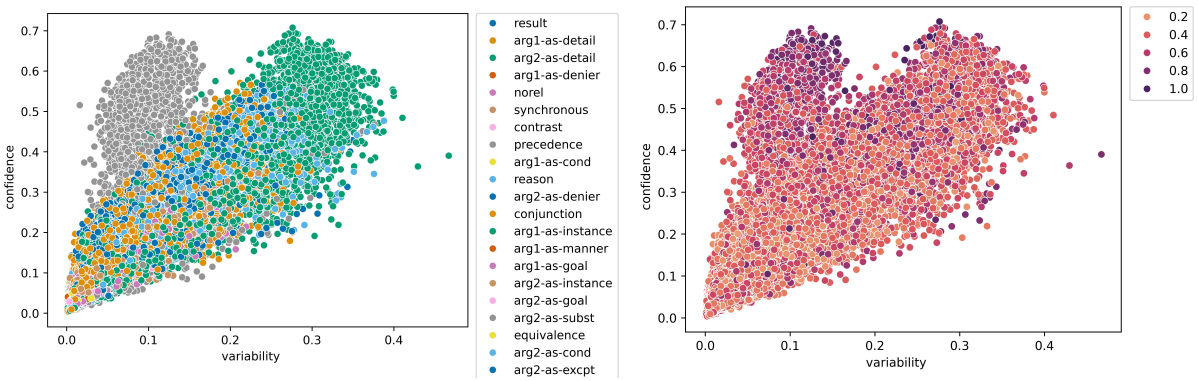

Figure 1: Data maps for the Annotator Embeddings model (Deng et al., 2023) under high sense granularity (level-3). X- and Y-axis correspond to **variability** and **confidence**, respectively. Plot colors, left: labels; right: agreement (ratio of the majority label). We note that high-agreement area is dominated by *precedence* (gray color).

to subjectivity and label imbalance, *hard-to-learn* IDRR data cannot be pruned indiscriminately, as it combines genuinely erroneous, subjective and merely infrequent labels.

At the same time, our experiments also help detect a persistent *easy-to-learn* subset that strongly affects model training and evaluation – namely, instances of the *precedence* relation supported by strong lexical or syntactic cues; e.g., a temporal dependent clause at the start of Arg2:

(2) Arg1: *...While it had devastated me and stopped me in my tracks, for Maxime, it had broken down barriers, released him from a straitjacket, and left him free to write his own story.* Arg2: After what happened, I was never the same...

This set is globally distinguishable, but more so when isolated: e.g., when training at levels 2 and 3 using the annotator-aware architectures; this accounts for annotators not following the cues.

Overall, we make the first attempt at studying crowdsourced IDRR data through the lens of training dynamics. We demonstrate that filtering the *hard-to-learn* data in DiscoGeM proves too difficult, but *easy-to-learn* instances expose heuristics that should be handled explicitly in data collection and in model training. Finally, we observe that perspectives and class granularity both determine the sharpness of class boundaries, which effectively enables improved confidence on granular labels.

2 Background

Dataset cartography was the first method in the larger training dynamics family; these methods track how specific examples affect the trained model across epochs (Swayamdipta et al., 2020). Cartography statistics describe each training example and derive from the predicted probability of the correct class, P_{corr}. This translates into several variables: **confidence** (mean P_{corr} across epochs), **variability** (standard deviation of P_{corr}), **correctness** (ratio of epochs where $P_{corr} > 0.5$). All training data fall into one of the three categories: *hard-to-learn* (low confidence & variability), *easy-to-learn* (high confidence, low variability), *ambiguous* (high variability). Crucially, Klie et al. (2023) found that data map **confidence** can reliably estimate annotation quality across a variety of NLP tasks, often outperforming other error detectors.

At the same time, Anand et al. (2024) found that data maps learned from majority labels tend to categorize items with high annotator disagreement – i.e., genuinely ambiguous data – as *hard-to-learn*. As a result, filtering based on this categorization ignores valuable information. This claim is in line with prior research showing that aggregation by majority voting can lead to information loss (Pavlick and Kwiatkowski, 2019; Larimore et al., 2021; Plank, 2022) and that models that learn from raw annotations may be able to mitigate these effects (Uma et al., 2021; Frenda et al., 2024).

Such disagreement-aware methods have also been shown to extend well to PDTB-style discourse parsing and even offer advantages over majority-label training (Yung et al., 2022; Pyatkin et al., 2023; Costa and Kosseim, 2024; Long et al., 2024; Yung et al., 2026). These findings can be grounded in the relevant theory, as PDTB allows for concurrent readings (Prasad et al., 2017), and, in practice, annotators tend to infer several concurrent relations (Rohde et al., 2018) or pick diverging, but mutually compatible labels (Yung et al., 2019).

Our analysis of the *easy-to-learn* subset deals with cue-based biases, which have been studied

in detail for natural language inference and SNLI specifically (Bowman et al., 2015). These include both lexical (Gururangan et al., 2018; Feng et al., 2019) and syntactic (McCoy et al., 2019) heuristics that the models acquire over the course of training.

3 Experimental setup

DiscoGeM 1.5 (Scholman et al., 2022; Yung and Demberg, 2025) is the largest crowdsourced corpus of PDTB-style implicit discourse relations that retains unaggregated crowd labels. The data collection procedure relied on the well-documented connective insertion method (Yung et al., 2019).

Version 1.5 focuses on English, the data amounting to 6489 text pairs labeled by 164 crowd workers with each pair having 10 crowd labels on the average. We focus on the canonical train subset of 4560 items.

We report our results at three levels of sense granularity in PDTB: level-3 (29 senses), level-2 (17 senses) and level-1 (5) (Webber et al., 2019). Disagreement is especially visible on level-3 ($\kappa = 0.404$), since the workers have to choose between semantically similar alternatives: for example, *precedence*, *succession* or *synchronous* within the coarse *temporal* group. Although the coarser levels are assumed to be easier to learn (Xiang and Wang, 2023), they also group heterogeneous data, as each granular class is associated with its own signals.

Our models for data mapping are based on DeBERTa-v3-base (He et al., 2023), an encoder-type model. For each of the three PDTB levels, we train two models: **SG**, a standard majority-label classifier, and **AE**, an instance of the Annotator Embeddings model (Deng et al., 2023), which learns annotator-specific labeling behavior from raw labels paired with annotator ids. We employ each model to obtain **confidence**, **correctness** and **variability**.

	Level-1	**Level-2**	**Level-3**
LowAgree.OR.SG	1.75	1.48	1.51
RareLabel.OR.SG	1.74	0.54	3.75
LowAgree.OR.AE	1.13	1.31	1.46
RareLabel.OR.AE	7.47	17.21	25.8

Table 1: Odds ratios for enrichment in the low-confidence (*hard-to-learn*) quartile.

4 Results

4.1 Cartography plots

Data maps project each training data point onto the **variability / confidence** coordinate plane. As the definitions suggest (see Section 2), *easy-to-learn* items occupy the upper left corner of the map, while *hard-to-learn* occupy the lower left part. Highly variable data, which belongs to neither category, forms the right part of the map. The data points can be colored to encode additional variables, such as their class labels or binned agreement scores, i.e., the proportion of votes that agrees with the majority.

The visual analysis of the obtained data maps reveals several trends. First, in Figures 1, 2, and 3, the data maps all exhibit the commonly seen wedge shape. Another global trend, characteristic of data maps (Swayamdipta et al., 2020), is the interaction between agreement and **confidence**, apparent in the *easy-to-learn* region. However, the maps for AE exhibit clearer *easy-* and *hard-to-learn* regions with low variability.

On the other hand, **confidence** visibly interacts with the class structure. Independently of the training method, *expansion* as the most frequent label dominates the *easy-to-learn* region at level-1, even surpassing the temporal relations. On the other hand, at levels 2 and 3, where *expansion* splits up into smaller classes, *asynchronous* and *precedence* consistently emerge as the easiest to learn and visibly coincide with high agreement.

For AE, the data maps offer an insight into the composition of the *hard-to-learn* region, as it can be seen that infrequent labels, such as *no relation* on level-1, largely have low **confidence** and **variability**. Comparing AE with the majority-label classifier, we note that the latter yields a sparsely populated *hard-to-learn* region that has no predom-

	Level-1	**Level-2**	**Level-3**
SG			
Agree coef.	0.210	0.240	0.219
Freq. coef.	0.420	−0.457	−0.445
R^2	0.219	0.244	0.137
AE			
Agree coef.	−0.101	0.334	0.338
Freq. coef.	1.900	1.061	2.362
R^2	0.575	0.470	0.455

Table 2: Regression coefficients and R^2 for OLS predicting confidence from agreement and label frequency. Both variables are significant across all levels with $p \leq 0.01$.

inant label at any of the 3 levels. Likewise, due to high **variability**, no item at levels 2 and 3 can be categorized as strictly *easy-to-learn*. At the same time, AE displays clearer *easy/hard-to-learn* subsets with fewer data points concentrated in the highly variable region. The quantitative results also reveal a drastic difference between SG and AE.

4.2 Cartography statistics

To complement the results of visual interpretation in Section 4.1, we quantify how **confidence** relates to item-wise agreement and label frequency, two properties that emerged as impactful in our visual analysis. In particular, we roughly define the *hard-to-learn* subset as the bottom quartile of **confidence** values and report enrichment as an odds ratio (OR) relating the prevalence of low agreement items/rare labels inside vs. outside that quartile.

As shown in Table 1, low agreement is only weakly enriched among low-confidence items across both model types, SG and AE (ORs 1.13–1.75), suggesting that disagreement alone does not define the *hard-to-learn* subset. In contrast, label rarity shows a much stronger effect, especially for AE. For SG, rare labels are not consistently concentrated in low confidence (OR = 1.74, 0.54, and 3.75 across levels 1-3), whereas for AE they are strongly enriched at every level, with the effect monotonically increasing with label granularity (OR = 7.47, 17.21, and 25.8). This pattern is also reflected in Figure 6, where rare labels occupy markedly lower confidence ranges, particularly for AE; the Mann-Whitney test shows a significant difference in how the **confidence** values are distributed.

Elaborating on this result, we further model **confidence** globally as a function of agreement and frequency using OLS regression (Table 2). Both predictors are significant across all levels, but their contribution differs sharply depending on model type. For SG, agreement has a small positive coefficient throughout, whereas frequency turns negative at levels 2 and 3, suggesting unstable interactions between majority-label training and fine-grained class structure. For AE, frequency is the dominant predictor, with large positive coefficients across all levels and substantially higher explained variance ($R^2 = 0.455$-0.575 vs. 0.137-0.244 for SG). Overall, low confidence in DiscoGeM appears to reflect sparsity and ambiguity more than annotation error.

Metric	**Level-1**	**Level-2**	**Level-3**
Acc.Maj	0.86 ± 0.02	0.90 ± 0.02	0.90 ± 0.01
F1.Maj	0.62 ± 0.03	0.60 ± 0.06	0.61 ± 0.02
Acc.Raw	0.87 ± 0.00	0.93 ± 0.00	0.93 ± 0.00
F1.Raw	0.67 ± 0.00	0.71 ± 0.01	0.68 ± 0.02

Table 3: Results of the lexical probe for *temporal*, *asynchronous*, and *precedence* for levels 1, 2, and 3 respectively. Metrics averaged across 5 cross-validation folds. Macro-F1 ≥ 0.5 surpasses the majority label baseline.

4.3 Auditing high-confidence labels

The cartography statistics point at temporal labels as the most consistent *easy-to-learn* cluster in DiscoGeM; particularly, *precedence* at level-3. Meanwhile, they offer no explanation for the trend. One likely reason is that models may exploit local lexical cues, and the same cues are also available to annotators.

In a further experiment, we address this hypothesis by training a lexical probe: a one-vs-all logistic regression over unigram and bigram features. We insert a special boundary token, **sep**, between Arg1 and Arg2, thus enabling the bigram model to capture Arg2-initial keywords. This setup is repeated for raw labels and majority-aggregated labels across 5 cross-validation folds. We report the results in Table 3.

Across the three granularity levels, the lexical probe retains high accuracy along with Macro-F1 that surpasses the majority label baseline (≈ 0.5 for all settings). The accuracy shifts across the three levels likely reflect the changes in label balance, as the *temporal* class is more populated than *asynchronous* or *precedence*: 13 vs 9%. On the other hand, the F1 score is greater or equal at granular levels, especially for unaggregated labels; this aligns with the clearer clustering of temporal items in the data maps, especially at finer sense levels. Overall, the metrics suggest that temporal relations in DiscoGeM are at least partly recoverable from surface features.

By inspecting the highest positive regression coefficients, we identify the ngrams that are most predictive of the temporal class: **sep after**, **sep when**, **sep suddenly**, **sep as**, and **after that** (see Figure 4 for an overview). Three of them are intra-argument explicit connectives that introduce a dependent clause within Arg2. As their scope is restricted to Arg2, the relation between Arg1 and Arg2 **remains implicit**. However, the annotators prompted to insert a connective between Arg1 and Arg2 were not instructed on such cases, which may

Subset	Var.	Conf.	Agreement
Non-temporal	0.306	0.730	0.559
Temporal, cue+	0.330	0.742	0.686
Temporal, cue-	0.339	0.637	0.560

Table 4: Cartography statistics for cue-rich and cue-light temporal instances as opposed to non-temporal instances.

be the reason why many of them converged on the *precedence* label.

To test whether these ngrams correspond to a distinct cartographic subset, we split temporal instances into cue-rich and cue-light groups. Cue-rich instances contain one of the boundary-adjacent ngrams at the start of Arg2, while cue-light instances do not. Table 4 shows that cue-rich temporal items have the highest agreement and slightly higher confidence than the non-temporal data, while cue-light temporal items have lower confidence and agreement. This suggests that the high-confidence temporal cluster is concentrated in a cue-rich subset.

We also compute the normalized pointwise mutual information (nPMI) between the relevant ngrams and 5-10 most frequent relation types for each level (see Appendix B, Figure 5). At each level, the ngrams are positively associated with the temporal target – *temporal*, *asynchronous*, and *precedence*; conversely, other labels have no positive association. This trend provides an explanation for the temporal subset coinciding with high model confidence and high annotator agreement, as the same local cues are available both to the model and to annotators.

5 Conclusion

Overall, our results point to two conclusions. First, *hard-to-learn* DiscoGeM instances cannot be reliably interpreted as erroneous, as agreement level and label frequency are two factors that strongly impact the **confidence** parameter while not directly translating into annotation quality. Rather, they point to item ambiguity, and filtering out *hard-to-learn* items would thus mean removing the ambiguous part, which dataset cartography normally suggests to keep for training.

On the other hand, the *easy-to-learn* region can pose a problem for robust learning, since for some classes, the right label can often be recovered from cues. nPMI and accuracy show that this trend is exacerbated on more granular levels. In general, this shows that models optimized for discourse labeling can be prone to acquired bias – similarly to NLI, where models are known to learn lexical or syntactic shortcuts during large-scale training (Feng et al., 2019; McCoy et al., 2019). In contrast to NLI, however, relatively few studies have focused on these artifacts and on the robustness of predictions. As data maps can reliably and interpretably identify some of such shortcuts, we hope this study can make a case for qualitative evaluation of models in discourse parsing, regardless of the specific annotation standard.

Lastly, future collection of IDRR data as well as model analysis may need to treat cue-rich arguments explicitly. DiscoGeM notably includes cue-light classes – particularly, *no relation*; as a result, cue-driven items may introduce an imbalance, since some relations require the model to interpret the underlying pragmatics and some can be inferred from strong lexical cues.

6 Acknowledgments

We thank the anonymous reviewers and Frances Yung for their helpful feedback on this work. This study is funded by NWO under the AINed Fellowship Grant NGF.1607.22.002 (“Dealing with Meaning Variation in NLP").

Limitations

This study is limited to a particular encoder-type model architecture (DeBERTa-v3-base). Consequently, the results of our work are not guaranteed to extend to different architecture types, e.g., decoder-type models, including most large language models (LLMs). Nevertheless, we argue that our results are still significant in the context of PDTB-style discourse parsing as well as discourse processing in general, as recent work on discourse relation classifiers has often relied on encoder-type architectures (Xiang and Wang, 2023; Costa and Kosseim, 2025; Chistova, 2025).

In this submission, we made use of generative AI tools for assisted programming and grammar correction.

References

Abhishek Anand, Negar Mokhberian, Prathyusha Naresh Kumar, Anweasha Saha, Zihao He, Ashwin Rao, Fred Morstatter, and Kristina Lerman. 2024. Don’t Blame the Data, Blame the Model: Understanding Noise and Bias When

Learning from Subjective Annotations. *arXiv preprint*. ArXiv:2403.04085 [cs].

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868.

Elena Chistova. 2025. Bridging discourse treebanks with a unified rhetorical structure parser. In *Proceedings of the 6th Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences (CODI 2025)*, pages 197–208.

Nelson Filipe Costa and Leila Kosseim. 2024. Exploring soft-label training for implicit discourse relation recognition. In *Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI 2024)*, pages 120–126, St. Julians, Malta. Association for Computational Linguistics.

Nelson Filipe Costa and Leila Kosseim. 2025. Multilingual implicit discourse relation recognition with multi-label hierarchical learning. In *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 48–61, Avignon, France. Association for Computational Linguistics.

Debopam Das, Manfred Stede, and Maite Taboada. 2017. The good, the bad, and the disagreement: Complex ground truth in rhetorical structure analysis. In *Proceedings of the 6th Workshop on Recent Advances in RST and Related Formalisms*, pages 11–19, Santiago de Compostela, Spain. Association for Computational Linguistics.

Naihao Deng, Xinliang Zhang, Siyang Liu, Winston Wu, Lu Wang, and Rada Mihalcea. 2023. You are what you annotate: Towards better models through annotator representations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12475–12498, Singapore. Association for Computational Linguistics.

Shi Feng, Eric Wallace, and Jordan Boyd-Graber. 2019. Misleading Failures of Partial-input Baselines. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5533–5538, Florence, Italy. Association for Computational Linguistics.

Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2024. Perspectivist approaches to natural language processing: a survey. *Language Resources and Evaluation*, pages 1–28.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation Artifacts in Natural Language Inference Data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. *arXiv preprint*. ArXiv:2111.09543 [cs].

Jan-Christoph Klie, Bonnie Webber, and Iryna Gurevych. 2023. Annotation Error Detection: Analyzing the Past and Present for a More Coherent Future. *Computational Linguistics*, 49(1):157–198.

Savannah Larimore, Ian Kennedy, Breon Haskett, and Alina Arseniev-Koehler. 2021. Reconsidering annotator disagreement about racist language: Noise or signal? In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 81–90, Online. Association for Computational Linguistics.

Wanqiu Long, N. Siddharth, and Bonnie Webber. 2024. Multi-label classification for implicit discourse relation recognition. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8437–8451, Bangkok, Thailand. Association for Computational Linguistics.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Rashmi Prasad, Bonnie Webber, and Aravind Joshi. 2017. The penn discourse treebank: An annotated corpus of discourse relations. In *Handbook of linguistic annotation*, pages 1197–1217. Springer.

Rashmi Prasad, Bonnie Webber, Alan Lee, and Aravind Joshi. 2019. Penn Discourse Treebank Version 3.0. Abacus Data Network.

Valentina Pyatkin, Frances Yung, Merel Scholman, Reut Tsarfaty, Ido Dagan, and Vera Demberg. 2023. Design choices for crowdsourcing implicit discourse relations: Revealing the biases introduced by task design. *Transactions of the Association for Computational Linguistics*, 11:1014–1032.

Hannah Rohde, Alexander Johnson, Nathan Schneider, and Bonnie Webber. 2018. Discourse Coherence: Concurrent Explicit and Implicit Relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2257–2267, Melbourne, Australia. Association for Computational Linguistics.

Merel Scholman and Vera Demberg. 2017. Crowdsourcing discourse interpretations: On the influence of context and the reliability of a connective insertion task. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 24–33, Valencia, Spain. Association for Computational Linguistics.

Merel Scholman, Tianai Dong, Frances Yung, and Vera Demberg. 2022. DiscoGeM: A crowdsourced corpus of genre-mixed implicit discourse relations. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3281–3290, Marseille, France. European Language Resources Association.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. *The Penn Discourse Treebank 3.0 annotation manual*. Philadelphia, University of Pennsylvania.

Wei Xiang and Bang Wang. 2023. A survey of implicit discourse relation recognition. *ACM Computing Surveys*, 55(12):1–34.

Frances Yung, Kaveri Anuranjana, Merel Scholman, and Vera Demberg. 2022. Label distributions help implicit discourse relation classification. In *Proceedings of the 3rd Workshop on Computational Approaches to Discourse*, pages 48–53, Gyeongju, Republic of Korea and Online. International Conference on Computational Linguistics.

Frances Yung and Vera Demberg. 2025. On crowdsourcing task design for discourse relation annotation. In *Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation*, pages 12–19, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Frances Yung, Vera Demberg, and Merel Scholman. 2019. Crowdsourcing discourse relation annotations by a two-step connective insertion task. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 16–25.

Frances Yung, Daniil Ignatev, Merel Scholman, Vera Demberg, and Massimo Poesio. 2026. Human Label Variation in Implicit Discourse Relation Recognition. *arXiv preprint*. ArXiv:2602.22723 [cs].

Mike Zhang and Barbara Plank. 2021. Cartography active learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 395–406. Association for Computational Linguistics.

A Training Parameters

We train SG models for 15 epochs and AE models for 5; in this manner, we ensure similar exposure to data, as AE models see each data instance several times per epoch, albeit with varying labels. We consistently use $lr = 1\mathrm{e}{-5}$.

In the SG setting, we use log-scaled class weights in training, so as to mitigate the effect of data imbalance; this modification is not applied to AE due to a different architecture.

B Illustrations

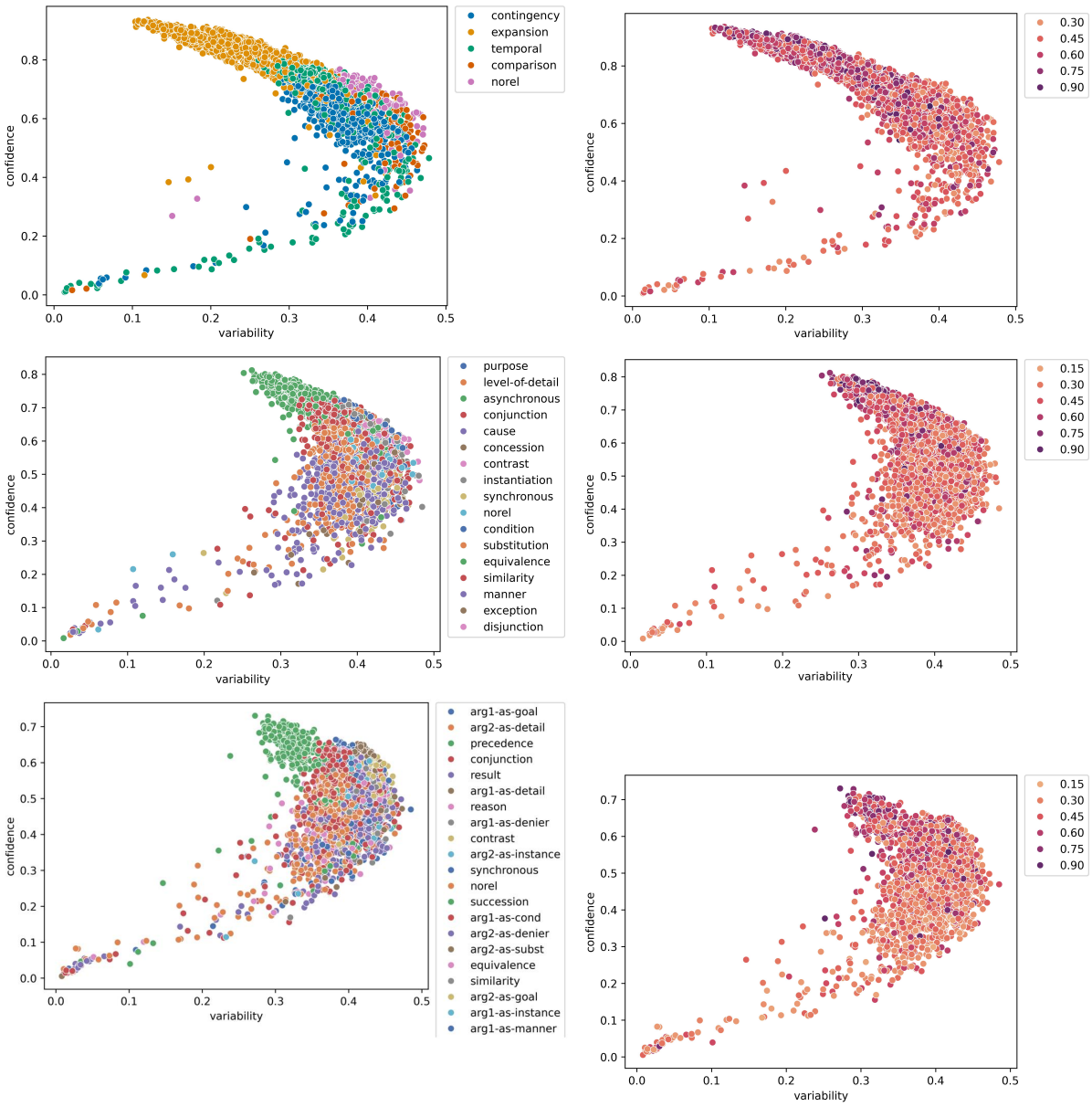

Figure 2: Data maps for the majority label classifier, DeBERTa-v3-base, granularity levels 1, 2, and 3 (top to bottom). X- and Y-axis correspond to **variability** and **confidence**, respectively. Left column: class-wise breakdown; right column: breakdown by agreement (ratio of the majority label per instance).

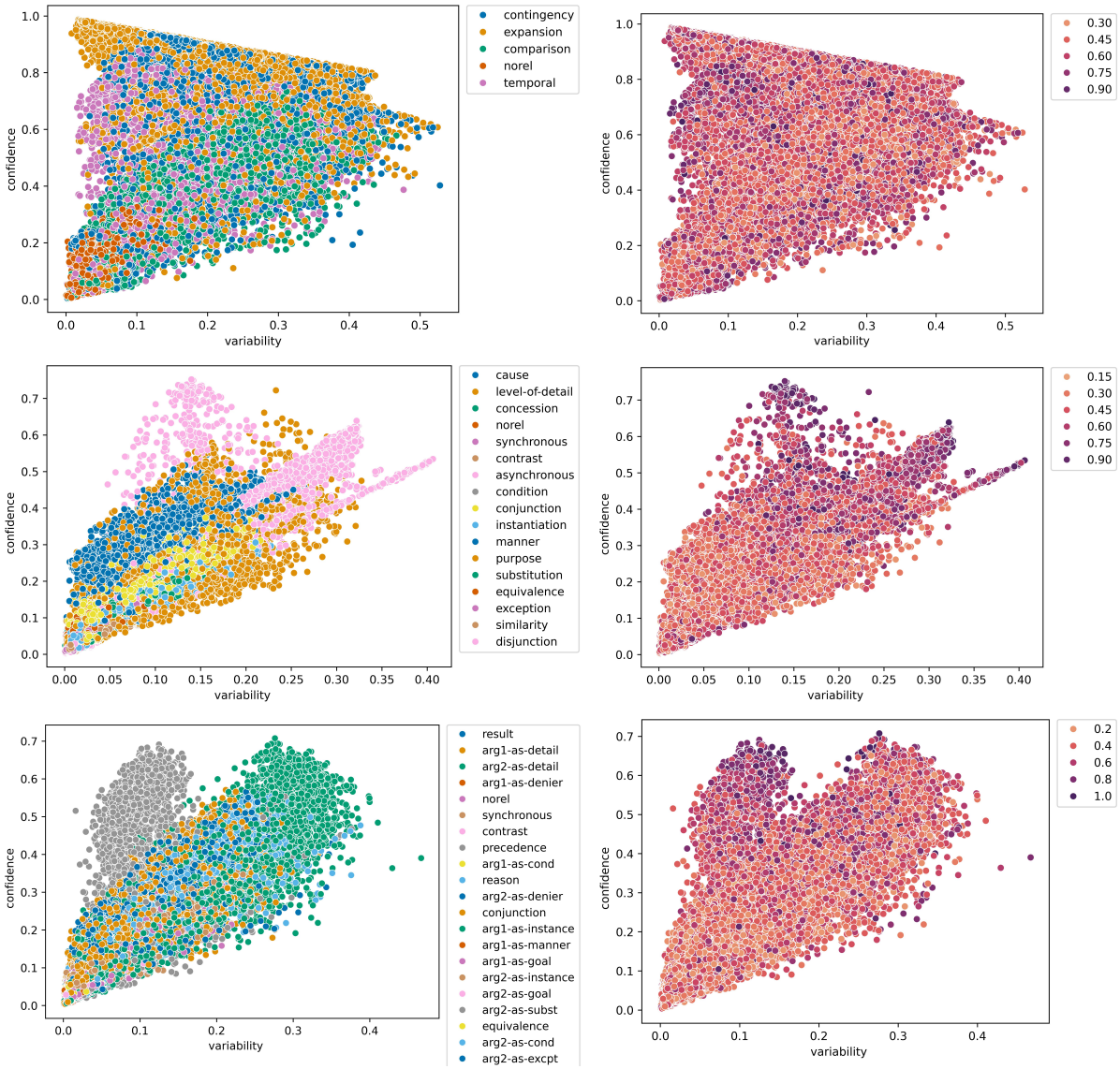

Figure 3: Data maps for the Annotator Embeddings model (Deng et al., 2023), DeBERTa-v3-base, granularity levels 1, 2, and 3 (top to bottom). X- and Y-axis correspond to **variability** and **confidence**, respectively. Left column: class-wise breakdown; right column: breakdown by agreement (ratio of the majority label per instance).

+1.305	sep after	+1.365	sep after	+1.270	sep after	+1.420	sep after	+1.405	sep after
+0.874	sep suddenly	+0.784	sep when	+0.856	sep at	+0.854	sep at	+0.946	sep when
+0.812	sep at	+0.759	sep suddenly	+0.729	sep suddenly	+0.854	sep when	+0.721	sep at
+0.760	sep when	+0.756	sep at	+0.687	sep when	+0.765	sep suddenly	+0.659	laughing sep
+0.653	programmes	+0.696	following	+0.653	following	+0.728	sep as	+0.659	sep clambered
+0.652	sep with	+0.622	had no	+0.626	sep as	+0.678	following	+0.659	sep suddenly
+0.652	following	+0.616	glanced	+0.623	leading	+0.658	became	+0.629	sep as
+0.615	suddenly	+0.611	became	+0.596	had no	+0.634	suddenly	+0.620	programmes
+0.601	became	+0.606	sep clambered	+0.582	muttered	+0.604	laughing sep	+0.595	lighter sep
... 30753 more positive ...		+0.606	laughing sep	*... 30857 more positive ...*		+0.604	sep clambered	*... 30483 more positive ...*	
... 69284 more negative ...		*... 30772 more positive ...*		*... 69180 more negative ...*		*... 30900 more positive ...*		*... 69554 more negative ...*	
-0.621	life	*... 69265 more negative ...*		-0.593	are	*... 69137 more negative ...*		-0.604	public
-0.636	are	-0.610	has	-0.594	so there	-0.611	are	-0.621	are
-0.642	sep her	-0.659	he was	-0.621	life	-0.622	however	-0.629	the most
-0.651	often	-0.666	so there	-0.640	sep her	-0.623	he was	-0.632	need
-0.660	has	-0.672	sep you	-0.658	the most	-0.628	often	-0.640	is
-0.699	he was	-0.683	are	-0.664	has	-0.648	need	-0.651	he was
-0.710	is	-0.700	is	-0.677	is	-0.743	is	-0.850	has
-0.721	sep you	-0.796	sep it	-0.682	sep you	-0.765	sep her	-0.861	sep it
-0.941	sep it	-0.806	sep her	-0.782	sep it	-0.775	sep it	-0.905	sep her
-1.391	sep	-1.343	sep	-1.361	sep	-1.418	sep	-1.380	sep
-2.068	<BIAS>	-2.057	<BIAS>	-2.115	<BIAS>	-2.035	<BIAS>	-2.134	<BIAS>

Figure 4: Top positive and negative logistic regression features for the *precedence* probe at level-3 across 5 validation folds. Each column corresponds to a different fold. The token **sep** marks the Arg1/Arg2 boundary, i.e., "sep after" indicates that Arg2 starts with "after".

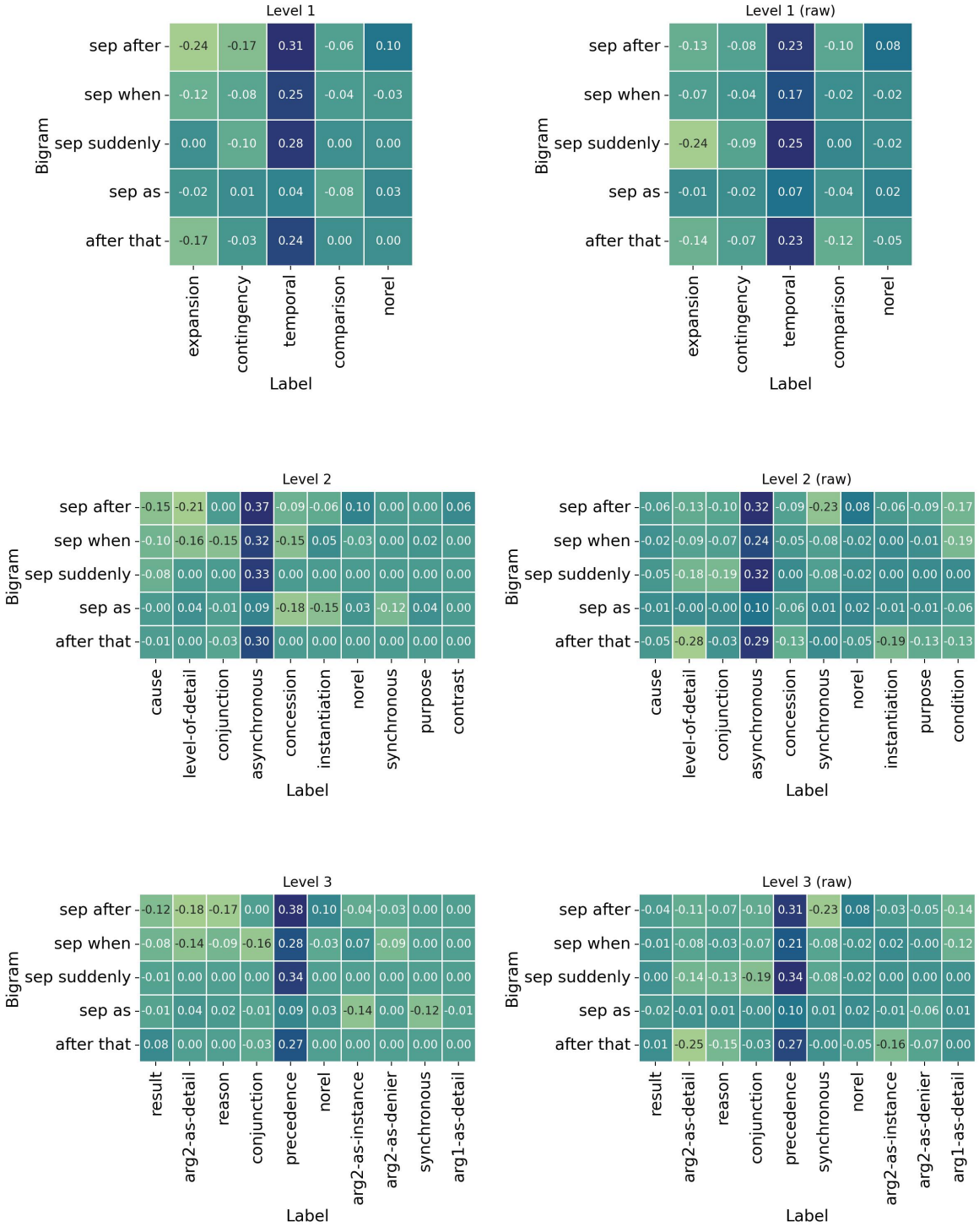

Figure 5: Normalized pointwise mutual information (nPMI) for *precedence*-related bigrams and most frequent labels for each sense level. Left: majority labels; right: raw labels. Top to bottom: levels 1, 2, and 3.

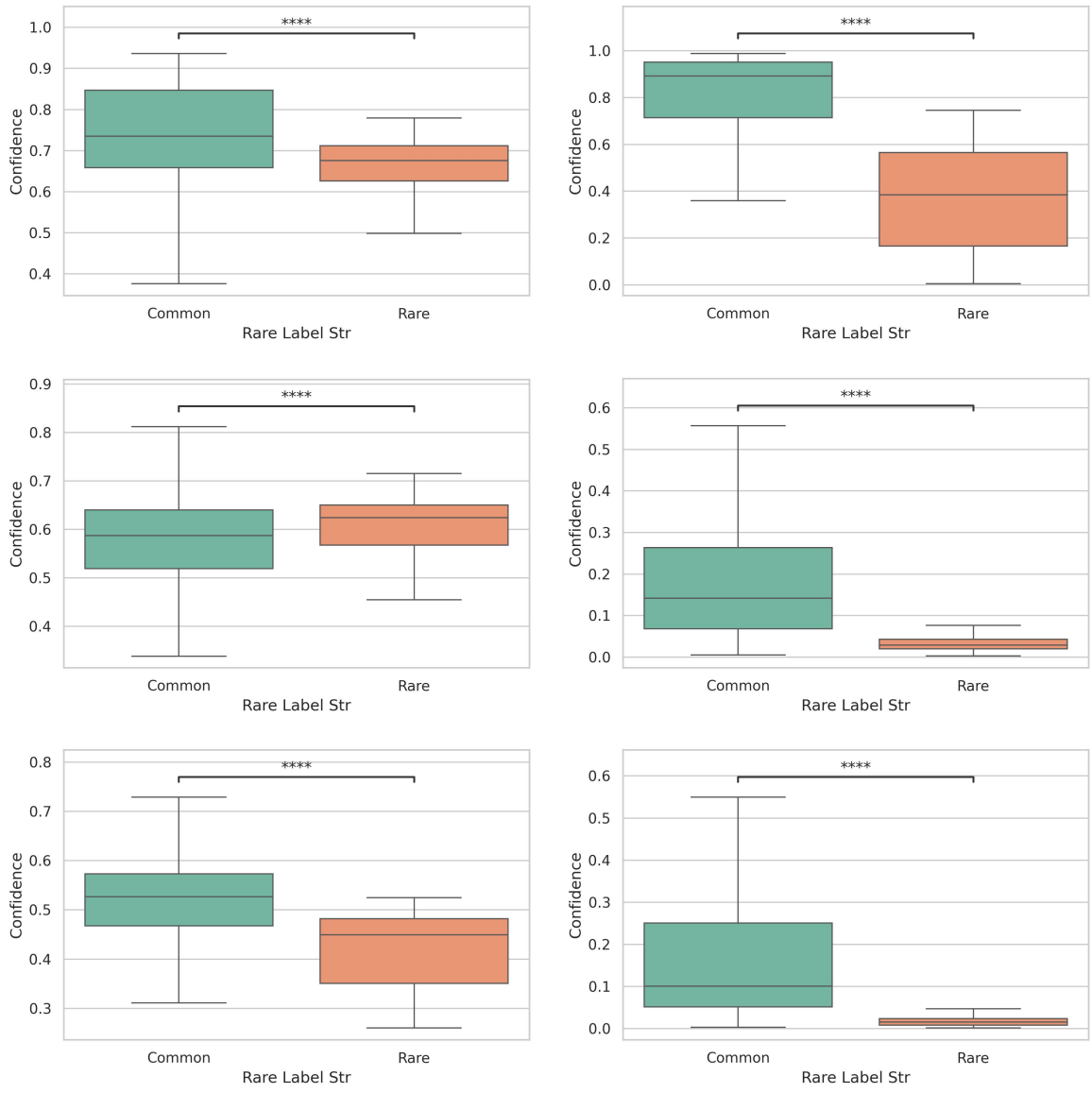

Figure 6: Confidence boxplots for frequent vs rare labels (relative frequency greater than /less than 0.33). '****' indicates statistical significance per the Mann-Whitney test with $p \leq 1e{-}4$. Left: SG; right: AE. Top to bottom: levels 1, 2, and 3.