Towards Explainable Implicit Discourse Relation Recognition: An SLM-LLM Collaboration Framework

Anonymous ACL submission

Abstract

Implicit discourse relation recognition (IDRR) aims to infer semantic connections between text spans without explicit connectives and remains a fundamental challenge in natural language understanding. Although recent studies have introduced explanations, existing methods still struggle with hard instances involving highly confusable relation categories, while the generated explanations are often remain ambiguous. To address these issues, we propose a SLM-LLM collaboration framework (SLCF) for explainable IDRR, where explanation-driven classification and classification-guided explanation refinement are jointly employed. Specifically, LLM first generates candidate explanations for all relation categories, allowing the small model to perform comparative reasoning and active selection. The predicted confusable relations are then fed back to LLM to generate discriminative rules and refine the explanations. Through this collaboration, the small model obtains stronger semantic cues for hard cases, while LLM produces more consistent and less ambiguous explanations. Experiments on PDTB show that our approach outperforms most baselines, while generating more reasonable and consistent explanations, demonstrating improved performance and explainability.

1 Introduction

Discourse relations capture the semantic connections between sentences or text spans and are fundamental to natural language understanding and generation (Li et al., 2020; Tang et al., 2021; Hu and Wan, 2023). Among them, implicit discourse relations are particularly challenging, as they lack explicit connectives (e.g., but, because) and require models to infer deep semantic relationships. This makes them difficult for both humans and machines. Despite the progress of large language models, implicit discourse relation recognition (IDRR) remains challenging due to its demand for fine-grained understanding and reasoning, and thus continues to be an important problem (Chan et al., 2024; Yung et al., 2024).

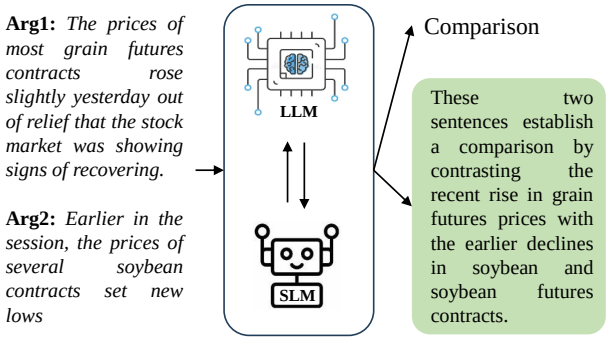

Figure 1: The small and large models collaborate to predict relations and generate explanations. The small model provides its prediction results as feedback to the large model, while the large model offers explanations and rules back to the small model. Rules capture distinctions between confusable relations, while explanations describe the rationale for a specific prediction. **Arg1** and **Arg2** are two arguments.

Recent advances in deep learning have substantially improved IDRR. Existing studies have enhanced relation prediction from different perspectives, including neural architectures for modeling argument interactions (Zhang et al., 2015; Bai and Zhao, 2018; Liu et al., 2020), hierarchy-aware learning that exploits dependencies among relation labels (Wu et al., 2022; Long and Webber, 2022), and prompt-based methods that reformulate IDRR as connective prediction to better leverage pre-trained language models (Xiang et al., 2022; Zhao et al., 2023; Zeng et al., 2024). These advances have consistently improved classification performance. Despite these advances, most methods still treat IDRR as a black-box classification problem, outputting only relation labels without explicit reasoning, which limits explainability and reliability. Recent work (Wang and Wu, 2026) introduces a classification–generation framework to in-

corporate natural language explanations via multi-task learning and knowledge distillation, achieving a trade-off between performance and explainability.

However, despite recent progress in explainable IDRR, two important challenges remain insufficiently explored. First, existing methods still struggle with hard instances involving highly confusable discourse relations. Implicit discourse relations often exhibit subtle semantic boundaries, making certain categories difficult to distinguish even when explanations are provided. Second, the generated explanations are often insufficiently discriminative. Although explanations may appear plausible, they frequently describe general semantic associations rather than highlighting the critical distinctions between competing relation categories.

To address these challenges, we propose a SLM-LLM collaboration framework for explainable IDRR. As shown in Figure 1, the framework is built on bidirectional interaction between large and small models, where explanation-driven classification and classification-guided explanation refinement are jointly employed. Specifically, LLM first generates candidate explanations for all relation categories, forming a candidate explanation reasoning process that enables the small model to perform comparative reasoning and active selection across competing relations rather than relying on a single explanation. For hard instances with highly confusable categories, the prediction results are further fed back to LLM to generate discriminative rules that explicitly characterize the semantic differences between competing relations. These rules are then used to refine both relation prediction and explanation generation.Through this collaborative process, the framework improves the model's ability to resolve hard cases while producing clearer and more discriminative explanations.

Beyond improving hard-instance reasoning and explanation quality, the collaborative framework offers two additional advantages. First, compared with conventional predict-then-explain paradigms, comparative reasoning and active selection help mitigate inconsistencies between predictions and explanations by tightly coupling explanation selection with relation prediction. Second, unlike traditional knowledge distillation methods that inevitably suffer from information loss, our framework directly utilizes candidate explanations as explicit inputs, enabling the small model to better exploit the rich reasoning capability of LLMs.

The main contributions of this work can be summarized as follows:

- We propose an SLM-LLM collaboration framework for explainable IDRR, where explanation-driven classification and classification-guided explanation refinement are jointly employed.

- We introduce a candidate explanation reasoning process that exposes the small model to the full explanation pool across all relation types, enabling comparative reasoning and active selection rather than passive conditioning on a single explanation.

- Experimental results on the PDTB dataset demonstrate that the proposed method not only improves classification performance but also enhances the quality of generated explanations.

2 Related Work

2.1 IDRR Methods

Most research on implicit discourse relation recognition has primarily focused on improving relation prediction performance. Early studies focused on designing task-oriented architectures to capture argument semantics and interactions (Zhang et al., 2018; Bai and Zhao, 2018; Zhang et al., 2021), and the introduction of pre-trained language models (PLMs) such as BERT and RoBERTa further enhanced semantic representation, yielding substantial performance gains (Liu et al., 2020). Subsequently, researchers incorporated label hierarchy information to exploit inter-label dependencies (Wu et al., 2022; Long and Webber, 2022; Jiang et al., 2023), while prompt-based learning emerged as a dominant paradigm by guiding PLMs to predict discourse connectives (Xiang et al., 2022; Zhou et al., 2022) and integrating hierarchical label information into prompt-tuning frameworks (Zhao et al., 2023; Chan et al., 2023; Zeng et al., 2024). To alleviate data scarcity, data augmentation strategies leveraging conditional variational autoencoders (Dou et al., 2021) and large language models (Omura et al., 2024) have also been explored. As classification performance has steadily improved, increasing attention has been paid to the explainability of IDRR models. In this context, Wang and Wu (2026) proposed the EIDRR framework, which distills the explanatory capability of

LLMs into lightweight models via a joint classification–generation framework that outputs both discourse relations and natural language explanations. Although this work represents an important step toward explainable IDRR, how to simultaneously optimize classification accuracy and explanation quality remains an open challenge.

2.2 Large–Small Model Collaboration

To balance performance and inference efficiency, recent studies have actively explored synergistic paradigms between large and small language models (Lv et al., 2025; Xu et al., 2025; Yan et al., 2025). One prominent line of research focuses on adaptive routing strategies, which dynamically allocate instances to models of appropriate capacity—either at the instance level based on prediction confidence (Chen et al., 2024a) or at the finer-grained token level via reinforcement learning (Zheng et al., 2025). Another well-established paradigm leverages LLMs as external reasoning or verification engines, where large models either validate candidate outputs from smaller counterparts (Chen et al., 2024b) or supply auxiliary reasoning signals for downstream decision-making (Xu et al., 2025). Knowledge distillation offers a complementary direction, transferring the sophisticated reasoning or generative capabilities of LLMs into lightweight models (Gu et al., 2024; Wang et al., 2025). Inspired by these collaborative paradigms, we explore a collaborative framework for explainable IDRR.

3 Method

As shown in Figure 2, in the SLM-LLM collaboration framework (SLCF), the large model is primarily responsible for generation tasks, including producing explanations and discriminative rules, while the small model focuses on relation classification. Specifically, in a multi-round interaction process, the small model predicts discourse relations based on the input instance, candidate explanations, and generated rules. The final prediction consists of the relation category $\tilde{y}$ and the corresponding explanation e, which is selected according to the predicted relation category.

3.1 Candidate Explanation Generation

Recent studies have shown that large language models possess strong capabilities in generating natural language explanations (Brown et al., 2020; Kroeger et al., 2023). Therefore, we expect them to produce

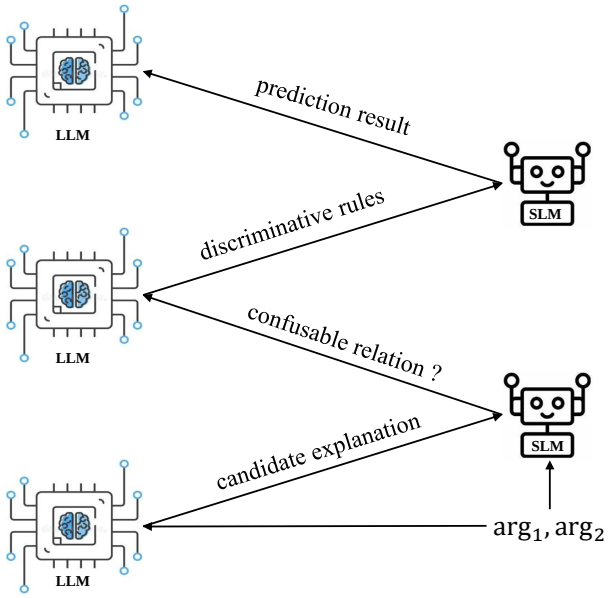

Figure 2: Overview of SLM-LLM collaborative inference process.

corresponding explanations for each possible discourse relation category given a discourse instance. However, preliminary experiments reveal a practical issue: when asked to generate an explanation for an incorrect candidate label, LLMs may rely on their own judgment to determine that the label is inappropriate and consequently refuse to produce an explanation.

To address this issue, we adopt a tailored prompting template to guide the model in generating plausible explanations under a given label. Following Wang and Wu (2026), the prompt consists of three components: (1) task instruction, (2) in-context learning examples, and (3) input template. The generated explanation also follows the same structured format, i.e., first restating the content of the two arguments and then providing the reasoning for the target relation. Unlike prior work, we explicitly state in the task instruction that the provided label may not be correct, and that the model is only required to generate a reasonable explanation. This design prevents the model from rejecting the given relation category due to its own judgment and ensures that it generates a corresponding explanation. Based on the above design, the explanation generation process is as follows:

$$e_i = \text{LLM}(T_L(arg_1, arg_2, label_i)), \qquad (1)$$

where $label_i$ denotes the possible relation between two arguments, e_i represents the explanation for a possible relation between the arguments, and T_L is

the template for explanation generation, which is provided in Appendix A.

3.2 Candidate Explanation-Enhanced IDRR

Given an input instance (arg_1, arg_2) and candidate explanations e_i generated by LLM for each relation category, an auto-encoding pre-trained language model is used as the backbone encoder. Specifically, the two arguments and their corresponding candidate explanations are jointly incorporated into a prompt template to guide the model in predicting the appropriate relation category. The detailed classification prompt template T_s is provided in Appendix A. Based on this design, relation classification is formulated as follows:

$$\begin{aligned} y_1 &= \text{Softmax}(o_1), \\ o_1 &= \text{LM_head}(h^1_{mask}), \\ h^1_{mask} &= \text{SLM}(T_s(arg_1, arg_2, e_i)) \end{aligned} \tag{2}$$

where h^1_{mask} represents the vector representation of the <mask> token, LM_head is a nonlinear layer, and y_1 denotes the predicted probability distribution.

To enhance the model's ability to distinguish confusable relation categories, we introduce an elimination reasoning strategy inspired by human cognitive selection. Since explanations associated with incorrect relation categories may introduce misleading semantic signals, directly considering all candidate explanations simultaneously can interfere with the final prediction. Therefore, the proposed strategy progressively filters out low-confidence relation categories based on the model's prediction distribution, allowing the model to focus on the remaining confusable candidates and their subtle semantic differences. Specifically, the model first performs an initial prediction over all candidate relation categories according to Equation 2. Based on the resulting prediction distribution, low-confidence relation categories are progressively eliminated by masking the corresponding explanation positions in the prompt, while the remaining potentially confusable candidates are retained for further discrimination. The refined prompt is then fed into the model for second-stage prediction:

$$\begin{aligned} y_2 &= \text{Softmax}(o_2), \\ o_2 &= \text{LM_head}(h^2_{mask}), \\ h^2_{mask} &= \text{SLM}(\text{Mask}(T_s(arg_1, arg_2, e_i))) \end{aligned} \tag{3}$$

where Mask represents the elimination operation that masks the explanation positions associated with low-confidence relation categories, h^2_{mask} denotes the vector representation corresponding to the <mask> token, LM_head is a nonlinear layer, and y_2 refers to the probability distribution of prediction in the second-stage prediction.

Recent studies have demonstrated the effectiveness of prompt-based learning for IDRR by predicting connective words and mapping them to predefined relation categories. Therefore, we integrate the proposed strategy into prompt learning, where the refined prompt is used to predict the connective (e.g., "but"), and the predicted connective is subsequently mapped to the corresponding relation category (e.g., "Comparison") according to predefined mapping rules (Xiang et al., 2022). The detailed mapping rules are provided in Appendix B.

3.3 Discriminative Rules Generation for Confusable Relation Categories

Due to the ambiguity of certain candidate explanations in some instances, the small model may still require LLM's deep reasoning assistance to accurately resolve complex instances. However, to avoid the prohibitive computational cost of invoking LLM for every single sample, we adopt a conditional on-demand interaction strategy. Specifically, we use the maximum probability from Equation 3 as a confidence measure. If this confidence is lower than a predefined threshold τ, the instance is regarded as a hard case and is further sent to LLM for assistance. Otherwise, SLM prediction is directly used as the final result, without additional LLM invocation.

For each hard case, SLM also provides the remaining confusable relation categories after elimination. Given these categories, LLM generates instance-specific discriminative rules, i.e., explicit criteria for distinguishing between competing relations under the current context. These rules differ from candidate explanations: candidate explanations describe why a relation may hold, whereas discriminative rules emphasize contrasts between confusable relations and highlight the contextual evidence that supports one relation over others. This design encourages LLM to focus on difficult decision boundaries rather than producing generic explanations. The rule generation process is formulated as:

$$\begin{aligned} r = \text{LLM}(T_R(arg_1, arg_2, c_i)), \\ \text{if } \max(y_2) < \tau \end{aligned} \tag{4}$$

where c_i are the confusable categories, and T_R is the template for discriminative rules generation, the specific discriminative rules generation prompt template can be found in Appendix A.

3.4 Discriminative Rules-Enhanced IDRR

After the corresponding rules are generated, they are used to enhance the small model's judgment on confusable categories. Specifically, the arguments and the discriminative rules are integrated into the same template. Based on the given discriminative rules enhanced classification prompt template (details can be found in Appendix A), the relation classification process is again implemented through prompt learning as follows:

$$\begin{aligned} r_logits &= \text{LM_head}(h^r_{mask}), \\ h^r_{mask} &= \text{SLM}(T_r(arg_1, arg_2, r)) \end{aligned} \tag{5}$$

where h^r_{mask} denotes the vector representation corresponding to the <mask> token, LM_head is a nonlinear layer. We then use a gate to dynamically combine the elimination-based prediction and the rules-enhanced prediction. The final prediction for hard cases is:

$$\begin{aligned} y_3 &= \text{Softmax}(o_2 + \lambda \cdot \text{Norm}(r_logits)), \\ \lambda &= \text{Sigmoid}(\text{H}(y_2)) \end{aligned} \tag{6}$$

where H(*) denote the entropy of the prediction. Norm denotes the normalization operation, and y_3 represents the final predicted probability distribution of connectives for hard examples. Therefore, the overall training loss consists of three components:

$$\begin{aligned} loss = \alpha * \text{CE}(y_1, y_c) + \beta * \text{CE}(y_2, y_c) \\ + \text{CE}(y_3, y_c), \end{aligned} \tag{7}$$

where α and β are the weight factors, CE denotes the cross-entropy loss function, y_1, y_2 and y_3 are the predicted connectives in the first round, the second round, and the final prediction respectively, and y_c is the ground-truth connective between the two arguments.

3.5 Discriminative Rules-Enhanced Explanation Refinement

Because the original explanations may remain ambiguous in some cases, the model may find it difficult to identify the corresponding relation category accurately using only these explanations. Therefore, it is necessary to further optimize and update the original explanations. To address this issue, we further refine the explanations by leveraging the final prediction, the identified confusable categories, and the generated discriminative rules.

Specifically, based on the designed template (the detailed explanation refinement prompt template can be found in Appendix A), LLM is used to update ambiguous explanations given the original arguments, the initial explanation, the predicted relation, confusable relation, and the corresponding rules. The process is defined as follows:

$$\begin{aligned} e_up_i = \text{LLM}(T_u(arg_1, arg_2, o_exp, \\ p_label, c_label, r)), \end{aligned} \tag{8}$$

where o_exp denotes the original explanation from the explanation candidate pool, p_label represents the relation predicted by the small model, c_label is the confusable relation category, and r is the generated discriminative rules.

4 Experiments

4.1 Datasets and Experimental Settings

The experiments are primarily conducted on the PDTB 2.0 dataset. Following Wu et al. (2022), we split the dataset into three subsets: a training set with 12,775 instances, a validation set with 1,183 instances, and a test set with 1,046 instances. The relation categories comprise four categories: *Expansion*, *Comparison*, *Contingency*, and *Temporal*. We generate explanations for each candidate discourse relation rather than restricting them to the gold label only. Moreover, unlike prior work, where explanations are not generated for the test set, we construct explanations for every test instance under all candidate relation categories. A total of 60,016 explanations were generated. More details of these datasets can be found in Appendix C.

The experiments are mainly conducted on an NVIDIA 3090 GPU. For a better comparison, RoBERTa-base is selected as the encoder backbone. The major hyperparameters are set as follows: the number of training epochs is 6, the learning rate is 5×10^{-6}, the dropout rates for the two prediction rounds are 0.2 and 0.4, respectively, and α, β are both set to 1. To ensure relative stability of the results, the temperature of LLM is set to 0 when generating discriminative rules. Following prior work, we evaluate using macro-averaged F_1 score and accuracy (Acc).

Table 1: Performance of discourse relation classification.

Method	$Acc(\%)$	$F_1(\%)$
ChatGPT	50.24	44.09
LLaMA-3B	65.11	57.11
LLaMA-8B	70.36	61.01
CVAE	70.17	65.06
LDSGM	71.18	63.73
GOLF	72.52	65.76
SCIDER	72.11	67.00
PCP-base	70.84	64.95
PCP-large	73.80	67.79
PEMI	71.13	64.05
Discoprompt	71.70	65.79
GHBUPP	73.89	67.83
EIDRR	73.14	68.01
SLCF	**75.43**	**69.41**
w/o rules	74.19	68.70
w/o rules & elimination	73.33	67.64
w/o all	71.80	66.87

4.2 Performance of IDRR

We compare our method (named as SLCF) with recent baselines, which primarily include the following:

LLM-based methods: ChatGPT (Chan et al., 2024), LLaMA-3B/8B: We formulate IDRR as a generation task and fine-tune with LoRA.

Non-prompt-based methods: CVAE (Dou et al., 2021), LDSGM (Wu et al., 2022), GOLF (Jiang et al., 2023), SCIDER (Cai et al., 2024).

Prompt-based methods: PCP (Zhou et al., 2022), PEMI (Zhao et al., 2023), Discoprompt (Chan et al., 2023), GHBUPP (Jiang et al., 2023), EIDRR (Wang and Wu, 2026).

According to Table 1, several conclusions can be drawn. First, directly prompting LLMs or fine-tuning lightweight LLMs yields suboptimal performance, suggesting that neither prompting nor parameter-efficient fine-tuning at modest scales is sufficient for capturing implicit discourse relations. Second, the proposed framework outperforms most prompt-based methods, verifying the effectiveness of our approach. This also demonstrates that rather than directly applying LLMs to IDRR under resource-constrained settings, leveraging them as auxiliary resources to supply explanations and discriminative rules for a small model proves to be a more effective strategy. Third, compared with the distillation-based method EIDRR, our framework achieves further improvements, demonstrating the advantage of the proposed collaboration mechanism over simple distillation. As further analyzed in Section 4.4, our method yields particularly pronounced gains on confusable relation categories and hard instances.

The ablation results show that removing any component consistently degrades performance, confirming that each module contributes positively to the overall model. Specifically, removing the discriminative rules-enhanced module (*w/o rules*) reduces accuracy by 1.24%, indicating that discriminative rules generated by LLM play an important role in distinguishing hard samples. Further removing elimination reasoning (*w/o rules & elimination*) further degrades performance, suggesting that without this strategy, explanations corresponding to incorrect relation categories may introduce noise into the small model's prediction, thereby hampering its ability to distinguish confusable relations. When candidate explanation enhancement is additionally removed (*w/o all*), the performance drop is the largest, with accuracy decreasing by 1.53%. This demonstrates the effectiveness of incorporating explanation information into the classification process and further confirms the necessity of explanatory research.

4.3 Evaluation of Explanation Quality

LLM generates explanations without access to ground-truth labels; that is, it produces explanations based solely on the input content regardless of prediction correctness. Therefore, the quality of these explanations requires further evaluation. Following the evaluation criteria and methods of Wang and Wu (2026), we similarly evaluate the explanations that are correctly classified in the test set, and a total of 789 instances are evaluated. Specifically, five annotators assess explanation quality from three aspects: interpretability, factuality, and fluency. Detailed evaluation guidelines and scoring criteria are provided in Appendix D. For ease of comparison, we further regroup the scores into three categories: *Reaso.* (reasonable, score > 3), *Accep.* (acceptable, score $= 3$ or 3.5), and *Unrea.* (unreasonable, score < 3). We additionally report *Incon.*, the proportion of explanations that are inconsistent with the predicted labels.

As shown in Table 2, the evaluation shows that 82% of the generated explanations are consistent with human judgments, indicating that our method can meet practical requirements in most cases and

Table 2: Comparison of explanation quality between EIDRR and SLCF.

Method	Reaso.	Accep.	Unrea.	Incon.
EIDRR	78%	10%	12%	10.84%
SLCF	82%	16%	2%	0%

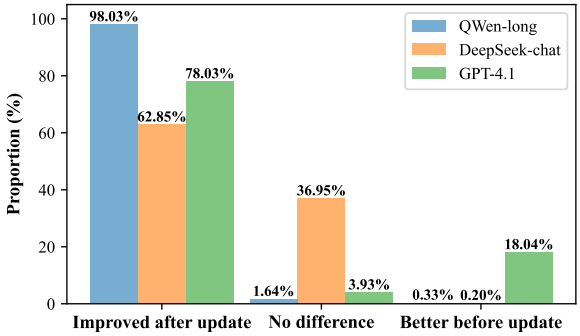

Figure 3: Comparison of evaluation results of large language models.

enhance user trust. Only 2% of the explanations exhibit clear deviations, mainly due to LLM occasionally introducing information not present in the original arguments. In addition, compared with EIDRR, SLCF generates more reasonable explanations and fewer unreasonable ones. This improvement mainly comes from two aspects. First, SLCF directly utilizes the explanation generation capability of LLMs, while EIDRR inevitably suffers from information loss during knowledge transfer, reducing explanation quality. Second, EIDRR is prone to explanation–label inconsistency: even when the predicted relation is incorrect, the model is still optimized using the gold explanation, resulting in an explanation–label mismatch rate of 10.84%. In contrast, SLCF aligns explanation selection with the predicted labels, effectively mitigating this inconsistency. These results demonstrate that the proposed collaborative framework not only reduces conflicts between explanations and labels but also improves overall explanation quality, enabling more effective utilization of LLM.

Following Thakur et al. (2025), which demonstrates that LLMs align more closely with human judgment on ranking tasks than on absolute scoring, we employ LLMs as evaluators to compare the updated explanations with the original ones. As shown in Figure 3, Qwen, DeepSeek and GPT judge the updated explanations to be superior in most instances, indicating that discriminative rules make the explanations clearer and more distinguishable. Although all three models show consistent trends, the degree of improvement varies slightly, likely due to model-specific evaluation preferences. In particular, Qwen is involved in both explanation refinement and evaluation, which may account for its relatively larger gains. Nevertheless, when DeepSeek and GPT are used as independent evaluators, the same conclusion still holds. This confirms that the measured advantages of our approach are consistent across different automated judges, minimizing potential model-specific evaluation bias.

Table 3: Effect of the hard-case threshold on performance and LLM calls. Samples denotes the proportion of instances requiring additional LLM calls; W.C denotes corrected wrong predictions after rules enhancement; C.PU and W.PU denote cases where the gold-label probability increases for correct and wrong predictions, respectively.

τ	Samples	W.C	C.PU	W.PU	Acc	F_1
0.7	33.74%	49	114	28	74.95	68.84
0.75	41.10%	52	159	29	75.05	69.06
0.8	47.89%	52	209	27	75.43	69.41

4.4 Effectiveness and Complexity Analysis of Discriminative Rule Enhancement

To evaluate the efficiency of the proposed on-demand collaboration mechanism, we further analyze the instances that trigger additional LLM calls during inference and their corresponding performance changes after LLM enhancement.

According to Table 3, two main observations can be drawn. First, varying the threshold has only a limited effect on overall performance. Although increasing the threshold from 0.7 to 0.8 substantially increases the number of LLM calls, the number of W.C samples remains nearly unchanged (around 50). This suggests that comparable gains can be achieved at lower computational cost, providing a favorable efficiency–performance trade-off. Second, LLM mainly benefits easily confusable samples. The probability statistics show that when class probabilities are close, LLM effectively enlarges the margin between the true and competing classes, regardless of whether the initial prediction is correct. This indicates that the collaborative mechanism mainly serves as a form of boundary refinement rather than simple global correction.

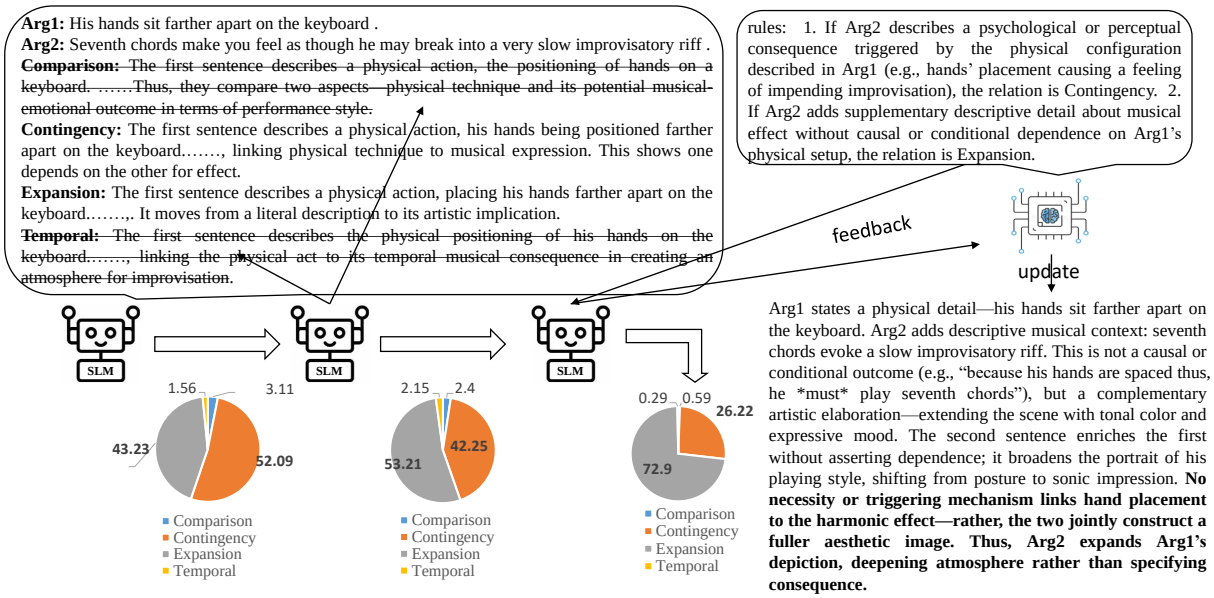

Figure 4: Case study. More examples are on Github.

4.5 Case Study

To better illustrate the prediction process of the small model, we present a representative case study. By tracing the model's intermediate decisions, we show how the prediction evolves from the initial input to the final output.

As shown in Figure 4, the instance is: *Arg1*: "His hands sit farther apart on the keyboard." *Arg2*: "Seventh chords make you feel as though he may break into a very slow improvisatory riff." The gold label is *Expansion*. Based on the candidate explanations generated by the large model, the small model initially favors *Contingency*, while assigning a similar probability to *Expansion*. After applying the elimination reasoning strategy, the *Comparison* and *Temporal* categories are filtered out, and the prediction shifts to the correct label *Expansion*. However, the confidence remains close to that of *Contingency*, indicating that this is a hard sample mainly involving confusion between these two relations. To further resolve this ambiguity, the instance is sent to LLM to generate discriminative rules for the competing categories. The rule-based output is then fused with the elimination-based prediction. After fusion, the model not only predicts the correct label but also assigns a substantially higher probability to *Expansion*, effectively reducing class confusion.

In terms of explanation, the original explanation correctly captures the *Expansion* relation but contains the phrase "implying how it affects his playing style," where the word *affects* introduces a causal implication that may bias the model toward *Contingency*. To address this issue, the discriminative rules explicitly distinguish the two relations: expressions such as "describes a psychological or perceptual consequence" are associated with *Contingency*, whereas "adds supplementary descriptive detail without causal or conditional dependence" corresponds to *Expansion*. Guided by these rules, the updated explanation further clarifies: "Thus, Arg2 expands Arg1's depiction, deepening atmosphere rather than specifying consequence," explicitly emphasizing the absence of causality. This refinement effectively removes the semantic ambiguity in the original explanation.

5 Conclusion

We proposed an SLM-LLM collaboration framework for explainable IDRR, where explanation-driven classification and classification-guided explanation refinement are jointly employed. By incorporating candidate explanations into SLM-based relation prediction for comparative reasoning and active selection, the framework provides explicit semantic cues for distinguishing subtle discourse relations. For hard cases, the LLM further generates discriminative rules to refine both prediction and explanation, reducing ambiguity between competing relations. Experiments on PDTB show that our method improves both relation classification and explanation quality, highlighting the potential of SLM-LLM collaboration for explainable discourse understanding.

Limitations

Despite the promising results, this work still has several limitations. First, the explanations generated in our framework are primarily task-oriented: they explain why a particular discourse relation holds between the two arguments, but do not explicitly reveal the internal reasoning process of the model itself. Therefore, the explainability provided by the current framework remains limited to the task level rather than the model level. Second, the evaluation of explanation quality still relies heavily on human judgments. Although human evaluation is important for assessing explanation plausibility, it is costly and may introduce subjectivity. Developing more reliable and automatic evaluation protocols for explanation quality remains an important direction for future work.

Acknowledgments

We sincerely thank all reviewers for their valuable efforts.

References

Hongxiao Bai and Hai Zhao. 2018. Deep Enhanced Representation for Implicit Discourse Relation Recognition. In *Proceedings of COLING*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Mingyang Cai, Zhen Yang, and Ping Jian. 2024. Improving implicit discourse relation recognition with semantics confrontation. In *Processings of LREC-COLING*.

Chunkit Chan, Cheng Jiayang, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. 2024. Exploring the Potential of ChatGPT on Sentence Level Relations: A Focus on Temporal, Causal, and Discourse Relations. In *Proceedings of EACL Findings*.

Chunkit Chan, Xin Liu, Jiayang Cheng, Zihan Li, Yangqiu Song, Ginny Wong, and Simon See. 2023. DiscoPrompt: Path Prediction Prompt Tuning for Implicit Discourse Relation Recognition. In *Proceedings of ACL Findings*.

Dong Chen, Yueting Zhuang, Shuo Zhang, Jinfeng Liu, Su Dong, and Siliang Tang. 2024a. Data shunt: Collaboration of small and large models for lower costs and better performance. In *Proceedings of AAAI*.

Wei Chen, Lili Zhao, Zhi Zheng, Tong Xu, Yang Wang, and Enhong Chen. 2024b. Double-checker: Large language model as a checker for few-shot named entity recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3172–3181.

Zujun Dou, Yu Hong, Yu Sun, and Guodong Zhou. 2021. CVAE-based Re-anchoring for Implicit Discourse Relation Classification. In *Proceeding of EMNLP Findings*.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2024. Minillm: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*.

Xinyu Hu and Xiaojun Wan. 2023. Exploring Discourse Structure in Document-level Machine Translation. In *Proceedings of EMNLP*.

Yuxin Jiang, Linhan Zhang, and Wei Wang. 2023. Global and Local Hierarchy-aware Contrastive Framework for Implicit Discourse Relation Recognition. In *Findings of ACL*.

Nicholas Kroeger, Dan Ley, Satyapriya Krishna, Chirag Agarwal, and Himabindu Lakkaraju. 2023. In-context explainers: Harnessing llms for explaining black box models. *arXiv preprint arXiv:2310.05797*.

Zhenwen Li, Wenhao Wu, and Sujian Li. 2020. Composing Elementary Discourse Units in Abstractive Summarization. In *Proceedings of ACL*.

Xin Liu, Jiefu Ou, Yangqiu Song, and Xin Jiang. 2020. On the Importance of Word and Sentence Representation Learning in Implicit Discourse Relation Classification. In *Proceeding of IJCAI*.

Wanqiu Long and Bonnie Webber. 2022. Facilitating Contrastive Learning of Discourse Relational Senses by Exploiting the Hierarchy of Sense Relations. In *Proceedings of EMNLP*.

Zheqi Lv, Tianyu Zhan, Wenjie Wang, Xinyu Lin, Shengyu Zhang, Wenqiao Zhang, Jiwei Li, Kun Kuang, and Fei Wu. 2025. Collaboration of large language models and small recommendation models for device-cloud recommendation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pages 962–973.

Kazumasa Omura, Fei Cheng, and Sadao Kurohashi. 2024. An Empirical Study of Synthetic Data Generation for Implicit Discourse Relation Recognition. In *Proceedings of COLING*.

Jialong Tang, Hongyu Lin, Meng Liao, Yaojie Lu, Xianpei Han, Le Sun, Weijian Xie, and Jin Xu. 2021. From Discourse to Narrative: Knowledge Projection for Event Relation Extraction. In *Proceedings of ACL*.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2025. Judging the judges: Evaluating alignment and vulnerabilities in LLMs-as-judges. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*.

Heng Wang and Changxing Wu. 2026. Improving implicit discourse relation recognition with natural language explanations from llms. In *Proceedings of AAAI*.

Linyong Wang, Lianwei Wu, Shaoqi Song, Yaxiong Wang, Cuiyun Gao, and Kang Wang. 2025. Distilling structured rationale from large language models to small language models for abstractive summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Changxing Wu, Liuwen Cao, Yubin Ge, Yang Liu, Min Zhang, and Jinsong Su. 2022. A Label Dependence-Aware Sequence Generation Model for Multi-Level Implicit Discourse Relation Recognition. In *Proceedings of AAAI*.

Wei Xiang, Zhenglin Wang, Lu Dai, and Bang Wang. 2022. ConnPrompt: Connective-cloze Prompt Learning for Implicit Discourse Relation Recognition. In *Proceedings of COLING*.

Ran Xu, Wenqi Shi, Yuchen Zhuang, Yue Yu, Joyce C Ho, Haoyu Wang, and Carl Yang. 2025. Collab-rag: Boosting retrieval-augmented generation for complex question answering via white-box and black-box llm collaboration. *arXiv e-prints*, pages arXiv–2504.

Youcheng Yan, Jinshuo Liu, Juan Deng, Junyan Li, Lina Wang, and Jeff Z Pan. 2025. Collaborate large and small language models for multi-modal emergency rumor detection. *Neural Networks*, 190:107625.

Frances Yung, Mansoor Ahmad, Merel Scholman, and Vera Demberg. 2024. Prompting Implicit Discourse Relation Annotation. In *Proceedings of the 18th Linguistic Annotation Workshop*.

Lei Zeng, Ruifang He, Haowen Sun, Jing Xu, Chang Liu, and Bo Wang. 2024. Global and Local Hierarchical Prompt Tuning Framework for Multi-level Implicit Discourse Relation Recognition. In *Proceedings of COLING*.

Biao Zhang, Jinsong Su, Deyi Xiong, Yaojie Lu, Hong Duan, and Junfeng Yao. 2015. Shallow Convolutional Neural Network for Implicit Discourse Relation Recognition. In *Proceedings of EMNLP*.

Biao Zhang, Deyi Xiong, Jinsong Su, and Min Zhang. 2018. Learning better discourse representation for implicit discourse relation recognition via attention networks. *Neurocomputing*, 275:1241–1249.

Yingxue Zhang, Fandong Meng, Peng Li, Ping Jian, and Jie Zhou. 2021. Context Tracking Network: Graph-based Context Modeling for Implicit Discourse Relation Recognition. In *Proceedings of NAACL*.

Haodong Zhao, Ruifang He, Mengnan Xiao, and Jing Xu. 2023. Infusing Hierarchical Guidance into Prompt Tuning: A Parameter-Efficient Framework for Multi-level Implicit Discourse Relation Recognition. In *Proceedings of ACL*.

Wenhao Zheng, Yixiao Chen, Weitong Zhang, Souvik Kundu, Yun Li, Zhengzhong Liu, Eric P Xing, Hongyi Wang, and Huaxiu Yao. 2025. Citer: Collaborative inference for efficient large language model decoding with token-level routing. In *Second Conference on Language Modeling*.

Hao Zhou, Man Lan, Yuanbin Wu, Yuefeng Chen, and Meirong Ma. 2022. Prompt-based Connective Prediction Method for Fine-grained Implicit Discourse Relation Recognition. In *Proceedings of EMNLP Findings*.

A Prompt Template

To improve reproducibility and clarify the interaction process between LLM and SLM, we provide the prompt templates used in our framework. Specifically, Figure 5 shows the template for candidate explanation generation, Figure 6 shows the explanation-enhanced IDRR template, Figure 7 shows the discriminative rule generation template , Figure 8 shows the discriminative rules-enhanced IDRR template, and Figure 9 shows the explanation refinement template.

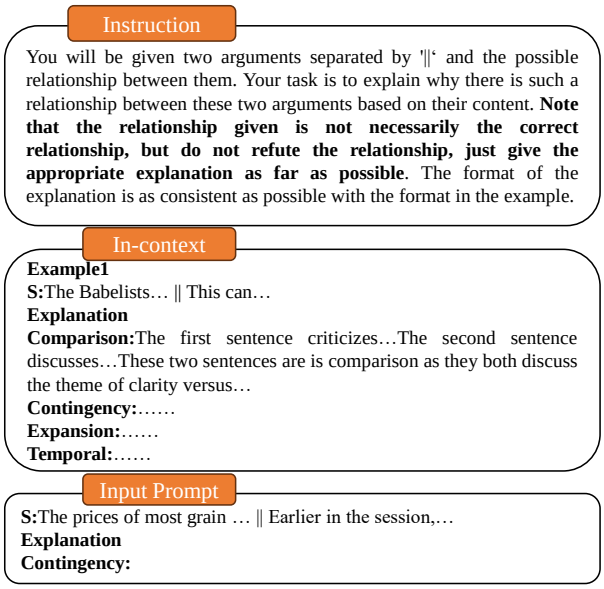

Figure 5: The Prompt Template of Candidate Explanation Generation T_L.

B Map Rules

Table 4 outlines the handcrafted mapping rules between connectives and relations, which are utilized in our verbalizer.

Arg1: arg_1. **Arg2:** arg_2. The conjunction between Arg1 and Arg2 is <mask>, so which of the following options represents the relationship between Arg1 and Arg2?
A.Comparison B.Contingency C.Expansion D.Temporal.
Reference explanation: </s>1.Comparison explanation </s>2.Contingency explanation </s>3.Expansion explanation </s>4.Temporal explanation.

Figure 6: The Prompt Template of Implicit Discourse Relation Recognition based on explanation enhanced T_s.

Below is an implicit discourse relation sample:
Arg1: $arg1 Arg2: $arg2
The two potential relation categories are: $class1 and $class2.
Your task: Generate only the instance-specific discrimination rules for distinguishing $class1 and $class2.
Follow all requirements below:
[STRICT REQUIREMENTS]
1. Output exactly 2 rules.
2. Each rule must explicitly differentiate two relation categories based on this specific instance (no generic category definitions).
3. Do NOT output explanations, introductions, conclusions, or any extra text.
4. The output MUST strictly follow format below, with no additional characters:
rules: 1... 2...

Figure 7: The Prompt Template of Discriminative Rules Generation T_R.

C Experimental Dataset

We split the PDTB corpus into training, validation, and test sets, as shown in Table 5.

D Criterion of Human Evaluation

We evaluate the generated explanations acorss three aspects: interpretability of explanation, factuality of arguments restatement, and fluency of language. Five human annotators were recruited for the evaluation. Each instance was independently scored by two annotators, and a third was involved if their scores differed by more than two points. The final score for each instance was computed as the average of the two closest scores. The detail scoring criteria are provided in Table 6.

Arg1: arg_1 . **Arg2:** arg_2. The conjunction between Arg1 and Arg2 is <mask> , so which of the following options represents the relationship between Arg1 and Arg2 ?
A. Comparison B. Contingency C. Expansion D. Temporal
The two confusing relationship categories are $confusion_class_1$ and $confusion_class_2$ The determination rules for these two relationship are as follows: rules

Figure 8: The Prompt Template of Implicit Discourse Relation Recognition Based on Discriminative Rules T_r.

Task: Improve the original explanation based on the provided materials. Your goal is not to determine the correct relation, but to generate a persuasive explanation for the given model prediction.
Output Requirements:
1. Format: Keep EXACTLY the same two-part structure as Original Explanation (Argument Restatement + Relation Justification)
2. Do NOT explicitly mention 'Rule 1/2' in the output and use the rules ONLY as internal reference for judgment
3. Word count: ≤120 words (English)
4. Writing style: Match the Original Explanation's tone and structure
5. Use the distinguishing rules only to strengthen the justification of the predicted relation and to implicitly refute the confusing relation.

Figure 9: The Template of Explanation Refinement T_u.

Relation	Connective
Comparison	although, nevertheless, but, however
Contingency	because, as, so, consequently, thus, since,
Expansion	instead, rather, or, and, also, furthermore, instance, example, first, indeed, specifically
Temporal	then, subsequently, previously, earlier, after, meanwhile

Table 4: Connective-to-Relation mapping rules.

Dataset	Training	Validation	Test
Total instances	12,775	1,183	1,046
Explanations	51100	4732	4184
Expansion	6,900	651	556
Comparison	1,898	191	146
Contingency	3,288	287	276
Temporal	689	54	68

Table 5: Statistics of the datasets for IDRR with explanations.

Aspect	Question	Score: 0	Score: 1	Score: 2
Interpretability of explanation	Does the rationale part justify the predicted label?	No (The relationship labels are misunderstood or incorrectly interpreted, and the explanation is unfaithful.)	Weak-Yes (The relationship labels are correctly understood, but the explanation is partially faithful.)	Yes (The relationship labels are correctly understood, and the explanation faithfully interprets the relationship between two arguments.)
Factuality of arguments restatement	Does the restatement part align with the facts stated in the arguments	No (The description contains significant errors and fails to reflect the original argument's meaning.)	Weak-Yes (The description is mostly accurate but contains minor errors or vague details.)	Yes (The description is entirely accurate and fully reflects the original argument's meaning.)
Fluency of language	Is the explanation grammatically correct and natural?	No (The text is not fluent, containing obvious grammatical errors or expressions that are difficult to understand.)	Yes (The text is fluent, and the expression is clear.)	-

Table 6: The criterion for human evaluation.