Linguistic Cues for LLM-based Implicit Discourse Relation Classification

Yi Fan and Michael Strube and Wei Liu[*]
Heidelberg Institute for Theoretical Studies
{yi.fan, michael.strube}@h-its.org, wei.liu.llm@gmail.com

Abstract

Large language models (LLMs) have achieved impressive success across many NLP tasks, yet implicit discourse relation classification (IDRC) is still dominated by encoder-only pretrained language models such as RoBERTa. This may be due to earlier reports that ChatGPT performs poorly on IDRC in zero-shot settings. In this paper, we show that fine-tuned LLMs can perform on par with, or even better than, existing encoder-based approaches. Nevertheless, we find that LLMs alone struggle to capture subtle lexical relations between arguments for the task. To address this, we propose a two-step strategy that enriches arguments with explicit lexical-level semantic cues before fine-tuning. Experiments demonstrate substantial gains, particularly in cross-domain scenarios, with F1 scores improved by more than 10 points compared to strong baselines.

1 Introduction

Discourse relations, such as *Cause* and *Contrast*, describe the logical connections between two text spans (Pitler et al., 2009). Accurately recognizing these relations is beneficial for many downstream NLP tasks, including coherence modeling (Lin et al., 2011; Liu and Strube, 2025a,b), reading comprehension (Mihaylov and Frank, 2019), argumentation mining (Habernal and Gurevych, 2017; Hewett et al., 2019), and text summarization (Liu et al., 2025). Discourse connectives (e.g., *but*, *as a result*) often signal the presence of a relation (Pitler and Nenkova, 2009). They can be explicit, as in (1), or implicit, as in (2):

(1) [I refused to pay the cobbler the full \$95]$_{\text{Arg1}}$ **because** [he did poor work.]$_{\text{Arg2}}$

(2) [It requires that "discharges of pollutants" into the "waters of the United States" be authorized by permits that reflect the effluent limitations developed under section 301.]$_{\text{Arg1}}$ (**Implicit=However**) [Whatever may be the problems with this system, it scarcely reflects "zero risk" or "zero discharge."]$_{\text{Arg2}}$

When connectives are explicit, classifying the relation is relatively easy. For example, Pitler and Nenkova (2009) show that using only connectives yields 85.8% accuracy on 4-way explicit classification in PDTB 2.0. In implicit cases, however, the absence of such signals makes classification more challenging (Zhou et al., 2010; Shi et al., 2017). As a result, most prior work on discourse relation classification has focused on implicit relations.

Method	Encoder	F1
(Zeng et al., 2024)	RoBERTa-base	73.89
(Cai et al., 2024)	RoBERTa-base	72.11
Llama-3-8B	Llama-3-8B	67.00
Llama-3-70B	Llama-3-70B	75.88

Table 1: Comparison between methods based on different encoders on PDTB 3.0 Level 1 test set.

Existing approaches to IDRC mainly focus on improving text encoders, from leveraging pretrained models such as RoBERTa (Liu et al., 2019) to incorporating artificially generated discourse connectives. These methods have achieved good results, with top-level classification accuracies in PDTB 3.0 exceeding 70% (Table 1).

With the advent of large language models (LLMs), the NLP community has witnessed their remarkable capabilities. LLMs now dominate leaderboards across a wide range of NLP tasks, sometimes even surpassing human performance. This shift has driven many research areas from encoder-only PLM-based approaches toward LLM-based ones. Yet, somewhat surprisingly, implicit discourse relation classification has remained dominated by PLM methods (see the most recent work in Table 1). A likely reason is the widely cited finding that ChatGPT performs poorly on this task

[*] Corresponding author.

Findings of the Association for Computational Linguistics: EACL 2026, pages 4585–4602
March 24-29, 2026 ©2026 Association for Computational Linguistics

in zero-shot settings, with accuracy dropping below 20% on PDTB 2.0 (Chan et al., 2024). While such results are valid, they paint an incomplete picture; in fine-tuned settings, LLMs can, in fact, perform strongly. As shown in Table 1, finetuned LLaMA achieves performance competitive with systems based on RoBERTa (Liu et al., 2019), suggesting that LLMs should be regarded as strong new baselines for this task.

Nevertheless, we observe that even fine-tuned LLMs sometimes fail in cases where clear lexical cues are present. For example, when handling Example (2), the model is likely confused by the concessive clause in Arg2 ("Whatever may be...") and defaults to a familiar but incorrect pattern, predicting it as a *Cause* relation despite strong lexical indicators such as "discharges of pollutants" and "effluent limitations" pointing toward *Contrast*. This raises the question: can linguistic features, specifically semantic relations between lexical pairs, further improve LLMs' performance?

To answer this, we suggest a two-step method that explicitly embeds lexicon-level semantic relation features into LLMs. Our experiments demonstrate that this approach yields significant improvements over strong baselines and greatly enhances performance on a cross-domain corpus. Our contributions are summarized as follows:

- Empirical analysis of LLMs for implicit discourse relation classification. We systematically evaluate the performance of finetuned LLMs on PDTB 2.0/3.0 and show that, contrary to previous zero-shot reports, LLMs are strong baselines when properly finetuned.
- Identification of limitations in LLM predictions. We find that even finetuned LLMs can make systematic errors in cases with strong lexical cues, highlighting the potential benefit of integrating linguistic knowledge.
- A novel lexicon-enhanced LLM framework. We propose a two-step approach that explicitly incorporates semantic relations between lexical pairs to enhance LLMs' understanding of discourse relations.
- Empirical validation of effectiveness. Extensive experiments demonstrate that our method consistently improves performance over existing PLM-based and LLM-based baselines, confirming the value of lexicon-level features.
- Robustness and generalization capabilities. We evaluate our model in a zero-shot cross-domain setting on a genre-mixed corpus comprising texts from political, literary, and encyclopedic domains, showing the robustness and adaptability of our approaches.

2 Related Work

Our work builds upon existing research in implicit discourse relation classification (IDRC), particularly studies that utilize linguistically informed features and neural networks.

Linguistic Feature Enhanced IDRC. Implicit discourse relation classification is a challenging aspect of shallow discourse parsing and has attracted significant attention since the release of PDTB 2.0. Early work focuses on the role of linguistically informed features derived from argument pairs. For instance, Pitler et al. (2009) investigate whether features such as lexicon, Levin verb classes, and verb phrase length could benefit the task. They find that these features are strong indicators of discourse relation types. Rutherford and Xue (2014) introduce Brown cluster pairs to represent discourse relations and incorporates coreference patterns to identify implicit discourse relation senses. Our work also leverages linguistically informed features, particularly the lexicon relations between arguments, for IDRC. However, we explore this approach within the context of Large Language Models (LLMs) and demonstrate that these features continue to enhance LLMs' performance.

Neural Approaches for IDRC. With the rise of deep learning, neural networks have become a popular method for IDRC. Zhang et al. (2015) introduce a shallow Convolutional Neural Network (CNN) for this task, while Rutherford et al. (2016) develop an LSTM-based method to automatically extract features from arguments without manual feature engineering. Qin et al. (2017) propose an adversarial model that transferred knowledge from a model with explicit connectives to one without.

Following the emergence of pre-trained models, much of the focus shifted to building classifiers based on these models. Shi and Demberg (2019) show that the next-sentence prediction task benefits IDRC. Kishimoto et al. (2020) proposes a multi-task framework based on BERT to predict discourse connectives and relations simultaneously. Liu and Strube (2023) and Liu et al. (2024) explore an annotation-inspired model that jointly generates discourse connectives between arguments and predicts a relation based on the gen-

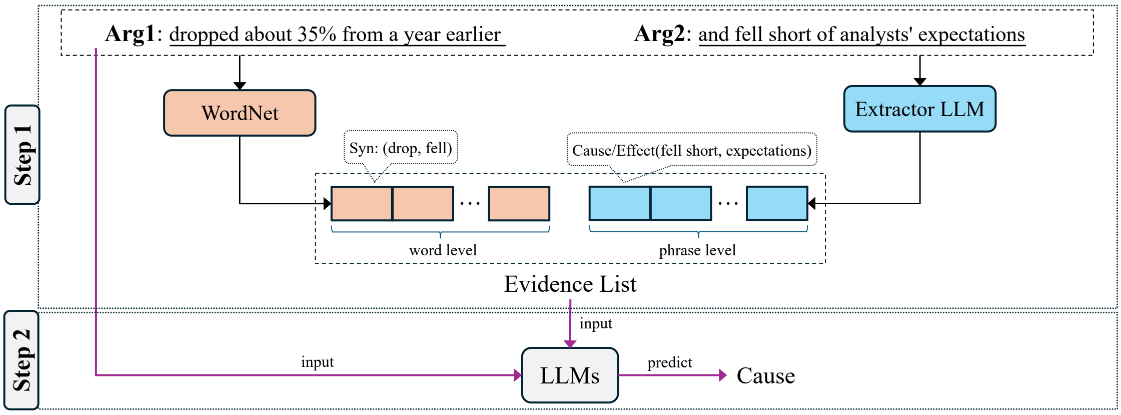

Arg1: dropped about 35% from a year earlier
Arg2: and fell short of analysts' expectations
Step 1
WordNet
Extractor LLM
Syn: (drop, fell)
Cause/Effect(fell short, expectations)
word level
phrase level
Evidence List
input
Step 2
input
LLMs
predict
Cause

Figure 1: An overview of our approach: In Step 1, we extract lexical-semantic and pragmatic features using WordNet and LLM. In Step 2, we predict the relation based on the input arguments and the extracted features.

erated connectives and arguments. Cai et al. (2024) introduce SCIDER, a RoBERTa-based approach for IDRC, designed to disentangle logical semantics from general semantics. Zeng et al. (2024) propose a hierarchical prompt-tuning framework for multi-level IDRC. Although these models have achieved impressive performance, they largely rely on pre-trained models like BERT and RoBERTa and have paid little attention to the performance of Large Language Models (LLMs) in this task.

Chan et al. (2024) is one of the few studies that investigates the performance of LLMs on IDRC. Their evaluation of ChatGPT's zero-shot performance finds that it struggles with the task. Another notable work is by Wang et al. (2025a), which employs GPT-3.5 to generate a subtext for argument pairs before fine-tuning LLaMA for IDRC. However, their results show that the LLaMA-based approach underperformed compared to RoBERTa-based models. Our work also uses LLaMA for IDRC, but we gain very different observations. We demonstrate that, with carefully designed prompts, LLaMA can achieve performance comparable to RoBERTa-based methods. Furthermore, when enhanced with lexical-semantic relations or scaled with larger LLMs, the LLaMA-based approach significantly outperforms RoBERTa, especially in cross-domain settings.

3 Method

To tackle the challenge of implicit discourse relation prediction, we introduce a two-step, feature-enhanced framework. The main idea of our method is first to clarify the latent semantic and pragmatic links between arguments, and then use these clarified signals to direct a language model in its final prediction. This process involves a Hybrid Semantic Feature Extraction step followed by a Feature-Guided Relation Prediction step. Figure 1 offers a broad overview of the entire process.

3.1 Hybrid Semantic Feature Extraction

The goal of the first step is to create a detailed and organized Evidence List that captures the various connections between the two arguments (Arg1 and Arg2). Acknowledging that different types of relations are represented through other methods, we use a hybrid approach that combines a traditional lexical knowledge base with the inferential capabilities of a large language model.

Lexical-Semantic Features from WordNet. To establish a foundation of structured lexical knowledge, we use WordNet (Miller, 1994), a large and widely-used lexical database of English. We query WordNet to gather a set of semantic relations between content words (nouns, verbs, adjectives, and adverbs) in the argument pair. The usefulness of these features is best shown through specific examples. Consider the following implicit relation, where the correct label is *Contingency.Cause*:

(3) [The company spent the last year aggressively developing its presence in the Asian market.]$_{\text{Arg1}}$ (**Implicit=Consequently**) [Its sales in the region grew by 30% in the most recent quarter.]$_{\text{Arg2}}$

In (3), WordNet provides an important indication by identifying a causal connection between the main verbs: `Causes(developing, grew)`. This feature offers a direct lexical basis for a causal inference, guiding the model to link "developing" in Arg1 and "grew" in Arg2.

Similarly, for a *Comparison.Contrast* relation:

(4) [The $833.6 million figure includes the new acquisitions.]$_{\text{Arg1}}$ (**Implicit=By comparison**)

[Excluding those businesses, earnings before interest, taxes and depreciation for 1988 would have been $728.5 million.]$_{\text{Arg2}}$

Here, WordNet highlights the strong and clear antonym pair `Antonym(includes, excluding)`. This explicit opposition acts as a solid starting point, guiding the model's focus on the natural contrast between the two concepts. As these examples show, features derived from WordNet connect the prediction task to concrete lexical evidence, offering clear and focused signals that are essential for the model's reasoning process.

Thus, the extracted relations include: `Synonyms and Antonyms` (Halliday and Hasan, 1976; Polanyi et al., 2004; Pitler et al., 2009; Das and Taboada, 2013), `Hypernyms and Hyponyms` (Halliday and Hasan, 1976), `Meronyms and Holonyms` (Bärenfänger et al., 2008), `Co-hyponyms` (Louis and Nenkova, 2011), `Verb Entailments and Causes` (Taboada and Das, 2013; Bott and Solstad, 2014) and `Similar To` (Pitler et al., 2009). Detailed reasons can be found in Appendix E. Together, these WordNet-derived features create a structured network of lexical links between the two arguments, offering a strong, knowledge-based foundation for the next step of our framework.

LLM-Extracted Semantic and Pragmatic Features. While WordNet provides a valuable foundation for lexical semantics, it is inherently limited to predefined, context-free word relationships. It cannot capture propositional logic, contextual subtleties, or the overall pragmatic functions crucial for interpreting implicit discourse relations. To fill this gap, we use Llama-3-70B as a versatile and comprehensive feature extractor, capable of understanding context and abstract structures to identify cues that are hidden from lexical databases.

The strength of this approach is best shown through examples where the discourse relation depends on such higher-level understanding. Consider the following *Comparison.Concession* relation:

(5) [Tanks currently are defined as armored vehicles weighing 25 tons or more that carry large guns.]$_{\text{Arg1}}$ (**Implicit=But**) [The Soviets complicated the issue by offering to include light tanks, which are as light as 10 tons.]$_{\text{Arg2}}$

Here, the model identifies a `Polarity Contrast` between the phrases "25 tons or more" and "as light as 10 tons." This feature explicitly reveals the latent opposition between the concepts of 'heavy' and 'light,' offering a strong contrast signal that goes beyond simple antonymy and would be overlooked by a tool like WordNet.

For an *Expansion.Level-of-detail* relation:

(6) [a country of about three million people with a relatively high soft-drink consumption rate]$_{\text{Arg1}}$ (**Implicit=Specifically**) [In Singapore, per-capita consumption is about one-third that of the U.S.]$_{\text{Arg2}}$

In this case, the model identifies the holistic `Generalization-Specification` structure. It recognizes that Arg1 makes a broad assertion ("a relatively high... consumption rate"), while Arg2 provides specific data to support and detail this claim. Understanding the pragmatic function of each argument is essential for correctly identifying the elaborative nature of the discourse.

These examples highlight the main contribution of our LLM-based extractor. The model is prompted to detect a wide range of these higher-level relationships, extending beyond individual words to include phrases, events, and the overall discourse structure. Although our prompt also asks the model to recognize lexical relations in WordNet (e.g., Antonyms, Synonyms), this serves two purposes: it helps the model identify these relations in multi-word expressions where WordNet's sense distinctions are too strict, and it consolidates all features into a single, unified Evidence List. The main contribution, however, comes from the following three categories of LLM-exclusive features:

Polarity and Degree Relations (Corston-Oliver, 1998), including `Polarity_Contrast` Zirn et al. (2011) and `Degree_Comparison` (Pitler et al., 2009).

Propositional and Event Relations, including `Cause-Effect`, `Condition/Consequence`, `Action/Purpose` (Asr and Demberg, 2012; Taboada and Das, 2013; Bott and Solstad, 2014) and `Temporal_Sequence` (Taboada and Das, 2013; Ning et al., 2018).

Structural and Pragmatic Relations, including `Generalization-Specification`, `Action-Manner`, `Statement-Elaboration` (Taboada and Das, 2013; Lin et al., 2009), `Claim-Justification` (Zhang et al., 2016), `Alternative_Choice`, `List_Continuation` (Pitler et al., 2009; Lin et al., 2009) and dialogic features like `Question-Answer` and `Rhetorical_Continuation` (Asr and Demberg, 2012; Webber et al., 2019).

The final output of this step is a single, semicolon-separated string called the Evidence List, which consolidates all features extracted from the LLM. This list provides the explicit, structured knowledge that will be used in the next step.

3.2 Feature-Guided Relation Prediction

After extracting extensive semantic and pragmatic cues in Step 1, the second step is supervised fine-tuning, where the model learns to use feature-enhanced input for final predictions. This step aims to overcome the weakness of naive fine-tuning: its reliance on superficial correlations instead of engaging with semantic cues. The WordNet hints and Evidence List serve as explicit linguistic scaffolding input to the model with the argument pair, so the model no longer needs to discover these connections from scratch. Its main task becomes to accurately predict the ground-truth discourse relation by learning to weigh the importance of different features we extracted within the context of the raw text. For example, it must understand that the presence of a `Cause-Effect` feature significantly increases the probability of a *Cause* relation, especially when the arguments themselves contain causal language. This process pushes the model beyond simple pattern matching and encourages it to engage with the linguistic foundations of the discourse.

We use LoRA fine-tuning techniques (Hu et al., 2022) to train the model for this task. This method shifts the model from being merely a probabilistic pattern matcher to a more intentional, linguistically informed reasoner, enabling it to leverage explicit semantic and pragmatic cues for more accurate and robust predictions.

4 Experiments

We evaluate our proposed method using standard benchmarks. Our experiments have three main goals. First, we aim to measure the impact of our method by comparing it to a standard fine-tuning baseline with the same LLMs on both in-domain and out-of-domain data. Second, we want to validate whether LLM is a suitable choice for IDRC by comparing it with previous models that use smaller encoder architectures, such as RoBERTa. Finally, through these comparisons, we aim to demonstrate the potential of LLMs for this task.

4.1 Experimental Settings

Datasets. Our experiments use three datasets: Penn Discourse TreeBank (PDTB) 2.0 and 3.0 (Prasad et al., 2008; Webber et al., 2019) and DiscoGeM (Scholman et al., 2022). The PDTB is the standard benchmark for IDRC. Following prior work (Ji and Eisenstein, 2015; Kim et al., 2020), we evaluate our models at the top-level (4-way) and second-level (11/14-way). We adopt the standard splits established by Ji and Eisenstein (2015). The models are fine-tuned and tested on both datasets to ensure a thorough in-domain performance assessment. Besides, to evaluate the robustness and domain generalization abilities of our framework, we use the DiscoGeM 1.0 corpus (Scholman et al., 2022). This dataset contains a diverse range of genres, including political speeches, literature, and encyclopedic texts. Notably, we conduct zero-shot cross-domain testing: the models fine-tuned only on PDTB 3.0 are tested on DiscoGeM 1.0 test set using the same split in Yung et al. (2022).

Implementation Details. Our experiments employ the Llama-3-8B-Instruct and Llama-3-70B-Instruct models (Grattafiori et al., 2024) from HuggingFace[1] to examine how our method affects models of different sizes. The implementation follows our two-step framework:

• Step 1 (Feature Extraction): To guarantee the highest quality of semantic and pragmatic cues, the Evidence List for all experiments (covering both 8B and 70B models) was created using the more powerful Llama-3-70B-Instruct model. This ensures a consistent and high-quality set of features for the second step.

• Step 2 (Fine-tuning and Prediction): During the main prediction step, we fine-tune both the Llama-3-8B-Instruct and Llama-3-70B-Instruct models. This setup enables a direct comparison of the performance of each model size with and without our feature-augmentation method. All models are fine-tuned with LoRA (Hu et al., 2022).

Evaluation Metrics. We adopt standard Accuracy and Macro-F1 as our evaluation metrics. Since some test instances are annotated with multiple labels, there are two different ways (Omura et al., 2024) to compute these metrics: (**M1**) Following the official setup of CoNLL 2015 (Xue et al., 2015), a prediction is counted as correct if it matches any of the annotated labels. For example, if the prediction is A while the gold labels are A and B, the

[1] `https://huggingface.co/`

	Model	PDTB-2				PDTB-3			
		Top Level		Second Level		Top Level		Second Level	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1
M1	(Wu et al., 2022)	71.18	63.73	60.33	40.49	-	-	-	-
	(Zhou et al., 2022)	70.84	64.95	60.54	41.55	-	-	-	-
	(Liu and Strube, 2023)	74.59	68.64	62.75	42.36	76.23	71.15	65.51	54.92
	(Jiang et al., 2023)	74.67	69.60	63.91	47.91	76.39	74.21	66.42	60.11
	Llama-3-8B-Instruct	69.02	60.44	55.82	34.22	72.12	67.00	59.89	48.64
	Ours$_{8b}$	76.39	71.30	64.49	48.37	77.82	73.34	67.74	60.00
	Ours$_{8B}$ w\o Semantic & Pragmatic	74.38	67.25	63.04	42.97	76.19	71.53	65.21	54.58
	Ours$_{8B}$ w\o WordNet Info	73.33	66.56	63.52	45.30	76.05	70.71	66.30	57.50
	Llama-3-70B-Instruct	75.14	69.02	65.06	47.41	79.72	75.88	69.99	58.40
	Ours$_{70B}$	**77.82**	**71.86**	**69.01**	**51.30**	**81.00**	**77.04**	**70.74**	**61.08**
	Ours$_{70B}$ w\o Semantic & Pragmatic	76.48	70.92	65.93	47.78	80.73	76.51	70.49	61.43
	Ours$_{70B}$ w\o WordNet Info	76.86	70.89	66.79	50.30	80.53	76.69	70.53	62.26
M2	(Long and Webber, 2022)	72.18	69.60	61.69	49.66	75.31	70.05	64.68	57.62
	(Chan et al., 2023)	73.80	67.79	61.41	44.04	-	-	-	-
	(Chan et al., 2023)	78.06	**75.34**	68.14	52.42	-	-	-	-
	(Zeng et al., 2024)	73.89	67.83	61.33	46.14	75.53	71.59	64.87	56.50
	(Cai et al., 2024)	72.11	67.00	59.62	-	-	-	-	-
	(Wang et al., 2025b)	78.20	71.14	62.46	46.38	76.93	73.33	64.35	55.04
	Ours$_{8b}$	77.15	73.89	65.61	53.17	78.61	73.88	68.86	62.29
	Ours$_{70B}$	**78.54**	74.55	**69.96**	**55.28**	**81.70**	**77.46**	**71.74**	**64.34**

Table 2: Main results on the PDTB 2.0 and PDTB 3.0 test sets for implicit discourse relation prediction. M1 and M2 indicate different evaluation metrics. We report accuracy (Acc) and Macro-F1 scores for both Top Level (4-way) and Second Level (11-way/14-way) classification. Llama-3-X-Instruct models serve as our standard fine-tuning baselines. Ours$_X$ refers to our full, feature-augmented models fine-tuned on the corresponding Llama-3 architecture. w/o (without) indicates ablation studies where a specific feature set was removed to assess its contribution. Best results in each column are in bold. Ours$_{8b}$ and Ours$_{70b}$ are statistically significantly better than the corresponding Llama baseline model, with a p-value of 0.05 based on five runs.

prediction is considered correct. (**M2**) In the alternative approach, a prediction that matches one of the annotated labels is counted as multiple correct predictions. For instance, if the prediction is A and the gold labels are A and B, this is counted as two correct matches: (pred = A, label = A) and (pred = B, label = B). This method may inflate the Macro-F1 score, since unpredicted labels can still be included in the pool of correctly predicted instances.

4.2 PDTB Test Results

The results, presented in Table 2, demonstrate that our feature-augmented models, Ours$_{8b}$ and Ours$_{70b}$, consistently establish strong performance on both PDTB-2 and PDTB-3 across two distinct evaluation settings (M1 and M2). Our analysis focuses on two key aspects: the significant impact of our framework on smaller models and the insights gained from the ablation studies.

The most notable finding is the beneficial effect our method has on the smaller Llama-3-8B model. While the standard fine-tuned Llama-3-8B-Instruct baseline struggles to compete with strong baseline models, our Ours$_{8b}$ model not only surpasses them but also outperforms the much larger Llama-3-70B-Instruct baseline. This result strongly supports our hypothesis that LLMs lack the innate ability to consistently focus on the subtle semantic cues necessary for IDRC. Besides, the ablation studies further dissect the sources of this performance gain. Removing Semantic & Pragmatic features, or WordNet Information, causes the performance drop. Besides, the performance drop upon feature removal is generally more pronounced for the 8B model than the 70B model, indicating that these explicit features act as crucial hints for smaller models, helping them notice the cues. We provide a more detailed ablation analysis in Section 5.

	Model	Europarl		Novel		Wiki	
		Acc	F1	Acc	F1	Acc	F1
Top Level	Llama-3-8B-Instruct	44.44	20.37	53.96	44.43	38.17	13.06
	Llama-3-70B-Instruct	59.07	42.04	54.44	45.93	73.28	48.92
	Ours$_{8b}$	55.93	37.92	58.00	49.41	71.76	42.18
	Ours$_{70B}$	60.74	47.21	56.54	47.78	75.57	51.48
Second Level	single. (Yung et al., 2022)	53.25	25.88	45.31	23.10	45.58	24.02
	Llama-3-8B-Instruct	46.64	24.65	43.74	23.16	49.59	37.87
	Llama-3-70B-Instruct	52.57	28.87	49.55	29.87	55.37	35.11
	Ours$_{8b}$	50.79	29.29	47.73	26.71	60.33	43.09
	Ours$_{70B}$	55.73	32.96	52.09	33.82	61.98	42.50

Table 3: Zero-shot cross-domain performance on the DiscoGeM 1.0 test set (Europarl, Novel, and Wiki) using the single label setting (Yung et al., 2022). All our models (Ours$_X$) and the Llama-3 baselines were fine-tuned only on PDTB 3 and evaluated directly on the three DiscoGeM domains without additional training. Ours$_X$ indicates our feature-augmented models. For comparison, we include single. (Yung et al., 2022), a strong baseline trained on DiscoGeM's in-domain training set. Our evaluation strictly follows the single-label setting from their work.

4.3 Cross-domain Test Results

The results on the DiscoGeM dataset, shown in Table 3, demonstrate the robustness and generalization abilities of our feature-augmented method. First, our models demonstrate generalization ability, achieving performance that is highly competitive with, and in some cases surpasses, a strong baseline that was trained on in-domain data. This result underscores that our method, by guiding the model with explicit semantic features, enables it to learn the fundamental, transferable principles of discourse, rather than simply memorizing the stylistic patterns of the news domain in which it was trained. Second, the most significant impact of our method is observed on the smaller model. This supports our key linguistic hypothesis: semantic and pragmatic relationships are universal principles of language that stay consistent across different domains. A smaller model, when fine-tuned naively, struggles to identify these principles independently. Our framework makes this universal knowledge accessible, ensuring stable and reliable reasoning even for smaller models in new contexts. Finally, these results suggest that for this complex task, choosing an effective method matters more than increasing the model size. The 70B baseline, despite its large size, still struggles with effective generalization, indicating that its semantic understanding is limited. While Ours$_{70b}$ model achieves the best overall performance, the Ours$_{8b}$ model consistently closes the performance gap and becomes competitive. This indicates that building a generalizable discourse parser is not just about scaling up. Instead, the key is using a method that requires the model to base its predictions on domain-invariant linguistic evidence rather than superficial, domain-specific cues, regardless of its size.

5 Analysis

The results show significant performance gains of our feature-augment method across various datasets. This section analyzes why, focusing on two questions. First, **what are the relative impacts of different types of semantic and pragmatic features on the model's performance?** We analyze this quantitatively through a series of ablation studies. Second, **how do these explicit linguistic cues help the model correct errors that occur with standard fine-tuning methods?** We explore this with a case study, highlighting where our model succeeds but the baseline fails. Since feature effects were consistent across models and the 8B model benefits greatly, results for Ours$_{8b}$ are presented.

5.1 Ablation Study

This subsection examines the contributions of our linguistic features. To assess the contributions of our feature categories, we conducted ablation studies by removing each group. As shown in Table 4, our Ours$_{8b}$ model outperforms the Llama-3-8B-Instruct baseline across all six high-frequency relations. Furthermore, the ablation results confirm that every feature category contributes positively to this success, as removing any single

Method	Concession	Cause	Conjunction	Level-of-Detail	Asynchronous	Instantiation
- Synonyms & Antonyms	-1.37	-1.80	0.10	-1.21	-0.95	-0.76
- Hypernyms & Hyponyms	-9.60	-2.46	-0.85	-0.82	-0.95	0.02
- Meronyms & Holonyms	-3.98	-1.59	-1.58	-1.91	-1.60	-2.10
- Co-hyponyms	-2.06	-1.78	-1.29	-0.36	0.03	-0.14
- Verb Entailments & Causes	-2.68	-1.34	-0.80	-2.14	-2.26	-1.30
- Similar To	-3.78	-1.66	-1.33	-2.13	-2.26	-2.36
- Polarity & Degree	-8.51	-2.51	-0.61	0.04	1.29	-0.62
- Propositional & Event Rel	-2.77	-1.93	-1.41	0.14	-0.95	-2.10
- Structural & Pragmatic Rel	-4.17	-1.58	-1.20	-1.44	1.64	-4.12

Table 4: Impact of feature category ablation on the F1 score of our Ours$_{8b}$ model, broken down by the six most frequent second-level relations on the PDTB-3 test set. Each cell shows the change in F1 score for a specific relation when the feature category in that row is removed. Red numbers indicate a decrease in performance, while blue numbers indicate an increase in performance.

group degrades performance. The analysis reveals strong, linguistically-grounded dependencies between specific relations and our feature types. For instance, performance on *Concession* is highly sensitive to `Hypernyms & Hyponyms` and `Polarity & Degree`. This is because concessionary logic is built on a conflict of expectation, which is often signaled by a Polarity_Contrast, or by contrasting a general rule with a specific exception (a `Hypernym-Hyponym` structure). Similarly, predicting *Cause* relations relies heavily on `Propositional & Event Relations`, which contain explicit `Cause-Effect` cues, and `Polarity & Degree`, which help characterize the evaluative nature of causal outcomes.

5.2 Case Study

To qualitatively analyze how these linguistic cues support the model's reasoning, we examine an example where our model succeeds while the baseline fails. In 94% of cases where the baseline misclassified a *Cause* relation and our model correctly predicted it, an explicit `Cause-Effect`, `Polarity Contrast`, or `Claim-Justification` feature appeared in the Evidence List. Consider the following example, correctly labeled as *Cause*:

(7) [Not to mention the incursion of imports.]$_{\text{Arg1}}$ (**Implicit=As**) [Japanese and European steelmakers are anxiously awaiting the lifting of trade restraints in 1992.]$_{\text{Arg2}}$
Evidence List: [Lexical(imports, trade restraints); Lexical(imports, steelmakers); Cause-Effect(incursion of imports, anxiously awaiting); Polarity_Contrast(Not to mention, anxiously awaiting); Generalization-Specification; Claim-Justification; Statement-Elaboration]

This case is hard due to the syntactic disconnect between the noun phrase in Arg1 and the full sentence in Arg2. The baseline model, unable to infer the deep connection, defaults to a simple *Conjunction*. Our model, however, correctly identifies the *Cause* relation by leveraging the provided Evidence List. First, features like `Lexical(imports, steelmakers)` establish a strong thematic bridge. More decisively, the `Cause-Effect(incursion of imports, anxiously awaiting)` feature makes the latent causal logic explicit: the eagerness of foreign steelmakers is the underlying cause for the import problem being a significant concern. These explicit cues enable our model to prioritize the core semantic relationship over the disconnected surface structure, resulting in accurate predictions.

6 Conclusions

In this paper, we introduced a two-step, feature-augmented framework that enriches model inputs with a comprehensive set of explicit lexical, semantic, and pragmatic features to capture subtle cues for implicit discourse relation prediction.

Our experiments demonstrate that our method not only outperforms strong baselines on the PDTB but also exhibits satisfactory zero-shot robustness on the out-of-domain DiscoGeM corpus. Crucially, this approach disproportionately benefits smaller models, elevating an 8B model to be competitive with 70B baselines. This key finding suggests that for complex inferential tasks, a methodology that provides explicit linguistic guidance is more impactful than brute-force model scaling alone.

Future work could expand and refine our linguistically inspired feature to enhance the performance of LLMs on this task. Additionally, investigating more languages and genres to understand the link between lexicon-level semantics and discourse relations, as well as uncovering nuanced linguistic phenomena, is also an exciting direction.

Limitations

While our feature-augmented framework shows significant improvements, it is essential to acknowledge its limitations to provide a comprehensive picture and inform future research. First, the quality of the semantic and pragmatic cues generated in our initial step is fundamentally limited by the capabilities of the feature extractor model. In this work, we used Llama-3-70B, a powerful open-source model. However, due to computational and financial constraints, we did not explore the use of even larger, state-of-the-art proprietary models. Future work could examine whether using more advanced models, such as GPT-4 or the Gemini family, for feature extraction might produce a more comprehensive and accurate "Evidence List," potentially leading to further performance gains. Second, our study is monolingual, with all experiments conducted solely on English datasets. The lexical, semantic, and pragmatic cues that we identify as beneficial for IDRC in English may not directly apply to other languages, especially those with different syntactic structures, cultural contexts, or discourse conventions. Therefore, the applicability of our findings beyond English is uncertain, and extensive cross-lingual research would be necessary to adapt and validate this approach for other languages.

Acknowledgements

The authors would like to thank the anonymous reviewers for their comments. This work has been funded by the Klaus Tschira Foundation, Heidelberg, Germany.

References

Fatemeh Torabi Asr and Vera Demberg. 2012. Implicitness of discourse relations. In *Proceedings of COLING 2012*, pages 2669–2684.

Maja Bärenfänger, Mirco Hilbert, Henning Lobin, and Harald Lüngen. 2008. Owl ontologies as a resource for discourse parsing. *Journal for Language Technology and Computational Linguistics*, 23(1):17–26.

Oliver Bott and Torgrim Solstad. 2014. From verbs to discourse: A novel account of implicit causality. In *Psycholinguistic approaches to meaning and understanding across languages*, pages 213–251. Springer.

Mingyang Cai, Zhen Yang, and Ping Jian. 2024. Improving implicit discourse relation recognition with semantics confrontation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8828–8839, Torino, Italia. ELRA and ICCL.

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 *Findings of the Association for Computational Linguistics: EACL 2024*, pages 684–721, St. Julian's, Malta. Association for Computational Linguistics.

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 *Findings of the Association for Computational Linguistics: ACL 2023*, pages 35–57, Toronto, Canada. Association for Computational Linguistics.

Simon H. Corston-Oliver. 1998. Beyond string matching and cue phrases: Improving efficiency and coverage in discourse analysis. In *The AAAI Spring Symposium on Intelligent Text Summarization*, pages 9–15.

Debopam Das and Maite Taboada. 2013. Explicit and implicit coherence relations: A corpus study. In *Proceedings of the 2013 annual conference of the Canadian Linguistic Association*. Victoria: University of Victoria.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179.

Michael Alexander Kirkwood Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. Routledge.

Freya Hewett, Roshan Prakash Rane, Nina Harlacher, and Manfred Stede. 2019. The utility of discourse parsing features for predicting argumentation structure. In *Proceedings of the 6th Workshop on Argument Mining*, pages 98–103, Florence, Italy. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Yangfeng Ji and Jacob Eisenstein. 2015. One vector is not enough: Entity-augmented distributed semantics for discourse relations. *Transactions of the Association for Computational Linguistics*, 3:329–344.

Yuxin Jiang, Linhan Zhang, and Wei Wang. 2023. Global and local hierarchy-aware contrastive framework for implicit discourse relation recognition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8048–8064, Toronto, Canada. Association for Computational Linguistics.

Najoung Kim, Song Feng, Chulaka Gunasekara, and Luis Lastras. 2020. Implicit discourse relation classification: We need to talk about evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5404–5414, Online. Association for Computational Linguistics.

Yudai Kishimoto, Yugo Murawaki, and Sadao Kurohashi. 2020. Adapting BERT to implicit discourse relation classification with a focus on discourse connectives. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1152–1158, Marseille, France. European Language Resources Association.

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing implicit discourse relations in the Penn Discourse Treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 343–351, Singapore. Association for Computational Linguistics.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2011. Automatically evaluating text coherence using discourse relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 997–1006, Portland, Oregon, USA. Association for Computational Linguistics.

Dongqi Liu, Xi Yu, Vera Demberg, and Mirella Lapata. 2025. Explanatory summarization with discourse-driven planning. *Transactions of the Association for Computational Linguistics*, 13:1146–1170.

Wei Liu and Michael Strube. 2023. Annotation-inspired implicit discourse relation classification with auxiliary discourse connective generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15696–15712, Toronto, Canada. Association for Computational Linguistics.

Wei Liu and Michael Strube. 2025a. Discourse relation-enhanced neural coherence modeling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4748–4762, Vienna, Austria. Association for Computational Linguistics.

Wei Liu and Michael Strube. 2025b. Joint modeling of entities and discourse relations for coherence assessment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 21910–21926, Suzhou, China. Association for Computational Linguistics.

Wei Liu, Stephen Wan, and Michael Strube. 2024. What causes the failure of explicit to implicit discourse relation recognition? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2738–2753, Mexico City, Mexico. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Wanqiu Long and Bonnie Webber. 2022. Facilitating contrastive learning of discourse relational senses by exploiting the hierarchy of sense relations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10704–10716, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Annie Louis and Ani Nenkova. 2011. Automatic identification of general and specific sentences by leveraging discourse annotations. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 605–613, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Todor Mihaylov and Anette Frank. 2019. Discourse-aware semantic self-attention for narrative reading comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2541–2552, Hong Kong, China. Association for Computational Linguistics.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994.*

Qiang Ning, Zhili Feng, Hao Wu, and Dan Roth. 2018. Joint reasoning for temporal and causal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2278–2288, Melbourne, Australia. Association for Computational Linguistics.

Kazumasa Omura, Fei Cheng, and Sadao Kurohashi. 2024. An empirical study of synthetic data generation for implicit discourse relation recognition. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1073–1085, Torino, Italia. ELRA and ICCL.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the Joint Conference* *of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 683–691, Suntec, Singapore. Association for Computational Linguistics.

Emily Pitler and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 13–16, Suntec, Singapore. Association for Computational Linguistics.

Livia Polanyi, Chris Culy, Martin van den Berg, Gian Lorenzo Thione, and David Ahn. 2004. A rule based approach to discourse parsing. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 108–117, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Lianhui Qin, Zhisong Zhang, Hai Zhao, Zhiting Hu, and Eric Xing. 2017. Adversarial connective-exploiting networks for implicit discourse relation classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1006–1017, Vancouver, Canada. Association for Computational Linguistics.

Attapol Rutherford and Nianwen Xue. 2014. Discovering implicit discourse relations through brown cluster pair representation and coreference patterns. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 645–654, Gothenburg, Sweden. Association for Computational Linguistics.

Attapol T. Rutherford, Vera Demberg, and Nianwen Xue. 2016. Neural network models for implicit discourse relation classification in english and chinese without surface features. *CoRR*, abs/1606.01990.

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.

Wei Shi and Vera Demberg. 2019. Next sentence prediction helps implicit discourse relation classification within and across domains. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5790–5796, Hong Kong, China. Association for Computational Linguistics.

Wei Shi, Frances Yung, Raphael Rubino, and Vera Demberg. 2017. Using explicit discourse connectives in translation for implicit discourse relation classification. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 484–495, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Maite Taboada and Debopam Das. 2013. Annotation upon annotation: Adding signalling information to a corpus of discourse relations. *Dialogue & Discourse*, 4(2):249–281.

Zhipang Wang, Yu Hong, Weihao Sun, and Guodong Zhou. 2025a. Using subtext to enhance generative IDRR. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 464–473, Vienna, Austria. Association for Computational Linguistics.

Zhipang Wang, Yu Hong, Weihao Sun, and Guodong Zhou. 2025b. Using subtext to enhance generative IDRR. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 464–473, Vienna, Austria. Association for Computational Linguistics.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The penn discourse treebank 3.0 annotation manual. *Philadelphia, University of Pennsylvania*, 35:108.

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 the AAAI Conference on Artificial Intelligence*, volume 36, pages 11486–11494.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Rashmi Prasad, Christopher Bryant, and Attapol Rutherford. 2015. The CoNLL-2015 shared task on shallow discourse parsing. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task*, pages 1–16, Beijing, China. Association for Computational Linguistics.

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.

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 the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7760–7773, Torino, Italia. ELRA and ICCL.

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 the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2230–2235, Lisbon, Portugal. Association for Computational Linguistics.

Fan Zhang, Diane Litman, and Katherine Forbes Riley. 2016. Inferring discourse relations from PDTB-style discourse labels for argumentative revision classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2615–2624, Osaka, Japan. The COLING 2016 Organizing Committee.

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 *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3848–3858, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhi Min Zhou, Man Lan, Zheng Yu Niu, Yu Xu, and Jian Su. 2010. The effects of discourse connectives prediction on implicit discourse relation recognition. In *Proceedings of the SIGDIAL 2010 Conference*, pages 139–146, Tokyo, Japan. Association for Computational Linguistics.

Cäcilia Zirn, Mathias Niepert, Heiner Stuckenschmidt, and Michael Strube. 2011. Fine-grained sentiment analysis with structural features. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 336–344, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

A More Case Studies

(8) [Despite traders' complaints, the links with the Chicago futures market worked as planned in Friday's rout to provide a cooling-off period]$_{\text{Arg1}}$(**Implicit=However**)[Of greater help, was the "natural circuit breaker" of the weekend,]$_{\text{Arg2}}$
WordNet Hints: CoHypo: (help, market); HyperHypo: (links < circuit), (period > weekend), (worked < was).
Evidence List: [Polarity_Contrast("Despite traders' complaints", "Of greater help")]; [Lexical(Synonyms_direct("cooling-off period", "circuit breaker"))]; [Generalization-Specification]; [Statement-Elaboration]; [Temporal_Sequence("Friday's rout", "the weekend")];

(9) [Instead, the rally only paused for about 25 minutes and then steamed forward as institutions resumed buying.]$_{\text{Arg1}}$ (**Explicit=In the end**) [The market closed minutes after reaching its high for the day of]$_{\text{Arg2}}$
WordNet Hints: Syn: (minutes,minutes); CoHypo: (day,minutes); HyperHypo: (institutions > high). Evidence List: [Temporal_Sequence(25 minutes, minutes after); Temporal_Sequence(paused, closed); Cause/Effect(resumed buying, steamed forward); Generalization-Specification; Statement-Elaboration;]

In example (8), the baseline model's misclassification of Conjunction probably results from a shallow interpretation. It recognizes two helpful mechanisms—the "cooling-off period" and the "circuit breaker"—which are correctly identified as near-synonyms by the [Lexical(Synonyms_direct)] feature, leading it to assume a simple additive relationship. Our model, on the other hand, correctly detects Concession by incorporating more subtle cues from the Evidence List. While the synonym feature confirms that two comparable items are being discussed, the [Polarity_Contrast("Despite traders' complaints", "Of greater help")] is the key factor. It captures the core comparative and concessive structure of the discourse by linking the concessive clause in Arg1 ("Despite...") to the explicit comparative judgment in Arg2 ("Of greater help"). This enables our model to recognize that the relationship is not merely a list, but a ranked comparison in which the weekend (Arg2) is presented as a more effective solution than the market links (Arg1).

In example (9), the baseline's mistake in predicting Conjunction comes from its failure to distinguish between a simple thematic relation and a structured chronological narrative. It correctly observes that both arguments describe the day's market activity, but without deeper guidance, it just lists them as parallel events. Our model's success, however, is due to its ability to use the Temporal_Sequence(paused, closed) and Temporal_Sequence(25 minutes, minutes after) features. This explicit cue links the two arguments, turning them from separate observations into a clear sequence: the market activity first 'paused' (Arg1), and then the market 'closed' (Arg2). This linguistic cue helps the model recognize the chronological order, allowing it to correctly identify the more precise Asynchronous relationship.

It is important to note that all examples in our case studies are taken directly from our test set.

B Dataset Information

Table 5 shows the datasets' statistics information we used in our work.

Dataset	Train	Val	Test
PDTB 2 (Prasad et al., 2008)	12632	1183	1046
PDTB 3 (Webber et al., 2019)	17085	1653	1474
DiscoGeM 1.0 (Scholman et al., 2022)	-	-	1288

Table 5: The datasets' information in this work.

C Experiment Details

All models were fine-tuned using the LoRA method for efficiency. For the Llama-3-70B-Instruct model, we set the learning rate to 3e-05, a dropout rate of 0.1, and trained for 2 epochs with 10 warmup steps; the LoRA configuration used a rank of 24 and an alpha of 48. The settings for the Llama-3-8B-Instruct model were a learning rate of 5e-05, a dropout of 0.05, and 20 warmup steps, also for 2 epochs, with a LoRA rank of 32 and an alpha of 64. All experiments were conducted on a single NVIDIA H200 GPU, with the 70B model taking approximately 3.5 hours to train and the 8B model taking approximately 30 minutes. We ran the experiments five times to calculate the statistical significance.

D More results

Table 6 and Table 7 confirm that while the features do provide a performance boost to the RoBERTa model, the relative improvement is less pronounced compared to the gains observed in our Llama architectures. We hypothesize that this disparity stems from Llama's reasoning capability, which allows it to use these external linguistic cues more effectively than the smaller encoder-based model. Contrary to early work that showed LLMs performing poorly on this task, our study highlights that the fundamental issue is not a lack of knowledge but a failure to capture subtle semantic and pragmatic features during fine-tuning. Our features act as a necessary attentional guidance mechanism that unlocks Llama's potential.

Table 8 shows detailed F1 scores for the ablation studies we conducted in Section 5, while Table 9 shows the overall performance.

Model	**PDTB-3**			
	Top Level		Second Level	
	Acc	F1	Acc	F1
RoBERTa (Baseline)	72.93	67.98	63.23	52.36
RoBERTa + Evidence List	74.33	69.40	64.26	53.67
Ours_{8B}	77.82	73.34	67.74	60.00
Ours_{70B}	81.00	77.04	70.74	61.08

Table 6: Performance comparison on PDTB-3 using RoBERTa and our proposed framework. The above performance is calculated using **M1**.

Model	**PDTB-3**			
	Top Level		Second Level	
	Acc	F1	Acc	F1
RoBERTa (Baseline)	73.90	68.29	64.38	54.24
RoBERTa + Evidence List	75.07	70.13	65.32	55.78
Ours_{8B}	78.61	73.88	68.86	62.29
Ours_{70B}	81.70	77.46	71.74	64.34

Table 7: Performance comparison on PDTB-3 using RoBERTa and our proposed framework. The above performance is calculated using **M2**.

Table 10 shows that the baseline model often misclassified specific relations as the simpler Conjunction, especially when implicit arguments lack clear lexical cues, defaulting to predict additive or sequential links instead of more complex logical structures.

In our method, explicit Cause-Effect, Polarity & Degree, or Claim-Justification features from the Evidence List appear in 92.9% of cases misclassified by the baseline model. Additionally, these critical features related to Cause appear in 92.9% and 100% of instances where the baseline misclassifies it as Concession and Level-of-Detail, respectively. This finding further confirmed our analysis in the Analysis section. Our method helps our model better recognize the underlying causal logic and correct these predictions. Moreover, it also shows that the reason why raw LLM or fine-tuned LLM cannot perform well on this task is that the model cannot capture the subtle semantic and pragmatic information conveyed by the input.

This analysis suggests that the baseline model relies on shallow heuristics, often defaulting to overly broad labels or being misled by superficial cues. Our feature-enhanced model succeeds because the Evidence List functions as a semantic framework, urging the model to engage with the deeper, linguistically grounded relationships between the arguments. It learns to verify the presence of causality, contrast, or specification before making a prediction, resulting in a more robust and accurate reasoning process.

E Feature Selection

For lexical-semantic features from WordNet, the relations we extract and the reasons are listed below:

- Synonyms and Antonyms: We extract direct synonyms (words sharing a synset) and antonyms. These provide the most direct signals for relations of *Expansion.Restatement* and *Comparison.Contrast* (Halliday and Hasan, 1976; Polanyi et al., 2004; Pitler et al., 2009; Das and Taboada, 2013), respectively, by identifying semantic equivalence or opposition.

- Hypernyms and Hyponyms: We identify "is-a" relationships, where a hyponym (e.g., car) is a type of a hypernym (e.g., vehicle). This feature is crucial for recognizing *Expansion.Instantiation* and *Expansion.Level-of-detail* (Halliday and Hasan, 1976), where one argument provides a specific instance of a general category mentioned in the other.

- Meronyms and Holonyms: Part-whole relationships (e.g., engine is a meronym of car)

Method	Concession	Cause	Conjunction	Level-of-Detail	Asynchronous	Instantiation
Ours$_{8b}$	63.68	75.00	63.18	60.06	68.60	72.79
Llama-3-8B-Instruct	48.25	66.51	56.65	52.04	59.07	67.83
- Synonyms & Antonyms	62.31	73.20	63.28	58.85	67.65	72.03
- Hypernyms & Hyponyms	54.08	72.54	62.33	59.24	67.65	72.81
- Meronyms & Holonyms	59.70	73.41	61.60	58.15	67.00	70.69
- Co-hyponyms	61.62	73.22	61.89	59.70	68.63	72.65
- Verb Entailments & Causes	61.00	73.66	62.38	57.92	66.34	71.49
- Similar To	59.90	73.34	61.85	57.93	66.34	70.43
- Polarity & Degree	55.17	72.49	62.57	60.10	69.89	72.17
- Propositional & Event Rel	60.91	73.07	61.77	60.20	67.65	70.69
- Structural & Pragmatic Rel	59.51	73.42	61.98	58.62	70.24	68.67

Table 8: Detailed ablation study results for Ours$_{8b}$ on the six most frequent Second-Level relations of the PDTB-3 test set. We report the F1 score for each relation class. The top two rows compare our full model against the standard fine-tuning baseline (Llama-3-8B-Instruct). Each subsequent row (- Feature Name) shows the performance after removing a specific feature category.

Method	Acc.	F1
Ours$_{8b}$	67.74	60.00
- Synonyms & Antonyms	66.92	57.36
- Hypernyms & Hyponyms	65.96	56.23
- Meronyms & Holonyms	65.35	56.23
- Co-hyponyms	66.71	56.91
- Verb Entailments & Causes	66.17	56.35
- Similar To	65.62	55.93
- Polarity & Degree	66.10	55.41
- Propositional & Event Rel	66.51	56.65
- Structural & Pragmatic Rel	66.51	57.25

Table 9: Ablation study results for Ours$_{8b}$ on the PDTB-3 Second-Level test set. The top row displays the full model's performance. Each following row (- Feature Name) shows the performance after removing a specific category of semantic or pragmatic features, highlighting its contribution to the final result.

are extracted primarily for nouns. This relation often signals an *Expansion.Level-of-detail* where a whole is introduced and its parts are subsequently described (Bärenfänger et al., 2008).

- `Co-hyponyms`: We also identify co-hyponyms—words that share a direct hypernym (e.g., apple and orange are co-hyponyms of fruit). This feature is a powerful indicator of a parallel list structure, often found in *Expansion.Specification* and *Expansion.Instantiation* relations (Louis and Nenkova, 2011).

- `Verb Entailments and Causes`: For verbs specifically, we extract entailment relations (e.g., snoring entails sleeping) and cause-to relationships (e.g., killing is a cause of dying). These are essential for identifying the logical core of *Contingency.Cause* and some *Temporal* relations where one action necessitates or directly causes another (Taboada and Das, 2013; Bott and Solstad, 2014).

- `Similar To` (for Adjectives and Adverbs): For adjectives and adverbs, we extract "Similar To" relations from WordNet. This provides a softer form of synonymy that is useful for identifying thematic similarity, comparison, and paraphrase across arguments (Pitler et al., 2009).

For LLM-exclusive features, the relations we extract and the reasons are listed below:

Polarity and Degree Relations: Polarity and Degree Relations: This category captures the evaluative and comparative tones often essential for understanding Comparison relations (Corston-Oliver, 1998).

- `Polarity_Contrast`: This feature is the reverse of Zirn et al. (2011), utilizing sentiment analysis (among others) and discourse relations. It is one of the most influential indicators for *Comparison.Contrast* and *Comparison.Concession.*

- `Degree_Comparison`: This identifies comparisons of scale or intensity (e.g., more, less,

Gold	Misclassify	Error Rate
Cause	Conjunction	41.8%
Level-of-Detail	Cause	40%
Cause	Concession	20.9%
Level-of-Detail	Conjunction	37.1%
Cause	Level-of-Detail	17.9%
Asynchronous	Conjunction	38.1%
Instantiation	Level-of-Detail	88.9%
Conjunction	Concession	38.1%
Concession	Conjunction	36.8%
Asynchronous	Cause	33.3%
Synchronous	Conjunction	77.8%

Table 10: Analysis of the most common error patterns of the Llama-3-8B-Instruct baseline that are correctly identified by our Ours$_{8b}$ on the PDTB-3 Second-Level test set. "Gold" indicates the ground-truth label, and "Misclassify" shows the incorrect label predicted by the baseline. The 'Error Rate' column shows how often a specific misclassification occurs, representing the percentage of that error type (e.g., Cause -> Conjunction) out of the total errors made by the baseline for that particular gold label. 41.8% of Cause relations misclassified by the baseline but correctly by Ours$_{8b}$ were labeled as Conjunction.

higher), offering direct evidence for Comparison relations (Pitler et al., 2009).

Propositional and Event Relations: This category explains the logical and temporal links between events or propositions in the arguments.

- The `Cause-Effect`, `Condition/Consequence`, and `Action/Purpose` features are the main indicators of the entire *Contingency* class (Asr and Demberg, 2012; Taboada and Das, 2013; Bott and Solstad, 2014). They clarify the underlying causal, conditional, or purposive logic, which is often the hardest part of the inference.

- The `Temporal_Sequence` feature shows the chronological order of events. It provides the evidence for *Temporal.Asynchronous* and *Cause* relations (if sequential) and can also support *Temporal.Synchronous* relations if events happen at the same time (Taboada and Das, 2013; Ning et al., 2018).

Structural and Pragmatic Relations: This category captures the comprehensive, functional connection between the two arguments as complete discourse units.

- Features like `Generalization-Specification`, `Action-Manner`, and `Statement-Elaboration` are essential for identifying the *Expansion* class (Taboada and Das, 2013; Lin et al., 2009). They recognize patterns where one argument offers specific details, examples, or a manner of execution for a more general statement in the other, directly indicating relations such as *Instantiation* and *Level-of-detail*.

- `Claim-Justification` serves as a strong indicator of argumentative structures, often pointing to a *Contingency* relation where one argument provides the reason for a claim in the other (Zhang et al., 2016).

- The `Alternative_Choice` and `List_Continuation` features identify parallel structures. `List_Continuation` strongly signals *Expansion.Conjunction* and *Temporal.Synchronous* relations, while `Alternative_Choice` indicates *Comparison* and *Expansion.Alternative* (Pitler et al., 2009; Lin et al., 2009).

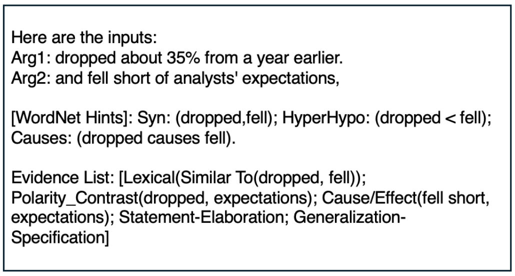
- Dialogic features like `Question-Answer` and `Rhetorical_Continuation` capture interactive patterns that are frequently overlooked by traditional analyses (Asr and Demberg, 2012; Webber et al., 2019).

F Prompt Details

Figure 4 shows the detailed instruction prompt used for the Semantic Feature Extraction task. This prompt guides the Llama-3-70B model to function as an objective feature extractor, identifying and listing all possible lexical, semantic, and pragmatic relationships between the arguments to create the Evidence List.

Figure 2 illustrates a specific example of the final, feature-enhanced input fed into our fine-tuned prediction model. This input combines the original arguments (Arg1, Arg2), the additional WordNet Hints, and the complete Evidence List produced during the first step. When we fine-tune the baseline models, all augmented features will be removed, and the instructions related to these features will also be deleted.

Figure 3 shows the prompt used for both fine-tuning and inference in the final prediction step. This instruction tasks the fine-tuned model with analyzing the full feature-augmented input and predicting the single most appropriate implicit discourse relation based on its thorough analysis.

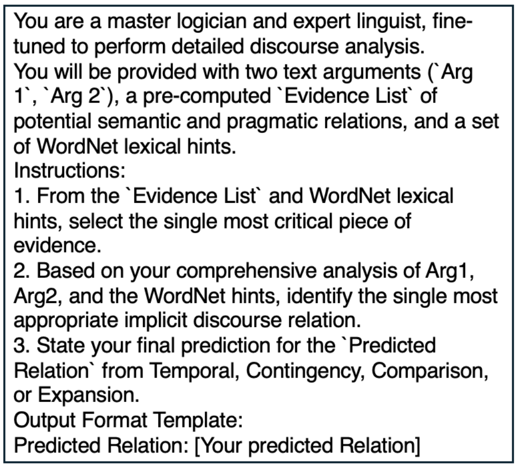


Here are the inputs:
Arg1: dropped about 35% from a year earlier.
Arg2: and fell short of analysts' expectations,

[WordNet Hints]: Syn: (dropped,fell); HyperHypo: (dropped < fell); Causes: (dropped causes fell).

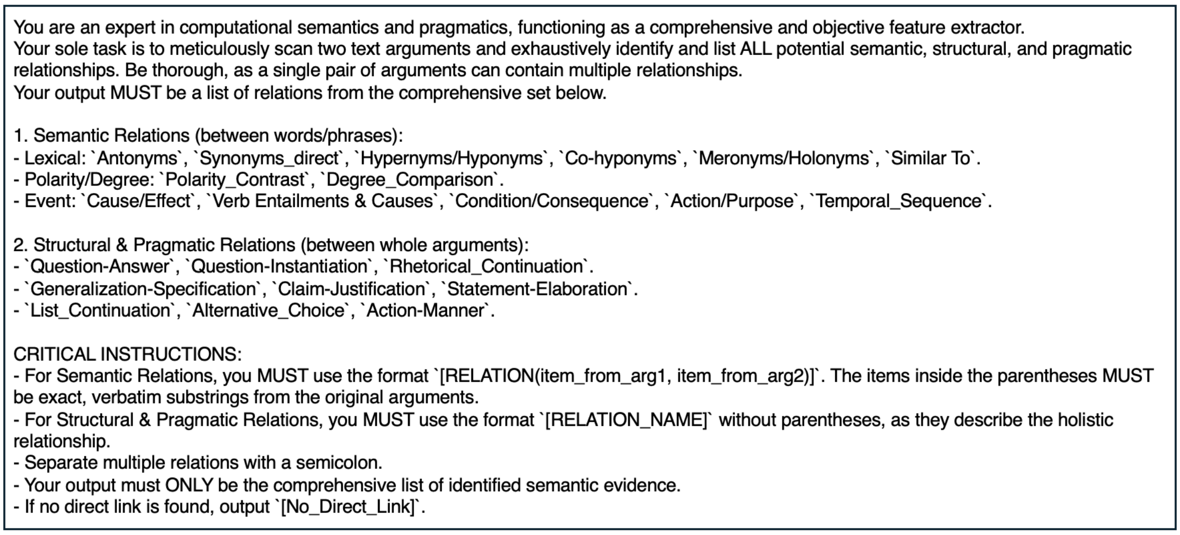
Evidence List: [Lexical(Similar To(dropped, fell)); Polarity_Contrast(dropped, expectations); Cause/Effect(fell short, expectations); Statement-Elaboration; Generalization-Specification]

Figure 2: An example of our input.



You are a master logician and expert linguist, fine-tuned to perform detailed discourse analysis.
You will be provided with two text arguments (`Arg 1`, `Arg 2`), a pre-computed `Evidence List` of potential semantic and pragmatic relations, and a set of WordNet lexical hints.
Instructions:
1. From the `Evidence List` and WordNet lexical hints, select the single most critical piece of evidence.
2. Based on your comprehensive analysis of Arg1, Arg2, and the WordNet hints, identify the single most appropriate implicit discourse relation.
3. State your final prediction for the `Predicted Relation` from Temporal, Contingency, Comparison, or Expansion.
Output Format Template:
Predicted Relation: [Your predicted Relation]

Figure 3: The prompt for fine-tuning the model.



You are an expert in computational semantics and pragmatics, functioning as a comprehensive and objective feature extractor.
Your sole task is to meticulously scan two text arguments and exhaustively identify and list ALL potential semantic, structural, and pragmatic relationships. Be thorough, as a single pair of arguments can contain multiple relationships.
Your output MUST be a list of relations from the comprehensive set below.

1. Semantic Relations (between words/phrases):
- Lexical: `Antonyms`, `Synonyms_direct`, `Hypernyms/Hyponyms`, `Co-hyponyms`, `Meronyms/Holonyms`, `Similar To`.
- Polarity/Degree: `Polarity_Contrast`, `Degree_Comparison`.
- Event: `Cause/Effect`, `Verb Entailments & Causes`, `Condition/Consequence`, `Action/Purpose`, `Temporal_Sequence`.

2. Structural & Pragmatic Relations (between whole arguments):
- `Question-Answer`, `Question-Instantiation`, `Rhetorical_Continuation`.
- `Generalization-Specification`, `Claim-Justification`, `Statement-Elaboration`.
- `List_Continuation`, `Alternative_Choice`, `Action-Manner`.

CRITICAL INSTRUCTIONS:
- For Semantic Relations, you MUST use the format `[RELATION(item_from_arg1, item_from_arg2)]`. The items inside the parentheses MUST be exact, verbatim substrings from the original arguments.
- For Structural & Pragmatic Relations, you MUST use the format `[RELATION_NAME]` without parentheses, as they describe the holistic relationship.
- Separate multiple relations with a semicolon.
- Your output must ONLY be the comprehensive list of identified semantic evidence.
- If no direct link is found, output `[No_Direct_Link]`.

Figure 4: The prompt for extracting the features.