Multi-Lingual Implicit Discourse Relation Recognition with Multi-Label Hierarchical Learning

Nelson Filipe Costa and Leila Kosseim
Computational Linguistics at Concordia (CLaC) Laboratory
Department of Computer Science and Software Engineering
Concordia University, Montréal, Québec, Canada
nelsonfilipe.costa@mail.concordia.ca
leila.kosseim@concordia.ca

Abstract

This paper introduces the first multi-lingual and multi-label classification model for implicit discourse relation recognition (IDRR). Our model, HArch, is evaluated on the recently released DiscoGeM 2.0 corpus and leverages hierarchical dependencies between discourse senses to predict probability distributions across all three sense levels in the PDTB 3.0 framework. We compare several pre-trained encoder backbones and find that RoBERTa-HArch achieves the best performance in English, while XLM-RoBERTa-HArch performs best in the multilingual setting. In addition, we compare our fine-tuned models against GPT-4o and Llama-4-Maverick using few-shot prompting across all language configurations. Our results show that our fine-tuned models consistently outperform these LLMs, highlighting the advantages of task-specific fine-tuning over prompting in IDRR. Finally, we report SOTA results on the DiscoGeM 1.0 corpus, further validating the effectiveness of our hierarchical approach.

1 Introduction

Discourse analysis explores how textual segments are connected through relations of coherence. One of the most widely adopted frameworks for computational discourse analysis is the Penn Discourse Treebank (PDTB) (Miltsakaki et al., 2004; Prasad et al., 2008a). In the PDTB framework, a discourse relation is established between two textual segments (referred to as arguments) either explicitly, through the presence of a discourse connective (e.g., *but* and *because*), or implicitly when no connective is present. The sense of the relation is then classified using a defined list of discourse senses, which are organized hierarchically across three levels (Webber et al., 2019). The task of identifying the sense of an implicit relation is known as implicit discourse relation recognition (IDRR).

Research on IDRR has mostly been conducted on the English annotated PDTB corpora (Prasad et al., 2008b, 2019). However, the subjective nature of discourse interpretation poses challenges for single-label annotation schemes, where each instance is assigned only one sense label (Stede, 2008; Scholman and Demberg, 2017; Hoek et al., 2021). The ambiguity in the annotation of implicit relations has been further highlighted by the challenges in mapping them across discourse frameworks (Demberg et al., 2019; Costa et al., 2023). As a response, recent works have advocated for multi-label annotation, recognizing that multiple discourse senses can often co-occur within a single relation (Yung et al., 2019; Pyatkin et al., 2020; Scholman et al., 2022a,b; Pyatkin et al., 2023; Yung et al., 2024b; Yung and Demberg, 2025). This shift in perspective has led to the development of the multi-label annotated DiscoGeM 1.0 corpus (Scholman et al., 2022a) for implicit discourse relations.

This perspective shift toward multi-label annotation has stirred an initial wave of research in multi-label IDRR (Costa and Kosseim, 2024; Long et al., 2024; Costa and Kosseim, 2025). However, as with the PDTB corpora, the DiscoGeM 1.0 corpus is limited to the English language and most work in the field of discourse continues to focus mostly on the English language. While single-label discourse annotated corpora following the PDTB framework exist in other languages, such as the German Potsdam Commentary Corpus 2.2 (Bourgonje and Stede, 2020) and the Czech Prague Discourse Treebank 3.0 (Synková et al., 2024), multilingual corpora remain scarce. For instance, the TED-MDB corpus (Zeyrek et al., 2024) provides parallel annotations of single-label discourse relations across seven languages. However, its size remains relatively small for the training of IDRR models - it contains 1,774 implicit relations. To address this gap and enable cross-linguistic research in IDRR, the recently released DiscoGeM 2.0 corpus (Yung et al., 2024b) introduces parallel multi-label annotations of implicit discourse relations

Proceedings of the 26th Annual Meeting of SIGDIAL, pages 48–61
Aug 25–27, 2025. ©2025 SIGDIAL

following the PDTB 3.0 sense hierarchy in four languages: English, German, French and Czech.

In this paper, we leverage the DiscoGeM 2.0 corpus to present the first multi-lingual and multi-label classification model for IDRR. Our main contributions are as follows:

- We propose a hierarchical multi-task architecture, HArch, that leverages dependencies across all three sense levels in the PDTB 3.0 framework, improving on prior multi-label approaches (Costa and Kosseim, 2025).
- We conduct the first evaluation of multi-label IDRR on the DiscoGeM 2.0 corpus in the English, German, French and Czech languages individually and in a multi-lingual setting.
- We compare different pre-trained encoder backbones and find that RoBERTa-HArch and XLM-RoBERTa-HArch achieve the best results in the English-only and multi-lingual settings, respectively.
- We evaluate our fine-tuned models against LLMs, including GPT-4o and Llama-4-Maverick, using few-shot prompting and show that our models consistently outperform them in multi-label IDRR.
- We benchmark our best-performing model on the English-only DiscoGeM 1.0 corpus and achieve new SOTA results.

2 Previous Work

The advent of transformer-based models (Vaswani et al., 2017) has led to substantial progress in natural language understanding. Nevertheless, IDRR remains one of the most challenging tasks in computational discourse analysis. Traditionally formulated as a single-label classification problem, most approaches have addressed IDRR in English either by fine-tuning (Long and Webber, 2022; Liu and Strube, 2023) or prompt-tuning (Chan et al., 2023; Zhao et al., 2023; Zeng et al., 2024; Long and Webber, 2024) pre-trained language models (PLMs). Another line of research has also explored directly prompting large language models (LLMs) through prompt-engineering to solve single-label IDRR (Chan et al., 2024; Yung et al., 2024a). However, results from both work indicate that zero-shot and few-shot prompting of LLMs continues to significantly underperform when compared to the results obtained through fine-tuning and prompt-tuning PLMs. The same outcome has been observed in other IDRR work focusing on different languages (Saeed et al., 2025; Ruby et al., 2025).

The recent move toward multi-label annotation in discourse data has led to initial efforts in multi-label IDRR. Long et al. (2024) used the 4.9% of implicit discourse relations in the PDTB 3.0 corpus that are annotated with two senses to build a model capable of predicting up to two senses per instance. However, given that most of PDTB 3.0 annotations remain single-label, their model predominantly produces single-label predictions. In contrast, Yung et al. (2022) and Costa and Kosseim (2024) used the multi-label DiscoGeM 1.0 corpus but convert its annotations into a single-label format during training. Finally, Costa and Kosseim (2025) present the first results in multi-label IDRR by training a model on DiscoGeM 1.0 that jointly predicts sense distributions across all three levels in the PDTB 3.0 framework. Rather than adapting multi-label data to a single-label setting or modifying a single-label corpus for multi-label use, they preserve the inherent multi-label structure of DiscoGeM 1.0 throughout training. However, like most previous work in IDRR, their model is limited to the English language.

3 Our Approach

We propose a hierarchical multi-task classification model, HArch, to predict probability distributions across all three sense levels in the PDTB 3.0 framework. We also propose a prompting framework to compare our HArch model against SOTA LLMs through few-shot learning.

3.1 HArch Model

Our HArch model improves on the model of Costa and Kosseim (2025) by explicitly modeling hierarchical dependencies across sense levels. Figure 1 shows the architecture of our model. The input of the model is a concatenated pair of discourse arguments (ARG1 + ARG2), which is encoded using a pre-trained language model. The resulting contextual representation is fed to a shared linear transformation followed by a dropout layer. This shared representation is then passed to three separate classification heads - one for each sense level. The level-1 classification head predicts a probability distribution over the 4 possible level-1 senses using a linear layer followed by a softmax activa-

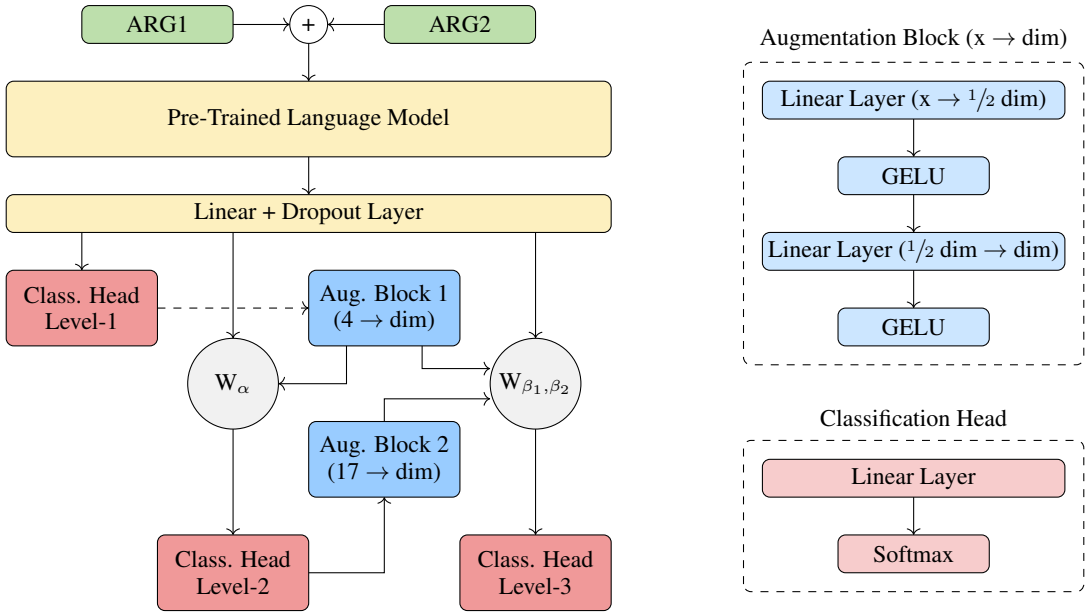

Figure 1: Architecture of our hierarchical multi-task model, HArch, for multi-label IDRR. The model takes as input a concatenated pair of discourse arguments and generates probability distributions across all three sense levels in the PDTB 3.0. The output of the lower-level classification heads is projected onto the embedding space of the encoder via augmentation blocks and then combined with the shared representation using learnable weighted sums. This architecture explicitly models hierarchical dependencies and enables joint learning across all sense levels.

tion - as shown in red on the bottom-right side of Figure 1. This 4-dimensional output vector is then projected onto the same embedding space of the pre-trained language model through an augmentation block. As shown in blue on the upper-right side of Figure 1, this augmentation is done in two steps: the output of the classification head is first mapped to a hidden dimension of half the size of the embedding space and then, in a second step, further projected to match the full embedding size. A GELU (Hendrycks and Gimpel, 2016) activation function after each incremental step introduces non-linearity, while helping with sparsity and preventing exploding activations in the dimensionality projection. Based on early experiments, this incremental dimensional projection ensured a smoother transition in feature space. The augmented level-1 output is then combined with the shared representation through a weighted sum defined by Equation 1, where α is a learnable parameter of the model, to yield the input of the level-2 classification head.

$$\mathbf{W}_\alpha = \alpha \cdot \mathbf{h}_{\text{Aug}_1} + (1-\alpha) \cdot \mathbf{h}_{\text{Dropout}} \quad (1)$$

The level-2 classification head then predicts a probability distribution over the 17 possible level-2 senses. Similar to level-1, the level-2 output is passed through a dimensionality augmentation block to project it into the same embedding space of the pre-trained language model. The augmented level-1 and level-2 outputs are then added to the shared representation through a second weighted sum defined by Equation 2, where β_1 and β_2 are learnable parameters. The output of this weighted sum is then fed into the level-3 classification head, which predicts a probability distribution over the level-3 senses.

$$\begin{aligned}\mathbf{W}_{\beta_1,\beta_2} =& \beta_1 \cdot \mathbf{h}_{\text{Aug}_1} + \beta_2 \cdot \mathbf{h}_{\text{Aug}_2} + \\ & (1-\beta_1-\beta_2) \cdot \mathbf{h}_{\text{Dropout}}\end{aligned} \quad (2)$$

This cascading of information ensures that the predictions of lower-level classification heads are used to help inform the prediction of higher-level classification heads that require more fine-grained information. Finally, we use the Adam optimization algorithm (Kingma and Ba, 2015) to minimize the loss of the model, which we calculate as the sum of the losses of each classification head. We use the Mean Absolute Error function to calculate the loss of each classification head as it was shown to lead to better performance in in multi-label classification (Costa and Kosseim, 2025). To evaluate our model, we use Jensen-Shannon (JS) distance to measure the similarity between the predicted and the reference probability distributions, similarly to

other multi-label classification works in NLP (Pyatkin et al., 2023; Yung et al., 2024b; van der Meer et al., 2024; Costa and Kosseim, 2025).

3.2 Few-Shot Prompting

Despite prior work showing poor results in IDRR through prompt engineering (Chan et al., 2024; Yung et al., 2024a), we compared our HArch model to LLMs in multi-label IDRR via direct prompting with few-shot learning. We tested GPT-4o (OpenAI, 2024) and Llama-4-Maverick with the prompt structure described in Appendix B. Depending on the language evaluated, the prompt was translated and examples were drawn from the corresponding language subset of DiscoGeM 2.0. When considering all languages simultaneously in a multi-lingual setting, we kept the structure of the prompt in English and provided examples in the different languages. All of the examples were taken from the training split of the corpus. To ensure consistency, we set the temperature to 0 and allowed up to 5 retries per instance in cases where prompt outputs did not conform to the expected format. Prompting costs were approximately 5.00$ for GPT-4o and $0.19 - 0.49$$ for Llama-4-Maverick per million input tokens.

The prompt was designed to closely replicate the annotation methodology of DiscoGeM 2.0. When crowdsourcing the annotation of the corpus, the authors provided the annotators a list of non-ambiguous connectives (Yung et al., 2024b) that represented each of the possible senses in the PDTB 3.0 framework. This list of connectives was adapted for each language. The annotators were then asked to choose the connective that best expresses the semantic relation between the arguments of the relation being annotated, irrespective of whether syntax would need to be adjusted. By mirroring this annotation process in our prompting template, we were able to directly evaluate the performance of both LLMs on multi-label IDRR using the test split of DiscoGeM 2.0.

4 Data Preparation

To train our model on multi-lingual IDRR, we use the recently released DiscoGeM 2.0 corpus (Yung et al., 2024b). The corpus comprises a total of 13,063[1] parallel annotated implicit discourse relations across four languages: 5,847 in English, 2,588 in German, 2,628 in French, and 2,000 in Czech. All annotations follow the PDTB 3.0 discourse framework (Webber et al., 2019). While its predecessor, DiscoGeM 1.0 (Scholman et al., 2022a), provided the first corpus of multi-label implicit discourse relations, DiscoGeM 2.0 is the first corpus to provide multi-lingual parallel annotation of multi-label implicit discourse relations.

Each implicit discourse relation in the corpus was annotated by at least 10 crowdworkers. Each annotator selected the sense they considered most appropriate for each relation through a proxy task and then all of the annotated senses of each relation were averaged to produce a multi-label sense distribution over 28 different senses in the PDTB 3.0 framework - the BELIEF and the SPEECH-ACT senses were not annotated. For each language, the authors of the corpus prepared a list of discourse connectives (one representing each of the available senses) and asked the annotators to select for each relation the discourse connective that best expresses the semantic relation between its two arguments. From the chosen connectives, the authors were able to infer the possible multiple senses of each implicit discourse relation.

To ensure reproducibility of our results, we split the DiscoGeM 2.0 corpus following the proposed training, validation and test splits (Yung et al., 2024b). Contrary to the common approach in traditional single-label IDRR, we follow the methodology of Costa and Kosseim (2025) and consider all three sense levels in the PDTB 3.0 framework when possible. Table 1 shows the distribution of level-1 senses and Table 6 in Appendix A shows the distribution of level-2 and level-3 senses per language in the DiscoGeM 2.0 corpus. We replaced the non-existent PDTB 3.0 level-3 senses with their corresponding level-2 sense.

5 Encoder Selection

In this section, we compare the performance of different PLMs as encoders in our HArch model (see Figure 1) for multi-label IDRR in both multilingual and English-only settings, using the DiscoGeM 2.0 corpus. Table 2 summarizes the performance of each configuration on the validation split of the corpus. Each model was trained for 10 epochs with a batch size of 16 and the results are reported as the average Jensen-Shannon (JS) distance across three independent runs. In addition to experimenting with different PLM encoders, we also

[1] In their paper they report a total of 12,834 implicit discoutse relations, out of which 5,618 are in English. However, we counted 5,847 discourse relations in English in the corpus.

Level-1	English	German	French	Czech	All
Temporal	556.8	438.4	400.7	458.8	1,854.6
Contingency	1,695.4	772.3	811.0	681.1	3,959.8
Comparison	796.3	346.5	525.0	227.1	1,894.9
Expansion	2,798.6	1,030.8	891.3	632.9	5,353.7
Total	5,847.1	2,588.0	2,628.0	1,999.9	13,063

Table 1: Distribution of level-1 senses per language in the DiscoGeM 2.0 corpus. Each value represents the sum of the corresponding sense in all the multi-label distributions of implicit discourse relations in the specific language.

Language		Model	Level-1	Level-2	Level-3	Params
Test	Train					
Eng	Eng	Costa and Kosseim (2025)	0.329 ± 0.005	0.498 ± 0.004	0.569 ± 0.005	125M
		RoBERTa-HArch	$\mathbf{0.327 \pm 0.004}$	$\mathbf{0.478 \pm 0.005}$	$\mathbf{0.541 \pm 0.005}$	125M
		ModernBERT-HArch	0.364 ± 0.004	0.507 ± 0.005	0.592 ± 0.006	149M
		XLM-RoBERTa-HArch	0.343 ± 0.004	0.487 ± 0.006	0.564 ± 0.005	125M
		GPT-4o	0.373 ± 0.002	0.559 ± 0.003	0.657 ± 0.003	200B
		Llama-4-Maverick	0.423 ± 0.004	0.641 ± 0.006	0.711 ± 0.007	400B
All	All	XLM-RoBERTa-HArch	$\mathbf{0.356 \pm 0.005}$	$\mathbf{0.531 \pm 0.006}$	$\mathbf{0.605 \pm 0.006}$	125M
		EuroBERT-HArch	0.414 ± 0.006	0.569 ± 0.007	0.641 ± 0.006	210M
		Flan-T5-HArch	0.364 ± 0.003	0.550 ± 0.003	0.623 ± 0.004	220M
		GPT-4o	0.438 ± 0.002	0.671 ± 0.002	0.709 ± 0.003	200B
		Llama-4-Maverick	0.534 ± 0.004	0.774 ± 0.004	0.839 ± 0.005	400B

Table 2: Experimental results using HArch with different PLM encoders. The results are the average JS distance of three different runs on the test split of DiscoGeM 2.0. Results are reported using the English language and all languages simultaneously. Lower scores indicate better performance. Values in bold show the best score at each level in each language setting. Rows shaded in gray present the results of LLMs with few-shot prompting. The final column reports the approximate number of parameters of each model.

report results obtained using LLMs via few-shot prompting. All of the code and prompt templates are available on GitHub[2].

To evaluate the impact of incorporating hierarchical learning in multi-label IDRR, we first compare our proposed HArch model against the non-hierarchical model proposed by Costa and Kosseim (2025). For a fair comparison, both models use the same pre-trained RoBERTa$_{\text{base}}$ encoder (Liu et al., 2019) and are trained under identical conditions with a learning rate of $1e^{-5}$. To our knowledge, Costa and Kosseim (2025) is the only existing approach modeling multi-label IDRR with probability distributions over the full set of PDTB 3.0 sense labels. Since their model was only trained in the English language, we compared the performance of our models using only the English subset of DiscoGeM 2.0. As shown in Table 2, our model RoBERTa-HArch, which incorporates hierarchical dependencies across sense levels, achieves lower JS distances at level-2 and level-3 compared to the non-hierarchical model of Costa and Kosseim (2025). At level-1, the performance of both models is comparable, which is to be expected as level-1 predictions do not benefit from lower-level inputs within the hierarchical structure.

Next, we explored the impact of using different PLMs as encoder backbones for our HArch hierarchical model. In particular, we experimented with ModernBERT$_{\text{base}}$ (Warner et al., 2024), a recently introduced encoder claimed to outperform RoBERTa, and the multi-lingual XLM-RoBERTa$_{\text{base}}$ (Conneau et al., 2020). Contrary to expectations, ModernBERT-HArch consistently underperformed across all three sense levels, as shown in Table 2. Even XLM-RoBERTa-HArch,

[2] https://github.com/nelsonfilipecosta/Multi-Lingual-Implicit-Discourse-Relation-Recognition

Model	Level-1	Level-2	Level-3
Costa and Kosseim (2025)	$\mathbf{0.299 \pm 0.002}$	0.446 ± 0.003	0.523 ± 0.002
RoBERTa-HArch	0.302 ± 0.003	$\mathbf{0.435 \pm 0.003}$	$\mathbf{0.506 \pm 0.004}$

Table 3: Final results showing the average JS distance of three different runs for each model on the test split of DiscoGeM 1.0. Lower scores indicate better performance. Values in bold show the best score at each level.

despite its multi-lingual design, achieved lower JS distances than ModernBERT-HArch - though, it still fell short of the performance achieved by RoBERTa-HArch. In addition to fine-tuned models, we also evaluated LLMs via few-shot prompting. As shown in Table 2, GPT-4o substantially outperforms Llama-4-Maverick in terms of JS distance across all levels, replicating a pattern already observed between GPT-3.5 and LLaMa (Touvron et al., 2023) in a similar setting (Yung et al., 2024a). However, despite its stronger performance in a few-shot setting, GPT-4o still does not surpass our smaller fine-tuned RoBERTa-HArch model, which remains the best-performing system overall in this multi-label IDRR task.

In the multi-lingual setting, we evaluated PLMs pre-trained on multiple languages as encoder backbones for HArch. Specifically, we experimented with XLM-RoBERTa$_{\text{base}}$, FLAN-T5$_{\text{base}}$ (Chung et al., 2024) and the recently released EuroBERT$_{\text{210m}}$ (Boizard et al., 2025). We also adapted the prompt template described in Appendix B to include examples in all four languages covered by DiscoGeM 2.0. Given the relatively limited number of annotated instances in German, French, and Czech within the DiscoGeM 2.0 corpus (see Table 1), we did not experiment with language specific PLMs for these languages. Instead, we infer model performance in these languages from the results in the multi-lingual setting. As shown in Table 2, using XLM-RoBERTa as the encoder backbone of HArch achieves the best overall performance in the multi-lingual configuration, even outperforming the more recent EuroBERT encoder model. Finally, we observe a surprising drop in performance in the LLM models when prompted in the multi-lingual setting compared to when prompted in the English language alone.

6 Results

In this section, we report the performance of our HArch models on the test splits of both DiscoGeM 1.0 (see Table 3) and DiscoGeM 2.0 (see Table 4) in a multi-lingual setting, where all languages are used simultaneously, and separately for each language. These evaluations allow us to assess the generalizability of our models across the different languages of the corpus.

We report results on DiscoGeM 1.0 to benchmark our approach with the only other existing model for multi-label IDRR that predicts probability distributions over the full set of PDTB 3.0 sense labels (Costa and Kosseim, 2025). Since DiscoGeM 1.0 contains only English data, we report the performance of our best-performing English model in Table 2 - RoBERTa-HArch. As shown in Table 3, RoBERTa-HArch achieves lower JS distances at level-2 and level-3 compared to the non-hierarchical model of Costa and Kosseim (2025), while maintaining comparable performance at level-1. These results provide further empirical support for our hypothesis that incorporating hierarchical dependencies improves the ability of the model to generate more accurate probability distributions at finer-grained sense levels.

Table 4 presents the results of our models across different language configurations. Since XLM-RoBERTa-HArch achieved the best overall performance in the multi-lingual setting with the validation split of DiscoGeM 2.0 (see Table 2), we report its results on the test split of the corpus when trained both on all languages and individually on each language. For English, we also include results from RoBERTa-HArch, which was fine-tuned exclusively on English data alone. In addition, we evaluate GPT-4o and Llama-4-Maverick using language-specific versions of the prompt detailed in Appendix B. As shown in Table 4, our HArch models consistently outperform both LLMs across all languages and all sense levels. The performance gap is particularly pronounced in all settings with the exception of the English language. Notably, XLM-RoBERTa-HArch achieves substantially lower JS distances than both LLMs at level-2 and level-3 in the multi-lingual and Czech language. Consistent with earlier results in Table 2, GPT-4o

Language		Model	Level-1	Level-2	Level-3
Test	Train				
All	All	XLM-RoBERTa-HArch	$\mathbf{0.353 \pm 0.005}$	$\mathbf{0.529 \pm 0.005}$	$\mathbf{0.606 \pm 0.04}$
	–	GPT-4o	0.434 ± 0.002	0.678 ± 0.003	0.705 ± 0.003
	–	Llama-4-Maverick	0.529 ± 0.004	0.769 ± 0.005	0.830 ± 0.005
Eng	Eng	RoBERTa-HArch	$\mathbf{0.331 \pm 0.004}$	$\mathbf{0.477 \pm 0.004}$	$\mathbf{0.548 \pm 0.005}$
	All	XLM-RoBERTa-HArch	0.349 ± 0.004	0.493 ± 0.006	0.568 ± 0.005
	Eng		0.347 ± 0.005	0.486 ± 0.006	0.561 ± 0.006
	–	GPT-4o	0.368 ± 0.001	0.551 ± 0.001	0.647 ± 0.002
	–	Llama-4-Maverick	0.420 ± 0.001	0.634 ± 0.002	0.702 ± 0.002
Ger	All	XLM-RoBERTa-HArch	$\mathbf{0.360 \pm 0.007}$	$\mathbf{0.566 \pm 0.006}$	$\mathbf{0.630 \pm 0.006}$
	Ger		0.393 ± 0.006	0.586 ± 0.006	0.656 ± 0.007
	–	GPT-4o	0.411 ± 0.002	0.644 ± 0.002	0.725 ± 0.002
	–	Llama-4-Maverick	0.509 ± 0.003	0.732 ± 0.006	0.809 ± 0.007
Fre	All	XLM-RoBERTa-HArch	$\mathbf{0.367 \pm 0.006}$	$\mathbf{0.553 \pm 0.005}$	$\mathbf{0.628 \pm 0.006}$
	Fre		0.382 ± 0.004	0.576 ± 0.006	0.655 ± 0.007
	–	GPT-4o	0.406 ± 0.001	0.611 ± 0.001	0.689 ± 0.002
	–	Llama-4-Maverick	0.473 ± 0.003	0.667 ± 0.003	0.748 ± 0.004
Cze	All	XLM-RoBERTa-HArch	$\mathbf{0.371 \pm 0.010}$	$\mathbf{0.555 \pm 0.009}$	$\mathbf{0.647 \pm 0.009}$
	Cze		0.404 ± 0.009	0.584 ± 0.011	0.689 ± 0.010
	–	GPT-4o	0.549 ± 0.011	0.820 ± 0.010	0.854 ± 0.012
	–	Llama-4-Maverick	0.613 ± 0.005	0.864 ± 0.004	0.890 ± 0.005

Table 4: Final results showing the average JS distance of three different runs for each model on the test split of DiscoGeM 2.0. Results are reported using all languages simultaneously and each language individually. Lower scores indicate better performance. Values in bold show the best score at each level in each tested language setting. Rows shaded in gray present the results of LLMs with few-shot prompting.

outperforms Llama-4-Maverick across all configurations, but still lags behind our smaller fine-tuned HArch models. These findings reinforce the idea that fine-tuning smaller encoder-based models can still outperform prompting large language models in specific tasks (Bucher and Martini, 2024; Chan et al., 2024; Yung et al., 2024a).

Focusing on our fine-tuned models, Table 4 shows that for German, French and Czech, XLM-RoBERTa-HArch achieves lower JS distances when trained on all languages compared to when trained on the test language alone. This suggests that multi-lingual training helps compensate for the limited annotated data available in these languages. In contrast, for English, XLM-RoBERTa-HArch performs slightly better when fine-tuned exclusively on the English subset of DiscoGeM 2.0. However, in both cases, it fails to outperform RoBERTa-HArch - which remains the strongest model for English overall. When comparing the performance of RoBERTa-HArch on DiscoGeM 1.0 in Table 3 and DiscoGeM 2.0 in Table 4, we observe a drop across all sense levels on the latter. This discrepancy can be attributed to the reduced label set used for evaluation on DiscoGeM 1.0. To allow a fair comparison with Costa and Kosseim (2025), we adopted the same adapted set of 14 level-2 sense labels proposed by Kim et al. (2020). This reduced set of level-2, and consequently level-3, sense labels lowers the complexity of the multi-label IDRR task on DiscoGeM 1.0 when compared to DiscoGeM 2.0, where we considered all sense labels on the corpus.

7 Multi-Task Contribution

To evaluate the benefit of jointly modeling all three sense levels in a multi-task setting, we conducted

Language		Model	Level-1	Level-2	Level-3	Time
Test	Train					
Eng	Eng	RoBERTa-HArch	$\mathbf{0.331 \pm 0.004}$	$\mathbf{0.477 \pm 0.004}$	$\mathbf{0.548 \pm 0.005}$	9h 09m
		RoBERTa-Individual	0.342 ± 0.004	0.513 ± 0.004	0.589 ± 0.005	21h 56m
All	All	XLM-RoBERTa-HArch	$\mathbf{0.353 \pm 0.005}$	$\mathbf{0.529 \pm 0.005}$	$\mathbf{0.606 \pm 0.04}$	20h 31m
		XLM-RoBERTa-Individual	0.365 ± 0.004	0.552 ± 0.005	0.641 ± 0.004	59h 14m

Table 5: Results showing the contribution of multi-task learning. The results are the average JS distance of three different runs on the test split of DiscoGeM 2.0. Results are reported using the English language and all languages simultaneously. Lower scores indicate better performance. Values in bold show the best score at each level in each language setting. The final column reports the total time required to compute all of the averaged results in each row.

an ablation study by training three individual models - each dedicated to predicting probability distributions for a single sense level. This study was conducted in the English and multi-lingual setting. To ensure a fair comparison, each individual model tries to replicate the architecture of the multi-task model (see Figure 1), with the exception that it includes only one classification head and no augmentation blocks. Specifically, each individual model encodes a concatenated pair of discourse arguments using the corresponding PLM, which it then passes to a linear transformation and dropout layer, before feeding the representation to a classification head — similar to the level-1 prediction path in the multi-task model. Table 5 compares the results of the three individual models for each sense level against RoBERTa-HArch in the English setting and XLM-RoBERTa-HArch in the multi-lingual setting, using the test split of DiscoGeM 2.0.

As shown in Table 5, the HArch model trained in a multi-task setting consistently outperforms the single-level models across all three sense levels. The performance gains are especially pronounced at level-2 and level-3, suggesting that these finer-grained senses benefit significantly from the shared representations and hierarchical cues provided by joint training. While level-1 predictions also improve slightly, the main advantage of multi-task learning emerges from its ability to propagate information downward through the sense hierarchy. This ablation study isolates the impact of our architecture and confirms that jointly modeling the hierarchy enhances multi-label IDRR performance. In addition to yielding better performance, HArch is also more efficient - as shown by the total time it took to calculate the averaged results in each row of Table 5 with a 32-core compute node with 512GB of RAM. This highlights both the effectiveness and efficiency of our proposed architecture.

8 Conclusion

In this work, we presented the first multi-lingual and multi-label classification model for IDRR. We modeled hierarchical dependencies between discourse senses to predict probability distributions across all three sense levels in the PDTB 3.0 framework. We introduced the first evaluation on the DiscoGeM 2.0 corpus across all languages and compared different pre-trained language models as encoder backbones within our proposed HArch model architecture. Results show that RoBERTa-HArch achieves the best performance in English, while XLM-RoBERTa-HArch performs best in the multi-lingual setting. We further evaluated our models against GPT-4o and Llama-4-Maverick using few-shot prompting across all language configurations and found that the HArch models consistently obtained lower JS distances. These findings demonstrate that fine-tuning smaller encoder-based models can outperform prompting large language models in the context of multi-label IDRR. Lastly, we benchmarked RoBERTa-HArch on the English-only DiscoGeM 1.0 corpus and achieved SOTA results compared to existing work.

These findings not only advance the field of IDRR, but also highlight the advantages of fine-tuning smaller encoder-based models over prompting large language models — particularly, given the significantly higher monetary and environmental costs associated with the latter. Looking ahead, we plan to investigate how discourse sense predictions vary across languages in DiscoGeM 2.0 for parallel discourse relations. In particular, we are interested in exploring how translation influences discourse reasoning and label distribution across these four languages. Additionally, we aim to compare the performance of using XLM-RoBERTa as the encoder in our HArch model in the German,

French and Czech languages against encoders pretrained specifically in these languages. This line of research may shed light on cross-linguistic challenges in discourse analysis and lead to more robust multi-lingual IDRR models.

9 Limitations

The PDTB 3.0 framework defines a total of 22 level-2 senses and 36 level-3 senses (after projection from level-2 senses). However, the BELIEF and SPEECH-ACT senses were not annotated in the DiscoGeM 2.0 corpus. As a result, our HArch model could only be trained on 17 level-2 and 28 level-3 senses (after projection from level-2 senses). Should these additional senses be annotated in future versions of the dataset, our model could be readily extended to incorporate them.

Another limitation of this work concerns the choice of PLMs used to generate embeddings for the pair of discourse arguments. Due to computational and time constraints, we did not experiment with larger models such as LLaMA 3 (Grattafiori et al., 2024), which might have led to improved performance. Although our experiments were conducted using a high-performance computing infrastructure, we relied on relatively smaller models, such as RoBERTa, to reduce resource demands. Nevertheless, the environmental impact of such experiments remains non-negligible. Future work should take into account the energy consumption associated with fine-tuning PLMs and prompting LLMs and consider it as an additional metric when evaluating model performance.

Acknowledgements

The authors would like to thank the anonymous reviewers for their comments. This work was financially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

Nicolas Boizard, Hippolyte Gisserot-Boukhlef, Duarte M. Alves, André Martins, Ayoub Hammal, Caio Corro, Céline Hudelot, Emmanuel Malherbe, Etienne Malaboeuf, Fanny Jourdan, Gabriel Hautreux, João Alves, Kevin El-Haddad, Manuel Faysse, Maxime Peyrard, Nuno M. Guerreiro, Patrick Fernandes, Ricardo Rei, and Pierre Colombo. 2025. EuroBERT: Scaling Multilingual Encoders for European Languages. *arXiv preprint arXiv:2503.05500*.

Peter Bourgonje and Manfred Stede. 2020. The Potsdam Commentary Corpus 2.2: Extending Annotations for Shallow Discourse Parsing. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC'20)*, pages 1061–1066, Marseille, France. European Language Resources Association (ELRA).

Martin Juan José Bucher and Marco Martini. 2024. Fine-Tuned 'Small' LLMs (Still) Significantly Outperform Zero-Shot Generative AI Models in Text Classification. *arXiv preprint arXiv:2406.08660*.

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 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL'24)*, pages 684–721, St. Julian's, Malta. Association for Computational Linguistics (ACL).

Chunkit Chan, Xin Liu, Jiayang Cheng, Zihan Li, Yangqiu Song, Ginny Y Wong, and Simon See. 2023. DiscoPrompt: Path Prediction Prompt Tuning for Implicit Discourse Relation Recognition. In *Findings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL'23)*, pages 35–57, Toronto, Ontario, Canada. Association for Computational Linguistics (ACL).

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, and 1 others. 2024. Scaling Instruction-Finetuned Language Models. *Journal of Machine Learning Research (JMLR)*, 25(70):1–53.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL'20)*, pages 8440–8451, Online. Association for Computational Linguistics (ACL).

Nelson Filipe Costa and Leila Kosseim. 2024. Exploring Soft-Label Training for Implicit Discourse Relation Recognition. In *Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI'24)*, pages 120–126, St. Julians, Malta. Association for Computational Linguistics (ACL).

Nelson Filipe Costa and Leila Kosseim. 2025. A Multi-Task and Multi-Label Classification Model for Implicit Discourse Relation Recognition. In *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL'25)*, Avignon, France. Association for Computational Linguistics (ACL).

Nelson Filipe Costa, Nadia Sheikh, and Leila Kosseim. 2023. Mapping Explicit and Implicit Discourse Relations between the RST-DT and the PDTB 3.0. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing (RANLP'23)*, pages 344–352, Varna, Bulgaria.

Vera Demberg, Merel CJ Scholman, and Fatemeh Torabi Asr. 2019. How compatible are our discourse annotation frameworks? Insights from mapping RST-DT and PDTB annotations. *Dialogue & Discourse*, 10(1):87–135.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian Error Linear Units (GELUs). *arXiv preprint arXiv:1606.08415*.

Jet Hoek, Merel C.J. Scholman, and Ted J.M. Sanders. 2021. Is there less annotator agreement when the discourse relation is underspecified? In *Proceedings of the 1st Workshop on Integrating Perspectives on Discourse Annotation*, pages 1–6, Online. Association for Computational Linguistics (ACL).

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 (ACL'20)*, pages 5404–5414, Online. Association for Computational Linguistics (ACL).

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR'15)*, pages 1–15, San Diego, California, USA.

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 (ACL'23)*, pages 15696–15712, Toronto, Ontario, Canada. Association for Computational Linguistics (ACL).

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 preprint arXiv:1907.11692*.

Wanqiu Long, N Siddharth, and Bonnie Webber. 2024. Multi-Label Classification for Implicit Discourse Relation Recognition. In *Findings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL'24)*, pages 8437–8451, Bangkok, Thailand. Association for Computational Linguistics (ACL).

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 (EMNLP'22)*, pages 10704–10716, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics (ACL).

Wanqiu Long and Bonnie Webber. 2024. Leveraging Hierarchical Prototypes as the Verbalizer for Implicit Discourse Relation Recognition. *arXiv preprint arXiv:2411.14880*.

Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2004. The Penn Discourse Treebank. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, pages 2237–2240, Lisbon, Portugal. European Language Resources Association (ELRA).

OpenAI. 2024. GPT-4o System Card. *arXiv preprint arXiv:2410.21276*.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008a. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 2961–2968, Marrakech, Morocco. European Language Resources Association (ELRA).

Rashmi Prasad, Alan Lee, Nikhil Dinesh, Eleni Miltsakaki, Geraud Campion, Aravind Joshi, and Bonnie Webber. 2008b. Penn Discourse Treebank Version 2.0. LDC2008T05. Web Download. Philadelphia: Linguistic Data Consortium.

Rashmi Prasad, Bonnie Webber, Alan Lee, and Aravind Joshi. 2019. Penn Discourse Treebank Version 3.0. LDC2019T05. Web Download. Philadelphia: Linguistic Data Consortium.

Valentina Pyatkin, Ayal Klein, Reut Tsarfaty, and Ido Dagan. 2020. QADiscourse - Discourse Relations as QA Pairs: Representation, Crowdsourcing and Baselines. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP'20)*, pages 2804–2819, Online. Association for Computational Linguistics.

Valentina Pyatkin, Frances Yung, Merel C. J. Scholman, Reut Tsarfaty, Ido Dagan, and Vera Demberg. 2023. Design Choices for Crowdsourcing Implicit Discourse Relations: Revealing the Biases Introduced by Task Design. *Transactions of the Association for Computational Linguistics (TACL)*, 11:1014–1032.

Ahmed Ruby, Christian Hardmeier, and Sara Stymne. 2025. Multimodal Extraction and Recognition of Arabic Implicit Discourse Relations. In *Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation*, pages 5415–5429, Abu Dhabi, UAE. International Committee for Computational Linguistics (ICCL).

Muhammed Saeed, Peter Bourgonje, and Vera Demberg. 2025. Implicit Discourse Relation Classification For Nigerian Pidgin. In *Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation*, pages 2561–2574, Abu Dhabi, UAE. International Committee for Computational Linguistics (ICCL).

Merel Scholman and Vera Demberg. 2017. Examples and Specifications that Prove a Point: Identifying Elaborative and Argumentative Discourse Relations. *Dialogue & Discourse*, 8(2):56–83.

Merel Scholman, Tianai Dong, Frances Yung, and Vera Demberg. 2022a. DiscoGeM: A Crowdsourced Corpus of Genre-Mixed Implicit Discourse Relations. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC'22)*, pages 3281–3290, Marseille, France. European Language Resources Association (ELRA).

Merel Scholman, Valentina Pyatkin, Frances Yung, Ido Dagan, Reut Tsarfaty, and Vera Demberg. 2022b. Design Choices in Crowdsourcing Discourse Relation Annotations: The Effect of Worker Selection and Training. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC'22)*, pages 2148–2156, Marseille, France. European Language Resources Association (ELRA).

Manfred Stede. 2008. Disambiguating Rhetorical Structure. *Research on Language and Computation*, 6(3):311–332.

Pavlína Synková, Jiří Mírovský, Lucie Poláková, and Magdaléna Rysová. 2024. Announcing the Prague Discourse Treebank 3.0. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING'24)*, pages 1270–1279, Torino, Italia. European Language Resources Association (ELRA) and International Committee for Computational Linguistics (ICCL).

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*.

Michiel van der Meer, Neele Falk, Pradeep K Murukannaiah, and Enrico Liscio. 2024. Annotator-Centric Active Learning for Subjective NLP Tasks. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP'24)*, pages 18537–18555, Miami, Florida, USA. Association for Computational Linguistics (ACL).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS'17)*, Long Beach, California, USA. Curran Associates, Inc.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. *arXiv preprint arXiv:2412.13663*.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The Penn Discourse Treebank 3.0 Annotation Manual. Technical report, University of Pennsylvania.

Frances Yung, Mansoor Ahmad, Merel Scholman, and Vera Demberg. 2024a. Prompting Implicit Discourse Relation Annotation. In *Proceedings of the 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 150–165, St. Julian's, Malta. Association for Computational Linguistics (ACL).

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 (CODI'22)*, pages 48–53, Gyeongju, Republic of Korea. International Conference on Computational Linguistics (ICCL).

Frances Yung and Vera Demberg. 2025. On Crowdsourcing Task Design for Discourse Relation Annotation. In *Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation*, pages 12–19, Abu Dhabi, UAE. International Committee for Computational Linguistics (ICCL).

Frances Yung, Vera Demberg, and Merel Scholman. 2019. Crowdsourcing Discourse Relation Annotations by a Two-Step Connective Insertion Task. In *Proceedings of the 13th Linguistic Annotation Workshop (LAW'19)*, pages 16–25, Florence, Italy. Association for Computational Linguistics (ACL).

Frances Yung, Merel Scholman, Sarka Zikanova, and Vera Demberg. 2024b. DiscoGeM 2.0: A Parallel Corpus of English, German, French and Czech Implicit Discourse Relations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING'24)*, pages 4940–4956, Torino, Italia. European Language Resources Association (ELRA) and International Committee for Computational Linguistics (ICCL).

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'24)*, pages 7760–7773, Torino, Italia. European Language Resources Association (ELRA) and International Committee for Computational Linguistics (ICCL).

Deniz Zeyrek, Giedrė Valūnaitė Oleškevičienė, and Amália Mendes. 2024. Multiple Discourse Relations in English TED Talks and Their Translation into Lithuanian, Portuguese and Turkish. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING'24)*, pages 125–134, Torino, Italia. European Language Resources Association (ELRA) and International Committee for Computational Linguistics (ICCL).

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 the 61st Annual Meeting of the Association for Computational Linguistics (ACL'23)*, pages 6477–6492, Toronto, Ontario, Canada. Association for Computational Linguistics (ACL).

A Data Statistics

Table 6 shows the distribution of level-2 and level-3 senses per language in the DiscoGeM 2.0 corpus. We replaced the non-existent PDTB 3.0 level-3 senses with their corresponding level-2 sense and marked them with * in the *Level-3* column. Each value represents the sum of the corresponding sense in all the multi-label distributions of implicit discourse relations in the specific language subset.

B Prompt Template

Figure 2 illustrates the prompt-response template used to evaluate GPT-4o and LLaMA-4-Maverick on multi-label IDRR using the English subset of DiscoGeM 2.0. For evaluations in other languages, we translated the prompt template and selected examples from the corresponding language subset of the corpus. In the multi-lingual setting, where all four languages are used simultaneously, we kept the prompt structure in English and provided examples in all four languages. Due to space constraints, Figure 2 only shows two examples of discourse relations in the initial *User Message* prompt. However, we included five examples when prompting the models. All of the examples were taken from the training split of DiscoGeM 2.0.

We designed the prompt to closely replicate the annotation methodology of DiscoGeM 2.0. When crowdsourcing the annotation of the corpus, the authors provided the annotators a list of non-ambiguous connectives (Yung et al., 2024b) that represented each of the possible senses in the PDTB 3.0 framework. This list of connectives was adapted for each language. The annotators were then asked to choose the connective that best expresses the semantic relation between the arguments of the relation being annotated, irrespective of whether syntax would need to be adjusted. By mirroring this annotation process in our prompting template, we were able to directly evaluate the performance of both LLMs on multi-label IDRR using the test split of DiscoGeM 2.0.

Level-2	English	German	French	Czech	All	Level-3	English	German	French	Czech	All
SYNCHRONOUS	99.4	191.2	139.8	254.6	685.0	SYNCHRONOUS*	99.4	191.2	139.8	254.6	685.0
ASYNCHRONOUS	457.4	247.1	260.9	204.2	1169.6	PRECEDENCE	420.7	222.7	240.0	193.7	1,077.1
						SUCCESSION	36.7	24.4	20.9	10.5	92.5
CAUSE	1,690.7	457.5	608.6	576.7	3,333.5	REASON	386.9	243.9	237.6	281.3	1,149.8
						RESULT	1,303.8	213.6	371.0	295.4	2,183.7
CONDITION	1.2	126.1	93.2	42.0	262.5	ARG1-AS-COND	0.1	109.4	87.9	25.1	222.4
						ARG2-AS-COND	1.1	16.7	5.4	16.9	40.1
NEG-CONDITION	1.1	14.0	14.0	12.9	42.0	ARG1-AS-NEGCOND	1.0	7.8	10.9	10.9	30.6
						ARG2-AS-NEGCOND	0.1	6.2	3.1	2.0	11.5
PURPOSE	2.3	174.8	95.2	49.5	321.8	ARG1-AS-GOAL	1.4	114.8	91.6	46.5	254.2
						ARG2-AS-GOAL	0.9	60.0	3.6	3.1	67.6
CONCESSION	505.1	161.3	261.3	160.9	1088.6	ARG1-AS-DENIER	167.5	61.9	91.3	39.9	360.5
						ARG2-AS-DENIER	337.7	99.4	169.9	121.0	728.1
CONTRAST	187.4	98.7	173.3	40.8	500.2	CONTRAST*	187.4	98.7	173.3	40.8	500.2
SIMILARITY	103.7	86.5	90.4	25.5	306.1	SIMILARITY*	103.7	86.5	90.4	25.5	306.1
CONJUNCTION	1,298.8	260.0	121.8	155.9	1,836.5	CONJUNCTION*	1,298.8	260.0	121.8	155.9	1,836.5
DISJUNCTION	2.7	6.5	3.0	3.1	15.3	DISJUNCTION*	2.7	6.5	3.0	3.1	15.3
EQUIVALENCE	19.0	154.7	101.5	57.4	332.6	EQUIVALENCE*	19.0	154.7	101.5	57.4	332.6
EXCEPTION	1.8	26.8	41.0	6.8	76.4	ARG1-AS-EXCEPTION	0.3	14.7	30.7	2.8	48.5
						ARG2-AS-EXCEPTION	1.4	12.2	10.3	4.0	27.8
INSTANTIATION	352.4	99.3	124.2	84.4	660.3	ARG1-AS-INSTANCE	16.5	0.0	35.6	25.9	78.0
						ARG2-AS-INSTANCE	335.9	99.3	88.6	58.5	582.3
LEVEL-OF-DETAIL	1079.1	393.2	363.1	278.8	2114.1	ARG1-AS-DETAIL	154.9	88.7	51.8	110.1	405.6
						ARG2-AS-DETAIL	924.1	304.5	311.3	168.6	1708.5
MANNER	4.5	33.7	124.0	37.2	199.4	ARG1-AS-MANNER	1.3	26.6	112.3	22.7	163.0
						ARG2-AS-MANNER	3.1	7.0	11.7	14.6	36.5
SUBSTITUTION	40.5	56.6	12.7	9.3	119.1	ARG1-AS-SUBSTITUTION	0.0	7.6	6.0	1.3	14.9
						ARG2-AS-SUBSTITUTION	40.5	49.0	6.7	8.0	104.1

Table 6: Distribution of level-2 and level-3 senses per language in the DiscoGeM 2.0 corpus. Each value represents the sum of the corresponding sense in all the multi-label distributions of implicit discourse relations in the specific language subset. Senses at level-3 marked with * represent level-2 senses that were projected to level-3 since they do not exist in the PDTB 3.0.

Prompt template in English

System Message:

You're an expert linguist with the task of identifying the best discourse connectives to fill in the gap between two arguments. Below you have the ordered list of 28 discourse connectives you can use.

["at the same time", "then", "after", "because", "as a result", "in that case", "if", "if not", "unless", "for that purpose", "so that", "even though", "nonetheless", "on the other hand", "similarly", "also", "or", "in other words", "other than that", "an exception is that", "this illustrates that", "for example", "in short", "in more detail", "thereby", "as if", "rather than", "instead"]

For each pair of arguments given to you, you have to choose the connectives from this list that could be used to connect the two arguments. Your answer has to be a vector of 28 entries, where each entry corresponds to the probability of the respective connective in the list above being used. The sum of all probabilities in the vector must be equal to 1.

User Message:

Here are a few examples.

Argument 1: There, comrades, is the answer to all our problems. It is summed up in a single word - Man.
Argument 2: Man is the only real enemy we have. Remove Man from the scene, and the root cause of hunger and overwork is abolished for ever.

Answer: [0.0, 0.1, 0.0, 0.3, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.1, 0.0, 0.0, 0.0, 0.0, 0.3, 0.2, 0.0, 0.0, 0.0, 0.0]

... (x3)

Argument 1: It had been a good one. There had been a flying motorcycle in it.
Argument 2: He had a funny feeling he'd had the same dream before. His aunt was back outside the door.

Answer: [0.0, 0.0, 0.0, 0.0, 0.111, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.111, 0.0, 0.0, 0.667, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.111, 0.0, 0.0, 0.0, 0.0]

Assistant Message:

Got it! Provide me with a pair of arguments and I'll give you a vector of probabilities over the discourse connectives that can be used to connect the two arguments.

User Message:

Argument 1: Everybody seemed to hear a rustling outside the door. There was no sound of footsteps.
Argument 2: It was like light silk sliding over the panel. Then it stopped.

Assistant Message:

[0.0, 0.0, 0.0, 0.0, 0.1, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.1, 0.1, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.2, 0.0, 0.0, 0.0, 0.5]

User Message:

Argument 1: It pretends to respect human rights. It pretends to persecute no one.
Argument 2: It pretends to fear nothing. It pretends to pretend nothing.

Assistant Message:

[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.3, 0.7, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0]

Figure 2: Prompt-response template for predicting multi-label IDRR in English using DiscoGeM 2.0.