Probing LLMs for Multilingual Discourse Generalization Through a Unified Label Set

Florian Eichin*, **Yang Janet Liu***, **Barbara Plank, and Michael A. Hedderich**
MaiNLP, Center for Information and Language Processing, LMU Munich, Germany
Munich Center for Machine Learning (MCML)
{feichin,yliu,bplank,hedderich}@cis.lmu.de

Abstract

Discourse understanding is essential for many NLP tasks, yet most existing work remains constrained by framework-dependent discourse representations. This work investigates whether large language models (LLMs) capture discourse knowledge that generalizes across languages and frameworks. We address this question along two dimensions: (1) developing a unified discourse relation label set to facilitate cross-lingual and cross-framework discourse analysis, and (2) probing LLMs to assess whether they encode generalizable discourse abstractions. Using multilingual discourse relation classification as a testbed, we examine a comprehensive set of 23 LLMs of varying sizes and multilingual capabilities. Our results show that LLMs, especially those with multilingual training corpora, can generalize discourse information across languages and frameworks. Further layer-wise analyses reveal that language generalization at the discourse level is most salient in the intermediate layers. Lastly, our error analysis provides an account of challenging relation classes.

1 Introduction

Many approaches to NLP primarily focus on sentence-level analyses (e.g. Heinzerling and Strube 2019; Pimentel et al. 2021; Mrini et al. 2020). However, there are many research questions which cannot be answered without considering sentences in a larger **discourse**: new meanings emerge from the relationships between sentences, and since more than one interpretation can be created, how do we determine the intended, most reasonable or justifiable meaning (Schiffrin et al., 2015)?

Despite significant progress in discourse processing (Webber et al., 2024; Zeldes et al., 2025; Stede, 2011), much of the research and resources remain constrained by theory-/framework-specific assumptions, limiting the generalizability of findings across languages, domains, and communicative intents (Liu and Zeldes, 2023). This leads to datasets that are tightly coupled to their respective frameworks and limits the development of generalizable discourse models. While there has been work on framework-dependent parsing that leverages resources from other frameworks (Braud et al., 2016) or languages (Braud et al., 2017; Liu et al., 2021), the reliance on framework-specific corpora, which are typically scarce and skewed towards high-resource languages, further exacerbates the challenge of multilingual discourse processing. Thus, we need a unified view and approach to investigating discourse generalization.

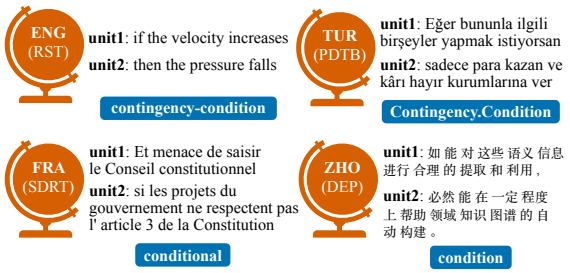


Figure 1: Examples of the core discourse relation CONDITION (Bunt and Prasad, 2016) annotated in different frameworks and languages using different labels.

From the **theory and data** perspectives, as argued in Bunt and Prasad (2016) and exemplified in Figure 1, despite differences between frameworks, there exists a set of 'core' discourse relations which are commonly found in existing approaches to discourse relations and their annotation. From the **model** perspective, there is growing evidence demonstrating that large language models (LLMs) learn and share generalizable abstraction across typologically diverse languages (e.g. Brinkmann et al. 2025; Peng and Søgaard 2024), but such capabilities remain underexplored in discourse.

In this work, we address **discourse generalization** across two dimensions using discourse relation

*Equal contribution.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 18665–18684

classification as a testbed: we first develop a unified discourse relation label set to enable cross-lingual and cross-framework discourse analysis on the the multilingual DISRPT benchmark (Braud et al., 2024). Then, we use *probing* (Alain and Bengio, 2018) to understand the internal discourse representations of 23 LLMs with varying sizes and multilingual capabilities. We investigate whether their representations capture generalizable discourse abstractions across typologically diverse languages or whether they are limited by dataset-specific biases. We hypothesize that multilingual models might encode universal representations of certain relations while adjusting to language-specific features.

We find that overall LLMs are able to generalize at the discourse level across languages and frameworks, and that multilingual training and larger model sizes both increase probe performance. Our layer-wise analyses show that language generalization at the discourse level is most salient in the intermediate layers, which are most predictive of multilingual discourse information. To our best knowledge, this is the first work to apply a unified label set to multilingual discourse relation classification at scale. Lastly, we discuss challenges and biases in LLM discourse representations, providing insights into the limitations and potential avenues for improving discourse modeling in multilingual and generalization settings. The implementations of our experiments are available on GitHub.[1]

2 Related Work

Unifying Discourse Relations. Discourse relations are fundamental to structuring coherent text and conveying meaning beyond the sentence level. Being able to identify and interpret these relations is crucial for many downstream NLP tasks, including machine comprehension (Narasimhan and Barzilay, 2015; Li et al., 2020), sentiment analysis (Huber and Carenini, 2019), question answering (Chai and Jin, 2004), and summarization (Durrett et al., 2016; Cohan et al., 2018; Xu et al., 2020; Adams et al., 2023). However, due to different approaches to discourse relations, such as the Rhetorical Structure Theory (RST, Mann and Thompson 1988), the Penn Discourse Treenbank (PDTB, Webber et al. 2019), the Segmented Discourse Representation Theory (SDRT, Asher and Lascarides 2003), Discourse Dependency Structure (DEP, Li et al. 2014; Morey et al. 2018), and the Cognitive Approach to Coherence Relations (CCR, Sanders et al. 1992), researchers have not reached a consensus on a unified set of discourse relations. There have been a few mapping proposals and examinations on existing annotations (Chiarcos, 2014; Benamara and Taboada, 2015; Bunt and Prasad, 2016; Rehbein et al., 2016; Sanders et al., 2021; Demberg et al., 2019), but they are either merely focused on two frameworks at a time (e.g. RST and SDRT in Benamara and Taboada 2015), or on high-resource languages and news-centric data such as mapping RST-DT and PDTB v2 by Demberg et al. (2019) and PDTB v3 by Costa et al. (2023).

In particular, while Bunt and Prasad (2016) identified a set of core discourse relations, it did not cover DEP and was limited to English and French only. The examined corpora in their work also did not cover discourse phenomena concerning pragmatics or textual organization, both of which are indispensable aspects in discourse analysis. For instance, BACKGROUND and MOTIVATION are two RST-style relations that are not present in the examined RST Discourse Treebank (RST-DT, Carlson et al. 2003). Both relations are expressed by various perlocutionary acts to affect readers' or speakers' attitude and beliefs. To address these limitations, we first conduct an extensive review of previous work on discourse relation mapping proposals and present a unified label set (§3) to enable empirical studies in a multilingual setting, which has not been systematically explored before.

The 2021 DISRPT Shared Task (Zeldes et al., 2021) introduced the first iteration of the discourse relation classification task in a unified format. It leveraged shared foundational assumptions across frameworks. However, no unified discourse relation labels were proposed, meaning that each dataset has its own label set, even for the ones that come from the same framework. We thus leverage this resource and propose a unified label set that is the first to be empirically tested in discourse relation classification across 13 languages, four frameworks, and 26 datasets, which cover various, modern genres, domains, and modalities.

Probing for Linguistic Representation and Generalization. A growing body of research has explored the extent to which pretrained language models (PLMs) and LLMs encode linguistic representations and exhibit generalizable abstraction (Hupkes et al., 2023). They primarily focus on probing morphology (Brinkmann et al., 2025), syn-

[1] https://github.com/mainlp/discourse_probes

tax (Conneau et al., 2018; Hale and Stanojević, 2024), semantic knowledge (Jumelet et al., 2021), and syntax-semantics comprehension through cognitive linguistics paradigms such as construction grammar (Weissweiler et al., 2022). While some studies demonstrate that LLMs share latent grammatical representations across diverse languages (Brinkmann et al., 2025), others also highlight key limitations in the semantic capabilities of LLMs (Scivetti et al., 2025).

Previous work has examined the ability of PLMs/LLMs to understand discourse (Gan et al., 2024; Saputa et al., 2024; Miao et al., 2024), but their investigations are either limited to framework-dependent representations, monolingual datasets, or focus on single domains. Specifically, Koto et al. (2021) examined a variety of PLMs through a set of framework-dependent probing tasks for discourse coherence by looking at the residual stream (i.e. token representations), while we approach discourse relation classification with a unified format and label set across various frameworks and 13 languages (§3) using attentions (§4), offering opportunities for investigating discourse generalization across languages and frameworks. Kurfalı and Östling (2021) extended discourse probing to multilingual PLMs such as the multilingual BERT and XLM-RoBERTa (Devlin et al., 2019; Conneau et al., 2020) to examine how well they transfer discourse-level knowledge across languages, but their evaluation of discourse coherence was also framework-dependent and was only performed on two English datasets in the news domain. Lastly, Kim and Schuster (2023) studied discourse understanding in LMs by probing their ability to track discourse entities, but their investigation is also limited to English.

3 A Unified Label Set

Building up on previous effort on mapping discourse relations across corpora and frameworks (Benamara and Taboada, 2015; Bunt and Prasad, 2016; Liu and Zeldes, 2023), we present a unified set of 17 discourse relation classes to facilitate empirical investigation that is not constrained by framework-dependent discourse representations. This unified label set is motivated by both theoretical groundings and empirical studies, and takes annotation guidelines into considerations. Specifically, the proposed unified label set adapts the top-level classes from the mapping proposal described in Benamara and Taboada (2015) and extends it to phenomena frequent in dialogues such as acknowledgment, interruption, and correction (Asher et al., 2016). Through a series of experiments, we demonstrate how our unified label set generalizes across languages and frameworks, providing a foundation for future empirical studies. Below we describe the top-level classes and include definitions and examples for all 17 relation labels in Appendix A.

TEMPORAL is mapped to framework-specific labels that establish a chronological sequence between events or states. Temporal relations indicate when one event occurs in relation to another such as *before*, *after*, or *simultaneously*. These relations help organize discourse by providing a timeline of events. RST's SEQUENCE, PDTB's TEMPORAL.ASYNCHRONOUS/SYNCHRONOUS, and SDRT's TEMPLOC and FLASHBACK (following Muller et al. 2012) all fall under this class.

STRUCTURING corresponds to fine-grained discourse relations that organize the structure of a text or conversation without necessarily conveying content-based meaning, connect discourse units of distinct context and equal prominence, and help guide the reader or listener through the discourse. RST-style relations such as LIST and TEXTUAL-ORGANIZATION, PDTB's EXPANSION.DISJUNCTION, and PARALLEL and ALTERNATION in SDRT are mapped to this class.

THEMATIC is a broad class which includes relations among the content of the propositions, according to Benamara and Taboada (2015). We adapt this top-level class to contain six subclasses: FRAMING, ATTRIBUTION, MODE, REFORMULATION, COMPARISON, and ELABORATION. In particular, we introduce REFORMULATION, which corresponds to relations by which one discourse unit re-expresses the meaning of another in a different form and/or from a different perspective to help reinforce understanding. RST's SUMMARY and RESTATEMENT and PDTB's EXPANSION.EQUIVALENCE are mapped to REFORMULATION.

CAUSAL-ARGUMENTATIVE contains subclasses that indicate a causal relation or involve rhetorical reasoning that shapes the coherence and persuasiveness of an argument: CAUSAL, ADVERSATIVE, EXPLANATION, EVALUATION, CONTINGENCY, and ENABLEMENT. ADVERSATIVE is defined as connecting discourse units for which some incompatibility is being highlighted and covers commonly used discourse relations such as CONCESSION and CONTRAST.

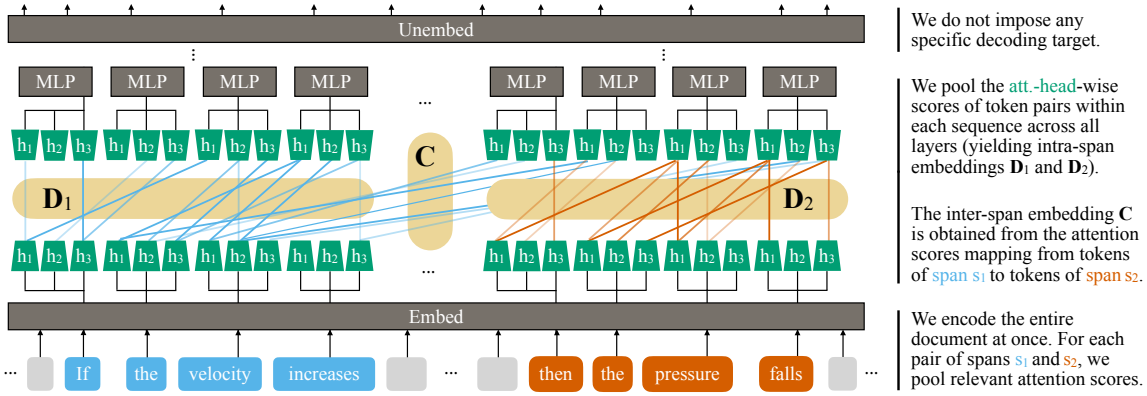


Figure 2: **Schematic visualization of how the attention representations are obtained.** We probe the concatenation of pooled representations $(\mathbf{D}_1, \mathbf{D}_2, \mathbf{C})$. Scores mapping from or to context-tokens are not considered. Note that we study decoder-only models where attention maps only to previous tokens.

TOPIC-MANAGEMENT contains three subclasses that cover discourse phenomena that involve interaction, topic shifts, and question-answer pairs: TOPIC-ADJUSTMENT, TOPIC-CHANGE, and TOPIC-COMMENT. TOPIC-ADJUSTMENT is primarily used for cases where a discourse segment modifies, redirects, or adjusts the ongoing topic of discussion such as interruption, which signal deviations from the expected discourse progression, often reflecting interactive or dynamic aspects of communication.

4 Probing LLMs for Discourse Relations

Discourse relation classification is the task of identifying the coherence relations that hold between different parts of a text, such as recognizing that one sentence specifies the cause of events in another, or that a subordinate clause indicates the condition of a main clause (Jurafsky and Martin, 2025), as exemplified in Figure 1. Successfully solving this task requires combining the information contained in different parts of these sequences. We aim to investigate whether current LLMs have access to and process this information. Since almost all current state-of-the-art open-source models are decoder-only Transformer models (Vaswani et al., 2017), we focus on this particular architecture.

In the computational graph of the Transformer, attention layers provide the only connections between the next token prediction and previous token positions of the input sequence. We therefore argue that generating predictions relying on discourse-level information necessarily has to involve the connections provided by the **self-attention** layers. To uncover the extent to which discourse information is represented, we thus propose to probe the attention scores between the tokens contained in the two input sequences for discourse relation classification. Note that we do not assume that the processing of discourse-level phenomena is exclusively located in the attention layers. Rather, observing attention scores provides a lightweight and scalable approach to our research question.

Probing requires a fixed length representation to enable training a classifier. Inspired by Alain and Bengio (2018) and Koto et al. (2021), we use maximum pooling to turn the attention scores of variable token-length sentences into a fixed-length relational representation. Our approach is illustrated in Figure 2. To be more precise, for a document d of length N tokens, we propose to compute the full attention matrix $\mathbf{X} \in \mathbb{R}^{L \times H \times N \times N}$ of attention scores where L is the number of layers and H is the number of attention heads in each layer. We do this by inputting the whole document d into the model at once computing a single forward pass. Since we are only interested in the model internals, we ignore the outputs and thus do not impose any specific decoding strategy. Let $I_1 = (i_1, i_2, ..., i_{N_1})$, $I_2 = (j_1, j_2, ..., j_{N_2})$ be the token indices of two sequences $s_1, s_2 \subset d$ where $I_1 < I_2$. The pooling step is carried out for each attention head in each layer. We thus have:

$$\begin{aligned} \mathbf{C} &= \left(\max(\{\mathbf{X}_{i,j,k,l} | k \in I_1, l \in I_2\})\right)_{i,j} \\ \mathbf{D}_m &= \left(\max(\{\mathbf{X}_{i,j,k,l} | k \in I_m, l \in I_m\})\right)_{i,j} \end{aligned} \quad (1)$$

In other words, $\mathbf{C}, \mathbf{D}_1, \mathbf{D}_2$ are the matrices of maximum-pooled attention scores between and within the two spans respectively (we only consider the lower half of the attention matrix, as for decoder models the upper half is masked). We also ablate other strategies such as mean pooling and

using subsets of the attention scores described (see Table 4 and Table 5 in Appendix D), and find that the proposed setup strikes the best balance between performance and size of the representations. We probe the flattened concatenation

$$\mathbf{a} = flatten((\mathbf{D}_1, \mathbf{D}_2, \mathbf{C})) \in \mathbb{R}^{3LH} \quad (2)$$

Here, we follow Tenney et al. (2019) and use a two-layer MLP with $\tanh$ and Sigmoid activations σ. Our probe is thus a classifier of the form

$$\mathbf{y} = \sigma(\mathbf{W}_2 \tanh(\mathbf{W}_1 \mathbf{a} + \mathbf{b}_1) + \mathbf{b}_2) \quad (3)$$

where $\mathbf{W}_1 \in \mathbb{R}^{D \times 3LH}, \mathbf{W}_2 \in \mathbb{R}^{C \times D}, \mathbf{b}_1 \in \mathbb{R}^D, \mathbf{b}_2 \in \mathbb{R}^C$ are parameters and $\mathbf{y}$ is the one-hot encoded target vector. The hidden size D is a hyperparameter (more training details in Appendix C).

Since the memory requirements of this approach scale quadratically with the number of tokens, we restrict the maximum document length to $N_{max} = 4000$, which is chosen to optimize the utilization of our GPUs. For documents with length $N > N_{max}$, we employ a moving window approach and encode slices of length N_{max} at every stride $S = \frac{N_{max}}{2}$. Relations that span over parts of the document that cannot be captured by any of these windows are discarded and encoded by the mean of the remaining relations. In practice, this affects a fraction of less than 0.2% of the DISRPT instances and thus has a negligible effect on the results.

Note that our approach requires only one forward pass per window encoding multiple relations in parallel and furthermore incorporates the document context into the representation.

We probe both the combined attention scores over all layers as well as the layer-wise representations. We hypothesize that the former will lead to better overall probe accuracy as the understanding of more complex discourse relations typically requires hierarchical processing steps. We include the latter layer-wise analysis to study the structural dynamics of discourse processing in the model.

5 Experimental Setup

Data. We use the DISRPT benchmark from the DISRPT 2023 shared task (Braud et al., 2023)[2], a multilingual, multi-domain, and cross-framework (RST, PDTB, SDRT, and DEP) dataset covering 13 languages from five language families (Dryer and Haspelmath, 2013), with $224,281$ discourse relation instances from 23 corpora (details in Appendix B). While all annotations from all frameworks have been represented in a unified format, i.e., a set of discourse unit pairs for which a discourse relation is known to apply (Zeldes et al., 2021), the set of discourse relation labels is corpus-specific, preventing multilingual discourse analysis in a directly comparable and generalizable manner. Thus, we map corpus-specific labels in DISRPT to our proposed unified label set.

Models. We study a wide range of decoder-only LLMs that reflect the diversity of the current state-of-the-art. Generally, we select models with publicly available weights along two dimensions: (1) the size of the model, and (2) the multilinguality of the training corpus. A full list of models including their advertised supported languages and parameter counts is shown in Table 2 in Appendix C.

Representing two of the most popular open-weights model families, we include most of the range of `Qwen2.5` (Qwen et al., 2025) and `Llama3` (Grattafiori et al., 2024) base models. The authors of both models report altered shares of multilingual data in their training corpora, with `Qwen2.5` 'supporting' 77% and `Llama3` 54% of the languages included in DISRPT. Covering 62% and 100%, we include `Mistral-Small-24B` (Mistral, 2025) and `Emma500` (Ji et al., 2024) as further recent multilingual models. `BLOOM` is another model family that targets multilinguality (Scao et al., 2023), covering 54% of the languages in DISRPT. We include the smaller versions as well as the more recent `bloomz-7b1` model trained on additional data. Covering many languages in DISRPT with 85%, we include the `Aya-23` (Aryabumi et al., 2024) and `Aya-Expanse` (Dang et al., 2024) model families. Finally, we also include `Phi-4` (Abdin et al., 2024) as a recent monolingual model.

DisCoDisCo. To better contextualize the performance of our simple probes, we train DisCoDisCo (Gessler et al., 2021), the 2021 DISRPT shared task winning system for the discourse relation classification task (Zeldes et al., 2021). It was not tested during the 2023 edition, but the reported scores are better than the 2023 winning system on common corpora (Braud et al., 2023), justifying its use as a reference system. DisCoDisCo is a Transformer-based model consisting of a sentence-pair classifier. To ensure comparability, we train a single DisCoDisCo model on the entire DISRPT benchmark using `xlm-roberta-base` (Conneau et al., 2020)

[2] https://github.com/disrpt/sharedtask2023

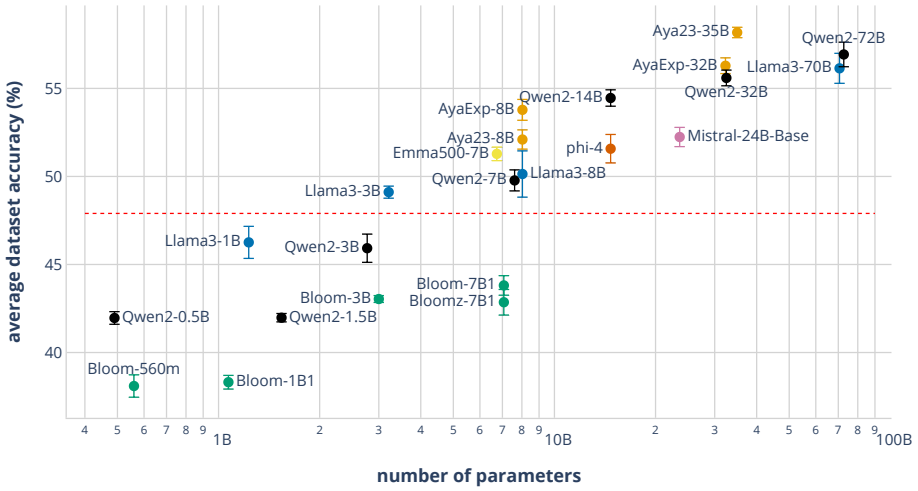


Figure 3: Mean accuracy over five runs of the probing classifiers trained on the entire DISRPT and full attention representations. The reference system DisCoDisCo achieved a mean accuracy of 47.9% (the red dashed line).

as a multilingual encoder. We also train dataset-specific models following the setup in Braud et al. (2023). For all scenarios, we report both the dataset-wise accuracy and the mean accuracy.

6 Results & Analysis

6.1 Overall Performance and Comparison

Figure 3 shows the performance of our probes on DISRPT using the unified label set for the 23 LLMs along with their sizes (see Table 2 in Appendix C for a detailed overview). Overall, a clear trend emerges wherein larger models generally achieve higher accuracy, but with notable deviations based on their language coverage. Interestingly, even the `Llama3-3B` probe is able to outperform the fine-tuned DisCoDisCo trained on all languages (49.1% vs 47.9%), while the larger model probes beat it by a margin of up to 10.3%. Despite their multilingual training, the `BLOOM` model probes underperform compared to other LLMs and DisCoDisCo. Several factors may be attributed to this including smaller training data, tokenizer effect (Toraman et al., 2023), and limited coverage of languages present in DISRPT (54%: French, Spanish, Portuguese, English, and Chinese). This is consistent and further supplements Dakle et al. (2023)'s findings on the evaluation of `BLOOM` models across a variety of syntactic and semantic tasks as well as their unsatisfying performance in multilingual settings. In addition, `Llama3` exhibits the lowest proportion of DISRPT languages covered, which does not seem to lead to a disadvantage as its probes outperform the similarly sized smaller versions of `Qwen2.5` and match its larger counterparts, reporting a higher coverage of 'supported' languages.

Examining scaling trends, we observe a log-linear increase in probe performance across both the `Llama3` and `Qwen2.5` model families. Interestingly, the English-only `Phi-4` lags behind the similarly sized `Qwen2.5-14B` model, suggesting that multilingual training plays a critical role in multilingual discourse analysis. This further reinforces the importance of language coverage in training data, beyond simple model scaling, in achieving robust performance in discourse relation classification. Furthermore, despite being a fine-tuned version of `Llama2-7B`, `Emma500`'s multilingual training appears to give it an edge over the more recent and larger `Llama3-8B`. Similarly, `Aya-23-35B`'s probe surpasses those of the largest `Qwen2.5-70B` and `Llama3-72B` models, despite operating with only half the number of parameters, emphasizing the efficiency of its learned representations. Given this evident edge in the overall probing performance, we focus the rest of our investigation into discourse generalization on `Aya-23-35B`.

6.2 Language Generalization

We evaluate `Aya-23-35B`'s performance on the DISRPT test sets across 13 languages and compare different training conditions: (1) monolingual training (**MONO-PROBE**), (2) multilingual training with languages from the same language family (**MULTI-LANG-PROBE**), and (3) multilingual training using instances from all languages (**MULTI-ALL-PROBE**). Figure 4 shows the performance of the `Aya-23-35B` model trained and tested on all and subsets of the data given different training regimes.

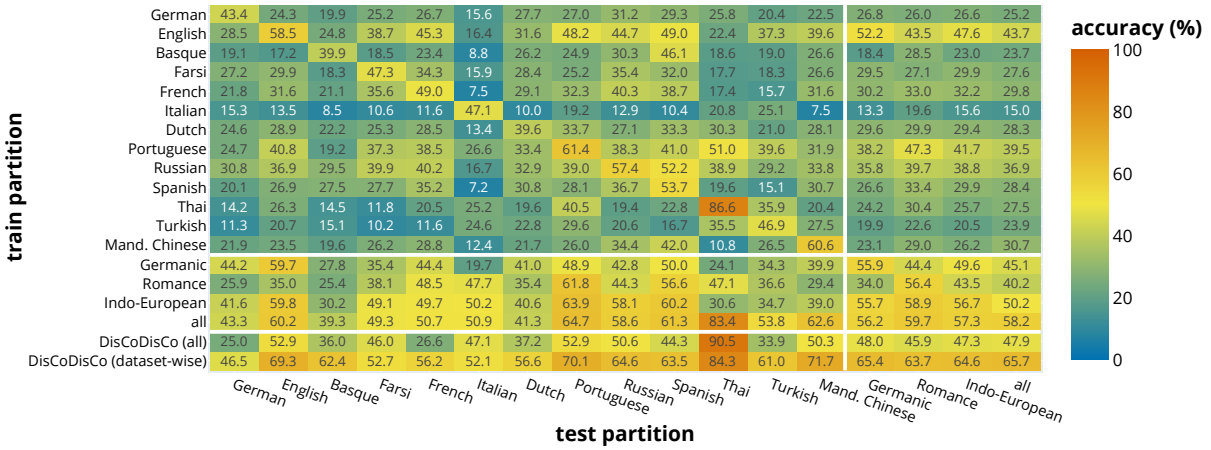
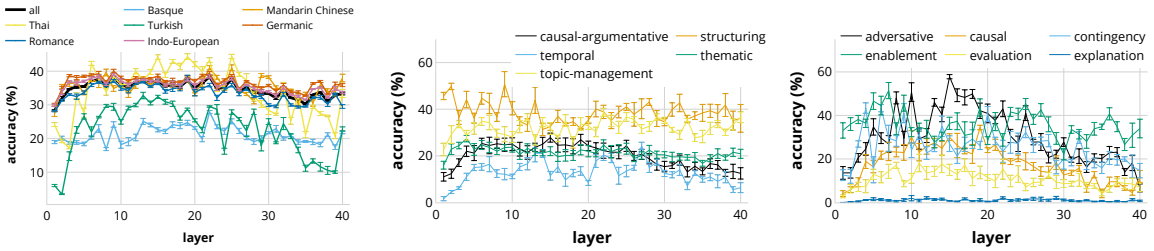


Figure 4: Mean accuracy over five runs of the `Aya-23-35B`-probe trained and tested on various partitions of DISRPT.

MULTI-ALL-PROBE & MONO-PROBE. The probes trained on the entire DISRPT dataset match or outperform the monolingually trained probes across most test partitions, including the combined test set with all languages. This suggests that multilingual training benefits discourse relation classification, leveraging shared discourse patterns across languages. This is the opposite to what we see in the reference system DisCoDisCo, where the dataset-specific models outperform the model trained on all the data by a margin of 17.8%. For Farsi and Russian, the best performance is observed in the MULTI-ALL-PROBE, with +2% and +1.2% over their respective MONO-PROBEs. The same generalization is observed for Turkish, Mandarin Chinese, and French, evidenced by an increase in accuracy of 6.9%, 2%, and 1.7% respectively.

There is one exception to this generalization trend: for the Thai dataset, the MONO-PROBE outperforms MULTI-ALL-PROBE with a margin of 3.2%. This is likely due to the fact that the Thai dataset is relatively large and contains only news data and was only annotated with explicit discourse relations (Prasertsom et al., 2024), which has been reported to be considerably easier than implicit discourse relations (Knaebel, 2021; Braud et al., 2023) due to the presence of connectives, which, while not unambiguous, help narrow down the likely senses of relations (Webber et al., 2019). The probes for Basque consistently underperform, with the MULTI-ALL-PROBE achieving the same performance as the Basque-only probe. This is likely due to the fact that Basque is considered a language isolate, a language that has no demonstrable genetic relationship with any other language (Campbell, 2010). Besides, most models do not include it in their list of supported languages (see Table 2 in Appendix C). The reference system DisCoDisCo also does not generalize well for Basque: its accuracy drops by 26.4% compared to the dataset-specific model. Moreover, the MONO-PROBEs for Basque and Turkish exhibit lower accuracy, which might mean that models trained on discourse tasks need more language-specific adaptations, particularly for morphologically rich and typologically different languages.

MULTI-LANG-PROBE. Training on discourse relation instances from related languages often leads to better generalization than training on a single language. For instance, the MULTI-LANG-PROBE for the Indo-European languages achieves reasonable performance across the Romance languages in DISRPT: there are improvements over the MONO-PROBE for Portuguese (+2.5%), Spanish (+6.5%), and Italian (+3.1%). For the Germanic and Romance language groups, we observe a similar generalization effect, though to a lesser extent. Notably, the Germanic MULTI-LANG-PROBE leads to the highest accuracy on the German test set (+0.8% over the MONO-PROBE). These are encouraging results as some of these languages do not have a large amount of instances in DISRPT: from the total number of data, Spanish covers 1.7%, German 1.19%, and Italian only 0.7% (see Table 1 in Appendix B). This suggests that leveraging discourse relation annotations from related language as well as the same underlying framework help with generalization. Adding to that, training on English only, which has the largest number of samples covering more than half of DISRPT, shows surprising generalization to other languages, with Portuguese and Spanish achieving an accuracy of 48.2% and 49% respectively, and French reaching 45.3%.



(a) languages and language groups (b) top-level classes (c) CAUSAL-ARGUMENTATIVE

Figure 5: Layer-wise probe performance by languages and relation classes. Mean accuracy over five runs.

6.3 Layer-wise Analysis

The layer-wise probes reveal additional insights into the structural processing of discourse information within the model. Firstly, Figure 5a shows the test set performance by language families. Here, we observe a trend by which the probes in lower to middle layers are the most performant as accuracy improves rapidly in the early layers and stabilizes around layer 10–15. In particular, for Turkish and Thai, the probe accuracy drops by almost 50% in higher layers. This effect is also observed for the Chinese probe but to a smaller extent. This shows that the discourse representations are best aligned across languages in the middle layers, which could be explained by the findings of Wendler et al. (2024) according to which multilingual computations are carried out in an English-aligned "concept space" in the intermediate layers. This result is also aligned with concurrent work by Skean et al. (2025) which finds that final-layer representations are consistently outperformed by intermediate representations across a range of tasks and architectures. Our work supplements this by extending their finding to linguistic tasks such as discourse relation classification.

Surprisingly, the final layer probes show another increase for Chinese, Turkish, and Thai, which is also reflected in the overall test set accuracy. This could indicate that their probes benefit from language-specific features, which might be more prevalent in layers close to the outputs if we assume a shared, compressed representation space in intermediate layers as proposed by Wendler et al. (2024) and Deletang et al. (2024).

For the MULTI-LANG-PROBEs of Romance, Germanic, and Indo-European, the corresponding performance is more constant, closely following the "all" curve (MULTI-ALL-PROBE). Here, the probes of the later layers show a drop of about 5-10% suggesting that their representations' alignment is more consistent throughout the model. This also indicates that both the higher and lower layers are involved in the discourse processing of LLMs.

Regarding relation classes, Figures 5b shows that STRUCTURING and TOPIC-MANAGEMENT achieve relatively higher accuracy across layers, peaking in the early-to-mid layers (around 5-15), whereas TEMPORAL and CAUSAL-ARGUMENTATIVE exhibit lower performance. Figure 5c zooms in on the more fine-grained classes: ADVERSATIVE and ENABLEMENT achieve the highest accuracy overall, with peaks around layers 10-20, while EVALUATION and EXPLANATION remain consistently low. Layer-wise plots for other relation classes are provided in Figure 8 in Appendix D. Overall, different relation classes appear to be encoded at varying layers, with some benefiting from middle-layer representations while others, particularly structure-oriented ones, maintain stable performance across layers. The drop in accuracy in later layers suggests that some discourse information might become less explicitly represented as the model progresses through deeper layers.

6.4 Error Analysis

To better understand the probe performance, we plot a confusion matrix in Figure 6 to study systematic errors. Overall, ELABORATION is the most frequent label. Here, our probe has a bias of confusing it with labels such as FRAMING and EXPLANATION. This suggests that the model struggles to differentiate between content expansion (ELABORATION) which sometimes overlaps semantically with EXPLANATION, and relations that involve additional information but with a primary focus on contextualization (FRAMING). In other words, ELABORATION is mainly used to give additional information about an entity or a proposition, while FRAMING provides information with a goal to increase

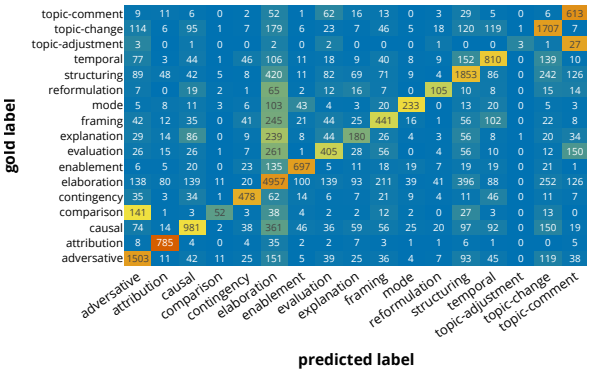


Figure 6: Confusion matrix of labels over MULTI-ALL-PROBE (colors normalized by row).

a reader's understanding of an entity or a proposition (Mann and Thompson, 1988; Carlson and Marcu, 2001). This corresponds to the distinction made by Hovy and Maier (1997): ELABORATION is considered an *ideational* class that is used to express states of affairs in the world, not including the interlocutors; on the other hand, FRAMING and EXPLANATION are considered *interpersonal* and are expressed by various perlocutionary acts to affect readers' attitude and beliefs. ELABORATION is also a highly prevalent relation in general, making it a probable default prediction in ambiguous cases.

In addition, a majority of the COMPARISON instances are predicted as ADVERSATIVE. A qualitative inspection reveals that this is due to some datasets not distinguishing similarities from differences, while the proposed unified label set does. We argue that similarity-based instances highlight commonalities between entities or situations and establish a shared property or behavior, reinforcing coherence by aligning elements. On the other hand, adversative-based cases emphasize differences, opposition, or unexpected alignments, which serve different pragmatic functions. Moreover, similarity-based comparisons often reinforce or extend prior discourse, leading to additive coherence; while adversative-based cases introduce shifts, which require the reader to re-evaluate assumptions and adjust interpretation. By distinguishing these two types, we think models can better capture rhetorical intent and argumentative structure.

7 Conclusion

Our study provides a comprehensive analysis of discourse generalization in LLMs, revealing important patterns in how these models encode discourse structures for cross-lingual transfer. To this end, we first present a unified discourse relation label set, which serves as a foundation for our probing experiments, allowing us to analyze discourse representations beyond individual frameworks. This is also the first work to apply a unified label set across frameworks and languages to multilingual discourse relation classification at scale.

Our probes exhibit generalization, whereby training across languages generally outperforms language–specific probes, which is not the case for the reference system DisCoDisCo. Through multi-faceted analyses, we find that model size alone does not lead to discourse probing success; instead, multilingual training, dataset composition, and language-specific factors play significant roles. While larger models generally perform better, discrepancies such as the under-performing `BLOOM` and the best-performing `Aya-23-35B`-probe emphasize the importance of training data quality and architectural optimizations. Surprisingly, we find that some of our probes generalize to languages unseen during probe training. For example, the English MONO-PROBE generalizes to Romance languages. Furthermore, our layer-wise analysis suggests that discourse representations are best aligned across languages in the intermediate layers, with later layers refining these representations for specific relation types. In addition, models struggle with relations requiring implicit reasoning such as EXPLANATION.

Overall, our findings highlight the interplay between multilinguality, scaling, and internal representations of LLMs for multilingual discourse processing and provide insights into cross-lingual alignment of discourse relations. Understanding which discourse relations are well-captured by LLMs and which are not could help improve discourse parsing models by highlighting gaps in current representation. In addition, our findings may be useful for designing better discourse-aware pretraining or fine-tuning strategies. Future work can improve cross-lingual alignment by refining representations of challenging relation classes. More generally, our method provides a systematic way to investigate discourse representation encoded in LLMs, making it a useful tool for answering linguistic questions that can be formulated into a consecutive span representation. By advancing our understanding of discourse generalization in LLMs, we contribute towards more interpretable and robust NLP systems capable of nuanced language comprehension.

Limitations

While this work offers a first step towards understanding discourse generalization in LLMs across languages and discourse annotation frameworks, we acknowledge the following limitations related to data, unified representation, and methodological approach.

Firstly, despite being multilingual, the DISRPT benchmark is imbalanced with regard to language coverage. English remains dominant (making up 53.5% of the DISRPT benchmark, as shown in Table 1), which may impact the generalizability of our findings to lower-resource languages. To maintain comparability with prior work, we did not stratify the training data based on language sample sizes. However, we mitigated this by training language-specific probes (MONO-PROBE) and MULTI-LANG-PROBEs based on language families and reporting their respective performance in Figure 4. Similarly, while DISRPT includes multiple domains and genres, some are underrepresented such as conversational data or dialogues. This results in a label imbalance such as TOPIC-ADJUSTMENT, limiting statistical robustness in such cases. However, this also highlights the need for the creation of more balanced multilingual discourse resources.

Secondly, our approach assumes a unified label set for discourse relations across languages and frameworks, which we present in §3. While this enables cross-linguistic and cross-framework discourse analysis, compromises were necessary, which to some extent simplify the complexity and granularity of discourse relations assumed by frameworks such as PDTB and RST. Fully standardizing and harmonizing discourse relations is inherently challenging, and finding a good trade-off between maintaining theoretical assumptions and ensuring practical applicability is crucial yet complex, as evidenced in previous mapping proposals and efforts (Benamara and Taboada, 2015; Demberg et al., 2019). In addition, segmentation differences exist across frameworks, which can have an impact on the performance of our probes for certain relations such as ATTRIBUTION. This work should thus be interpreted with an awareness of the theoretical and practical difficulties in creating an informed and unified taxonomy suitable for both theoretical studies and computational research. This work should also be viewed as facilitating the development of a unified discourse relation representation for computational discourse modeling such as the effort and initiatives that have been put forward for dependency parsing (Universal Dependencies, de Marneffe et al. 2021), empirical study of anaphora (Universal Anaphora, Poesio et al. 2024), and semantic parsing (Uniform Meaning Representation, Bonn et al. 2024).

Lastly, rather than using highly optimized architectures, we employed relatively simple probing methods, which aligns with our interest in assessing the intrinsic capabilities of LLMs for discourse processing. While achieving state-of-the-art performance was not our primary goal, better performance could likely be achieved by fine-tuning the LLMs. Future work would need to assess the trade-off between generalization and task-specific optimization.

Acknowledgments

We would like to thank the members of the MaiNLP lab for their valuable feedback, especially to Philipp Mondorf, Siyao Peng, and Shijia Zhou. We would also like to express our gratitude to Amir Zeldes for the constructive and valuable feedback. We thank the reviewers and the meta-reviewer for their feedback and suggestions. Lastly, we recognize the support for Yang Janet Liu and Barbara Plank through the ERC Consolidator Grant DIALECT 101043235.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. Phi-4 Technical Report. *Preprint*, arXiv:2412.08905.

Griffin Adams, Alex Fabbri, Faisal Ladhak, Noémie Elhadad, and Kathleen McKeown. 2023. Generating EDU extracts for plan-guided summary re-ranking. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2680–2697, Toronto, Canada. Association for Computational Linguistics.

Guillaume Alain and Yoshua Bengio. 2018. Understanding intermediate layers using linear classifier probes. *Preprint*, arXiv:1610.01644.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick

Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, and 2 others. 2024. Aya 23: Open Weight Releases to Further Multilingual Progress. *Preprint*, arXiv:2405.15032.

Nicholas Asher, Julie Hunter, Mathieu Morey, Benamara Farah, and Stergos Afantenos. 2016. Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2721–2727, Portorož, Slovenia. European Language Resources Association (ELRA).

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press, Cambridge.

Farah Benamara and Maite Taboada. 2015. Mapping different rhetorical relation annotations: A proposal. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 147–152, Denver, Colorado. Association for Computational Linguistics.

Julia Bonn, Matthew J. Buchholz, Jayeol Chun, Andrew Cowell, William Croft, Lukas Denk, Sijia Ge, Jan Hajič, Kenneth Lai, James H. Martin, Skatje Myers, Alexis Palmer, Martha Palmer, Claire Benet Post, James Pustejovsky, Kristine Stenzel, Haibo Sun, Zdeňka Urešová, Rosa Vallejos, and 4 others. 2024. Building a broad infrastructure for uniform meaning representations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2537–2547, Torino, Italia. ELRA and ICCL.

Chloé Braud, Maximin Coavoux, and Anders Søgaard. 2017. Cross-lingual RST discourse parsing. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 292–304, Valencia, Spain. Association for Computational Linguistics.

Chloé Braud, Yang Janet Liu, Eleni Metheniti, Philippe Muller, Laura Rivière, Attapol Rutherford, and Amir Zeldes. 2023. The DISRPT 2023 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 1–21, Toronto, Canada. The Association for Computational Linguistics.

Chloé Braud, Barbara Plank, and Anders Søgaard. 2016. Multi-view and multi-task training of RST discourse parsers. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1903–1913, Osaka, Japan. The COLING 2016 Organizing Committee.

Chloé Braud, Amir Zeldes, Laura Rivière, Yang Janet Liu, Philippe Muller, Damien Sileo, and Tatsuya Aoyama. 2024. DISRPT: A multilingual, multi-domain, cross-framework benchmark for discourse processing. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4990–5005, Torino, Italia. ELRA and ICCL.

Jannik Brinkmann, Chris Wendler, Christian Bartelt, and Aaron Mueller. 2025. Large language models share representations of latent grammatical concepts across typologically diverse languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6131–6150, Albuquerque, New Mexico. Association for Computational Linguistics.

Harry Bunt and Rashmi Prasad. 2016. ISO DR-Core (ISO 24617-8): Core Concepts for the Annotation of Discourse Relations. In *Annual Meeting of the Association for Computational Linguistics*.

Lyle Campbell. 2010. Language isolates and their history, or, what's weird, anyway? In *Annual Meeting of the Berkeley Linguistics Society*, pages 16–31.

Lynn Carlson and Daniel Marcu. 2001. Discourse tagging reference manual. Technical Report ISI-TR-545, University of Southern California Information Sciences Institute.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In *Current and New Directions in Discourse and Dialogue*, Text, Speech and Language Technology 22, pages 85–112. Kluwer, Dordrecht.

Joyce Y. Chai and Rong Jin. 2004. Discourse structure for context question answering. In *Proceedings of the Workshop on Pragmatics of Question Answering at HLT-NAACL 2004*, pages 23–30, Boston, Massachusetts, USA. Association for Computational Linguistics.

Christian Chiarcos. 2014. Towards interoperable discourse annotation. discourse features in the ontologies of linguistic annotation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4569–4577, Reykjavik, Iceland. European Language Resources Association (ELRA).

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

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*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

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*, pages 344–352, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Parag Pravin Dakle, SaiKrishna Rallabandi, and Preethi Raghavan. 2023. Understanding BLOOM: An empirical study on diverse NLP tasks. *Preprint*, arXiv:2211.14865.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya Expanse: Combining Research Breakthroughs for a New Multilingual Frontier. *Preprint*, arXiv:2412.04261.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics*, 47(2):255–308.

Gregoire Deletang, Anian Ruoss, Paul-Ambroise Duquenne, Elliot Catt, Tim Genewein, Christopher Mattern, Jordi Grau-Moya, Li Kevin Wenliang, Matthew Aitchison, Laurent Orseau, Marcus Hutter, and Joel Veness. 2024. Language modeling is compression. In *The Twelfth International Conference on Learning Representations*.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online (v2020.4)*. Zenodo.

Greg Durrett, Taylor Berg-Kirkpatrick, and Dan Klein. 2016. Learning-based single-document summarization with compression and anaphoricity constraints. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1998–2008, Berlin, Germany. Association for Computational Linguistics.

Yujian Gan, Massimo Poesio, and Juntao Yu. 2024. Assessing the capabilities of large language models in coreference: An evaluation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1645–1665, Torino, Italia. ELRA and ICCL.

Luke Gessler, Shabnam Behzad, Yang Janet Liu, Siyao Peng, Yilun Zhu, and Amir Zeldes. 2021. DisCoDisCo at the DISRPT2021 shared task: A system for discourse segmentation, classification, and connective detection. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 51–62, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

John T. Hale and Miloš Stanojević. 2024. Do LLMs learn a true syntactic universal? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17106–17119, Miami, Florida, USA. Association for Computational Linguistics.

Benjamin Heinzerling and Michael Strube. 2019. Sequence tagging with contextual and non-contextual subword representations: A multilingual evaluation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 273–291, Florence, Italy. Association for Computational Linguistics.

Eduard H Hovy and Elisabeth Maier. 1997. Parsimonious or Profligate: How Many and Which Discourse Structure Relations. *Discourse Processes*.

Patrick Huber and Giuseppe Carenini. 2019. Predicting discourse structure using distant supervision from sentiment. 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 2306–2316, Hong Kong, China. Association for Computational Linguistics.

Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, Dennis Ulmer, Florian Schottmann, Khuyagbaatar Batsuren, Kaiser Sun, Koustuv Sinha, Leila Khalatbari, Maria Ryskina, Rita Frieske, Ryan Cotterell, and Zhijing Jin. 2023. A taxonomy and review of generalization research in NLP. *Nature Machine Intelligence*, 5(10):1161–1174.

Shaoxiong Ji, Zihao Li, Indraneil Paul, Jaakko Paavola, Peiqin Lin, Pinzhen Chen, Dayyán O'Brien, Hengyu Luo, Hinrich Schütze, Jörg Tiedemann, and Barry Haddow. 2024. EMMA-500: Enhancing Massively Multilingual Adaptation of Large Language Models. *Preprint*, arXiv:2409.17892.

Jaap Jumelet, Milica Denic, Jakub Szymanik, Dieuwke Hupkes, and Shane Steinert-Threlkeld. 2021. Language models use monotonicity to assess NPI licensing. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4958–4969, Online. Association for Computational Linguistics.

Daniel Jurafsky and James H. Martin. 2025. Discourse coherence. In *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd edition, chapter 24, pages 531–552. Online manuscript released January 12, 2025.

Najoung Kim and Sebastian Schuster. 2023. Entity tracking in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3835–3855, Toronto, Canada. Association for Computational Linguistics.

René Knaebel. 2021. discopy: A neural system for shallow discourse parsing. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 128–133, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. Discourse probing of pretrained language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3849–3864, Online. Association for Computational Linguistics.

Murathan Kurfalı and Robert Östling. 2021. Probing multilingual language models for discourse. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 8–19, Online. Association for Computational Linguistics.

Jiaqi Li, Ming Liu, Min-Yen Kan, Zihao Zheng, Zekun Wang, Wenqiang Lei, Ting Liu, and Bing Qin. 2020. Molweni: A challenge multiparty dialogues-based machine reading comprehension dataset with discourse structure. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2642–2652, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Sujian Li, Liang Wang, Ziqiang Cao, and Wenjie Li. 2014. Text-level discourse dependency parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25–35, Baltimore, Maryland. Association for Computational Linguistics.

Yang Janet Liu and Amir Zeldes. 2023. Why can't discourse parsing generalize? a thorough investigation of the impact of data diversity. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3112–3130, Dubrovnik, Croatia. Association for Computational Linguistics.

Zhengyuan Liu, Ke Shi, and Nancy Chen. 2021. DMRST: A joint framework for document-level multilingual RST discourse segmentation and parsing. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 154–164, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

William C Mann and Sandra A Thompson. 1988. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Yisong Miao, Hongfu Liu, Wenqiang Lei, Nancy Chen, and Min-Yen Kan. 2024. Discursive socratic questioning: Evaluating the faithfulness of language models' understanding of discourse relations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6277–6295, Bangkok, Thailand. Association for Computational Linguistics.

AI Team Mistral. 2025. Mistral Small 3. https://mistral.ai/en/news/mistral-small-3.

Mathieu Morey, Philippe Muller, and Nicholas Asher. 2018. A dependency perspective on RST discourse parsing and evaluation. *Computational Linguistics*, 44(2):197–235.

Khalil Mrini, Franck Dernoncourt, Quan Hung Tran, Trung Bui, Walter Chang, and Ndapa Nakashole. 2020. Rethinking self-attention: Towards interpretability in neural parsing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 731–742, Online. Association for Computational Linguistics.

Philippe Muller, Stergos Afantenos, Pascal Denis, and Nicholas Asher. 2012. Constrained decoding for text-level discourse parsing. In *Proceedings of COLING 2012*, pages 1883–1900, Mumbai, India. The COLING 2012 Organizing Committee.

Karthik Narasimhan and Regina Barzilay. 2015. Machine comprehension with discourse relations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1253–1262, Beijing, China. Association for Computational Linguistics.

Qiwei Peng and Anders Søgaard. 2024. Concept space alignment in multilingual LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5511–5526, Miami, Florida, USA. Association for Computational Linguistics.

Tiago Pimentel, Maria Ryskina, Sabrina J. Mielke, Shijie Wu, Eleanor Chodroff, Brian Leonard, Garrett Nicolai, Yustinus Ghanggo Ate, Salam Khalifa, Nizar Habash, Charbel El-Khaissi, Omer Goldman, Michael Gasser, William Lane, Matt Coler, Arturo Oncevay, Jaime Rafael Montoya Samame, Gema Celeste Silva Villegas, Adam Ek, and 39 others. 2021. SIGMORPHON 2021 shared task on morphological reinflection: Generalization across languages. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–259, Online. Association for Computational Linguistics.

Massimo Poesio, Maciej Ogrodniczuk, Vincent Ng, Sameer Pradhan, Juntao Yu, Nafise Sadat Moosavi, Silviu Paun, Amir Zeldes, Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtský, and Daniel Zeman. 2024. Universal anaphora: The first three years. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17087–17100, Torino, Italia. ELRA and ICCL.

Ponrawee Prasertsom, Apiwat Jaroonpol, and Attapol T. Rutherford. 2024. The Thai Discourse Treebank: Annotating and Classifying Thai Discourse Connectives. *Transactions of the Association for Computational Linguistics*, 12:613–629.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. Qwen2.5 Technical Report. *Preprint*, arXiv:2412.15115.

Ines Rehbein, Merel Scholman, and Vera Demberg. 2016. Annotating discourse relations in spoken language: A comparison of the PDTB and CCR frameworks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1039–1046, Portorož, Slovenia. European Language Resources Association (ELRA).

Ted J. M. Sanders, Wilbert P.M. Spooren, and Leo G.M. Noordman. 1992. Towards a taxonomy of coherence relations. *Discourse Processes*, 15:1–35.

Ted J.M. Sanders, Vera Demberg, Jet Hoek, Merel C.J. Scholman, Fatemeh Torabi Asr, Sandrine Zufferey, and Jacqueline Evers-Vermeul. 2021. Unifying dimensions in coherence relations: How various annotation frameworks are related. *Corpus Linguistics and Linguistic Theory*, 17(1):1–71.

Karol Saputa, Angelika Peljak-Łapińska, and Maciej Ogrodniczuk. 2024. Polish coreference corpus as an LLM testbed: Evaluating coreference resolution within instruction-following language models by instruction–answer alignment. In *Proceedings of The Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 23–32, Miami. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, and 373 others. 2023. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *Preprint*, arXiv:2211.05100.

Deborah Schiffrin, Deborah Tannen, and Heidi E. Hamilton. 2015. Introduction to the first edition. In *The Handbook of Discourse Analysis*, chapter 00, pages 1–7. John Wiley & Sons, Ltd.

Wesley Scivetti, Melissa Torgbi, Austin Blodgett, Mollie Shichman, Taylor Hudson, Claire Bonial, and Harish Tayyar Madabushi. 2025. Assessing language comprehension in large language models using construction grammar. *Preprint*, arXiv:2501.04661.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. 2025. Layer by Layer: Uncovering Hidden Representations in Language Models. *CoRR*, abs/2502.02013.

Manfred Stede. 2011. Discourse Processing. *Synthesis Lectures on Human Language Technologies*, 4(3):1–165.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Cagri Toraman, Eyup Halit Yilmaz, Furkan Şahinuc, and Oguzhan Ozcelik. 2023. Impact of tokenization on language models: An analysis for turkish. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(4).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Bonnie Webber, Rashmi Prasad, and Aravind Josh. 2024. Reflections on the Penn Discourse TreeBank and its relatives. *Computational Linguistics*.

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

Leonie Weissweiler, Valentin Hofmann, Abdullatif Köksal, and Hinrich Schütze. 2022. The better your syntax, the better your semantics? probing pretrained language models for the English comparative correlative. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10859–10882, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do Llamas Work in English? On the Latent Language of Multilingual Transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.

Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Discourse-aware neural extractive text summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5021–5031, Online. Association for Computational Linguistics.

Amir Zeldes, Tatsuya Aoyama, Yang Janet Liu, Siyao Peng, Debopam Das, and Luke Gessler. 2025. eRST: A Signaled Graph Theory of Discourse Relations and Organization. *Computational Linguistics*, 51(1):23–72.

Amir Zeldes, Yang Janet Liu, Mikel Iruskieta, Philippe Muller, Chloé Braud, and Sonia Badene. 2021. The DISRPT 2021 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 1–12, Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Unified Label Set: Definitions and Examples

The proposed unified label set adapts the mapping proposal described in Benamara and Taboada (2015) and extends it to be applicable to phenomena frequent in dialogues. In total, there are 17 labels corresponding to the core discourse relations identified in Bunt and Prasad (2016). Below we provide definitions and list typical framework-specific discourse relations that are mapped to the unified label set.

TEMPORAL is used to map to framework-specific labels that establish a chronological sequence between events or states. RST's SEQUENCE, PDTB's TEMPORAL.ASYNCHRONOUS/SYNCHRONOUS, and SDRT's TEMPLOC and FLASHBACK are mapped to this class.

STRUCTURING is mapped to RST-style relations such as LIST, PREPARATION, DISJUNCTION, ORGANIZATION-HEADING, JOINT, and TEXTUAL-ORGANIZATION. SDRT-style relations such as ALTERNATION, CONTINUATION, and PARALLEL are mapped to this class. PDTB's EXPANSION.DISJUNCTION is mapped to this class.

ATTRIBUTION informs about the source of information, which is useful and crucial for many real-world applications such as misinformation detection and fact-checking. It is considered a discourse relation by RST, SDRT, and DEP. PDTB's attribution annotation is considered a separate discourse annotation type and is not included in DISRPT. It is worth pointing out that ATTRIBUTION is not considered a core discourse relation by Bunt and Prasad (2016), but it is one of the most frequent discourse relations in most frameworks.

COMPARISON is used to group fine-grained relations that highlight similarities rather than differences between spans or entities such as PDTB's COMPARISON.SIMILARITY and ANALOGY in RST-style corpora.

ELABORATION provides additional information about an entity or a proposition. Framework-specific relations that contain ELABORATION or are prefixed with ELAB (e.g. ELAB-PROCESS_STEP, ELAB-ENUMEMBER, and Q_ELAB) are all mapped to this class. PROGRESSION used in DEP-style corpora is also mapped to this class.

FRAMING is used for framework-specific relations that provide a framework for understanding the content of the situation described in the discourse units (Benamara and Taboada, 2015) such as FRAME, BACKGROUND, and CIRCUMSTANCE

MODE is used to supply information about *how* events happens. Commonly mapped fine-grained relations are MANNER and MEANS in RST-style corpora as well as PDTB's EXPANSION.MANNER.

REFORMULATION corresponds to relations by which one discourse unit re-expresses the meaning of another in a different form and ensure coherence by providing alternative expressions of the same idea. SUMMARY and RESTATEMENT in RST-style corpora and PDTB's EXPANSION.EQUIVALENCE are mapped to REFORMULATION.

ADVERSATIVE highlights incompatibility and covers commonly used discourse relations in all frameworks such as CONCESSION and CONTRAST. PDTB's EXPANSION.EXCEPTION/SUBSTITUTION are also mapped to this subclass given their definitions in Webber et al. (2019).

CAUSAL is used to indicate a cause-and-effect relationship. Fine-grained relations that signal that one event, state, or proposition (the cause) leads to or explains another event, state, or proposition (the effect) is mapped to this class. This is one of the most core discourse relation types recognized in all frameworks.

CONTINGENCY is used to map condition-based relations such as CONDITIONAL, UNLESS, UNCONDITIONAL, and CONTINGENCY.NEGATIVE-CONDITION.

ENABLEMENT is used to connect discourse units where one enables the other. Framework-specific relations such as GOAL and PURPOSE are mapped to this class.

EXPLANATION is used when the situation described by one argument provides the reason, explanation, or justification for the situation described by the other (Webber et al., 2019).

EVALUATION is used where one discourse unit provides an assessment, judgment, or commentary on the content of another unit. Framework-dependent relations such as COMMENT and INTERPRETATION are mapped to this class.

TOPIC-CHANGE involves a shift or drift in topic that links large textual units (Carlson and Marcu, 2001). This is used for fine-grained relations that connect multiple, non-contrasting discourse units that are of equal prominence such as JOINT-OTHER in the `eng.rst.gum` corpus and PDTB's EXPANSION.CONJUNCTION.

TOPIC-COMMENT is used for framework-specific relatiosn that involve question-answer pairs or problem-solution pairs, commonly in RST-style and SDRT-style corpora. PDTB's HYPOPHORA is also mapped to this class.

TOPIC-ADJUSTMENT is primarily used for cases where a discourse unit modifies, redirects, or adjusts the ongoing topic of discussion such as CORRECTION and INTERRUPTED, which signal deviations from the expected discourse progression.

B Data

Table 1 provides an overview of the DISRPT benchmark we use for our experiments. The data produced and used in this paper is in accordance with the original licenses of the underlying resources, as specified in the repository of DISRPT.[3]

dataset	language	language family	# of relation instances	% of total instances
deu.rst.pcc	German	Indo-European, Germanic	2665	1.19%
eng.dep.covdtb	English		4985	2.22%
eng.dep.scidtb			9904	4.42%
eng.pdtb.pdtb			47851	21.34%
eng.pdtb.tedm			529	0.24%
eng.rst.gum			24688	11.01%
eng.rst.rstdt			19778	8.82%
eng.sdrt.stac			12235	5.46%
eus.rst.ert	Basque	Language Isolate	3825	1.71%
fas.rst.prstc	Farsi	Indo-European, Iranian	5191	2.31%
fra.sdrt.annodis	French	Indo-European, Romance	3338	1.49%
ita.pdtb.luna	Italian		1544	0.69%
nld.rst.nldt	Dutch	Indo-European, Germanic	2264	1.01%
por.pdtb.crpc	Portuguese	Indo-European, Romance	11330	5.05%
por.pdtb.tedm			554	0.25%
por.rst.cstn			4993	2.23%
rus.rst.rrt	Russian	Indo-European, Slavic	34566	15.41%
spa.rst.rststb	Spanish	Indo-European, Romance	3049	1.36%
spa.rst.sctb			692	0.31%
tha.pdtb.tdtb	Thai	Tai-Kadai, Kam-Tai	10865	4.84%
tur.pdtb.tdb	Turkish	Altaic, Turkic	3185	1.42%
tur.pdtb.tedm			577	0.26%
zho.dep.scidtb	Mandarin	Sino-Tibetan, Chinese	1298	0.58%
zho.pdtb.cdtb			5270	2.35%
zho.rst.gcdt			8413	3.75%
zho.rst.sctb			692	0.31%

Table 1: Overview of datasets in DISRPT 2023 for the discourse relation classification task.

C Models and Probe Training

LLMs. Table 2 provides an overview of the examined LLMs in this paper, along with their number of parameters, multilingual capabilities, and the proportion of languages in DISRPT covered by the advertised supported languages.

Hyperparameters. We train our probes using AdamW (Loshchilov and Hutter, 2018) with a batch size of 64, learning rate of 0.0001, and weight decay of 0.0001. To mitigate the class imbalance that is inherent to discourse relation data, we use a weighted cross entropy loss, where each class' loss is weighted by the inverse square root of the number of samples in the respective class. For the hidden layer, we choose a dimension of D=512. For the input as well as the hidden layer, we regularize adding a Dropout of 0.2. Furthermore, we use layer normalization in the hidden layer. For the probes over all model attention scores, depending on the number of samples contained in the train dataset, we train for 60 epochs and increase that number to ensure at least 10000 gradient update steps on smaller datasets. For the layer-wise probes, we reduce the number of epochs to 20 as the probes converge much faster on these smaller representations. Because token-length of encoded sequence quadratically scales the GPU memory required to process the attention matrix, we had to lower the maximum window length for the larger models. Namely, we reduced the maximum window sizes from N_{max}=4000 to N_{max}=3800 for `Aya-Expanse-32B` and to N_{max}=3400 for `Aya-23-35B`, `Llama3-70B`, and `Qwen2.5-72B`.

Compute. For each model, we compute one pass over the dataset computing all the attention representations which we cache for the probing experiments. For the smaller models, a cluster equipped with eight Nvidia A100 GPUs was used for around 80 hours. For the large models of size 70B and 72B, we used a cluster equipped with four Nvidia H200 GPUs for about 30 hours.

Code License. As specified in the code repository,[4] we release our code under MIT license.

[3] https://github.com/disrpt/sharedtask2023

[4] https://github.com/mainlp/discourse_probes

model name	model family	# of params	# languages supported	deu	eng	eus	fas	fra	nld	por	rus	spa	tur	zho	tha	ita	fraction supported
Qwen2.5-0.5B	Qwen 2.5	0.49B	29	✓	✓	✗	✓	✓	✗	✓	✓	✓	✗	✓	✓	✓	0.769
Qwen2.5-1.5B		1.54B		✓	✓	✗	✓	✓	✗	✓	✓	✓	✗	✓	✓	✓	
Qwen2.5-3B		2.77B		✓	✓	✗	✓	✓	✗	✓	✓	✓	✗	✓	✓	✓	
Qwen2.5-7B		7.61B		✓	✓	✗	✓	✓	✗	✓	✓	✓	✗	✓	✓	✓	
Qwen2.5-14B		14.7B		✓	✓	✗	✓	✓	✗	✓	✓	✓	✗	✓	✓	✓	
Qwen2.5-32B		32.5B		✓	✓	✗	✓	✓	✗	✓	✓	✓	✗	✓	✓	✓	
Qwen2.5-72B		72.7B		✓	✓	✗	✓	✓	✗	✓	✓	✓	✗	✓	✓	✓	
Llama-3.2-1B	Llama 3	1.23B	8	✓	✓	✗	✗	✓	✗	✓	✗	✓	✗	✗	✓	✓	0.538
Llama-3.2-3B		3.21B		✓	✓	✗	✗	✓	✗	✓	✗	✓	✗	✗	✓	✓	
Llama-3.1-8B		8.03B		✓	✓	✗	✗	✓	✗	✓	✗	✓	✗	✗	✓	✓	
Llama-3.1-70B		70.6B		✓	✓	✗	✗	✓	✗	✓	✗	✓	✗	✗	✓	✓	
bloom-560m	Bloom	0.56B	46	✗	✓	✓	✓	✓	✗	✓	✗	✓	✗	✓	✗	✗	
bloom-1b1		1.07B		✗	✓	✓	✓	✓	✗	✓	✗	✓	✗	✓	✗	✗	
bloom-3b		3B		✗	✓	✓	✓	✓	✗	✓	✗	✓	✗	✓	✗	✗	
bloom-7b1		7.07B		✗	✓	✓	✓	✓	✗	✓	✗	✓	✗	✓	✗	✗	
bloomz-7b1		7.07B		✗	✓	✓	✓	✓	✗	✓	✗	✓	✗	✓	✗	✗	
aya-expanse-8b	Aya	8.03B	23	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	0.846
aya-expanse-32b		32.3B		✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	
aya-23-8B		8.03B		✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	
aya-23-35B		35B		✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	
phi-4	Phi	14.7B	1	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	0.077
emma-500-llama2-7b	Emma	6.74B	546	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	1
Mistral-Small-24B-Base-2501	Mistral	23.6B	10	✓	✓	✗	✗	✓	✗	✓	✓	✓	✗	✓	✗	✓	0.615

Table 2: Overview of the included LLMs and their multilingual capabilities for languages in DISRPT 2023.

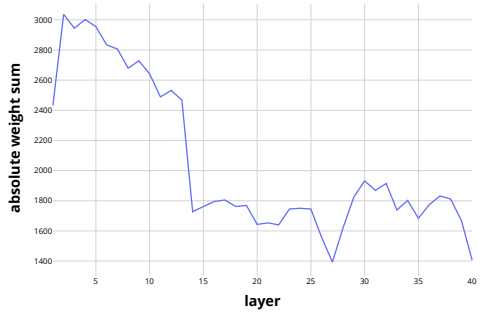


Figure 7: Sum of the absolute dot product between the weight matrices of the probe by layer.

D Additional Results and Analysis

Table 3 provides dataset accuracy scores for all examined LLM probes as well as the reference system DisCoDisCo's performance.

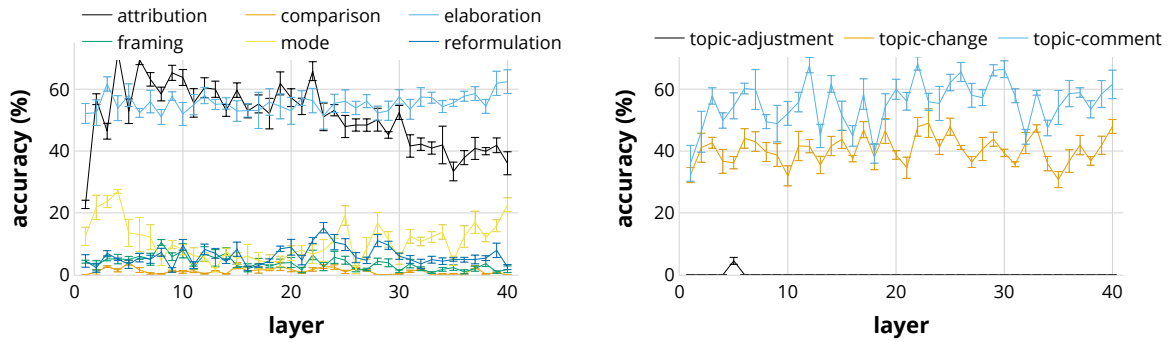
Figure 7 shows the layer-wise sums of the dot-products of the weight matrices. The magnitude of these scores can be interpreted as a feature importance and confirm that earlier layers play a crucial role in predictions, although the higher layers also show higher scores from layers 30 to 38, indicating a progressive refinement of discourse representations.

E Use of AI Assistants

The implementation of this work has been written with the support of code completions of an AI coding assistant, namely GitHub Copilot. Completions were mostly single lines up to a few lines of code and were always checked carefully to ensure their functionality and safety. Furthermore, we did our best to avoid accepting code completions that would be incompatible with the license of our code or could be regarded as plagiarism. We also include this statement in the `README.md` of the codebase.

dataset	Llama-3.1-70B	Llama-3.1-8B	Llama-3.2-1B	Llama-3.2-3B	Mistral-Small-24B-Base-2501	Qwen2.5-0.5B	Qwen2.5-1.5B	Qwen2.5-14B	Qwen2.5-32B	Qwen2.5-3B	Qwen2.5-72B	Qwen2.5-7B	aya-23-35B	aya-23-8B	aya-expanse-32b	aya-expanse-8b	bloom-1b1	bloom-3b	bloom-560m	bloom-7b1	bloomz-7b1	emma-500-llama2-7b	phi-4	DisCoDisCo (all)	DisCoDisCo
deu.rst.pcc	41.1	35.9	31.1	33.5	39.6	24.8	24.2	36.3	37.5	26.6	41.4	30.5	43.3	39.4	38.3	38.8	23.2	27.8	19.4	22.5	23.1	36.2	35.2	25	46.5
eng.dep.covdtb	61.8	56.4	54.7	56.9	57.9	51.1	49.2	60.9	61.6	51.9	62.9	55.6	65.9	58.5	62.5	60.5	49.9	53.8	50.2	53.8	54.7	56.6	58.2	59.9	74.1
eng.dep.scidtb	71.8	63.2	59.5	62.7	64.0	54.4	54.4	66.2	66.8	57.8	67.3	60.5	72.4	63.9	69.5	66.3	54.5	59.1	55.0	58.8	58.9	64.2	63.9	70.7	82.7
eng.pdtb.pdtb	67.4	61.6	56.7	60.6	63.8	51.3	51.7	65.6	67.1	58.5	68.4	61.5	68.8	65.3	67.9	65.3	49.4	56.2	49.4	57.1	56.6	62.0	63.4	59.1	79.9
eng.pdtb.tedm	48.9	53.2	49.0	47.1	52.5	43.8	45.0	52.3	55.7	49.6	53.0	52.6	51.9	58.2	54.6	56.9	40.1	46.4	45.6	47.0	46.4	53.3	52.5	40.7	58.1
eng.rst.gum	54.1	47.4	43.9	45.9	50.5	40.9	40.8	54.4	54.5	45.4	56.3	48.5	55.6	49.7	55.5	52.8	40.9	45.6	40.5	46.5	45.8	49.3	50.2	44.7	64
eng.rst.rstdt	56.0	51.7	51.2	51.7	54.7	48.6	47.6	57.6	57.5	51.1	56.5	53.7	57.6	53.3	55.9	55.2	48.1	51.7	46.7	51.0	50.5	50.1	55.1	49.7	66.1
eng.sdrt.stac	48.0	44.0	41.6	43.9	44.7	40.1	41.7	46.3	46.9	42.7	48.6	42.0	49.1	45.4	47.7	45.2	39.9	41.5	38.7	41.6	41.6	43.1	45.6	45.4	60.4
eus.rst.ert	40.4	29.1	27.4	31.5	35.7	22.9	20.7	36.6	36.8	24.1	38.4	29.6	39.3	30.0	36.5	32.1	27.2	29.4	25.7	30.2	29.4	34.2	34.2	36	62.4
fas.rst.prstc	49.2	43.8	43.1	46.9	47.1	36.7	39.6	49.4	49.0	43.1	51.9	46.8	49.3	48.9	49.5	48.2	37.7	40.2	33.6	40.4	38.6	47.4	46.1	46	52.7
fra.sdrt.annodis	48.4	43.3	41.1	42.5	45.7	41.5	41.9	53.4	52.9	40.9	51.2	47.1	50.7	44.3	47.6	47.6	37.8	42.4	39.3	43.1	42.8	43.2	49.2	26.6	56.2
ita.pdtb.luna	45.7	36.1	32.7	35.4	40.2	18.2	21.1	39.8	41.7	28.7	43.7	34.6	50.9	39.8	44.5	43.9	12.9	23.4	16.6	26.7	24.2	39.9	35.0	47.1	52.1
nld.rst.nldt	40.0	31.9	34.9	33.5	34.5	32.8	33.5	40.7	42.4	33.4	44.9	35.4	41.3	34.9	40.6	35.0	27.4	30.1	25.0	31.3	31.5	34.5	38.0	37.2	56.6
por.pdtb.crpc	65.9	58.8	54.7	57.5	61.5	52.4	50.9	64.9	66.2	54.9	68.0	60.1	67.4	60.7	64.4	63.7	47.0	54.0	50.0	53.6	53.3	63.0	63.8	55.9	75.4
por.pdtb.tedm	57.1	51.6	43.0	47.4	49.3	39.3	41.1	55.2	57.0	45.1	58.7	49.6	61.4	53.2	57.4	55.4	39.7	43.9	39.8	46.7	45.7	50.4	54.1	49.2	66.2
por.rst.cstn	63.9	57.9	55.6	57.8	59.7	55.6	52.9	62.4	61.9	54.8	62.5	57.2	65.3	58.9	62.2	59.9	49.0	55.2	49.1	56.5	54.8	57.2	59.0	53.7	68.8
rus.rst.rrt	56.0	52.4	50.4	51.3	53.9	47.6	47.1	57.2	57.9	50.6	59.2	53.7	58.6	55.7	58.5	55.3	41.8	45.1	42.9	45.2	42.3	54.6	54.5	50.6	64.7
spa.rst.rststb	48.4	40.0	39.7	39.5	46.9	35.2	34.9	44.6	47.7	39.2	47.6	43.7	52.9	47.5	50.8	48.4	31.7	38.0	31.4	40.4	38.7	46.4	40.9	42	61
spa.rst.sctb	62.9	66.9	63.5	63.5	65.0	60.4	56.2	60.9	61.9	62.0	68.3	63.1	69.7	64.2	68.9	66.3	51.2	55.7	50.6	55.5	56.0	64.3	59.6	46.5	66
tha.pdtb.tdtb	85.2	76.3	65.0	72.1	77.6	55.2	56.7	81.5	82.1	69.3	83.4	72.7	83.4	75.7	81.9	76.4	27.3	37.1	25.1	39.0	34.4	77.6	78.4	90.6	84.3
tur.pdtb.tdb	52.8	47.7	42.8	44.5	48.3	39.2	38.0	54.2	56.5	43.6	56.5	48.5	57.9	51.7	56.3	49.4	33.2	34.8	30.5	36.1	32.6	50.8	50.2	40.3	68.7
tur.pdtb.tedm	48.2	43.2	39.2	43.0	48.1	32.9	32.2	49.6	50.2	35.9	52.6	43.4	49.7	44.2	50.2	46.3	17.9	24.2	20.5	23.8	23.7	48.1	42.7	27.5	53.3
zho.dep.scidtb	60.7	48.3	36.3	51.6	48.4	32.8	30.8	49.8	53.9	39.4	52.8	42.0	61.6	48.0	56.8	53.8	33.9	37.4	32.9	43.3	42.0	46.7	47.3	53.5	72.1
zho.pdtb.cdtb	75.9	71.2	70.6	72.6	75.4	62.7	65.4	76.3	77.1	66.2	78.0	71.8	76.3	74.5	78.4	74.3	57.2	68.7	55.8	66.3	66.8	65.6	73.7	68.3	93.8
zho.rst.gcdt	52.4	46.1	41.6	43.1	48.5	35.5	35.6	54.9	53.4	44.1	53.9	47.9	56.4	50.0	54.9	51.5	36.0	41.6	37.5	43.8	43.0	51.9	49.5	45.5	64.3
zho.rst.sctb	57.5	45.4	33.5	40.8	44.8	35.2	38.0	44.8	49.2	35.2	54.0	41.9	56.1	38.9	51.9	50.1	41.5	35.8	38.9	39.1	36.9	42.5	40.6	34	56.6
average	56.1	50.1	46.3	49.1	52.2	42.0	42.0	54.5	55.6	45.9	56.9	49.8	58.2	52.1	56.3	53.8	38.3	43.0	38.1	43.8	42.9	51.3	51.6	47.9	65.7

Table 3: **Dataset accuracy scores by LLM probe averaged over five runs.** The last two columns refer to the DisCoDisCo reference system trained on all languages (all) and trained a multiple models with different encoders per dataset.



(a) THEMATIC (b) TOPIC-MANAGEMENT

Figure 8: **Layer-wise probe performance by relation classes.** Mean accuracy over five runs.

	Llama-3.1-70B	Llama-3.1-8B	Llama-3.2-1B	Llama-3.2-3B	Mistral-Small-24B-Base-2501	Qwen2.5-0.5B	Qwen2.5-1.5B	Qwen2.5-14B	Qwen2.5-32B	Qwen2.5-3B	Qwen2.5-72B	Qwen2.5-7B	aya-23-35B	aya-23-8B	aya-expanse-32b	aya-expanse-8b	bloom-1b1	bloom-3b	bloom-560m	bloom-7b1	bloomz-7b1	emma-500-llama2-7b	phi-4
$\mathbf{C}$ (inter)	53.8	45.5	41.5	43.4	45.5	35.1	35.6	49.1	52.9	40.8	55.8	42.4	53.7	47.6	54.3	47.0	32.4	39.9	32.3	40.3	40.1	44.4	46.9
$\mathbf{D}_1, \mathbf{D}_2$ (intra)	46.0	38.2	35.2	37.3	40.1	33.3	32.9	43.5	44.5	33.4	47.4	38.3	45.5	36.0	39.9	40.3	28.5	30.8	29.0	31.9	31.3	40.4	41.3
$\mathbf{D}_1, \mathbf{D}_2, \mathbf{C}$ (all)	55.9	50.8	46.2	48.6	52.5	41.6	42.2	54.2	55.8	46.0	56.8	50.3	58.2	52.6	56.0	54.0	37.8	42.6	37.0	43.5	43.9	51.4	51.9

Table 4: **Overall accuracy scores by LLM probe averaged over five runs training on the entire DISRPT (probe representation ablation).** We ablate different types of attention representations used for probing and find that using all described matrices $\mathbf{D}_1, \mathbf{D}_2, \mathbf{C}$ yields the best results.

	Llama-3.1-70B	Llama-3.1-8B	Llama-3.2-1B	Llama-3.2-3B	Mistral-Small-24B-Base-2501	Qwen2.5-0.5B	Qwen2.5-1.5B	Qwen2.5-14B	Qwen2.5-32B	Qwen2.5-3B	Qwen2.5-72B	Qwen2.5-7B	aya-23-35B	aya-23-8B	aya-expanse-32b	aya-expanse-8b	bloom-1b1	bloom-3b	bloom-560m	bloom-7b1	bloomz-7b1	emma-500-llama2-7b	phi-4
mean	55.6	48.7	45.2	47.8	48.6	39.4	40.5	50.5	51.7	42.9	54.4	46.4	54.2	48.6	53.1	50.1	35.1	40.0	34.6	40.1	39.9	49.7	48.3
max	55.9	50.8	46.2	48.6	52.5	41.6	42.2	54.2	55.8	46.0	56.8	50.3	58.2	52.6	56.0	54.0	37.8	42.6	37.0	43.5	43.9	51.4	51.9
mean,max	56.1	51.0	46.5	48.6	52.8	42.6	42.8	54.0	56.0	47.1	56.1	50.8	58.4	52.8	56.9	54.2	38.5	43.7	38.4	43.2	44.0	51.5	52.0

Table 5: **Overall accuracy scores by LLM probe averaged over five runs training on the entire DISRPT (pooling strategy ablation).** We ablate different attention-head-wise pooling strategies, namely mean pooling, maximum pooling, and the concatenation of both on the attention score matrices $\mathbf{D}_1, \mathbf{D}_2, \mathbf{C}$. We find that concatenating both generally yields the best results, though usually only by a margin of less than 1%. To keep the size of the representations shorter, we thus opt to only keep the maximum pooling.