DIAM: Discourse Interaction and Addressee Modeling for Malaysia Hansard

Kun Shun Goh[1], Mohan Raj Chanthran[2], Lay-Ki Soon[1,2], Wern Han Lim[1], Ian K. T. Tan[3],
[1]School of Information Technology, Monash University Malaysia
[2]Action Lab, Monash University Malaysia
[3]Heriot-Watt University Malaysia
Correspondence: soon.layki@monash.edu

Abstract

Parliamentary debates are vital records of democratic processes, yet official transcripts such as the Hansard provide only a sequential account of speeches, omitting the underlying interactional structure. Although conducted in a formal setting, debates can become highly dynamic, making it difficult to determine who addresses whom and how utterances relate within the broader discussion. To address these gaps, we introduce DIAM, the first manually annotated discourse resource for the Malaysian Parliament, jointly supporting EU-level Dialogue Discourse Parsing (DDP) and multi-addressee Addressee Recognition (AR) over long, formal, unscripted spoken multi-party debate. Unlike existing DDP datasets, DIAM provides substantially longer conversations and explicit multi-addressee labels. Our benchmark experiments validate that DIAM supports both DDP and AR, while showing that existing discourse parsing methods cannot be directly applied without adaptations such as context-window strategies, and that addressee recognition is challenging for prompted LLMs, though fine-tuning a smaller model with LoRA substantially improves robustness to the large number of participants per sitting. We release DIAM and our code to facilitate future research on formal discourse parsing and addressee modeling: https://github.com/xKHUNx/DIAM

1 Introduction

Parliamentary debates are central to democracy. Official records of these debates, known as Hansard, document how political decisions are made, party positions evolve, and politicians debate laws and policies while representing their constituents interests (Sousa and Lopes Cardoso, 2024; Vladu et al., 2024). Parliamentary debate follows strict procedural norms, including formal honorifics and constituency-based references, with each utterance serving a specific role such as questioning, explaining, or clarifying. However, debates remain highly dynamic, with interruptions, rebuttals, and rapid exchanges that make it difficult to track interactions and the relationship between utterances. Recent work has also highlighted the importance of developing Malaysian specific annotated resources for downstream language understanding tasks (Chanthran et al., 2024).

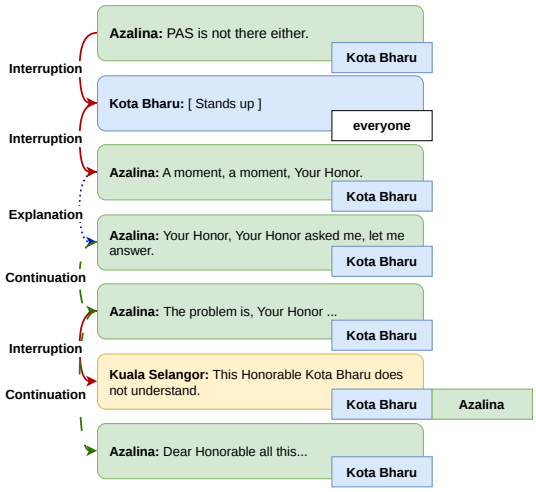


Figure 1: Sample DIAM dialogue excerpt with three participants, showing EUs, their addressees, and discourse relations. Frequent interruptions are visible. Participant names are abbreviated for clarity.

Despite capturing every utterance verbatim, the official Hansard presents parliamentary proceedings as a sequential record of speeches and actions. It does not explicitly encode the interactional structure of debates, such as the contextual function of individual **Elementary Units (EUs)** (speech segments and non-verbal events) or their intended addressees. Consequently, two critical aspects of parliamentary discourse remain implicit and must be inferred computationally:

1. Dialogue Discourse Parsing (DDP): Identifying relationships between EUs, such as question–answer, clarification, and elabora-

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 240–253
November 6-10, 2026 ©2026 Association for Computational Linguistics

tion, to capture how Members of Parliament (MPs) interact during debates.

2. Addressee Recognition (AR): Identifying the intended recipient(s) of each **EU**. Speakers may address an individual MP, a group of MPs, or the entire House, often without explicit references, making multi-addressee interactions difficult to detect.

To address this gap, we present **DIAM**, a dataset for **D**iscourse **I**nteraction and **A**ddressee Modeling for **M**alaysia Hansard, derived from the Malaysian Parliament Hansard. In Malay, the word *diam* means "be quiet" a fitting reference to the primary language of the Malaysian Parliament and a playful nod to the occasional heated debates that take place within it. To the best of our knowledge, this is the first dataset for the Malaysian Parliament that provides manual annotations for both discourse relations and addressees. The dataset comprises 40 agendas from 31 sittings from 19/12/2022 to 30/11/2023, containing a total of 9,792 elementary units (EUs), where each EU is either an Elementary Discourse Unit (EDU, a verbal segment) or an Elementary Event Unit (EEU, a non-verbal action), all annotated with addressees, and 10,494 discourse relations between the EUs. Through this work, we provide baseline models for DDP and AR, establishing a benchmark for future research on the computational modeling of interactions within the Malaysian Parliament. An excerpt of the DIAM dataset is visualized in Figure 1. The dataset and baseline are available in: https://github.com/xKHUNx/DIAM

2 Related Work

In this section, we review the existing landscape of datasets for Dialogue Discourse Parsing (DDP) (Section 2.1) and Addressee Recognition (AR) (Section 2.2). Our review analyzes these datasets in terms of their theory frameworks, structure and conversation context that helps to plan and build DIAM within the current research landscape.

2.1 Existing Dialogue Discourse Parsing Dataset

Several datasets exist for dialogue discourse parsing, varying in their environmental grounding and theoretical framework. The STAC corpus (Asher et al., 2016) is derived from an online version of the board game *The Settlers of Catan* and consists of strategic, multiparty dialogues involving two to six speakers. Grounded in Segmented Discourse Representation Theory (SDRT), STAC represents dialogue at the level of Elementary Discourse Units (EDUs) and models non-linguistic game actions as Elementary Event Units (EEUs). It supports complex discourse units (CDUs) and non-tree discourse graphs, and its set of 16 discourse relations has laid the foundation for subsequent research (Asher et al., 2012).

Building on this foundation, Molweni (Li et al., 2020) was constructed from the Ubuntu Chat Corpus to provide a larger-scale resource for multiparty Internet Relay Chat (IRC) conversations. Unlike STAC, Molweni targets non-situated dialogue and thus omits EEUs. While it retains the 16 relations from STAC, it employs a simplified, dependency-style version of SDRT that removes CDUs to improve parsing tractability. Similarly, the Minecraft Situated Dialogue Corpus (MSDC) (Thompson et al., 2024) explores situated interaction but focuses on dyadic, task-oriented dialogue where an Architect instructs a Builder. In MSDC, EEUs correspond to Builder actions, and CDUs are used strictly to compress action sequences rather than linguistic material.

Several discourse parsing datasets are related to our work but adopt different theoretical frameworks. The GUM Corpus (Zeldes, 2017) follows Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) and covers a broad range of genres, including interviews from Wikinews. Notably, GUM also provides speaker and addressee annotation for its spoken and interactional genres; however, these are annotated at the sentence level rather than at the EDU/EU level and are not integrated with an SDRT-style multi-party discourse graph. Similarly, the LUNA Corpus (Tonelli et al., 2010) and DiscoGeM (Scholman et al., 2022) adopt a Penn Discourse Treebank (PDTB)-style annotation scheme (Webber et al., 2019). Additionally, many datasets are either limited to a subset of a broader collection (such as GUM and DiscoGeM) or are built in different languages, like LUNA in Italian. Notably, DiscoGeM includes parliamentary speeches from Europarl, including English texts and translations from other European languages, making it thematically related to our work. However, it does not model structured multi-party parliamentary interaction under an SDRT framework.

In summary, existing SDRT-based corpora (STAC, Molweni) target informal online settings,

while broader corpora (GUM, DiscoGeM) either annotate addressees only at the sentence level or do not jointly model deep multiparty interaction with discourse structure. DIAM extends SDRT-style annotation to situated, formal, face-to-face parliamentary interaction, pairing EU-level multi-addressee labels with discourse relations and non-linguistic events.

2.2 Existing Addressee Recognition Dataset

In multi-party conversations, knowing who a speaker is talking to is much harder than in two-person chats. While this is a key task, very few datasets actually provide labels for group or subgroup targets that simply assume there is only one addressee.

Early research focused on physical meetings. For example, the AMI Meeting Corpus provides manual labels for individuals, subgroups, the whole group, or unknown targets (Jovanovic et al., 2005). These labels were created using extra clues like where the speaker was looking (gaze) and their body language. Similarly, the STAC corpus provides group and individual labels for an online board game setting, but without the extra clues as it is of online interaction (Asher et al., 2012).

Other researchers have used large-scale technical chats, like the Ubuntu-based corpora. However, these datasets are more limited. The Ubuntu Multiparty Conversation Corpus uses a simple rule to guess the addressee based on name mentions in the text (Ouchi and Tsuboi, 2016). Another version, used in the Graph-Structured Network (GSN) model, builds a conversation graph using "@" mentions but does not have explicit human made labels for group addressing (Hu et al., 2019).

In summary, while AMI and STAC offer the best labels for group addressing, they are limited to small meetings or games. This leaves a gap for a dataset like DIAM, which must handle formal parliamentary addressing where there are no "gaze" clues or "@" mentions, and speakers often address specific constituencies or the entire house.

3 Data Collection and Annotation

This section describes the development of the DIAM corpus. We detail our data collection and processing approach in Section 3.1, while Section 3.2 outlines the annotation of addressees and discourse relations.

3.1 Data Collection and Preparation

We collected 40 agendas from 31 sittings of the Malaysian Parliament, covering proceedings from December 2022 to November 2023. The data, comprising verbatim reports of Dewan Rakyat proceedings, is publicly available on the official Parliament of Malaysia website[1] . Unlike DiscoGeM (Scholman et al., 2022), which uses professionally translated Europarl proceedings (Koehn, 2005; Cartoni and Meyer, 2012), our source texts require machine translation into English. We first translated the raw Malay transcripts to English at the passage level, prior to any segmentation, using Mesolitica's `translation-t5-small-standard-bahasa-cased-v2` model,[2] a T5-based (Raffel et al., 2020) translation model that reports a chrF2++ score of 64.53 for standard Malay-to-English translation. Translating before segmentation preserves full sentential context for the translation model and avoids propagating Malay–English structural divergences into unit boundaries. We preserve the original Malay source text aligned to its English translation at the sentence level only. Because EDU segmentation is performed on the English text, no finer Malay-to-EDU alignment is retained. These source sentences can be easily reincorporated into the dataset if required. We selected this open, reproducible model because it handles Malay and code-mixed regional varieties while preserving source structure, and produced qualitatively better translations of our parliamentary data than alternative services without relying on proprietary APIs. Only after translation were non-verbal events (e.g., applause or laughter) extracted as standalone EEUs, and the remaining English text split into sentences and segmented into EDUs using SegBot (Li et al., 2018), which identifies minimal coherent units of dialogue conveying a single proposition or communicative function. SegBot's automatic segmentation was subsequently validated by our annotators. During annotation, an annotator who judged a split to be incorrect flagged the two oversegmented EDUs by linking them with an empty-label relation, and during adjudication the adjudicator decided whether to merge them back, and could likewise split EDUs manually using the adjudication tools. This flag-and-merge procedure, docu-

[1] https://www.parlimen.gov.my

[2] https://huggingface.co/mesolitica/translation-t5-small-standard-bahasa-cased-v2

mented in our annotation guidelines,[3] provides a degree of human quality control over the segmentation. Each agenda was treated as a self-contained dialogue, resulting in 40 agenda-level documents ready for annotation.

3.2 Data Annotation Process

We hired three Malaysian annotators who are fluent in both English and Malay. We chose local annotators because they are familiar with the Malaysian context and the topics discussed in Parliament. The annotators were selected after an interview assessing their linguistic background and familiarity with parliamentary discourse, followed by a two-hour training phase covering the annotation guidelines with worked examples. Two were assigned as main annotators, while the third served as an adjudicator. Annotators were compensated at RM 60 (Malaysian Ringgit; approximately USD 13) per agenda (document). All annotators provided informed consent prior to the start of the study. They were fully briefed on the annotation guidelines, the intended use of the data, and the public release of the final dataset for research purposes. We did not conduct a separate quantitative evaluation of translation quality on a held-out subset. Instead, quality control was embedded directly into the annotation workflow. The main annotators independently annotated all dialogues and cross-checked the translated EUs against the original Malay Hansard to ensure accuracy. All annotators and the adjudicator are fluent in both Malay and English and, as Malaysians, possess the domain and cultural knowledge needed to catch nuances that machine translation may miss. Because the mapping to the original Malay text was preserved (Section 3.1), the main annotators cross-referenced the source sentences and corrected broken translations during annotation, and the adjudicator served as a second validation pass, revising translations where necessary. As translation is not the primary objective of this corpus, this intervention targeted semantic accuracy rather than stylistic polish. Minor grammatical errors or awkward phrasing were left unchanged, and corrections were applied only when the machine translation altered the core meaning or was outright wrong. The adjudicator resolved disagreements between the main annotators and verified the overall completeness and integrity of the annotations, including ensuring the absence of orphaned nodes or clusters, the absence of backward relations, and that every EU was assigned at least one addressee.

3.2.1 Discourse Relation Annotation

Discourse relation annotation follows the Segmented Discourse Representation Theory (SDRT) framework, as adopted in the STAC corpus and subsequent dialogue discourse parsing datasets. As a form of situated dialogue, our dataset consists of both EDUs and EEUs. EEUs correspond to non-verbal actions such as laughter, audience reactions, or technical events (e.g., microphone being turned off), and are explicitly demarcated with square brackets (e.g., `[ Stands up ]`) following the convention of the original Malaysian Hansard, which distinguishes them from verbal EDUs. Following standard practice in dialogue discourse parsing, we do not annotate Complex Discourse Units (CDUs).

The main annotators were tasked with identifying and labeling discourse relations between pairs of EUs. For each pair, annotators identified the presence of a discourse relation and assigned the appropriate label from our predefined taxonomy. If no meaningful relation exists, no link is annotated. We extend the standard relation set by introducing an "Interruption" relation, motivated by the frequent occurrence of interruptions in parliamentary proceedings. This relation captures abrupt breaks in discourse flow and is necessary to maintain connectivity in the resulting discourse structure. Detailed definitions for all 17 relations are provided in Table 10 (Appendix A).

Consistent with SDRT, we allow EUs to have multiple parents and multiple children, but disallow self-referential links. The resulting annotations form a directed acyclic discourse graph, where nodes correspond to EUs and edges correspond to labeled discourse relations.

3.2.2 Addressee Annotation

In addition to discourse relations, annotators labeled the intended addressee for each EU. For every EU produced by a speaker, annotators identified the participant(s) to whom the EU was directed. Addressees were selected from the set of active participants in the conversation, defined as speakers who produced at least one verbal or non-verbal contribution within the agenda. An EU could not designate its own speaker as the addressee. This annotation captures interaction structure in multi-party par-

[3] https://github.com/xKHUNx/DIAM/blob/main/annotation_guidelines/Annotation_Guideline.pdf

liamentary dialogue and complements discourse relation labeling.

3.2.3 Inter-Annotator Agreement Results

To assess the reliability of our annotations, we calculated the Inter-Annotator Agreement (IAA) using two standard metrics. Both tasks were double-annotated by the same two main annotators; the choice of metric reflects the structure of each label rather than any difference in the number of annotators. For discourse relations, each unit pair receives a single categorical judgment (the presence of a link and, if present, one relation label), for which we report Cohen's Kappa (κ) (Cohen, 1960). Addressee annotation is instead set-valued: a single EU may be directed at multiple participants or at the entire House, which Cohen's κ cannot represent. We therefore use Krippendorff's Alpha (α) (Krippendorff, 2011) with a set-based distance, which accommodates multi-label and overlapping annotations.

Table 2 summarizes IAA over 39 dialogue documents. One document was excluded because it was annotated by a subject expert to serve as the gold standard for annotator reference, rather than independently double-annotated, and therefore is not suitable for inter-annotator agreement computation. For both the Kappa and Alpha metrics, values range from 0 to 1, where higher scores indicate stronger consensus. For addressee annotation, the high α score indicates strong consistency among annotators in identifying the intended recipients of utterances and non-verbal events. Meanwhile, for discourse relation annotation, agreement is high for link detection and moderately high for relation labeling, reflecting the complexity of fine-grained discourse relations.

To contextualise our IAA scores, we compare them with previously reported results where available. For discourse relation annotation, link-level agreement in DIAM ($\kappa \approx 0.77$) is comparable to STAC ($\kappa \approx 0.72$) while falling behind Molweni ($\kappa \approx 0.91$); label-level agreement ($\kappa \approx 0.59$) is similarly comparable to STAC ($\kappa \approx 0.58$) and Molweni ($\kappa \approx 0.56$). When both annotators agreed a link existed, label disagreements centred on Continuation being used as a default relation, with the most frequent conflicts being Continuation vs. Elaboration (n=229) and Acknowledgment vs. Continuation (n=228). The latter typically involved floor-management moves across speaker turns, where one annotator marked a social receipt and the other a structural continuation.

For addressee annotation, we obtain Krippendorff's $\alpha = 0.80$, indicating strong agreement. We refrain from a direct numerical comparison with prior corpora such as AMI (Jovanovic et al., 2005), which report Cohen's κ under a single-addressee formulation and are therefore not directly comparable to our set-valued α. Disagreements concentrated on non-verbal events (*[Applause]*, *[Laughs]*) and formal parliamentary motions, where annotators differed on whether the Speaker or the whole chamber was the intended recipient, reflecting both domain regularities and the inherent complexity of multi-party dialogue.

4 Corpus Statistics and Analysis

This section provides a quantitative analysis of the DIAM corpus. Section 4.1 provides a Dataset Overview, including a comparison with existing datasets (Section 4.2). This is followed by an analysis of Relation Label Distributions (Section 4.3), the Number of Addressees (Section 4.4), and Relation Distance Statistics (Section 4.5). We further examine Non-Tree Structural Properties, Relation-Specific Distance Analysis, and Speaker-Role Discourse Patterns in Appendix B, C, and D, respectively.

4.1 Overview

The DIAM dataset statistics are summarized in Table 3, consisting of 40 agendas with a total of 9,792 EDUs and 10,494 discourse relations. We randomly split the corpus into experimental sets, allocating 28 agendas for training and 6 agendas each for the development and test sets. Although we have a much smaller number of dialogues, the number of EUs and relations are comparable to dataset like STAC. This is because parliamentary debates are usually much longer than online chats, as speakers must follow a formal structure.

4.2 Dataset Comparison and Statistics

Following the multi-party corpus taxonomy of Mahajan and Shaikh (2021), DIAM is a *Spoken Unscripted* corpus, in contrast to the *Written* synchronous-chat settings of STAC, Molweni, and MSDC. This difference in setting is reflected in DIAM's structural statistics. As shown in Table 1, although DIAM contains only 40 dialogues, each dialogue comprises an average of 244.8 elementary units (EUs) and 262.35 discourse relations,

Category	Statistic	DIAM	STAC†	Molweni	MSDC
Setting‡	Modality	**Spoken Unscripted**	Written	Written	Written
	Register	**Formal**	Casual	Casual	Casual
Corpus Size	# Dialogues	40	1,173	**10,000**	540
	# EUs	9,792	12,867	**88,303**	28,697
	# Relations	10,494	12,476	**78,245**	34,529
Speakers	Mean # Speakers	**9.88**	2.93	3.52	2.00
	Min # Speakers	**3**	2	2	2
	Max # Speakers	**18**	6	9	2
Per-Dialogue Avg.	Avg. # EUs	**244.80**	10.97	8.83	53.14
	Avg. # Relations	**262.35**	10.64	7.82	63.94
	Avg. # EEUs	7.78	0.00	0.00	**21.57**
EU Length (char.)	Mean Length	**58.07**	16.49	51.70	23.47
	Min Length	3	1	**8**	1
	Max Length	**554**	100	223	384
Relation Distance	Mean Distance	1.88	**2.02**	1.61	1.88
	Min Distance	1	1	1	1
	Max Distance	**234**	60	12	63

† Linguistic-only version (Shi and Huang, 2018); situated STAC is omitted.
‡ Following the multi-party corpus taxonomy of Mahajan and Shaikh (2021).

Table 1: Comprehensive comparison of DIAM against existing DDP dataset.

Annotation	Metric	Mean Score
Discourse Relation	κ_{link}	0.773
	κ_{label}	0.594
Addressee	$\alpha_{\text{addressee}}$	0.803

Table 2: Mean inter-annotator agreement scores over 39 documents, grouped by annotation task, where κ denotes Cohen's Kappa and α denotes Krippendorff's Alpha.

	Total	Train	Val	Test
# Dialogues	40	28	6	6
# EUs	9,792	6,792	1,535	1,465
# Relations	10,494	7,271	1,647	1,576

Table 3: Key statistics of the DIAM dataset, showing the distribution of dialogues, Elementary Units (EUs), and relations across experimental splits.

which is several times higher than in other widely used datasets. DIAM also provides a comparable number of EEUs per dialogue to other situated dialogue datasets that include EEU annotations. In fact, 3.18% of EUs in DIAM are actually EEUs.

Moving from volume to linguistic depth, Table 1 shows that EUs in DIAM are generally longer than in most other datasets, with a mean length of 58.07 characters per EDU and a maximum of 554 characters. This is likely due to DIAM being derived from a formal, face-to-face setting of unscripted spoken debate, where utterances are more elaborate and contain more information per EU compared to written chat. In contrast, datasets such as STAC and MSDC, which comprise written, synchronous in-game chat, exhibit shorter EUs on average, reflecting their concise nature.

Label	%	Label	%
Continuation	30.33	Elaboration	17.51
Acknowledgement	9.90	Result	7.19
Interruption	7.14	Contrast	4.71
Explanation	4.60	Comment	3.32
QAP	3.02	Conditional	2.66
Parallel	2.46	Background	2.04
Clarification Q.	1.91	Narration	1.78
Correction	0.70	Alternation	0.46
Q-Elab	0.29		

Table 4: Distribution of discourse relation labels in DIAM. QAP denotes Question-Answer Pair.

Beyond the length of individual units, Table 1 summarizes the relation distance and number of speakers across DIAM and other comparison datasets. DIAM exhibits a higher maximum relation distance (234) compared to other datasets, reflecting longer-range discourse dependencies within dialogues. It also involves more participants on average, with a mean of 9.88 speakers per dialogue, whereas STAC and Molweni typically include fewer participants, while MSDC is strictly a two-party dialogue. These statistics highlight the multiparty nature and structural complexity of DIAM dialogues relative to other datasets.

4.3 Distribution of Relation Labels

Table 4 presents the distribution of discourse relation types in DIAM. The most frequent relation is

# Addressees	Count	Percentage
1	7,947	81.2%
2	436	4.5%
>2	6	0.1%
Everyone†	1,403	14.3%
Total	9,792	100%

Table 5: Distribution of the number of addressees per EU in DIAM. Everyone† refers to EUs directed at the entire assembly.

Distance	Count	Percentage
1	8,350	79.57%
2	1,126	10.73%
3	370	3.53%
4	159	1.52%
5	98	0.93%
6–10	206	1.96%
11–99	174	1.66%
≥100	11	0.10%

Table 6: Distribution of relation distances in the DIAM corpus. Distance is defined as the numerical gap between the indices of the source and target units.

Continuation (30.33%), followed by *Elaboration* (17.51%) and *Acknowledgement* (9.90%). *Interruption* accounts for around 7% of the corpus, indicating its notable presence in parliamentary discourse. Overall, the distribution reflects a dominance of *Continuation* and *Elaboration*, which frequently occur during long speeches or extended explanations by MPs. *Acknowledgement*, on the other hand, highlights the turn-taking nature of our corpus, as MPs request permission to speak and the Speaker grants it, producing repeated short signals of understanding or acceptance.

4.4 Number of Addressees

Table 5 shows the distribution of the number of addressees per EU in the DIAM corpus. Most EUs are directed to a single participant (81.2%), and 4.5% of the EUs are directed to two specific addressees, while a smaller fraction addresses the entire Parliament (14.3%). EUs targeting multiple specific addressees (>2) are extremely rare (0.1%). Despite their rarity, predicting addressees is challenging in DIAM because each session involves a large pool of potential targets (averaging about 10 speakers per session, and up to 18), meaning even single-addressee EUs must be correctly assigned from many candidates. This makes multi-addressee resolution more complex than in most other dialogue corpora.

4.5 Relation Distance Statistics

Table 6 shows the distribution of discourse relation distances in the DIAM corpus. Most relations (79.57%) connect to adjacent EUs (distance 1), indicating that discourse relations are predominantly local. Relations spanning two EUs account for about 10.73%, while longer-distance relations are relatively rare. Only a small fraction of relations (0.10%) connect EUs separated by 100 or more positions, reflecting the largely sequential and locally connected nature of parliamentary discourse. More detailed analysis of the distances for each relation label is provided in Appendix C.

5 Benchmarking

5.1 Dialogue Discourse Parsing

5.1.1 Experimental Setup

We benchmark DIAM using three approaches: (1) SDDP (Chi and Rudnicky, 2022), a task-specialized, fully supervised structured parser (non-LLM) predicting links and labels over 20-EU windows (stride 10); (2) GPTforDDA (Fan et al., 2024), a prompt-based LLM (ChatGPT-3.5-Turbo) using the full dialogue; and (3) Seq2Seq-DDP (Li et al., 2024), an open, fully fine-tuned T0-3B encoder–decoder over 8-EU windows, evaluated in end-to-end and transition modes. These baselines span the three dominant DDP paradigms: a task-specialized supervised structured parser (SDDP), an open fine-tuned generative model (Seq2Seq-DDP/T0-3B), and a prompted closed LLM (GPTforDDA). Two of the three are fully open and reproducible. For SDDP and Seq2Seq-DDP, windowing is necessary because DIAM dialogues are lengthier, containing more EUs per dialogue than typical datasets, which poses a unique challenge for fitting the input into GPU memory. Full model configurations, data splits, and evaluation scripts are released for reproducibility (Appendix E).

5.1.2 Results and Analysis

Table 7 reports overall performance. SDDP achieves the highest link F1-Score, confirming the effectiveness of structured modeling. Seq2Seq-DDP (end-to-end) remains competitive on joint link+label prediction, benefiting from access to gold labels of prior EUs within the 8-EU window,

Method	Window	*F1 (Link)	*F1 (L+L)
SDDP	20 (stride=10)	**0.7859**	**0.5212**
GPTforDDA	Full	0.6754	0.1268
Seq2Seq-DDP (end-to-end)	8	0.6352	0.4332
Seq2Seq-DDP (transition)	8	0.0205	0.0145

Table 7: Dialogue discourse parsing baseline performance on the DIAM dataset. *F1 refers to F1-Score. L+L denotes Link and Label agreement.

while GPTforDDA performs worst, and Seq2Seq-DDP (transition) collapses due to error propagation.

Link detection versus relation labeling. To explain the large gap between link-only and joint link+label performance, we examine per-relation scores. GPTforDDA attains competitive link-level F1 on frequent relations (Interruption 0.90, Acknowledgement 0.80, Continuation 0.73) but collapses once labels are required (0.09, 0.17, and 0.23 respectively), accounting for its low overall L+L score. SDDP, by contrast, retains most of its link performance after labeling (e.g., Interruption $0.85 \rightarrow 0.64$). This indicates that the prompted LLM baseline can often detect *that* two units are related but cannot reliably identify *which* discourse relation holds in the formal parliamentary register. Across models, *Comment* and *QAP* remain the hardest relations (best L+L F1 of 0.22 and 0.34, respectively), consistent with their comparatively non-local nature (19.5% and 51.4% of instances spanning $d \geq 4$).

Distance analysis shows SDDP excels on local links (d=1), whereas Seq2Seq-DDP surprisingly predicts some long-distance links (d=4–8, d $\geq$ 9) despite its 8-EU window (see Table 13 in Appendix F). Appendix G illustrates an example where distant parent EDUs outside the context window are correctly predicted. This highlights a trade-off: structured models provide stable local modeling, while generative models can capture long-range dependencies but are sensitive to accumulated errors.

Full-context versus windowed parsing. To assess whether DIAM's dialogue length affects LLM-based parsing, we evaluate a vanilla Llama-3.1-8B-Instruct (no fine-tuning) under two regimes: the full dialogue in a single pass versus 20-EU windows (stride 10), with a generation budget large enough to emit a relation for every EU in either regime (Table 8). Link detection is essentially regime-independent: the small test-set gap is within variance and reverses on the development split. Relation labeling, however, improves robustly under windowing, with link+label F1 roughly doubling and windowed parsing outperforming full-context on every one of the 12 development and test documents. Thus, even without task-specific fine-tuning, focused windows do not reliably aid link identification but substantially help a model assign the correct relation label, though absolute labeling quality remains low. For DIAM users, windowing is therefore a useful strategy for relation labeling on these long dialogues, complementing the structural observation that 0.88% of relations (and 9% of QAPs) connect EUs farther apart than a 20-EU window.

Setting	*F1 (Link)	*F1 (L+L)
Full-context	0.7048	0.1321
Windowed (20, stride=10)	0.7415	0.2364

Table 8: Vanilla Llama-3.1-8B (no fine-tuning) on the test set, varying only the context regime. The link-F1 difference is within variance (it reverses on the development split); only relation labeling (L+L) improves consistently, across all 12 documents.

5.2 Addressee Recognition

5.2.1 Experimental Setup

To benchmark addressee recognition on DIAM, we predict the addressee set for each EU given the dialogue and candidate participants, and evaluate two complementary baselines. First, as a general-purpose reference, we adopt a prompt-based approach using ChatGPT (gpt-5.2) (OpenAI, 2026). The full dialogue and participant list are provided as input, and the model is instructed to output addressee predictions in a structured JSON format, with no task-specific fine-tuning. This provides a useful general reference point, particularly given previous research indicating that ChatGPT's performance can vary on Malaysian language data (Chanthran et al., 2023). This design is motivated by the lack of prior work on explicit multi-addressee prediction, as most previous studies focus on the single-addressee setting (Ouchi and Tsuboi, 2016; Zhang et al., 2018; Le et al., 2019; Gu et al., 2021, 2023). Second, to assess whether task-specific adaptation helps, we fine-tune Qwen3-8B (Qwen

Team, 2025) with Low-Rank Adaptation (LoRA) (Hu et al., 2022). Rather than predicting all addressees in a single pass, the fine-tuned model predicts the addressee(s) of each target EU individually, conditioned on a 20-EU sliding context window together with the candidate participant list. The prompt template is provided in Appendix H.

5.2.2 Results and Analysis

Table 9 reports both baselines. The prompted GPT-5.2 baseline achieves only moderate overall performance (EM = 0.361, F1 = 0.368), high when the addressee is everyone but dropping sharply for single- and multi-addressee cases and degrading as the candidate pool grows, indicating sensitivity to search-space complexity.

Fine-tuning substantially improves performance. The LoRA-fine-tuned Qwen3-8B outperforms the prompted GPT-5.2 baseline overall, raising EM from 0.361 to 0.839. Performance also remains robust as the candidate pool grows: whereas GPT-5.2 falls to 0.138 EM with 19 candidates, Qwen3-8B maintains EM between 0.760 and 0.898 across the larger pools, indicating that task-specific adaptation greatly reduces sensitivity to search-space complexity. Multi-addressee recognition nonetheless remains challenging: fine-tuning lifts multi-addressee F1 from 0.149 to 0.393, but exact recovery of full addressee sets is still difficult (EM = 0.125), a result to read cautiously given only 40 such test instances. GPT-5.2 retains a slight edge only on the everyone scenario (0.790 vs. 0.716). Echoing our discourse-parsing findings (Section 5.1), restricting each prediction to a local 20-EU context window rather than the full agenda is also beneficial for addressee recognition on these long dialogues.

Category	Count	GPT-5.2 EM / F1	Qwen3-8B EM / F1
Overall			
Total EUs	1,465	0.361 / 0.368	**0.839 / 0.852**
By Addressee Scenario			
everyone	229	**0.790 / 0.790**	0.716 / 0.716
single	1,196	0.291 / 0.294	**0.886 / 0.894**
multi	40	0.000 / 0.149	**0.125 / 0.393**
By Candidate Pool Size			
4	141	0.858 / 0.858	0.858 / 0.858
7	111	0.568 / 0.568	**0.892 / 0.892**
10	276	0.359 / 0.373	**0.808 / 0.844**
12	262	0.164 / 0.167	**0.878 / 0.890**
16	313	0.489 / 0.491	**0.898 / 0.898**
19	362	0.138 / 0.151	**0.760 / 0.777**

Table 9: Addressee-recognition performance on DIAM, reported overall, by addressee scenario, and by candidate pool size. GPT-5.2 is prompted (zero-shot); Qwen3-8B is fine-tuned with LoRA using a 20-EU sliding context window. Results are Exact Match (EM) / F1; best per row in bold.

6 Conclusion

We introduced DIAM, the first manually annotated discourse resource for the Malaysian Parliament. Although prior corpora such as GUM provide sentence-level speaker and addressee annotation for some interactional genres, DIAM is, to our knowledge, the first to provide EU-level multi-addressee annotations jointly with SDRT-style discourse relations over long, formal, agenda-driven multi-party debate. Its novelty lies in pairing substantially longer dialogues than existing DDP corpora with explicit multi-addressee labels, enabling detailed analysis of speaker-target interactions in formal proceedings. Our benchmark experiments validate that DIAM is usable for both tasks, while revealing that existing discourse parsing methods cannot directly handle it without strategies such as context windows, and that addressee recognition is challenging in the zero-shot prompted setting, though fine-tuning a smaller LLM with LoRA substantially improves it. For practitioners building on DIAM, our analysis indicates that windowing aids relation labeling on these long dialogues and that, while prompted LLMs remain sensitive to the candidate-pool size, task-specific fine-tuning largely overcomes this, pointing to clear directions for future modeling.

Beyond serving as an NLP benchmark, DIAM enables empirical study of how the Malaysian Parliament functions as an interactional body. Because every EU carries both an addressee and its discourse relations, the corpus makes explicit who engages whom and how, supporting analyses of exchange networks, contestation, and procedural control by speaker role. Appendix D illustrates this, showing that the Chair produces more *Acknowledgement* and *Interruption* relations while MPs dominate *Elaboration* and *Contrast*.

We hope DIAM serves as a foundation for future research on formal discourse parsing and addressee modeling.

Limitations

Despite its contributions, DIAM has several limitations. First, the data is restricted to Hansard

sessions from December 2022 to November 2023, limiting temporal coverage from earlier or more recent sessions. Second, while the Malay-to-English translations were carefully cross-checked and manually corrected by annotators, translation imperfections may still persist. Finally, the baseline experiments are not comprehensive, we evaluate only a selected set of representative methods, and broader benchmarking with additional models may provide a more comprehensive understanding of task performance.

Ethics Statement

According to institutional guidelines, this study did not require formal ethics approval as it involved annotation of publicly available documents and did not involve sensitive personal data.

Acknowledgements

The annotations in this research were funded by the Malaysian Fundamental Research Grant Scheme (FRGS) FRGS/1/2022/ICT02/MUSM/02/2. We would like to express our appreciation for their support.

References

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. European Language Resources Association (ELRA).

Nicholas Asher, Vladimir Popescu, Philippe Muller, Stergos Afantenos, Anais Cadilhac, Farah Benamara, Laure Vieu, and Pascal Denis. 2012. Manual for the analysis of settlers data. `https://www.irit.fr/STAC/stac-annotation-manual.pdf`.

Bruno Cartoni and Thomas Meyer. 2012. Extracting directional and comparable corpora from a multilingual corpus for translation studies. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2132–2137, Istanbul, Turkey. European Language Resources Association (ELRA).

Mohan Raj Chanthran, Lay-Ki Soon, Huey Fang Ong, and Bhawani Selvaretnam. 2023. How well ChatGPT understand Malaysian English? an evaluation on named entity recognition and relation extraction. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 372–397, Singapore. Association for Computational Linguistics.

Mohan Raj Chanthran, Lay-Ki Soon, Huey Fang Ong, and Bhawani Selvaretnam. 2024. Malaysian English news decoded: A linguistic resource for named entity and relation extraction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10999–11022, Torino, Italia. ELRA and ICCL.

Ta-Chung Chi and Alexander Rudnicky. 2022. Structured dialogue discourse parsing. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 325–335. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Yaxin Fan, Feng Jiang, Peifeng Li, and Haizhou Li. 2024. Uncovering the potential of ChatGPT for discourse analysis in dialogue: An empirical study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16998–17010, Torino, Italia. ELRA and ICCL.

Jia-Chen Gu, Zhenhua Ling, Quan Liu, Cong Liu, and Guoping Hu. 2023. GIFT: Graph-induced fine-tuning for multi-party conversation understanding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11645–11658. Association for Computational Linguistics.

Jia-Chen Gu, Chongyang Tao, Zhenhua Ling, Can Xu, Xiubo Geng, and Daxin Jiang. 2021. MPC-BERT: A pre-trained language model for multi-party conversation understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3682–3692, Online. Association for Computational Linguistics.

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

Wenpeng Hu, Zhangming Chan, Bing Liu, Dongyan Zhao, Jinwen Ma, and Rui Yan. 2019. GSN: A graph-structured network for multi-party dialogues. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 5010–5016. International Joint Conferences on Artificial Intelligence Organization.

Natasa Jovanovic, Rieks op den Akker, and Anton Nijholt. 2005. A corpus for studying addressing behavior in multi-party dialogues. In *Proceedings of the 6th SIGdial Workshop on Discourse and Dialogue*, pages 107–116. Special Interest Group on Discourse and Dialogue (SIGdial).

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86.

Klaus Krippendorff. 2011. Computing Krippendorff's alpha-reliability.

Ran Le, Wenpeng Hu, Mingyue Shang, Zhenjun You, Lidong Bing, Dongyan Zhao, and Rui Yan. 2019. Who is speaking to whom? learning to identify utterance addressee in multi-party conversations. 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 1909–1919. Association for Computational Linguistics.

Chuyuan Li, Yuwei Yin, and Giuseppe Carenini. 2024. Dialogue discourse parsing as generation: A sequence-to-sequence LLM-based approach. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 1–14. 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.

Jing Li, Aixin Sun, and Shafiq Joty. 2018. SegBot: A generic neural text segmentation model with pointer network. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4166–4172.

Khyati Mahajan and Samira Shaikh. 2021. On the need for thoughtful data collection for multi-party dialogue: A survey of available corpora and collection methods. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 338–352. Association for Computational Linguistics.

William Mann and Sandra Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8:243–281.

OpenAI. 2026. Chatgpt (gpt-5.2) [large language model]. `https://chat.openai.com/`. Accessed: 2026-02-14.

Hiroki Ouchi and Yuta Tsuboi. 2016. Addressee and response selection for multi-party conversation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2133–2143. Association for Computational Linguistics.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

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

Zhouxing Shi and Minlie Huang. 2018. A deep sequential model for discourse parsing on multi-party dialogues. *Preprint*, arxiv:1812.00176 [cs].

Afonso Sousa and Henrique Lopes Cardoso. 2024. PTPARL-V: Portuguese parliamentary debates for voting behaviour study. In *Proceedings of the IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora (ParlaCLARIN) @ LREC-COLING 2024*, pages 38–42, Torino, Italia. ELRA and ICCL.

Kate Thompson, Julie Hunter, and Nicholas Asher. 2024. Discourse structure for the Minecraft corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4957–4967, Torino, Italia. ELRA and ICCL.

Sara Tonelli, Giuseppe Riccardi, Rashmi Prasad, and Aravind Joshi. 2010. Annotation of discourse relations for conversational spoken dialogs. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA).

Adina I. Vladu, Elisa Fernández Rei, Carmen Magariños, and Noelia García Díaz. 2024. Gender, speech, and representation in the Galician parliament: An analysis based on the ParlaMint-ES-GA dataset. In *Proceedings of the IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora (ParlaCLARIN) @ LREC-COLING 2024*, pages 21–29, Torino, Italia. ELRA and ICCL.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The penn discourse treebank 3.0 annotation manual.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Rui Zhang, Honglak Lee, Lazaros Polymenakos, and Dragomir Radev. 2018. Addressee and response selection in multi-party conversations with speaker interaction rnns. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.

A Relation Labels

Relation	Definition
Acknowledgement	B shows understanding or acceptance of A, or signals to move on.
Alternation	Logical OR operation; marks a disjunction between A and B.
Background	B provides context or stage-setting for the event in A (or vice versa).
Clarification_question	B seeks clarification regarding the content of A.
Contrast	A and B share structure but express opposing ideas or negate expectations.
Correction	B identifies and fixes an error in A.
Comment	B provides an opinion or evaluation of the content associated with A.
Conditional	Connects a hypothesis to a consequence (If A, then B).
Continuation	B extends or adds to the content of the previous utterance A.
Elaboration	B provides additional details to A via examples or sub-event listing.
Explanation	B provides the cause or reasoning behind what happened in A.
Narration	Events in A and B occur in a sequential temporal order.
Parallel	A and B share a common theme, property, or semantic structure.
Q-Elab	B is a follow-up question to a prior question A to elicit more info.
Question-Answer Pair (QAP)	A poses a question, and B provides the corresponding answer.
Result	Shows a cause-effect relationship (A, therefore B).
Interruption	B interrupts A to shift the topic or seize conversational control.

Table 10: Definitions of the 17 discourse relations annotated in the DIAM corpus.

B Non-Tree Structural Properties

Table 11 shows several structural properties of DIAM. A total of 746 nodes (7.62%) have more than one parent, meaning that a single unit can be linked to multiple earlier units. Additionally, 1,779 nodes (18.17%) have multiple children, indicating that one unit can connect to several subsequent units. There are also 41 node pairs (0.42%) connected by more than one relation, although this is relatively rare. Overall, these statistics show that DIAM is not organized as a simple tree structure. Instead, the discourse graph forms a directed acyclic graph (DAG), where units may have multiple incoming or outgoing connections.

Structural Property	Count	Percentage
Multi-parent nodes	746	7.62%
Multi-child nodes	1,779	18.17%
Node pairs with multiple relations	41	0.42%

Table 11: Distribution of non-tree structural properties in the DIAM dataset.

C Relation Specific Distance Analysis

Table 12 presents the distribution of relation distances broken down by relation type. Consistent with the overall statistics, most relations exhibit strong locality, with a median distance of 1 and the majority of instances occurring at distance 1. However, several relation types show comparatively higher proportions of non-local links ($d \geq 4$). In particular, QAP stands out with a median distance of 4 and 51.4% of instances occurring at $d \geq 4$, reflecting the fact that answers in parliamentary discourse may be delayed or embedded within extended turns. Other relation types with noticeable long-distance proportions include Clarification_question (20.5%), Comment (19.5%), Correction (9.6%), Elaboration (8.4%), and Acknowledgement (8.3%). These patterns suggest that while the discourse structure is predominantly local, interactive and dialogically responsive relations are more likely to span larger distances.

D Speaker-Role Discourse Patterns

Parliamentary debate is moderated, inducing a structural asymmetry between the Chair (the Speaker of the House) and ordinary MPs. Partitioning source EUs by speaker role (2,270 Chair vs. 8,224 MP), the Chair produces proportionally far more *Acknowledgement* (16.3% vs. 8.1%) and *Interruption* (10.1% vs. 6.3%) relations, consistent with a floor-management role of granting turns and curbing digressions. Conversely, MPs dominate the argumentative relations *Elaboration* (19.8% vs. 9.2%) and *Contrast* (5.4% vs. 2.4%), reflecting extended position-building turns. This role-conditioned distribution is a distinctive property of formal moderated debate, absent in the unmoder-

Relation	Median	% $d = 1$	% $d = 2$	% $d = 3$	% $d \geq 4$
Interruption	1	96.7	3.1	0.3	0.0
Conditional	1	93.9	5.7	0.4	0.0
Alternation	1	93.8	4.2	2.1	0.0
Continuation	1	89.0	8.0	2.2	0.8
Background	1	88.8	6.5	2.3	2.3
Explanation	1	85.3	10.4	2.9	1.4
Contrast	1	80.6	14.2	3.0	2.2
Acknowledgement	1	80.1	8.3	3.4	8.3
Q-Elab	1	80.0	13.3	3.3	3.3
Parallel	1	76.0	14.0	4.3	5.8
Elaboration	1	75.1	11.6	4.8	8.4
Narration	1	73.8	16.0	4.8	5.3
Correction	1	61.6	13.7	15.1	9.6
Result	1	61.4	25.9	5.7	7.0
Comment	1	57.8	16.7	6.0	19.5
Clarification_question	1	51.5	15.5	12.5	20.5
QAP	**4**	**32.5**	**10.4**	**5.7**	**51.4**

Table 12: Discourse relation distances by type in DIAM. Distance (d) represents the numerical gap between linked EDUs. Relations are sorted by locality (% $d = 1$). While most relations are predominantly local, QAPs are a significant outlier, with a median distance of 4 and over half of all instances spanning 4 or more units.

ated peer-to-peer settings of STAC and Molweni.

E Reproducibility and Implementation Details

We release the dataset, train/development/test splits, all model configurations, and evaluation scripts at the repository linked in Section 1. For dialogue discourse parsing, (1) SDDP (Chi and Rudnicky, 2022) is trained over 20-EU windows (stride 10), (2) Seq2Seq-DDP (Li et al., 2024) is a fully fine-tuned T0-3B over 8-EU windows, and (3) GPT-forDDA (Fan et al., 2024) uses ChatGPT-3.5-Turbo over the full dialogue. The vanilla LLM probe uses Llama-3.1-8B-Instruct without fine-tuning, evaluated both on the full dialogue and over 20-EU windows (stride 10), with greedy decoding and a generation budget of 32,768 tokens (required so that a relation can be emitted for every EU in the full-dialogue regime). Addressee recognition uses two baselines: (1) a prompted ChatGPT (gpt-5.2) over the full dialogue with the prompt in Appendix H, and (2) a Qwen3-8B (Qwen Team, 2025) model fine-tuned with LoRA (Hu et al., 2022) that predicts the addressee(s) of each target EU over a 20-EU sliding context window. It is fine-tuned with LoRA for 3 epochs at a learning rate of 2×10^{-4} on a single NVIDIA RTX 5090 (32GB) GPU. The full configuration is released with our code. All discourse-parsing results use the set-based link and link+label F1 described in Section 5.

Model	d=1	d=2-3	d=4-8	d≥9
SDDP	0.588	0.285	0.084	0.000
Seq2Seq-DDP	0.532	0.207	0.136	0.236

Table 13: Distance-wise F1-Score (Link + Type) for SDDP and Seq2Seq-DDP. Seq2Seq-DDP shows stronger performance for longer-distance relations despite a smaller context window.

F Distance-wise F1-score for SDDP and Seq2Seq-DDP

G Example of Long-Distance Relation Prediction

The following excerpt illustrates a long-distance prediction case. Several consecutive EDUs elaborate on edu44, which lies outside the visible 8-EDU input window for Seq2Seq-DDP:

```
[edu61] ...
...
[edu64] ... is Elaboration of [edu44];
[edu65] ... is Elaboration of [edu44];
[edu66] ... is Elaboration of [edu44];
[edu67] ... is Elaboration of [edu44];
[edu68] ... is Elaboration of [edu44];
Predicted answer:
[edu69] is Elaboration of [edu44]
```

H Addressee Prediction Prompt Template

The prompt template used for multi-addressee prediction is shown in Figure 2.

```
You are an assistant specializing in dialogue analysis. Given a conversation,
    identify the addressee(s) for EVERY utterance (EDU).

Candidates Pool (possible addressees):
{candidates_str}

Conversation:
{edus_str}

Instructions:
1. For each EDU in the conversation (identified by [edu0], [edu1], etc.),
    identify who the speaker is addressing.
2. The addressee(s) MUST be EXACTLY one or more items from the Candidates Pool
    above.
3. If an utterance is directed to everyone, use "everyone".
4. Multiple addressees are allowed (e.g., ["Speaker A", "Speaker B"]).
5. Return a JSON object with a key "addressees" containing a list of lists of
    strings.

Output format:
{
  "addressees": [
    ["everyone"],
    ["Name 1"],
    ["Name 2", "Name 3"]
  ]
}
```

Figure 2: Prompt template for addressee identification.