In Search of the Lost Arch: The Dependency Dialogue Acts Corpus

Jon Z. Cai[*] **Brendan King**[†] **Peyton Cameron**[*] **Susan Brown**[*]
Miriam Eckert[*] **Dananjay Srinivas**[*] **George A. Baker**[*] **Victoria K. Everson**[*]
Martha Palmer[*] **James H. Martin**[*] **Jeffrey Flanigan**[†]
[*]University of Colorado Boulder [†]University of California Santa Cruz

Abstract

Understanding the structure of multi-party conversation and the intentions and dialogue acts of each speaker remains a significant challenge in NLP. While a number of corpora annotated using theoretical frameworks of dialogue have been proposed, these typically focus on either utterance-level labeling of speaker intent, missing wider context, or the rhetorical structure of a dialogue, losing fine-grained intents captured in dialogue acts. Recently, the Dependency Dialogue Acts (DDA) framework has been proposed for modeling both the fine-grained intents of each speaker and the structure of multi-party dialogues (Cai et al., 2023). However, there is not yet a corpus annotated with this framework available for the community to study. To address this gap, we introduce a new corpus of 33 English language dialogues with over 9,000 utterance units, densely annotated using the Dependency Dialogue Acts (DDA) framework.Our dataset spans four genres of multi-party conversations from different modalities: (1) physics classroom discussions, (2) engineering classroom discussions, (3) board game interactions, and (4) written online game chat logs. Each session is doubly annotated and adjudicated to ensure high-quality labeling. We present a description of the dataset and annotation process, an analysis of speaker dynamics enabled by our annotation, and a baseline evaluation of LLMs as DDA parsers. We discuss the implications of this dataset for understanding dynamics between speakers and for developing more controllable dialogue agents.

1 Introduction

Understanding and representing speaker intention has long been a fundamental challenge in dialogue analysis, spanning multiple disciplines, including linguistics, philosophy, cognitive science, and artificial intelligence (Austin, 1975; Searle, 1979; Mann and Thompson, 1988). Accurately modeling and interpreting speaker intentions is crucial for dialogue system development, particularly in areas such as explainable AI, human-computer interaction, and conversational AI safety. Effective speaker intention modeling requires capturing not only explicit dialogue acts but also implicit communicative goals, social dynamics, and discourse structures that shape how interactions unfold.

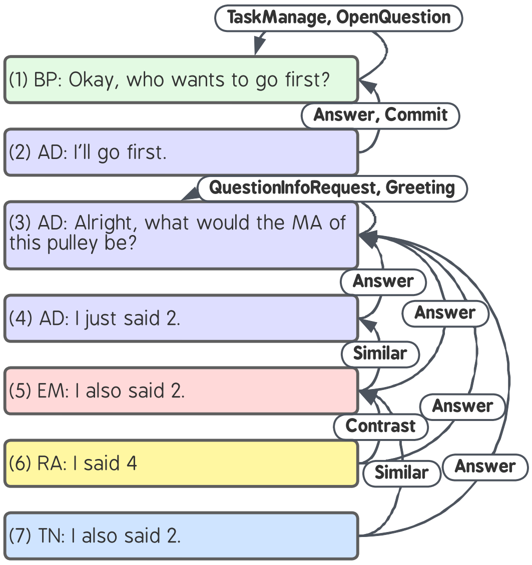
(1) BP: Okay, who wants to go first?
(2) AD: I'll go first.
(3) AD: Alright, what would the MA of this pulley be?
(4) AD: I just said 2.
(5) EM: I also said 2.
(6) RA: I said 4
(7) TN: I also said 2.

Figure 1: An example dialogue snippet from the Pulley K12 data annotated with Dependency Dialogue Acts (DDA) structure

Over the years, various annotation frameworks and corpora have been developed to analyze discourse structure and speaker intentions in dialogue. Switchboard DAMSL (SWBD; Jurafsky 1997) pioneered utterance-level annotation by categorizing dialogue acts such as questions, statements, and back-channels, in two-person phone conversations. Meanwhile, Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) and Penn Dis-

[*]jon.z.cai@colorado.edu
[†]bking2@ucsc.edu

Findings of the Association for Computational Linguistics: ACL 2025, pages 20135–20149
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

course TreeBank (PDTB) (Prasad et al., 2008) focused on structural text coherence by identifying discourse relations such as elaboration, contrast, and causality, laying the groundwork for understanding how speakers build meaning beyond isolated utterances. Segmented Discourse Representation Theory (SDRT) (Lascarides and Asher, 2007) extended this approach to dialogue, modeling multi-turn dependencies and coherence relations crucial for capturing conversational flow.

While these frameworks and corpora have significantly advanced computational dialogue understanding, they face challenges in multi-party conversations, where interactions are more fragmented, overlapping, and dynamic. Where two speakers on the phone might share uninterrupted joint attention, real-world multi-party dialogues are situated in complex tasks and contain deviations whose structure could be missed in frameworks like SWBD, which focus on utterance-level labels. Alternatively, frameworks focused on rhetorical structure might miss subtle distinctions in a speaker's communicative intent, such as whether a speaker is 'proposing' someone take an action, which is open to rejection, or 'directing' they do, which is not. To better capture the complexities encountered in multi-party dialogue, Cai et al. (2023) propose the Dependency Dialogue Act (DDA) framework, a general purpose framework for capturing multi-party speaker intentions and structure. A visualization of the DDA structure is shown in Figure 1. To our knowledge, no publicly available corpus has previously been annotated with the DDA framework, limiting its adoption and evaluation.

In this work, we present the first publicly released corpus of multi-party dialogues annotated with the DDA framework, capturing deep, structured representations of speaker intentions. Our corpus spans four domains, including both spoken and written (chat) modalities across 33 English language multi-party dialogues.

Our primary goal is to develop benchmark data that enables the evaluation of AI systems in collaborative multi-party conversations. Unlike dyadic exchanges, multi-party dialogues present additional challenges, such as overlapping speech, conflicting intentions, implicit coordination strategies, and emotional dynamics. Human communicators are not always perfectly rational, meaning interactions are often shaped by personal biases, misunderstandings, and differing levels of assertiveness. As such, an AI system capable of accurately identifying speaker intentions and behavioral patterns can help reduce unnecessary tension, promote effective collaboration, and support more structured group discussions. Despite the importance of this goal, existing frameworks and benchmarks for analyzing AI systems' understanding of multi-party dialogue remain limited. Our corpus provides a deep, structured approach for evaluating an AI system's understanding of multi-party dialogue.

Structured representations like DDA also provide an opportunity for controllable generation. While the rise of large language model-based (LLM) AI chat agents has reshaped the dialogue systems landscape, current LLMs still require elaborate prompt engineering to produce desirable behaviors consistently, even after fine-tuning (Addlesee et al., 2023). Their dialogue actions are not inherently controllable, which means that modifying an agent's behavior often requires trial-and-error adjustments to the prompts rather than direct intervention at the intent and strategy level. By equipping conversational AI systems with the ability to reason about complex speaker intentions within the DDA framework, we move closer to developing AI agents that can navigate collaborative scenarios, facilitate productive discussions, and promote synergy between human participants. This structured framework will help AI agents understand, anticipate, and appropriately respond to multi-party dialogue dynamics, ultimately improving explainability, controllability, and safety in conversational AI applications.

The contributions of this paper:

- We present the DDA Corpus: the first multi-party dialogue corpus annotated with the Dependency Dialogue Acts (DDA) annotation scheme, spanning 4 domains and including both spoken and written (chat) modalities.

- We release a codebook for the DDA framework with detailed definitions and software tools supporting annotation[1] to facilitate community use and customization.

- We analyze the annotations by measuring inter-annotator agreement across different granularity levels, present the statistical characteristics of our DDA Corpus, and demonstrate how our representation captures multi-party speaker dynamics.

[1] Available at: https://github.com/NSF-iSAT/DDA-corpus

- We conduct a baseline evaluation of LLMs as DDA parsers to assess their understanding of intention and structure in multi-party dialogue.

2 Related Work

The study of dialogue acts and speaker intention modeling has driven the development of various annotation schemes and corpora, serving as benchmarks for dialogue systems. In this section, we review key annotation frameworks, multi-party dialogue corpora, and recent evaluations of LLMs' capabilities in multi-party dialogue settings.

2.1 Dialogue Act Annotation Schemes

DAMSL and SWBD-DAMSL DAMSL (Dialogue Act Markup in Several Layers; Core and Allen 1997) and its variant Switchboard DAMSL (Jurafsky, 1997) are two of the pioneering dialogue act annotation schemes, in which individual utterances are labeled with classes corresponding to their function within a conversation. The DAMSL family of schemes builds on previous approaches by allowing multiple labels to be applied to each utterance; however, they are limited in that they are only designed with two-party (dyadic) dialogues in mind, and so make no attempt to capture reply or rhetorical structure as it is assumed that each utterance corresponds to its immediate predecessor, forming a simple adjacent pair.

ISO 24617-2 In addition to dialog acts, the ISO standard (Bunt et al., 2012) supports the annotation of various relationships *between* units of dialog as backward pointing edges: rhetorical relations, which describe how two dialog units are logically or structurally related; and dependence relations, in which one unit's semantic content depends on another's. These relational annotations capture more complex dialog structures than previous schema, but are limited in that dialog act annotations still correspond to dialog units, and so the schema cannot handle cases in which a dialog unit has multiple different relations to multiple previous dialog units (as there would be no way of disambiguating which relations correspond to which units).

Dialogue Dependency Acts Cai et al. (2023) propose Dialogue Dependency Acts (DDA), which places labels on response edges rather than the dialog units themselves. This allows more complex, multi-relational dialog structures to be captured; as a consequence, annotators can label all possible relations between dialog units as opposed to only the most prominent one. The compact dependency edges function like highways, streamlining the annotation process by making dependencies more intuitive. Each relation can be naturally read as an English statement reflecting the speaker's intention and attention, making the scheme easier to learn, customize and apply.

2.2 Multi-Party Dialogue Corpora

Here we describe multi-party dialogue corpora relevant to our intended use-case: the facilitation of collaborative, goal-oriented interactions. We thereby limit the discussion to corpora that are unscripted, synchronous, and of a collaborative nature. For a broader discussion of multi-party dialogue corpora we refer the reader to Mahajan and Shaikh (2021).

Pulley K12 Puntambekar et al. (2021) collect a conversational dataset examining K-12 students' learning behaviors in physical and virtual lab environments.[2] The dataset includes interactions between students and with their teacher.

Sensor Immersion K12 Cao et al. (2023) collect a dataset from a U.S. public middle school STEM classroom over four class periods. Students engaged in a STEM curriculum unit focused on developing digital sensor skills. The recordings primarily capture student discussions during collaborative problem-solving activities related to engineering challenges.

TEAMS The TEAMS corpus (Litman et al., 2016) is a dataset containing textual transcriptions and audio recordings from sessions of the cooperative board game Forbidden Island, played by groups of three to four adult speakers. The stated purpose of TEAMS is to study entrainment (when group members begin to speak more like one another) and participation dominance. Analysis in the original work generally focuses on acoustic features such as pitch and volume, finding that many of these features converge over the course of the play session.

STAC The STAC corpus (Asher et al., 2016) consists of textual chat and gameplay logs from online sessions of the game *The Settlers of Catan*. In contrast to the TEAMS dialogues which were

[2] Referred to as"Pulley" because the primary topic of the physics class in this study is pulleys and forces.

Source	**Pulley K12** (Puntambekar et al., 2021)	**Sensor Immersion K12** (Cao et al., 2023)	**TEAMS** (Litman et al., 2016)	**STAC** (Asher et al., 2016)
Source Type	Spoken	Spoken	Spoken	Written
# of Dialogues	1	26	4	2
Total # of Slash Units	1123	4765	1922	1267
Avg. # of Speakers	7.0	4.58	3.0	3.5
Avg. Edge Dist.	1.26	1.31	1.48	2.46
Release Terms	IRB required	IRB required	GNU-GPL	CC BY-NC-SA 4.0
Scenario	Physics Class	Engineering Class	In-Person Board Game	Online Board Game
Multimodal Access	transcript	video access	audio access	chat log
Demographics info	K12 students	K12 students	adults	N/A

Table 1: Contents of the Studied and Released Corpus. Source Type is from the taxonomy of Mahajan and Shaikh (2021).

spoken and transcribed, STAC's dialogues come from textual messages, and contain misspellings and non-standard language. The board game is also competitive instead of the collaborative. STAC also contains situated grounding messages originating from the game itself, which describe gameplay events and on which many player utterances depend rhetorically. The original STAC corpus is annotated with dialogue acts specific to in-game negotiation and discourse relations in the style of Segmented Discourse Coherence Theory (Asher and Lascarides, 2003). Our annotations provide a richer set of dialogue acts, covering the frequent non-negotiation dialogues.

2.3 Multi-Party Capabilities of LLMs

A number of recent works find that successful participation in multi-party conversation by LLMs depends highly on understanding discourse structure. Tan et al. (2023) evaluate GPT-3.5 and GPT-4 on tasks related to multi-party conversations (MPCs), including prediction of speakers and addressees of utterances, emotion detection, and generation of appropriate responses. They find that the inclusion of a simple speaker-addressee structure in prompts generally improves MPC understanding and generation capabilities. Gu et al. (2021) introduce MPC-BERT, a variant of the BERT language model trained in a multitask setting with the auxiliary tasks of reply-to-utterance prediction, speaker prediction, and speaker-addressee prediction. They find that the incorporation of these MPC tasks improves performance substantially on the Ubuntu IRC dataset's response selection task. Addlesee et al. (2023) study several language models' goal-tracking and intent-recognition capabilities in a multi-party hospital memory clinic setting. Their findings suggest that goal-tracking and intent-recognition performance change drastically in comparison to previous work using dyadic conversation. Our work provides a comprehensive labeled structure for understanding multi-party conversations which can be applied to these settings.

3 Coding Scheme

We closely follow the DDA framework definition and provide our domain-aware customizations of the coding scheme of DDA.

Slash Unit (SU) as Atomic Unit of Analysis: In the DDA framework, a Slash Unit (SU) is the fundamental unit of analysis and annotation. It approximates the minimal functional segment of an utterance, ensuring that each unit corresponds to a meaningful communicative act.

Dialogue Act (DA) Categories: Each SU is assigned one or more Dialogue Act (DA) labels to represent its communicative function.

Response Dependency Structure: Dependency edges encode response relations between SUs. Each edge is directed from the dependent SU (the responding unit) to its head SU (the unit it responds to), effectively capturing: Functional dependencies (e.g., an Answer depends on a Question), Rhetorical dependencies (forming discourse pairs between discourse units), and Content dependencies (e.g. a confusing statement raising a question).

Label Set Hierarchy: To improve annotation consistency and agreement on DDA's Discourse Relation label set, we adopt a hierarchical labeling strategy. When annotators encounter ambiguity or struggle to find a fine-grained label that accurately captures the speaker's intention, they are encouraged to revert to a coarser-grained label. This approach ensures greater reliability in labeling while preserving interpretability. Additionally, we analyze agreement and model performance across different hierarchy levels in Sec. 5 and 7.2. We provide the full hierarchy in Appendix B.

3.1 Changes

Over- and Under-segmentation Perfect segmentation is challenging, and we adapt the annotation scheme to accommodate both over-segmented and under-segmented units in our dataset.

Example 1: Merging Over-Segmented Units

```
(1) A: I think that [long pause]
(2) A: we can just connect the two pins.

(2) --(empty)--> (1)
```

In this case, the two Slash Units clearly form a single meaningful unit. To avoid unnecessary fragmentation, we introduce a connection edge to merge them into a single unit, allowing them to share the same functional and rhetorical labels.

Example 2: Annotating Within-SU Relations

```
(3) B: Anyone has an ore?
(4) C: I have one, but you're almost achieving a
      monopoly, so ...

(4) --Conceded--> (4)
```

In SU (4), C states “I have one”, but then concedes “you’re almost achieving a monopoly.” We would like to annotate this ‘Concession’, but standard discourse relation annotation requires at least two distinct SUs. To capture this structure without forcing additional segmentation, we extend the DDA scheme to support within-SU dependencies through self-pointing edges with discourse labels. This approach preserves critical speaker intentions while deferring further segmentation to future processing.

3.2 Domain Specific Conventions

Certain intentions and behaviors fall outside the original scope of the DDA scheme. Below, we summarize the most notable customized coding conventions observed in our corpus:

Reading Aloud: Reading action differs significantly from a regular statement, as the intention is not merely to share information but also to process it or lend authority to an argument. To capture this behavior, we use a combined label of *Statement* + *Self Talk*, with additional underlying intentions encoded based on context. For example, when paired with *Action Directive*, the annotation reflects the intent of using the read material as a suggestion for others’ actions.

Pointing Out Objects: Indicating the location of objects that participants interact with is a common dialogue action in collaborative settings. To represent this, we use a combination of *Action Directive* + *Give Details*, which reflects the speaker’s intent to direct the listener’s attention.

Microphone Testing: Another frequently observed dialogue action in our recordings is microphone testing. We assign the labels *Exclamation* + *Self-Talk* to encode the speaker’s intent when producing noise for this purpose.

For additional discussion and resolutions collected during the annotation process, we refer readers to the discussion forum.[3]

4 Dataset Formation and Quality Control

We manually annotated multi-party conversational data across various domains. In this section, we show the annotation process, an overview of the core characteristics of the dataset, distributions of the DDA labels and how our annotation process ensures the quality of the annotation.

As previously mentioned, all annotations were performed on manually transcribed conversations. However, where available, multimodal cues such as speaker expressions, gestures, gaze patterns, and vocal tones were taken into account to ensure a more comprehensive representation of speaker intentions and actions. We construct DDA through a web-based DDA annotation application.[4] Each session is doubly annotated by trained annotators with linguistics backgrounds.

We employ a systematic adjudication process as our primary strategy for quality control and to address the initially moderate inter-annotator agreement. This process not only resolves minor discrepancies but also enhances overall annotation quality, even though in-depth interpretation and annotation of dialogue data remain inherently challenging. The adjudication process has evolved from a group discussion and voting approach to a more asynchronous collaboration method, improving efficiency while maintaining annotation quality. Initially, group discussions required at least two annotators to deliberate and reach consensus. As annotators became more familiar with the DDA coding scheme, the process transitioned to an asynchronous workflow.

[3] https://github.com/NSF-iSAT/DDA-corpus/discussion-forum

[4] https://dda.colorado.edu/v2/

In the asynchronous adjudication process, once independent annotations are completed, the two annotators attempt to justify their disagreements on DDA edges per SU. This is done through: First, Densely Paraphrasing – REwriting the original utterance to make dialogue acts and discourse structure more explicit while preserving the intended meaning. Second, Documenting Decision Rationale on edges with disagrement. A third annotator then reviews these justifications and casts a deciding vote in the event of a tie. Among all dialogue sessions, Pulley K12 conversations and 3267 out of 4765 SUs in the Sensor Immersion K12 datatset were adjudicated via real-time group discussion, while the reminder of the released dataset was adjudicated asynchronously.

Table 1 presents an overview of the DDA structural patterns and the distribution of labels within the transcribed dialogues. Furthermore, to provide a visualization of the annotated dataset, Figure 2 illustrates the distribution of Dialogue Acts (DA) across the four domains we analyzed. Visualization of discourse relation distributions can be found in Figure 3.

These annotations and visualizations serve as a foundation for further analysis of dialogue flow, discourse coherence, and conversational engagement patterns within multi-party interactions.

5 Inter-Annotator Agreement

Since DDA is not a straightforward classification task but still relies on a finite set of labels, we adapt Krippendorff's Alpha (Krippendorff, 2004) to evaluate inter-annotator agreement for our annotations. Each dialogue consists of a set of SUs, denoted as $\mathcal{U} = \{u_1, \ldots, u_N\}$. Each annotator $a \in \mathcal{A}$ provides an annotation $x_{i,a}$ for each u_i. Here $x_{i,a}$ is a set of edges or labels, depending on the aspect we consider.

We define a distance function $\delta(x, y) \geq 0$ on the annotation domain Ω. Krippendorff's alpha can be defined as:

$$\alpha = 1 - \frac{D_o}{D_e}$$

where

$$D_o = \sum_{u_i \in \mathcal{U}} \sum_{\substack{a,a' \in \mathcal{A} \\ a<a'}} w_{u_i,a,a'} \, \delta\big(x_{u_i,a},\, x_{u_i,a'}\big)$$

stands for observed disagreement with $w_{u_i,a,a'}$ denote a weight parameter, which we assign as 1.

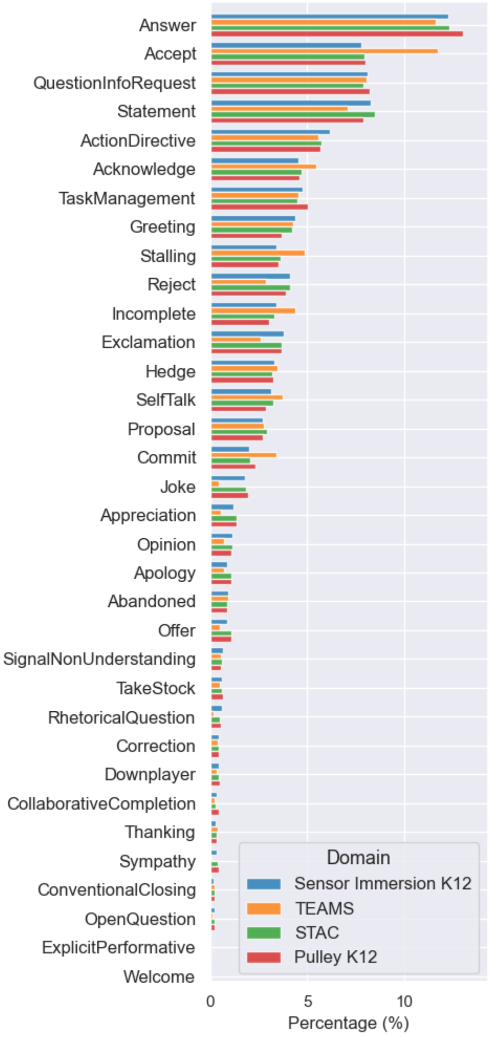

Figure 2: Distribution of Dialogue Act Categories across the four domains. Labels are arranged in descending order of frequency in the corpus.

$$D_e = \sum_{v \in \Omega} \sum_{w \in \Omega} p(v)p(w)\delta(v, w)$$

where $p(v)$ is the empirical proportion of the annotation value $v \in \Omega$ across all (u_i, a) pairs. Formally this is:

$$p(v) = \frac{\sum_{u_i \in \mathcal{U}} \sum_{a \in A} \mathbf{1}[x_{u_i,a} = v]}{\sum_{u_i \in \mathcal{U}} \sum_{a \in A} \mathbf{1}[x_{u_i,a} \text{exists}]}$$

here we consider each annotator's annotation as one "value" from Ω. An annotation can be then defined as:

$$x_{u_i,a} \subseteq \Omega_{\text{joint}} = \{(f, l, t) \mid f, t \in Z, \\ f \geq t, l \in L_{\text{DDALabel}}\}$$

$$\delta_{\text{joint}}(x, y) = 1 - \frac{|x \cap y|}{|x \cup y|}$$

Domain	Edge α	Label $\alpha^{(l_3)}$	Label $\alpha^{(l_2)}$	Label $\alpha^{(l_1)}$	Joint $\alpha^{(l_3)}$	Joint $\alpha^{(l_2)}$	Joint $\alpha^{(l_1)}$
Sensor Immersion K12	0.671	0.513	0.539	0.603	0.384	0.410	0.480
TEAMS	0.638	0.550	0.572	0.648	0.369	0.389	0.456
STAC	0.732	0.572	0.618	0.648	0.458	0.500	0.604
Pulley K12	0.710	0.533	0.548	0.576	0.427	0.443	0.514
overall	0.669	0.532	0.559	0.630	0.388	0.414	0.486

Table 2: Inter-annotator agreement (α) for different domains and different granularity levels of DDA labels. With l_3 being the finest level and l_1 being the coarsest level

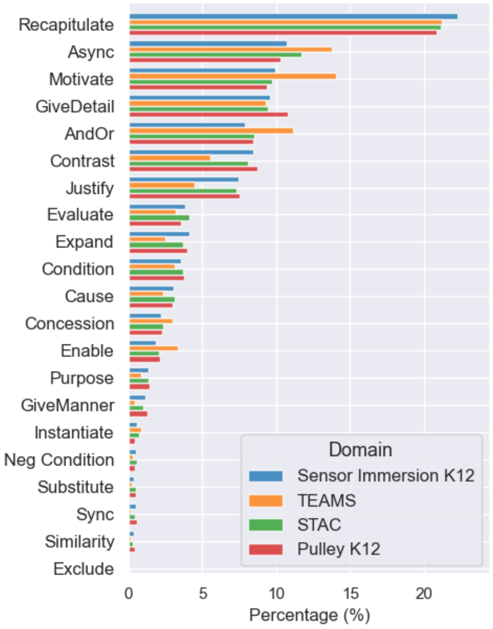

Figure 3: Distribution of Discourse Relation Categories across the four domains. Labels are arranged in descending order of frequency in the corpus.

For joint annotations, we set $\delta_{\text{joint}} = 0$ for exact matches, $\delta_{\text{joint}} = 1$ when both directional and label slots mismatch, and $0 < \delta_{\text{joint}} < 1$ when there are partial matches of either label or edge direction. We use Jaccard set distance for δ. In general, a higher positive α value indicates a greater level of agreement beyond chance between annotators.

To have a better understanding of the intricate annotation task, we incorporate a label hierarchy within the DDA label set. Table 2 presents α values across different domains and levels of label granularity.

6 Downstream Use Cases

The development of the DDA corpus is driven not only by theoretical interest but also by its practical applications. One key use case of the DDA framework is in collaborative learning analysis, where it can be used to assess students' interaction dynamics in group discussions. By aggregating DDA edges across a dialogue session, we can construct a holistic representation of speaker interaction behaviors. Figure 4 illustrates this approach, where each node represents a speaker in a given session, and incoming edges denote DDA response edges directed toward the speaker's utterance, while outgoing edges indicate the opposite. This speaker dynamics graph captures interaction patterns throughout an entire session, both when considering all response dependencies or particular subsets. For example in Figure 4 (a),(b) and (c), considering all response dependencies shows more frequent communication between particular speakers in a group for different domains. In (d1) and (d2), we can isolate subsets of relations to consider, such as those with Forward communicative function (asking a 'Question', making a 'Proposal', etc.) or Backward communicative function (providing an 'Answer', 'Accept', or 'Acknowledgment'). Where (c) shows significant communication between the Pilot and Messenger or Pilot and Engineer, (d1-2) highlight that the Pilot is driving most of the discussion in each case.[5] With an automated DDA parser, such analyses could be generated dynamically, enabling real-time analysis of group discussions.

Another important application of the DDA structure is in dialogue policy development, particularly for improving AI-driven conversational agents. By leveraging DDA annotations, AI systems can infer speaker intentions more effectively, even when users do not explicitly verbalize their thought processes. The structured representation of speaker relations provides a richer context for response generation, facilitating a more transparent, controllable, and interpretable dialogue system.

		Joint			Label			Edge		
Model		F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.
					Sensor Immersion K12					
Llama-3.1-8B-Instruct	*(zero-shot)*	**0.06**	0.06	**0.06**	**0.62**	**0.70**	0.55	**0.58**	0.56	**0.60**
	(few-shot)	0.05	**0.07**	0.04	0.59	0.63	**0.56**	0.53	**0.61**	0.46
Human(reference)		0.63	0.67	0.59	0.86	0.92	0.81	0.89	0.91	0.87
					TEAMS					
Llama-3.1-8B-Instruct	*(zero-shot)*	0.10	0.10	0.09	0.73	0.80	0.66	0.67	0.64	0.69
	(few-shot)	0.06	0.07	0.04	0.75	0.74	**0.76**	0.59	0.67	0.54
gpt-4o	*(zero-shot)*	**0.18**	**0.18**	**0.19**	**0.79**	**0.83**	0.74	**0.75**	**0.67**	**0.86**
Human(reference)		0.66	0.68	0.63	0.90	0.96	0.85	0.88	0.88	0.88
					STAC					
Llama-3.1-8B-Instruct	*(zero-shot)*	0.10	0.12	0.09	0.66	0.80	0.56	0.48	0.46	0.50
	(few-shot)	0.08	0.12	0.06	0.68	0.74	0.64	0.49	0.57	0.43
gpt-4o	*(zero-shot)*	**0.27**	**0.29**	**0.25**	**0.73**	**0.78**	**0.69**	**0.76**	**0.71**	**0.81**
Human (reference)		0.69	0.71	0.68	0.89	0.91	0.87	0.90	0.91	0.89

Table 3: Results for LLM parsing of Dependency Dialogue Acts (DDA) structures, in the Sensor Immersion K12, TEAMS, and test STAC domains, with the best model performance in **bold**. An estimate of human performance computed from our annotation process is given as reference. We find gpt-4o strongly out-performs other methods but falls well short of the human annotator reference.

7 DDA Parsing with LLMs

We evaluate whether LLMs can analyze conversational structure and intents using the DDA coding scheme. To do this, we prompt a variety of proprietary and open-source models with the task of incrementally constructing a DDA parse.

7.1 Experimental Setup

We describe the prompts, models, and evaluation metrics used in our parsing experiments.

Datasets For all of our experiments, we use the dialogues from the Pulley K12 domain as development data or as a source for few-shot examples. Using the remaining domains as test data allows us to measure generalization across shifts in setting and dialogue modality. Specifically, we evaluate on near-domain data from Sensor Immersion K12 classroom dialogues, a different domain in the same spoken modality (TEAMS), and a distant domain in a written modality (STAC).

Prompting Approaches We consider prompting approaches in an incremental setting, inspired by the LLaMIPA SDRT discourse parser (Thompson et al., 2024). For each new set of incoming utterance(s) from a speaker, we predict new edges to add to an on-going DDA parse.[6] We use a chat-formatted prompt which provides general instructions for DDA Parsing as a text-to-code problem, as we find text-to-code helps ensure a structurally valid parse. The system prompt includes the definition of each DDA relation as a simple python class, where all relations have a source and target to indicate the edge direction. We evaluate models ability to parse DDA under two prompt settings. For a 'zero-shot' approach, we give only a definition of each relation type and instructions for forming a parse. For a 'few-shot' approach, we additionally provide a fixed set of $k = 5$ demonstrations randomly sampled from the Pulley K12 domain. To prevent label bias in our sample of examples, we ensure each example demonstrates a unique combination of DDA labels when sampling. Details for each prompt are available in Appendix A.

Models We run experiments using LLaMa 3.1 8B Instruct (Grattafiori et al., 2024) and gpt-4o (OpenAI et al., 2024). To satisfy data use constraints, we evaluate with gpt-4o only on the openly released domains: TEAMS & STAC. For the same reason, we run only the 'zero-shot' approach, as the 'few-shot' setting requires demonstrations from the Pulley K12 domain.

Evaluation Metrics As an overall assessment of parse quality, we compute the F1 over the set of labeled edges in each graph, which we refer to as **Joint F1**. Only a perfect parse scores a Joint F1 of 1. We are also interested in the ability of a parser to

[5] Note that some relations have neither a forward nor backward communicative function. Table 5 details the relations in each group.

[6] To prevent prompting with incomplete dialogue context, we add adjacent slash units if they are from the same speaker.

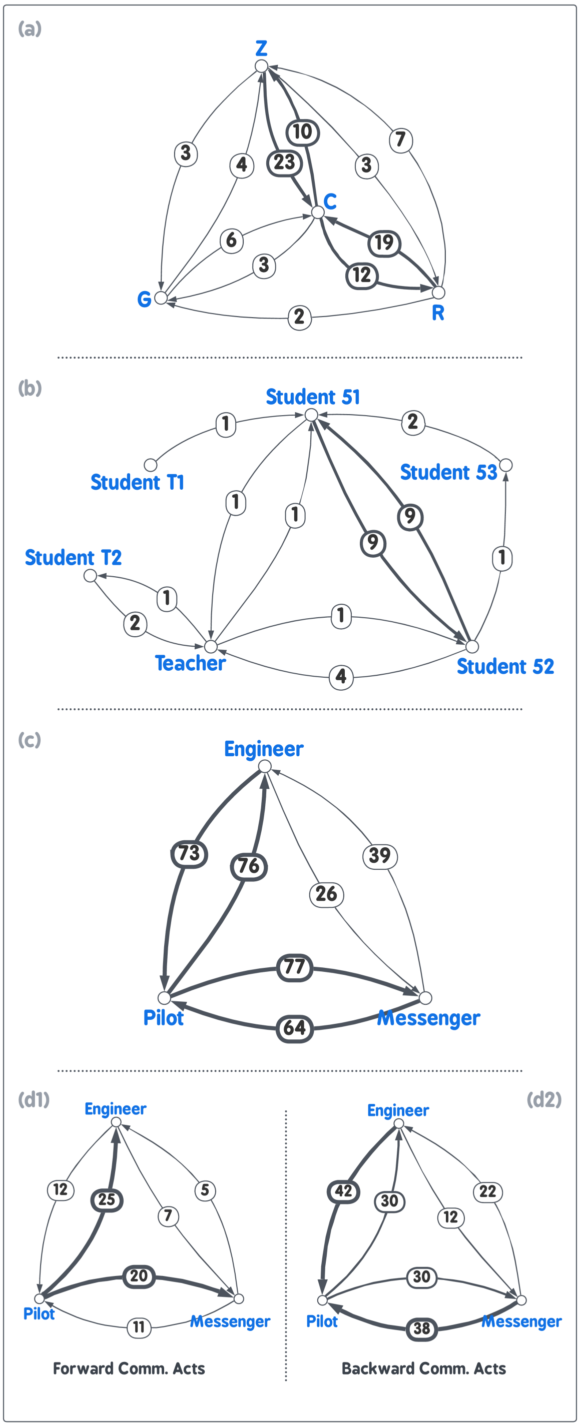

Figure 4: Sample speaker dynamics graphs from the (a) STAC corpus, (b) Sensor Immersion K12, and (c-d) TEAMs corpus arranged from top to bottom

classify the speaker intent of a slash unit, regardless of how it is attached to the graph. To assess this, we compute a **Label F1** over the out-going labels for each slash unit. Finally, to assess structure without consideration of labels, we compute an **Edge F1** over the targets of the unlabeled edges. For each of these F1 measures, we also report the precision and recall. For reference, we estimate the performance of a human annotator from our annotation process as the average of each F1 measure when comparing each annotator's labeling to the gold adjudicated result.[7]

7.2 Results

Table 3 presents our parsing results across the three evaluation domains for each model, compared with our human reference. We find the 'zero-shot' gpt-4o approach significantly outperforms both 'zero-shot' and 'few-shot' LLaMa 3.1. Remarkably, we find the inclusion of few-shot examples does not help, and in some cases even hurts performance. Overall, LLM performance for all approaches falls well short of our estimated human performance. We think a number of factors might explain this. First, human annotators are able to use dialogue video and audio as context when annotating, giving key insights into understanding the dialogue that are not captured in a transcript. Second, annotators communicate with each other about the annotation process, establishing norms and conventions for phenomena discovered in each dialogue domain that are not captured in relation definitions. We leave development of an improved DDA parser that could leverage these features to future work.

8 Conclusion and Future Work

While we produce DDA resources for structured speaker intention modeling and their characteristics and potential impact, several directions remain open for our future exploration. One of our concurrent efforts is to further evaluate the impact of DDA as an intention representation in facilitating dialogue policy learning and response generation. We aim to assess whether structured intention representation improves dialogue coherence and response controllability and will compare DDA-driven models against dialogue act tagging based approaches.

Additionally, we plan to expand DDA annotations to additional datasets, both within our existing domains and in new conversational settings. This expansion will enable cross-domain analysis, helping to identify commonalities and variations in speaker intention patterns across educational, collaborative, and social dialogue contexts.

[7] It is possible this over-estimates human performance since each annotator contributes their draft labeling to the adjudication process, though an un-biased estimate would require costly triple annotation. Despite this short-coming, we think this estimate can be useful for comparisons.

Limitations

The DDA annotation process, though intuitive, is time-intensive and requires trained annotators, limiting scalability for large datasets. While it captures utterance-level dependencies, it lacks higher-order discourse modeling, making it less suited for multi-session dialogues. Enhancing efficiency may require hierarchical discourse structures and context-aware AI integration. Additionally, data access control imposes privacy and compliance constraints, restricting corpus expansion and broader deployment. Future work should explore secure data-sharing protocols to balance accessibility and privacy while maintaining annotation quality.

Ethics Statement

Our research is designed to support deeper evaluation of speaker intention, a process that inherently involves interpretive analysis. Throughout this work, we uphold strict standards of data privacy and ethical responsibility to ensure that all human subject research is conducted with no risk of harm to participants. All data used in this study were collected with explicit informed consent, and we are committed to protecting the privacy and rights of every individual involved.

We further commit to fairness and the conscientious minimization of bias in both our methodology and interpretation. We encourage the broader research community to adopt similarly rigorous standards in the pursuit of safe and responsible language technology development.

Acknowledgment

This research was made possible through support from the NSF National AI Institute for Student-AI Teaming (iSAT), Grant DRL 2019805. Any opinions, findings, or conclusions expressed are solely those of the authors and do not necessarily reflect the views of the NSF. We sincerely thank Dr. Jie Cao and Dr. Marilyn Walker for their generous contributions of time and expertise, which played a significant role in shaping the direction of this project.

References

Angus Addlesee, Weronika Sieińska, Nancie Gunson, Daniel Hernandez Garcia, Christian Dondrup, and Oliver Lemon. 2023. Multi-party goal tracking with LLMs: Comparing pre-training, fine-tuning, and prompt engineering. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 229–241, Prague, Czechia. Association for Computational Linguistics.

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.

John Langshaw Austin. 1975. *How to do things with words*. Harvard university press.

Harry Bunt, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Volha Petukhova, Andrei Popescu-Belis, and David R Traum. 2012. Iso 24617-2: A semantically-based standard for dialogue annotation. In *LREC*, pages 430–437.

Jon Z Cai, Brendan King, Margaret Perkoff, Shiran Dudy, Jie Cao, Marie Grace, Natalia Wojarnik, Ananya Ganesh, James H Martin, Martha Palmer, et al. 2023. Dependency dialogue acts–annotation scheme and case study. *arXiv preprint arXiv:2302.12944*.

Jie Cao, Ananya Ganesh, Jon Cai, Rosy Southwell, E. Margaret Perkoff, Michael Regan, Katharina Kann, James H. Martin, Martha Palmer, and Sidney D'Mello. 2023. A comparative analysis of automatic speech recognition errors in small group classroom discourse. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '23, page 250–262, New York, NY, USA. Association for Computing Machinery.

Mark G Core and James Allen. 1997. Coding dialogs with the damsl annotation scheme. In *AAAI fall symposium on communicative action in humans and machines*, volume 56, pages 28–35. Boston, MA.

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, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky

Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The Llama 3 Herd of Models. *arXiv preprint*. ArXiv:2407.21783 [cs].

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.

Dan Jurafsky. 1997. Switchboard swbd-damsl shallow-discourse-function annotation coders manual. *Institute of Cognitive Science Technical Report*.

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

Klaus. Krippendorff. 2004. *Content analysis : an introduction to its methodology*, 2nd edition. edition. Sage, Thousand Oaks, Calif.

Alex Lascarides and Nicholas Asher. 2007. Segmented discourse representation theory: Dynamic semantics with discourse structure. In *Computing meaning*, pages 87–124. Springer.

Diane Litman, Susannah Paletz, Zahra Rahimi, Stefani Allegretti, and Caitlin Rice. 2016. The teams corpus and entrainment in multi-party spoken dialogues. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1421–1431, Austin, Texas. Association for Computational Linguistics.

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, Singapore and Online. Association for Computational Linguistics.

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.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

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

Sadhana Puntambekar, Dana Gnesdilow, Catherine Dornfeld Tissenbaum, N. Hari Narayanan, and N. Sanjay Rebello. 2021. Supporting middle school students' science talk: A comparison of physical and virtual labs. *Journal of Research in Science Teaching*, 58(3):392–419. Publisher Copyright: © 2020Wiley Periodicals, LLC.

John R Searle. 1979. *Expression and meaning: Studies in the theory of speech acts*. Cambridge University Press.

Chao-Hong Tan, Jia-Chen Gu, and Zhen-Hua Ling. 2023. Is ChatGPT a good multi-party conversation solver? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4905–4915, Singapore. Association for Computational Linguistics.

Kate Thompson, Akshay Chaturvedi, Julie Hunter, and Nicholas Asher. 2024. Llamipa: An incremental discourse parser. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6418–6430, Miami, Florida, USA. Association for Computational Linguistics.

A Parsing Prompts

We present our prompts for incremental DDA parsing as text-to-code. The system prompt is presented in Figure 5, and the instance prompt is presented in Figure 6. The system prompt describes the task and defines all DDA relations and demonstrates the expected output format. The instance prompt contains details for the current dialogue turn, including the dialogue history and the turn for the incremental parse.

level 1	level 2	level 3
Temporal	Asynchronous	Before
		After
	Synchronous	Synchronous
Contingency	Cause	Causing
		Caused
	Motivation	Motivating
		Motivated
	Justify	Justifying
		Justified
	Condition	Conditioning
		Conditioned
	Negative-Condition	NegConditioning
		NegConditioned
	Purpose	Purposing
		Purposed
	Enablement	Enabling
		Enabled
	Evaluation	Evaluating
		Evaluated
Comparison	Contrast	Contrast
	Similarity	Similar
	Concession	Conceding
		Conceded
Expansion	Instantiation	Instantiating
		Instantiated
	Equivalence	Restate/Equal
	Level-of-Detail	GivingDetails
		GivenDetails
	Disjunction	Alternative
	Exception	Excluding
		Excluded
	Conjunction	Conjunction
	Manner	GivingManner
		GivenManner
	Substitution	Substituting
		Substituted

Table 4: Discourse relations adapted from Penn Discourse Treebank 3.0 (Kim et al., 2020)

B DDA Relation Label Hierarchy

Both dialogue acts and rhetorical relations in DDA are hierarchically organized, allowing annotators to use a coarser grain label when a slash unit is ambiguous. This hierarchy also permits leveled analyses of agreement, where finer-grain labels are summarized by their higher-level categories, as in Table 2.

In Table 5 we present the hierarchy of dialogue acts, and in Table 4 we present the hierarchy for discourse relations.

level 1	level 2	level 3
Backward Communicative Function	Answer	Answer
	Agreement	Accept
		Reject
	Understanding	Collaborative Completion
		Appreciation
		Downplayer
		Sympathy
		Acknowledge
		Signal non -understanding
		Correction
communicative status	communicative status	Abandoned
		Stalling
		Self-talk
Forward Communicative Function	Statements	Statement
		Opinion
	Commiting speaker future action	Offer
		Commit
	Influencing addressee future action	Proposal
		Action Directive
		Question Info-request
		Open Question
		Rhetorical Question
	Other forward function	Apology
		Thanking
		Exclamation
		Explicit performative
Information Level	Communication Management	Greeting (conventional opening)
		Conventional closing
		Welcome
	Task	TakeStock
		Task Management
Other	Other	Incomplete
		Hedge
		Joke

Table 5: Dialogue Acts adapt from Switchboard DAMSL scheme

~~~
You are an assistant with an expertise in annotating multi-party conversations using "Dependency Dialogue Acts" or DDA, a new framework combining
       rhetorical relations and dialogue acts. The framework unifies rhetorical relations from PDTB and dialogue acts from Switchboard-DAMSL. You are
       tasked with annotating portions of ongoing conversations. A DDA relation can be defined by a dataclass with a `source` and `target` attribute. `
       source` indicates the index of the speaker's utterance, and `target` indicates the index of the dependent utterance.

```python
@dataclass
class DDARelation:
    source: int # index of the source utterance
    target: int # index of the target utterance
~~~

Each relation is a unique subclass of `DDARelation`. Here is the list of relations and their simplified definitions:

{relations_definitions}

Always be sure to adhere to the following facts and rules:
1. Annotations in DDA are edges which connect a speaker's utterance to a previous utterance using one of the relations. This is called a dependency.
2. An utterance may relate to more than one previous utterance, so be sure to capture all edge dependencies.
3. Dependency edges are directed from speaker (`source`) to each dependent utterance (`target`). Self-pointing edges have `source` == `target`.
4. Edges always point to a previous utterance in the history, or to the source for a self-pointing edge. Therefore, `target` <= `source`.
5. An utterance may relate to the same previous utterance in multiple ways. Each rhetorical or dialogue-act relationship is a separate edge dependency.
6. Be specific: carefully consider the intent of the speaker and choose the relations that best capture this function.
7: Be thorough: speakers often convey multiple intentions in a single utterance, especially when applying rhetoric to facilitate a communicative
       function, requiring multiple edges.
8. Utterances which start a new conversation thread can be annotated with a self-pointing edge (`source` == `target`).
9. Format: You will produce a list of DDA dependency edges in a code-block, prefixed with 'Parse:', for each new set of utterances in the conversation.
10. Format: Use ONLY the above relation labels, do NOT create your own.
11. You will provide no information that does not adhere to this format.

In each turn, you'll be given the next utterance(s) in the conversation, and some contextual information about the ongoing parse. You need to give a
       parse for the new utterances in the dialogue in a code-block prefixed
with 'Parse:', which will be processed by the system to incrementally update the ongoing parse.
DO NOT include edges from other messages! Only include the new edges that are introduced in the new set of utterances.
In the code block, set the variable `edges` to the new edges of your incremental parse. For example:

Parse:
~~~python
edges = [
    QuestionInfoRequest(source=0, target=0)
]
```{example_str}
~~~

Figure 5: System prompt used in parsing experiments. The `relation_definitions` is replaced with a line for each relation and its definition. The `example_str` is replaced with the in-context examples. We found adding in-context examples as historic 'messages' in the prompt degraded performance.

```
Here is a part of a conversation between {n_speakers} speakers. They are {speakers_description}
discussing {conversation_topic}. Please produce a DDA parse:

History:
{history}

Next Turn:
{current_turn}
```

Figure 6: The prompt used in our parsing experiments for each problem instance, following the system prompt. The history is the $k = 12$ most recent utterances in the dialogue and their speakers. The current turn provides new utterances to parse. We format each prompt with short descriptives derived from the domain. For example, all dialogues in the Sensor Immersion K12 domain are parsed with `speakers_description` as 'middle school students' and `conversation_topic` as 'a lab on sensors'