

Discursive Socratic Questioning: Evaluating the Faithfulness of Language Models' Understanding of Discourse Relations

Anonymous ACL submission

Abstract

While large language models have significantly enhanced the effectiveness of discourse relation classifications, it remains unclear whether their comprehension is faithful and reliable. We provide DiSQ, a new method for evaluating the faithfulness of understanding discourse based on question answering. We first employ in-context learning to annotate the reasoning for discourse comprehension, based on the connections among key events within the discourse. Following this, DiSQ interrogates the model with a sequence of questions to assess its grasp of core event relations, its resilience to counterfactual queries, as well as its consistency to its previous responses.

We then evaluate language models with different architectural designs using DiSQ, finding: (1) DiSQ presents a significant challenge for all models, with the top-performing GPT model attaining only 42% of the ideal performance; (2) Open-source models generally lag behind their closed-source GPT counterparts, with notable exceptions being those enhanced with chat and code/math features which demonstrate zero-shot capabilities in discourse comprehension; (3) Our analysis reveals that LLMs exhibit asymmetries in grasping particular relations, validate the effectiveness of having explicitly signalled discourse connectives, the role of contextual information, and the benefits of integrating historical QA data.

1 Introduction

While language models can generate coherent and seemingly human-like text, their true grasp of discourse relations remains unclear. Traditionally, discourse relation prediction has been evaluated using accuracy scores from classification tasks. However, task accuracy may not reflect a reliable understanding, as a high score might not reflect sound reasoning or consistent comprehension of discourse semantics. Drawing inspiration from Socrates' method of examining his students' understanding

Discourse relation: Contingency.Cause.Result

Arg1: When I want to buy, they run from you -- *they keep changing their prices.*

Arg2: *It's very frustrating.*

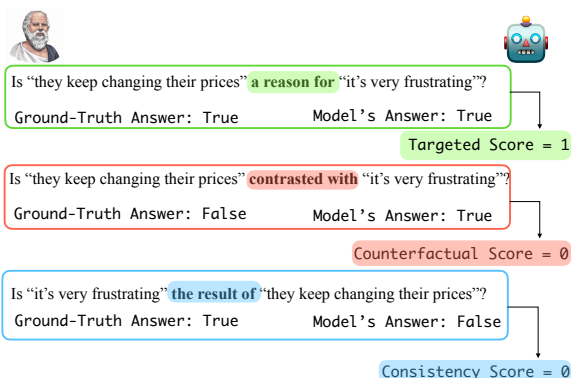


Figure 1: DiSQ combines three discourse-relevant scores: (1) Targeted Score, gauging responses to key events; (2) Counterfactual Score, assessing robustness against irrelevant queries; (3) Consistency Score, measuring logical coherence to equivalent questions.

through a series of questions, we introduce Discursive Socratic Questioning (DiSQ), a new method that assesses a model's understanding of discourse relations by requiring systematic accuracy over multiple questions, rather than just a single accurate prediction (Figure 1).

While QA-based evaluation is well-researched (Fabbri et al., 2022; Hu et al., 2023), DiSQ addresses unique, discourse-centric challenges: **What to Ask:** While many discourse spans can form questions, not all provide *salient* insights into discourse understanding. Previous research like e-SNLI (Camburu et al., 2018) rely on labor-intensive human annotations to extract relevant signals for natural language inference, we advocate the use of in-context learning (ICL). This harnesses the power of large language models to annotate salient discourse signals efficiently. **How to Ask:** The manner in which questions are framed is essential for assessing three key attributes of a model's faithfulness: (1) *Responsiveness to Targets:* We gener-

ate questions centered on pivotal discourse spans with accurate semantics, to which the model should respond affirmatively (e.g., affirming a cause in a contingency relation). (2) *Robustness to Counterfactuals*: We design questions based on counterfactual semantics, expecting the model to abstain or negate the query (e.g., dismissing contrast in a contingency relation). (3) *Logical Consistency*: We formulate Converse questions with equivalent semantics, anticipating the model to deliver consistent responses (e.g., aligning answers to a result question with its corresponding reason). A study (Berglund et al., 2023) finds that LLMs struggle with the “reversal curse” (finding it hard to infer “B is A” from “A is B”). We introduce this facet as a discourse-centric assessment to the broader LLM research, emphasizing models’ logical consistency.

We base our evaluation on the widely-recognized PDTB corpus (Prasad et al., 2008). Initially, we select 11 second-level discourse senses and employ in-context learning to select salient evidence for questioning. Subsequently, we invite human experts to validate our chosen evidence, underscoring the soundness of our questions.

We apply DISQ to a range of models encompassing various architectures and sizes. Notably, many models demonstrate zero-shot capabilities, even without training on discourse-specific data. This suggests that the prevailing training paradigms yield emergent ability to understand discourse semantics. Our findings show that while larger, closed-source models excel in responsiveness, they also struggle with the “reversal curse”, indicating a largely probabilistic approach to discourse semantics without full logical consistency. We highlight the benefits and limits of using linguistic features like discourse connectives, context, and historical QA to enhance comprehension faithfulness¹.

2 Question Bank for DISQ

We detail “*what to ask*” in DISQ by identifying key events and use in-context learning to identify salient evidence. We then perform human verification to guarantee the quality of these questions.

2.1 Preliminaries

What counts as discourse understanding? Organized text makes sense as discourse elements link the text together. Such linking elements are

¹DISQ’s software and data will be publicly available after acceptance.

referred to as cohesive devices (Halliday, 1976), including reference, ellipsis, and lexical cohesion. Formally, two textual spans s_1 and s_2 are linked by the relation r . We define (s_1, s_2, r) as an evidence triple to understand the discourse. Concretely,

Definition 1. (s_1, s_2, r) is an evidence triple to understand the discourse, where Arg_1 and Arg_2 are two given discourse arguments participating in a discourse relation R , and two contiguous spans $s_1 \in Arg_1$ and $s_2 \in Arg_2$ link the two arguments into a coherent discourse with semantic relation r .

We argue that a model understands discourse when it reliably identifies such triples. As shown in Table 1, the event triple (s_{13}, s_{21}, r) is the salient signal for the causal semantics. A model must identify them to understand the *Contingency* discourse relation (R).

Discourse relation (R): Contingency.Cause.Result
Arg_1 : When I want to buy, they run from you – <u>they keep changing their prices</u>
Arg_2 : <u>It’s very frustrating</u>
s_{11} : I want to buy;
s_{12} : they run from you;
s_{13} : <u>they keep changing their prices</u>
s_{21} : <u>It’s very frustrating</u>
Salient signals: (s_{13}, s_{21}, r) , r is “the reason for”.
Targeted questions: Is s_{13} the reason for s_{21} ?
Counterfactual question: Does s_{13} contrast against s_{21} ?
Converse question: Is s_{21} the result of s_{13} ?

Table 1: DISQ formalizes discourse understanding as question answering (QA).

Define a proxy for discourse understanding: We approach the notion of understanding by questioning. We interrogate the model with a set of questions concerning different semantic relations and text spans. If a model is said to understand, it must answer questions in a manner under three criteria: (1) **Responsiveness** to targeted questions (e.g., providing affirmative “True” answers, without abstaining); (2) **Robustness** against counterfactual queries (e.g., responding with “False” or abstaining to answer); (3) **Consistency** across consecutive responses (e.g. consistently saying “True” (or “False”) to converse questions).

2.2 Annotating Salient Signals Using ICL

As illustrated in Table 1, not all spans serve as salient signals for understanding the discourse. For instance, the span “I want to buy” (s_{11}) lacks a causal connection with “It’s very frustrating” (s_{21}). It is important to filter unessential ones to make

our evaluation reliable. Earlier research in explainable NLP, such as e-SNLI (Camburu et al., 2018), adopts a similar span-based reasoning formalization. In their corpus, humans are tasked with annotating key spans pivotal to understanding the NLI labels. However, this method is labor-intensive and lacks generalizability to other tasks.

To address this challenge, we introduce a new annotation method requiring minimal human intervention. **(1) Candidate extraction:** We first extract m spans from Arg_1 and n spans from Arg_2 , resulting in $m \times n$ evidence triple (s_1, s_2, r) that may act as evidence for discourse understanding. These spans, characterized by a self-contained subject-verb-object (s-v-o) structure (identified by semantic role labeling (Gardner et al., 2018)), can be regarded as *events*. However, only some of them are salient indicators of discourse relation. **(2) Select salient pairs:** Subsequently, we leverage In-context Learning (ICL) (Brown et al., 2020; Liu et al., 2022) to identify these pivotal evidence triples. The underlying premise is that by providing the model with exemplars of salient versus non-salient triples, it can distinguish them in new instances. Formally, we ask model to predict if r holds between events (s_1, s_2) in a discourse Arg_1, Arg_2, R with in-context learning:

$$P(y|C, x) = \prod_{t=1}^T P(y_t|D, x, y_{<t}) \quad (1)$$

where y is the output sequence of length T , x is the test input, and $D = \{x_1, y_1, \dots, x_k, y_k\}$ is the demonstration containing k in-context examples.

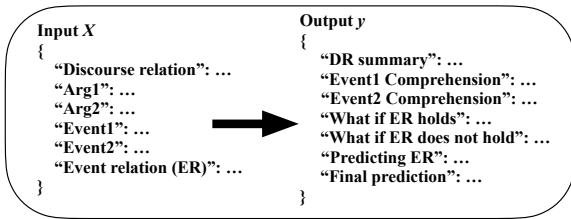


Figure 2: The input and output for in-context learning for selecting salient signals.

Figure 2 presents the in-context learning (ICL) template in JSON, designed for step-by-step reasoning to predict event relation (ER). The ICL performs a binary classification on whether ER holds. **The output y** mirrors human reasoning and includes: “*DR Summary*”, condensing the discourse relation (DR) in model-specific terms; “*Event1 comprehension*”, linking Event1 to argument Arg_2

(and Event2 to Arg_1) and examining their discourse roles; “*What if ER holds*” and “*What if ER does not hold*” exploring event relation (ER)’s influence on the discourse; and “*Predicting ER*” followed by the “*Final prediction*”, synthesizing the analysis to predict ER between events. We implement ICL using the LLaMA2-13B model, employing just one positive and one negative example for each discourse relation (Appendix A.1). Concretely, for 11 Level-2 discourse relations, we create 22 examples in total.

2.3 Dataset Statistics

Discourse relation (R)	Event relation (r)	Ratio	# of Q
Comparison.Concession	deny or contradict with	0.617	1,392
Comparison.Contrast	contrast with	0.596	708
Contingency.Reason	reason of	0.69	5,124
Contingency.Result	result of	0.735	2,268
Expansion.Conjunction	contribute to the same situation	0.964	4,512
Expansion.Equivalence	equivalent to	0.791	408
Expansion.Instantiation	example of	0.792	2,148
Expansion.Level-of-detail	provide more detail about	0.739	3,696
Expansion.Substitution	alternative to	0.714	180
Temporal.Asynchronous	happen before/after	0.696	1,236
Temporal.Synchronous	happen at the same time as	0.771	768
Total / Avg		0.737	22,400

Table 2: **Dataset Statistics:** Ratio of confirmed event relations, event pair counts, and number of resultant questions for each discourse category.

DiSQ integrates the 11 Level-2 relations from the PDTB-3.0 sense taxonomy, with tailored adjustments in the Contingency relation through merging minor groups (see Appendix A.2). We annotated all implicit discourse in the test set (Sections 21 and 22) following the guidelines from (Webber et al., 2019). For counterfactual analysis, we chose 5 irrelevant r not pertaining to a particular discourse R . Additionally, both targeted and counterfactual questions include converse inquiries, amounting to 22,400 questions in total. Table 2 shows a high satisfaction ratio for Expansion.Conjunction event relations, likely due to most pairs aligning with a common theme. In contrast, Comparison and Contingency relations exhibit lower ratios, suggesting that only central events effectively highlight causal or contrasting semantics in these categories.

2.4 Human Verification

To guarantee the reliability of DiSQ evaluations, we verify the salient event pairs identified by ICL

with humans. Importantly, our verification design contrasts with many evaluation metrics studies which align a system’s end output score with human evaluations (Fabbri et al., 2021a). Rather, we directly evaluate the inherent logic of the benchmark. Upon verification of the questions, subsequent steps are transparent, deterministic, and verifiable — all important characteristics for any measure.

	A1&A2	A1&ICL	A2&ICL
Agreements	85.2%	85.2%	83.7%

Table 3: The agreements between two annotators (A1 and A2) and the semi-automated ICL method.

Verification Results Annotators evaluated event pairs identified as significant by ICL, analyzing a discourse instance $(Arg1, Arg2, R)$ with two events s_1 and s_2 , and a potential relation r , to verify if r exists between s_1 and s_2 . For this task, two NLP-specialized graduate students annotated 61 event relation instances. To facilitate their comprehension, instances requiring extensive domain knowledge, like finance, were excluded after random sampling. Following a basic discourse semantics tutorial, they started annotation (detailed in Appendix B). The alignment between our model’s responses and human annotator agreement was encouragingly similar in Table 3, affirming our ICL method’s effectiveness in selecting salient event pairs for discourse understanding.

3 Discursive Socratic Questioning for Evaluation

Having confirmed the validity of salient events in discourse comprehension, we now consider our measure’s core as established. Now we outline our systematic approach to “how to ask”: generating questions, querying models with these questions, and subsequently computing the scores.

3.1 Question Generation

Type	Formalization	Expected Answer	Score
Targeted	$Q_t = \{QG(s_1, s_2, r)\}$	True	s_t
CF	$Q_c = \{QG(s_1, s_2, r')\}$	False	s_{cf}
Converse	$\tilde{Q}_t = \{QG(s_2, s_1, \tilde{r})\}$	Equals to original	s_{con}

Table 4: Formalization of three question types and their yielding scores: Targeted Score s_t , Counterfactual Score s_{cf} , and Consistency Score s_{con} .

We generate three types of questions (Table 4):
(1) Targeted Questions: Ground truth answers for Q_t are always affirmative, as they tap into the salient signals. We employ a rule-based question generator (QG) to weave events into a cohesive query. As an example, for Contingency.Result, where r denotes “the reason for”, a typical question might be “Is s_1 the reason for s_2 ?”
(2) Counterfactual (CF) Questions: These gauge model robustness, as their answers are negative, due to the event relation being altered into a counterfactual r' . For instance, “Is s_1 contrasted against s_2 ?” is unrelated to contingency discourse.
(3) Converse Questions: These test a model’s response consistency to the logically-equivalent converse question. For example, “Is s_2 the result of s_1 ?” (\tilde{r}) corresponds to the earlier question about *reason* (r). We anticipate consistent responses from LMs.

3.2 Question Answering

Algorithm 1 DISQ interrogates a language model.

- 1: **Input:** Discourse d and its corresponding question series Q .
- 2: $\mathcal{H} = \{\emptyset\}$ \triangleright The history is initialized.
- 3: **Stage 1: Targeted and Counterfactual QA**
- 4: **for** q_i in Q_t and Q_c **do**
- 5: $a_i = \text{LM}(q = q_i, c = d)$ \triangleright The model performs QA. The context c is the discourse d .
- 6: $\mathcal{H} \leftarrow (q_i, a_i)$ \triangleright The history is updated.
- 7: **end for**
- 8: **Stage 2: Converse QA**
- 9: **for** (q_i, a_i) in \mathcal{H} **do**
- 10: $\tilde{q} = \text{Lookup}(q, \{\tilde{Q}_c, \tilde{Q}_t\})$ \triangleright Look up the converse question in converse question sets.
- 11: $\tilde{a}_i = \text{LM}(q = \tilde{q}_i, c = d, (q_i, a_i) \in \mathcal{H})$ \triangleright The model executes QA on the converse question, \tilde{q}_i , optionally utilizing the previous response (q_i, a_i) as supplemental context.
- 12: $\mathcal{H} \leftarrow (\tilde{q}_i, \tilde{a}_i)$ \triangleright The history is updated.
- 13: **end for**
- 14: **Output:** \mathcal{H}

Questioning is divided into two stages (Algorithm 1). In the first stage, DISQ interrogates the model with targeted questions Q_t (expecting a positive answer) and counterfactual questions Q_c (expecting negative). The context is the discourse d : only Arg_1 and Arg_2 . Note that we do not inform the model of the discourse relation; the model must infer the discourse relation to answer correctly. The answer to each question updates the history \mathcal{H} . In

the second stage, DISQ performs converse QA to test the model’s consistency. For each converse \tilde{q}_i , we look for model’s response to the original question and can choose to reuse it to promote the consistency (we find that this choice affects performance significantly; see §4.2).

3.3 DISQ Score

We gauge a model’s overall proficiency by its **DISQ Score**², which combines three scores, as coefficients of a product: $s_{disq} = s_t \times s_{cf} \times s_{con}$.

DISQ then comprises of (1) **Target Score** (s_t); (2) **Counterfactual Score** (s_{cf}) — assessing the accuracy of the model’s answers to targeted and counterfactual questions; and (3) **Consistency Score** (s_{cs}) — evaluating the model’s consistent responses to a question and its converse:

$$s_t = \frac{1}{N} \sum_{i=1}^N \mathbb{1}[a_i = True], q_i \in \{Q_t, \tilde{Q}_t\} \quad (2)$$

$$s_{cf} = \frac{1}{N} \sum_{i=1}^N \mathbb{1}[a_i = False], q_i \in \{Q_c, \tilde{Q}_c\} \quad (3)$$

$$s_{con} = \frac{1}{N} \sum_{i=1}^N \mathbb{1}[a_i = \tilde{a}_i], q_i \in Q, \tilde{q}_i \in \tilde{Q} \quad (4)$$

DISQ uses a product since we favor balanced individuals scores, which may not occur with a sum aggregate — c.f., (0.6, 0.6, 0.6) vs. (0.9, 0.9, 0).

4 Evaluations

Models: While any language model can be assessed, our evaluation targets two of the current strongest large language models (LLMs): (1) **Closed-Sourced Models:** GPT-4 and GPT-3.5-turbo. To manage costs, we limit evaluations to 20% of our test samples. Despite this constraint, we noted stable performance throughout our experiment. (2) **Open-Sourced Models:** LLaMA-2 (Touvron et al., 2023) and its variations, Vicuna (Chiang et al.) and Wizard (Xu et al., 2023). Vicuna is a distilled model that refines LLaMA-2 using user interactions with GPT models; and Wizard is specifically designed for complex code and math instruction adherence, which we felt might help with discourse explanation. We used 13B sized models for all of these but also investigated a smaller 7B LLaMA-2 model for scaling effects.

²Named after our method DISQ, this term is also used to denote our measurement score when unambiguous.

Inference: We initially adopt a zero-shot approach for inference to simulate natural evaluation settings. Following Zhao et al. (2021), we prompt models with concise instructions, context, and questions (Appendix C.3), assessing their immediate reactions. Noticing challenges in task comprehension for some models (hallucinate irrelevant tokens), we also conduct two-shot experiments to acquaint them with the task’s format. We cap a maximum of two shots to prevent models from learning the task.

4.1 Main Results

Our main experiments focus on two key questions: (1) How do models perform on DISQ’s three scores overall? (2) What impact do different discourse relations have on model performance?

Figure 3 displays the overall performance of various models in a zero-shot setting, leading to the following observations on Question 1: **(A) GPT Models Show Room for Improvement** (1st row): None of the models achieve an ideal score (all three scores at 1.0), indicating room for growth. Closed-source models outperform, highlighting a significant development opportunity for open-source language models and affirming DISQ as a challenging benchmark. GPT-4 and GPT-3.5, while similar in DISQ, display distinct behaviors. GPT-4 tends toward more negative responses, hence a higher Counterfactual Score, whereas GPT-3.5 shows a more balanced approach. This difference may be due to varying alignment strategies with human reasoning, yet their comparable DISQ suggest similar levels of discourse understanding. **(B) Only LLaMA’s Chat Variants have (zero-shot) Discourse Capacity** (2nd row): The standard LLaMA 7B and 13B models perform poorly, scoring high in Target and Consistency Scores, but nearly zero in Counterfactual Score. This indicates a tendency to respond affirmatively without comprehending the task from instructions, thus not answering the questions faithfully. The chat-enhanced versions of LLaMA show improvement, with 13B-Chat outshining 7B-Chat, but still falling short of GPT models. **(C) Vicuna and Wizard Models Show Mixed Results** (3rd row): Even though finetuned on user’s shared conversation and complex instructions, Vicuna and Wizard models have the same problem with their LLaMA base model that always tend to answer affirmatively (or not). Wizard-Code and -Math are notable exceptions, presenting zero-shot

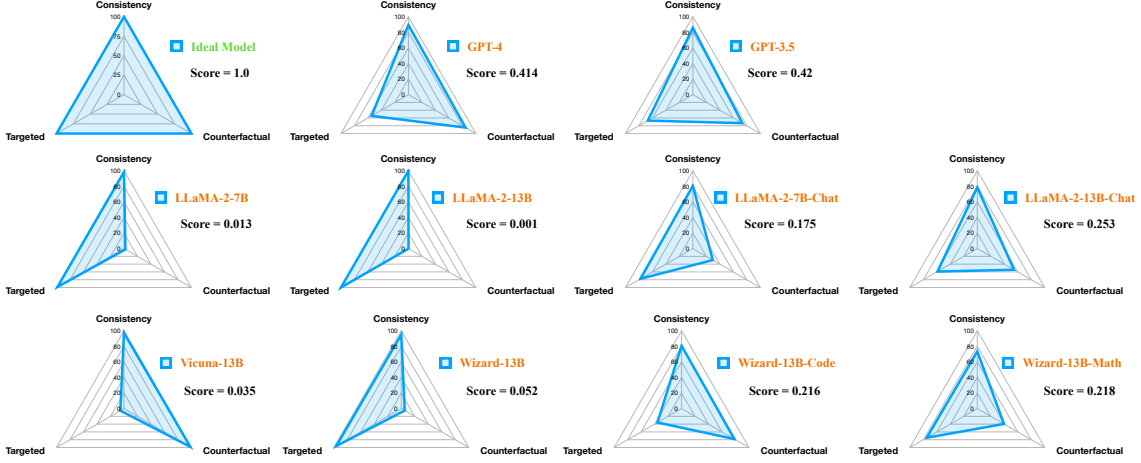


Figure 3: **Overall performance on DiSQ:** This table showcases the overall performance of state-of-the-art GPT models (1st row), LLaMA-2 models with 7B/13B parameters and their chat variants (2nd row), and specialized Vicuna and Wizard models that are further fine-tuned from the LLaMA architecture (3rd row).

	Models	Overall	Comp.Conce	Comp.Contra	Cont.Reason	Cont.Result	Exp.Conj	Exp.Equiv	Exp.Inst	Exp.Detail	Exp.Subst	Temp.Async	Temp.Sync
0 shot	Baseline 1. GPT-3.5	0.42	0.214	0.869	0.21	0.4	0.62	0.239	0.529	0.43	0.476	0.61	0.355
	GPT-4	99%	25%	65%	57%	88%	98%	80%	125%	112%	89%	113%	0%
	2A. LLaMA2-7B	3%	5%	0%	8%	5%	3%	0%	2%	3%	5%	1%	3%
	3A. LLaMA2-7B-Chat	42%	108%	50%	63%	44%	35%	44%	23%	35%	42%	18%	11%
	4A. LLaMA2-13B	0%	0%	1%	0%	1%	0%	0%	0%	0%	0%	1%	0%
	5A. LLaMA2-13B-Chat	60%	89%	55%	61%	43%	47%	66%	61%	87%	61%	32%	7%
	6A. Vicuna-13B	8%	5%	4%	6%	11%	4%	0%	20%	8%	0%	10%	0%
	7A. Wizard	12%	53%	10%	17%	11%	10%	17%	11%	6%	12%	2%	15%
	8A. Wizard-Code	51%	12%	26%	76%	68%	17%	3%	53%	74%	38%	90%	29%
	9A. Wizard-Math	52%	72%	33%	89%	55%	31%	36%	47%	78%	42%	26%	33%
2 shot	2B. LLaMA2-7B	10%	0%	1%	0%	1%	2%	0%	2%	28%	11%	37%	4%
	3B. LLaMA2-7B-Chat	35%	66%	25%	56%	37%	29%	24%	25%	34%	33%	23%	29%
	4B. LLaMA2-13B	60%	61%	20%	79%	64%	56%	50%	41%	44%	22%	90%	82%
	5B. LLaMA2-13B-Chat	47%	126%	35%	65%	39%	34%	52%	39%	54%	71%	27%	26%
	6B. Vicuna-13B	67%	160%	60%	87%	67%	47%	13%	54%	70%	62%	63%	54%
	7B. Wizard	24%	67%	17%	40%	23%	18%	26%	19%	21%	19%	15%	39%
	8B. Wizard-Code	72%	40%	15%	129%	92%	39%	7%	80%	86%	46%	78%	129%
	9B. Wizard-Math	1%	4%	1%	4%	3%	1%	1%	1%	1%	1%	0%	3%

Table 5: **DiSQ Score s_{disq} vs. GPT-3.5:** We present models’ relative performance (in percentage) to GPT-3.5 baseline (reported as its DiSQ Score). Open-source models enhanced with two-shot inferences for task adaptation (mean of 3 random sample sets is reported). Scores above 80% of GPT-3.5’s are in **bold**.

capacity. It is potentially due to their training in coding and math, which may impart logical reasoning skills.

To answer Question 2, Table 5 provides performance per discourse relation. These results yields several intriguing insights: **(1) Persistent Challenge of Minority Classes for LLMs:** Historically, minority classes have posed difficulties for supervised methods (Kim et al., 2020), and this trend continues with LLMs. Classes like Comp.Concession, Exp.Equivalence, and Temp.Synchronous remain challenging, evidenced by DiSQ score at or below 0.3, suggesting that merely increasing model and data samples do not yield comprehensive discourse understanding. **(2) Benefits of Code and Math Training for Discourse Understanding:** A significant performance

boost is observed in Wizard-Code and -Math models under zero-shot conditions. Notably, Wizard-Code (Row 8B) reaches a DiSQ of **72%** of GPT-3.5 under the 2-shot setting, outperforming all other open-source models. **(3) Varied Impacts of Two-Shot Demonstrations:** While intended to acclimate models to the task format, two-shot demonstrations yield mixed results. Some models, like Vicuna and Wizard-Code, show considerable improvement over zero-shot settings, closing the gap with GPT-3.5. Conversely, LLaMA chat variants (Rows 3B and 5B) and Wizard-Math (Row 9B) see performance dips, possibly due to their existing QA strengths and unintended biases introduced by new examples. For example, comparing Rows 9A and 9B, we see uniform performance degradation, from 52% to 1% of the baseline GPT-3.5. **(4) Task**

Difficulty Asymmetry: An intriguing pattern is Contingency. Reason consistently outscoring Contingency. Result across all models, despite both addressing causality. Similar trends are noted in other fine-grained relations like Temp.Async.Precedence and Temp.Async.Succession (Appendix C.4), indicating a potential asymmetry in semantic processing by language models.

4.2 Discussions

We are driven to find what can improve the faithfulness of discourse understanding, especially for open-source models. We examine the following discussion questions (DQs):

DQ1: Is the presence of a discourse connective beneficial? Discourse connectives transform implicit discourse into explicit forms, enhancing comprehension (Kurfali and Östling, 2021). Does this extend to LLM comprehension? In short, yes.

	Models	Overall	Exp.	Cont.	Comp.	Temp.
Baseline	GPT-3.5	0.420	0.510	0.268	0.308	0.565
w/o Conn	LLaMA2-13B-Chat	60%	63%	55%	95%	26%
	Wizard-Code	51%	42%	78%	27%	69%
w/ Conn	LLaMA2-13B-Chat	65%	64%	70%	107%	27%
	Wizard-Code	62%	48%	100%	60%	68%

Table 6: **DQ1:** Models’ DiSQ Scores with the help of discourse connective (reported as percentage relatives to GPT-3.5’s vanilla result).

We selected LLaMA2-13B-Chat and Wizard-Code as representative open-source models. We then inserted PDTB connectives at the start of Arg2, and conducted DiSQ. Table 6 reveals that connectives benefit both models, increasing DiSQ by 5% to 11%. For example, it elevates Wizard’s overall performance from 51% to 62%. They prove particularly advantageous for Contingency and Comparison relations, with some scores surpassing GPT-3.5’s standard results (which lack connectives). This suggests that correctly inferring connectives significantly enhances the accuracy of LMs’ discourse comprehension.

DQ2: Does Broader Context Enhance Comprehension? Next, we assess the influence of surrounding context on discourse comprehension. Does it enhance LLM’s faithfulness? In short, yes.

In the PDTB corpus, texts are segmented into paragraphs. Accordingly, we provide the models with the local paragraph surrounding the discourse arguments. Table 7 demonstrates that models exhibit overall performance enhancements when contextual information is integrated. For instance,

	Models	Overall	Exp.	Cont.	Comp.	Temp.
Baseline	GPT-3.5	0.420	0.510	0.268	0.308	0.565
w/o Context	LLaMA2-13B-Chat	60%	63%	55%	95%	26%
	Wizard-Code	51%	42%	78%	27%	69%
w/ Context	LLaMA2-13B-Chat	70%	73%	84%	64%	29%
	Wizard-Code	58%	46%	99%	45%	60%

Table 7: **DQ2:** Models’ DiSQ Scores with the help of context (GPT-3.5’s vanilla result as baseline).

LLaMA’s overall performance has improved from 60% to 70%. We observe that for discourse relations showing improved performance, the primary factor is a significant increase in Targeted Score, accompanied by only minor changes in the Counterfactual Score, as detailed in Appendix C.5. This suggests that models indeed benefit from additional context, particularly in positively responding to targeted questions.

DQ3: Is QA History beneficial for consistency? Open-source LMs typically achieve an 80% Consistency Score without accessing their own QA history. We explore whether models exhibit greater consistency when referencing their previous QA interactions. In this process, while posing converse questions, we include the history of corresponding targeted and counterfactual questions, along with the model’s responses, in the input. The ideal outcome is for the model to make consistent predictions.

	w/o history	w/ history
LLaMA2-13B-Chat	78.6	70.1
Wizard-Code	81.6	99.8

Table 8: **DQ3:** Models’ Consistency Scores with the insertion of QA history.

Table 8 reveals mixed results: Wizard-Code nearly achieves perfect consistency when referencing its history, whereas LLaMA2-13B-Chat shows a decrease in consistency.

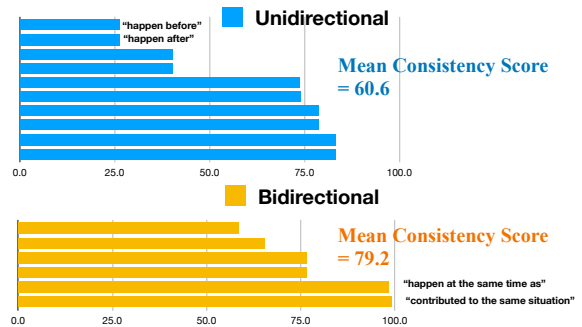


Figure 4: **DQ3:** LLaMA’s consistency score w.r.t. relations. Bidirectional relations have higher scores.

Figure 4 offers an intuitive breakdown, interpreting LLaMA’s Consistency Scores by question type. A noticeable trend emerges: questions with lower Consistency Scores involve unidirectional relations (e.g., “happen before”), while those with higher Consistency Scores tend to be bidirectional (e.g. “happen at the same time as”, though also similarly a temporal relation). This pattern suggests that LLaMA-13B-Chat might focus mainly on literal keywords, lacking in deeper reasoning abilities for question answering. In contrast, Wizard-Code’s code-based training appears to have bolstered its logical reasoning skills, contributing to its improved performance.

5 Related Work

QA for NLP Tasks: While question answering (QA) has been widely applied in various NLP tasks, its primary use has been in annotating golden data and parsing, which . Notable efforts include QASRL (FitzGerald et al., 2018), QANorm (Klein et al., 2020), QADiscourse (Pyatkin et al., 2020), QASem (Klein et al., 2022), QA for reference/ellipsis resolution (Hou, 2020; Aralikatte et al., 2021), and QA for commonsense reasoning (Shwartz et al., 2020). The recent Question Under Discussion (QUD) framework (Ko et al., 2022a; Wu et al., 2023) also employs QA for discourse parsing. However, their focus is on parsing rather than utilizing QA for evaluating faithfulness.

Evaluation Methods in NLP: Recent developments in evaluating and interpreting LMs include several approaches: (1) **Probing paradigm** takes out the representation of LMs and train a model to predict whether one linguistic property is captured by the representation (Tenney et al., 2019; Wallace et al., 2019; Li et al., 2021). (2) **Behavior analysis and post-hoc interpretation** produce fine-grained interpretation of model’s output. The common practice is to perturb the text to reveal the decision boundary or unwanted bias of the model (Belinkov et al., 2020; Ribeiro et al., 2016; Poliak et al., 2018; Rudinger et al., 2018). But the creation of the perturbation usually requires manual efforts. (3) **QA-based Evaluation** offers a transparent and granular approach (Hu et al., 2023; Fabbri et al., 2021b), yet its application in evaluating discourse faithfulness remains unexplored.

Discourse Modeling and Evaluation: (1) **Discourse Modeling:** Language Models (LMs) serve

as the core for custom neural networks to predict discourse relations (Liu et al., 2020; Jiang et al., 2021; Zhou et al., 2022; Xiang et al., 2022; Chan et al., 2023b; Wang et al., 2023). These approaches show improvements over traditional feature-based methods (Pitler et al., 2009; Rutherford and Xue, 2014) but lack in interpretability. Additionally, LMs are applied in coherence modeling (Mohiuddin et al., 2018; Jwalapuram et al., 2022) and hierarchical discourse parsing (Huber and Carenini, 2022; Ko et al., 2022b), yet they often fall short in interpretability and robustness evaluation. (2) **Discourse Evaluations:** Recent benchmarks have moved beyond traditional treebanks like PDTB (Webber et al., 2019). DiscoEval by Chen et al. (2019) assesses sentence embeddings across various discourse tasks. Wu et al. (2023) evaluate QUD parsers, and Chan et al. (2023a) analyze ChatGPT’s capabilities in diverse discourse tasks. However, none of these studies focus on the faithfulness aspect of LMs.

6 Conclusion and Future Work

LMs have been very successful at surface-level tasks, it is important to study deeper tasks like discourse comprehension. Due to the lack of attention, little progress has been made towards evaluating how they understand discourse in a faithful manner. We contribute DISQ, the first systematic evaluation for faithful discourse comprehension. To guarantee the question quality, we leverage in-context learning, utilizing LLMs’ potential to annotate salient event evidence. After human verification of our questions, we extensively test various LLMs against DISQ. DISQ assesses three detailed scores: Targeted Score, Counterfactual Score, and Consistency Score. We posit that a model must excel in all three aspects to demonstrate true discourse comprehension. Our findings reveal that even the most advanced models like GPT-4 and GPT-3.5 are not flawless, and their open-source counterparts lag further behind. We discover that discourse connectives and contextual information can boost model performance, and only models with enhanced logical reasoning benefit from referencing their QA history. In the future, we aim to extend our analysis to longer-range discourse, distilling knowledge from larger models to benefit smaller models, and incorporate additional discourse annotation frameworks beyond PDTB.

Ethical Considerations and Limitations

Our annotation of human verification is approved by institutional review board (IRB). IRB reviewed on our experimental design and research procedures, to ensure that the research involves no more than minimal risks to the research participants. In particular, we ensure that none of the phrases involve sensitive topics or would not elicit strong negative responses. We also ensure that research participants' privacy and the confidentiality of their research data will be protected.

When performing DISQ, we note that output answers may be offensive in certain contexts, because the model can response True/False to any question. This is a common concern for all LMs to overcome, not specific to DISQ. But according to our pilot study, we have not found any cases of such offensive Q&A pairs.

DISQ also has particular limitations. (1) We only use the behavior of the model given a set of questions as a proxy for understanding. It is not a causal analysis. We can causally study the role of individual neuron or subnetwork for discourse function in the future, similar to a recent study about individual neuron's role for factual knowledge (Meng et al., 2022). (2) We have only studied standard English corpora. It is meaningful to apply DISQ to LMs' understanding of discourse on other English corpora with language variations and to corpora in other languages. (3) Our research has been exclusively focused on the Penn Discourse Treebank (PDTB) for discourse annotation. Nevertheless, adapting our workflow to other frameworks, such as Rhetorical Structure Theory (RST) is also feasible.

References

- Rahul Aralikatte, Matthew Lamm, Daniel Hardt, and Anders Søgaard. 2021. Ellipsis resolution as question answering: An evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 810–817.
- Yonatan Belinkov, Sebastian Gehrmann, and Ellie Pavlick. 2020. Interpretability and analysis in neural nlp. In *Proceedings of the 58th annual meeting of the association for computational linguistics: tutorial abstracts*, pages 1–5.
- Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2023. The reversal curse: LLMs trained

on "a is b" fail to learn "b is a". *arXiv preprint arXiv:2309.12288*.

- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31.
- Chunkit Chan, Jiayang Cheng, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. 2023a. Chatgpt evaluation on sentence level relations: A focus on temporal, causal, and discourse relations. *arXiv preprint arXiv:2304.14827*.
- Chunkit Chan, Xin Liu, Jiayang Cheng, Zihan Li, Yangqiu Song, Ginny Y Wong, and Simon See. 2023b. Discoprompt: Path prediction prompt tuning for implicit discourse relation recognition. *arXiv preprint arXiv:2305.03973*.
- Mingda Chen, Zewei Chu, and Kevin Gimpel. 2019. Evaluation benchmarks and learning criteria for discourse-aware sentence representations. *arXiv preprint arXiv:1909.00142*.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. 2023. URL <https://lmsys.org/blog/2023-03-30-vicuna>, 1(2):3.
- Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021a. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.
- Alexander R Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2021b. Qafacteval: Improved qa-based factual consistency evaluation for summarization. *arXiv preprint arXiv:2112.08542*.
- Alexander Richard Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. Qafacteval: Improved qa-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601.
- Nicholas FitzGerald, Julian Michael, Luheng He, and Luke Zettlemoyer. 2018. Large-scale qa-srl parsing. *arXiv preprint arXiv:1805.05377*.
- Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018.

672	Allegnlp: A deep semantic natural language processing platform. arXiv preprint arXiv:1803.07640 .	726
673		727
674	MAK Halliday. 1976. Cohesion in english. Longman .	728
675	Yufang Hou. 2020. Bridging anaphora resolution as question answering. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics , pages 1428–1438.	729
676		730
677		731
678		
679	Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. 2023. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. arXiv preprint arXiv:2303.11897 .	732
680		733
681		734
682		735
683		736
684	Patrick Huber and Giuseppe Carenini. 2022. Towards understanding large-scale discourse structures in pre-trained and fine-tuned language models. NAACL .	737
685		738
686		
687	Feng Jiang, Yaxin Fan, Xiaomin Chu, Peifeng Li, and Qiaoming Zhu. 2021. Not just classification: Recognizing implicit discourse relation on joint modeling of classification and generation. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing , pages 2418–2431.	739
688		740
689		741
690		742
691		
692		
693	Prathyusha Jwalapuram, Shafiq Joty, and Xiang Lin. 2022. Rethinking self-supervision objectives for generalizable coherence modeling. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) , pages 6044–6059.	743
694		744
695		745
696		746
697		
698		
699	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.	747
700		748
701		749
702		
703		
704		
705	Ayal Klein, Eran Hirsch, Ron Eliav, Valentina Pyatkin, Avi Caciularu, and Ido Dagan. 2022. Qasem parsing: Text-to-text modeling of qa-based semantics. arXiv preprint arXiv:2205.11413 .	750
706		751
707		752
708		753
709	Ayal Klein, Jonathan Mamou, Valentina Pyatkin, Daniela Stepanov, Hangfeng He, Dan Roth, Luke Zettlemoyer, and Ido Dagan. 2020. QANom: Question-answer driven SRL for nominalizations . In Proceedings of the 28th International Conference on Computational Linguistics , pages 3069–3083, Barcelona, Spain (Online). International Committee on Computational Linguistics.	754
710		755
711		756
712		
713		
714		
715		
716		
717	Wei-Jen Ko, Cutter Dalton, Mark Simmons, Eliza Fisher, Greg Durrett, and Junyi Jessy Li. 2022a. Discourse comprehension: A question answering framework to represent sentence connections. EMNLP .	757
718		758
719		759
720		760
721	Wei-Jen Ko, Yating Wu, Cutter Dalton, Dananjay Srinivas, Greg Durrett, and Junyi Jessy Li. 2022b. Discourse analysis via questions and answers: Parsing dependency structures of questions under discussion. arXiv preprint arXiv:2210.05905 .	761
722		762
723		763
724		764
725		765
		766
		767
		768
		769
		770
		771
		772
		773
		774
		775
		776
		777
		778
		779
		780

- Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In NAACL-HLT (2).
- Attapol Rutherford and Nianwen Xue. 2014. Discovering implicit discourse relations through brown cluster pair representation and coreference patterns. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 645–654.
- Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with self-talk. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 4615–4629.
- Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4593–4601.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.
- Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. 2019. Do nlp models know numbers? probing numeracy in embeddings. 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 5307–5315.
- Chenxu Wang, Ping Jian, and Mu Huang. 2023. Prompt-based logical semantics enhancement for implicit discourse relation recognition. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 687–699.
- Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The penn discourse treebank 3.0 annotation manual. Philadelphia, University of Pennsylvania, 35:108.
- Yating Wu, Ritika Mangla, Greg Durrett, and Junyi Li. 2023. Qudeval: The evaluation of questions under discussion discourse parsing. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 5344–5363.
- Wei Xiang, Zhenglin Wang, Lu Dai, and Bang Wang. 2022. Connprompt: Connective-cloze prompt learning for implicit discourse relation recognition. In Proceedings of the 29th International Conference on Computational Linguistics, pages 902–911.
- Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhao Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. arXiv preprint arXiv:2304.12244.
- Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In International Conference on Machine Learning, pages 12697–12706. PMLR.
- Hao Zhou, Man Lan, Yuanbin Wu, Yuefeng Chen, and Meirong Ma. 2022. Prompt-based connective prediction method for fine-grained implicit discourse relation recognition. EMNLP.

A Details for Question Bank Preparation

A.1 In-Context Learning for Salient Event Identification

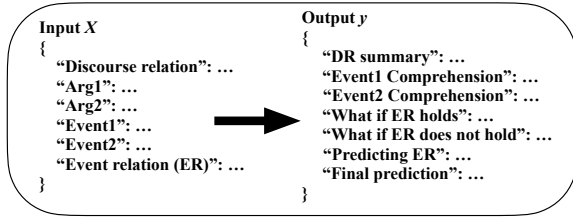


Figure 5: The input and output for in-context learning for selecting salient signals.

In Section 2.2, we briefly introduced our In-Context Learning (ICL) approach; here, we offer a more detailed explanation. Figure 5 elucidates the ICL template, further exemplified by Table 5. The ‘*DR summary*’ section encapsulates the discourse using the model’s specific terminology. In ‘*Event1 comprehension*’, Event1 is linked with *Arg2*, exploring its role in discerning the discourse relation. A similar analysis is conducted for Event2 in ‘*Event2 comprehension*’. This stage prompts LMs to begin reasoning, as demonstrated in Example 2, where the model identifies the actual object of denial, such as ‘shipping the card’. The sections ‘*What if ER holds*’ and ‘*What if ER does not hold*’ present hypothetical scenarios of ER’s presence or absence, exploring their implications for the given DR. The model is encouraged to offer explanations (like “It suggests to the audience that IBM’s actions are inconsistent and perhaps not well-planned”). ‘*Predicting ER*’ synthesizes the preceding rationale to predict an ER between Event1 and Event2, leading to the ‘*Final prediction*’ that provides the definitive conclusion. While our focus is on using ICL to identify salient signals in discourse understanding, we have not fully explored the potential of prompt engineering.

A.2 Detailed Dataset Statistics

In Table 10, we present an in-depth analysis of our dataset statistics. It reveals that the Expansion and Temporal classes exhibit a notably higher satisfaction rate, leading to an increased number of event pairs per instance. Our dataset mirrors the distribution pattern of the PDTB dataset, predominantly featuring Expansion and Contingency cases. Furthermore, we observe that the original Level-2 relation Contingency.Cause+Belief have an insufficient number of instances and shared a causal

meaning. Consequently, we merge its Reason and Result subclasses into Contingency.Cause.Reason and Contingency.Cause.Result, respectively. This leads to the formation of the customized classes Contingency.Reason and Contingency.Result.

B Annotation Details for Human Verification

B.1 Annotator Recruitment

Following Institutional Review Board (IRB) approval, we enlisted two graduate students specializing in Natural Language Processing (NLP) to conduct our annotations. These individuals possess strong English proficiency and academic expertise, equipping them with the necessary skills to comprehend our discourse task effectively. They have consented to the use of their anonymized data, with the assurance that their identities will remain confidential.

B.2 Annotator Training

Discourse and events

```

Positive case:
{
  "Discourse relation": "Expansion.Level-of-detail.Arg2-as-detail, which means Arg2 provides more detail about Arg1",
  "Arg1": "An international group approved a formal ban on ivory trade despite objections from southern African governments, which threatened to find alternative channels for selling elephant tusks",
  "Arg2": "The move by the Convention on Trade in Endangered Species, meeting in Switzerland, places the elephant on the endangered-species list",
  "Event1": "An international group approved a formal ban on ivory trade",
  "Event2": "The move places the elephant on the endangered - species list",
  "Event relation": "Event2 provides more detail about Event1"
}

```

(a) Training slides screenshot: Positive sample (negative sample omitted).

Your task

Given a discourse, predict whether Event1 and Event2 satisfy event relation (ER).

Step 1: Read the discourse and the events.

Step 2: Judge ER.

(b) Task instructions for annotators.

Figure 6: **Training slides screenshot:** Simplistic instruction to perform the task.

An author led the training for the two annotators, starting with a tutorial on discourse semantics by referring to PDTB annotation guideline (Webber et al., 2019). This session then equipped annotators with the skills to discern positive and negative

Example 1 (Positive)	<p>Input:</p> <pre>{ "Discourse relation": "Comparison.Concession.Arg2-as-denier, which indicates that Arg2 denies or contradicts something in Arg1", "Arg1": "In late September, IBM began shipping a plug-in card that converts its PS/2 model 70-A21 from a 80386 machine to an 80486 machine", "Arg2": "the company told customers Monday about the bugs and temporarily stopped shipping the product", "Event1": "IBM began shipping a plug-in card", "Event2": "The company temporarily stopping the shipping of the product", "Event relation": "Event1 is denied by (or contrasted with) Event2" }</pre> <p>Output:</p> <pre>{ "DR Summary": "The discourse revolves around IBM's actions concerning a specific product. The discourse relation is of 'Comparison.Concession.Arg2-as-denier' type, which implies that Arg2 denies or contradicts a point made in Arg1. The key elements are IBM's initial action of shipping a plug-in card and the subsequent action of halting the shipping due to issues.", "Event1 Comprehension": "Event1 ('IBM began shipping a plug-in card') is a positive action, highlighting a company's forward momentum in its business strategies. It contributes to the DR by being the initial statement that gets contradicted by Arg2. This helps establish the 'denier' aspect in the discourse relation.", "Event2 Comprehension": "Event2 ('The company temporarily stopping the shipping of the product') is a corrective action that contradicts the assertion made in Arg1. It adds nuance to the discourse by showing that IBM's initial action had flaws, thereby satisfying the 'denier' role in the discourse relation.", "What if ER Holds": "If the Event Relation directly holds, Event1 (beginning the shipping) is directly denied by or contrasted with Event2 (stopping the shipping). This amplifies the impact of Arg2 as a 'denier', confirming the failure or problems in the initial action represented by Arg1. It suggests to the audience that IBM's actions are inconsistent and perhaps not well-planned, potentially affecting the company's reputation.", "What if ER Does Not Hold": "If the Event Relation does not directly hold, it implies that the act of IBM stopping shipping may not necessarily contradict or negate their initial action of shipping the plug-in card. This could lead to a range of interpretations. It might suggest that the company had valid, perhaps unrelated, reasons for halting shipments, such as supply chain issues or broader quality checks, without implicating the plug-in card specifically.", "Predicting ER": "Given the analyses in the 'What if ER Holds' and 'What if ER Does Not Hold' sections, we see two contrasting implications. If ER directly holds, the inconsistency and potential poor planning in IBM's actions become the focal point, amplifying Arg2's role as a 'denier' in the discourse. On the other hand, if ER does not directly hold, the 'denier' role of Arg2 gets weakened, and other factors might be considered as reasons for stopping the shipping, thereby reducing the impact on IBM's reputation. In this specific discourse, the halting of shipping appears to directly negate IBM's initial action of beginning the shipping, supporting the 'Comparison.Concession.Arg2-as-denier' relation. Therefore, the prediction is that 'ER Directly Holds'." }</pre>
Example 2 (Negative)	<p>Input:</p> <pre>{ "Discourse relation": "Comparison.Concession.Arg2-as-denier, which indicates that Arg2 denies or contradicts something in Arg1", "Arg1": "In late September, IBM began shipping a plug-in card that converts its PS/2 model 70-A21 from a 80386 machine to an 80486 machine", "Arg2": "the company told customers Monday about the bugs and temporarily stopped shipping the product", "Event1": "Plug-in card converting PS/2 model 70-A21 from a 80386 machine to an 80486 machine", "Event2": "The company temporarily stopping the shipping of the product", "Event relation": "Event1 is denied by (or contrasted with) Event2" }</pre> <p>Output:</p> <pre>{ "DR Summary": "The discourse relation here is 'Comparison.Concession.Arg2-as-denier,' suggesting that Arg2 should negate or contradict Arg1. Arg1 discusses IBM's initial act of shipping a specific plug-in card, while Arg2 presents the company halting this action due to bugs.", "Event1 Comprehension": "Event1 highlights the functionality of the plug-in card, focusing specifically on its ability to convert a PS/2 model from one specification to another. This is more of a technical detail than a core message in the discourse.", "Event2 Comprehension": "Event2 is about IBM stopping the shipping of a product, presumably the plug-in card. This action seems to contradict or deny Arg1's claim that IBM began shipping the card.", "What if ER Holds": "If ER directly holds, the discourse would imply that the specific functionality of the plug-in card (conversion of machine specifications) is directly negated by IBM stopping its shipping. This would make the halt in shipping an explicit refutation of the plug-in card's value or functionality.", "What if ER Does Not Hold": "If ER doesn't directly hold, stopping the shipping may not directly deny the specific conversion capability of the plug-in card. The focus shifts from the functionality of the plug-in card to the broader issue of IBM's decision-making. It might suggest problems beyond the card's functionality, such as quality control, that led to the halt in shipping.", "Predicting ER": "Given the analyses in the 'What if ER Holds' and 'What if ER Does Not Hold' sections, it seems that the discourse isn't centered around denying the specific functionality of the plug-in card. Rather, Arg2 denies the broader action of shipping the product. Given this nuance, the prediction is 'ER Indirectly Holds or Does Not Hold'." }</pre>

Table 9: Examples for In-Context Learning for saleint event relation prediction.

Event Relation (ER) cases. Our task instructions were straightforward, focusing on binary classification.

B.3 Annotation Interface

```
{
  "Discourse relation": "Contingency.Cause.Result, it means when Arg1 gives the reason, explanation or justification, while Arg2 gives its effect",
  "Arg1": "If they do good credit analysis, they will avoid the hand grenades",
  "Arg2": "the market is in good shape",
  "Event 1": "they do good credit analysis",
  "Event 2": "the market is in good shape",
  "Event relation": "Event1 is the reason for Event2"
}
```

Does event relation hold under this discourse?

☐ Yes

☐ No

Figure 7: **Annotation Interface:** The interface guides annotators in making binary judgments, focusing on discourse arguments, two distinct events, and their potential event relation.

Our annotation interface (Figure 7) adopts a min-

imalist style, straightforwardly prompting the annotators to determine whether an Event Relation (ER) is present in the given scenario.

C Experiment Details

C.1 Model Details

We list the models being evaluated in Table 11, using APIs and weights hosted on Huggingface. We also use AllenNLP (Gardner et al., 2018) for semantic role labeling toolkit.

C.2 Computing Resource and AI Tools

We use one NVIDIA A40 GPU to perform our experiment. For in-context learning for ER prediction, it takes around 30 seconds for each instance due to long reasoning to be decoded. It takes less than one day to finish all predictions. For the evaluating models against DiSQ, since it only needs to decode a short answer, it takes around 0.1 seconds

Discourse Relation	Event Relation	Ratio	# of Event Pairs	# of PDTB Instance	# of Q
Comparison.Concession	deny or contradict with	0.617	1.35	86	1392
Comparison.Contrast	contrast with	0.596	1.31	45	708
Contingency.Reason	reason of	0.69	2.64	162	5124
Contingency.Result	result of	0.735	1.67	113	2268
Expansion.Conjunction	contribute to the same situation	0.964	1.96	192	4512
Expansion.Equivalence	equivalent to	0.791	1.31	26	408
Expansion.Instantiation	example of	0.792	1.49	120	2148
Expansion.Level-of-detail	provide more detail about	0.739	1.71	180	3696
Expansion.Substitution	alternative to	0.714	1.07	14	180
Temporal.Asynchronous	happen before/after	0.696	1.84	56	1236
Temporal.Synchronous	happen at the same time as	0.771	2	32	768
Total / Avg		0.737	1.67	1026	22,400

Table 10: **Comprehensive Statistics of Our Dataset:** This table delineates the corresponding event relations for each discourse relation. We include the satisfaction ratio of event pairs and their resultant quantities. The table also enumerates the number of discourse instances in the PDTB test set and the total count of questions generated from these events.

Model	Resource
GPT-3.5-turbo	API (Accessed on Dec 2023)
GPT-4	API (Accessed on Dec 2023)
LLaMA2-7B	https://huggingface.co/meta-llama/Llama-2-7b-hf
LLaMA2-7B-Chat	https://huggingface.co/meta-llama/Llama-2-7b-chat-hf
LLaMA2-13B	https://huggingface.co/meta-llama/Llama-2-13b-hf
LLaMA2-13B-Chat	https://huggingface.co/meta-llama/Llama-2-13b-chat-hf
Vicuna-13B	https://huggingface.co/lmsys/vicuna-13b-v1.5
Wizard	https://huggingface.co/WizardLM/WizardLM-13B-V1.2
Wizard-Code	https://huggingface.co/WizardLM/WizardCoder-Python-13B-V1.0
Wizard-Math	https://huggingface.co/WizardLM/WizardMath-13B-V1.0

Table 11: **Model Detail:** For the GPT models, access is provided through their APIs, as these are closed-source. In contrast, for open-source models, we utilize their weights hosted on Huggingface.

for one instance. It takes around 2-3 hours to finish evaluation of one model against DISQ.

We employ GitHub Copilot as a coding assistant, primarily to complete specific lines of code once the core functions are established. Additionally, we use GPT for grammar checking, but all the writing is conducted independently by us.

C.3 Task template

We adopt the approach outlined by Zhao et al. (2021), employing a straightforward instruction template. Initially, a succinct instruction is provided, followed by the context information for the model. Recognizing that the model may not be well-versed in discourse semantics, we use the term “sentence” in place of “argument” for clarity. Subsequent to the question, we include an “Answer: ” prompt, guiding the model to respond with either “True” or “False” tokens.

C.4 Asymmetric Results

Table 13 presents the DISQ Scores for four groups of converse relations. Apart from the first group (Comparison), the remaining three groups exhibit a noticeable performance disparity. For instance,

Respond to a true-or-false question derived from a two-sentence discourse, comprising Sentence 1 (Sent1) and Sentence 2 (Sent2), linked by a relationship type like causal, temporal, expansion, contrasting, etc. The question targets two events within this discourse, and your task is to evaluate if these events exhibit the specified relationship. Answer with 'True' or 'False' based on your analysis.

Sent1: “When I want to buy, they run from you – they keep changing their prices.” **Sent2:** “It’s very frustrating.”

Question: Is “It’s very frustrating. (event 2)” the result of “they keep changing their prices (event 1)” ? True or False?

Answer:

Table 12: **Instruction Template** begins with a concise task instruction for the Language Model (LM) (1st line), followed by the provision of context (2nd line), and culminates with posing the question (3rd line).

		Comparison		Cont		Exp		Temp	
	all	Comp.Arg1-as-denier	Comparison.Arg2-as-denier	Cont.Reason	Cont.Result	Exp.Arg1-as-detail	Exp.Arg2-as-detail	Temp.Succession	Temp.Precedence
gpt-4	0.414	0.0	0.058	0.119	0.351	0.222	0.519	0.0	0.756
gpt-3.5-turbo	0.42	0.281	0.203	0.21	0.4	0.232	0.464	0.0	0.794
13b	0.001	0.0	0.0	0.0	0.003	0.0	0.002	0.0	0.004
13bchat	0.253	0.13	0.197	0.128	0.173	0.289	0.384	0.122	0.202
7b	0.013	0.0	0.013	0.016	0.021	0.015	0.01	0.0	0.009
7bchat	0.175	0.149	0.239	0.132	0.174	0.067	0.159	0.025	0.116
vicuna-13b	0.035	0.033	0.008	0.013	0.044	0.0	0.039	0.085	0.061
wizard	0.052	0.111	0.114	0.036	0.042	0.012	0.029	0.0	0.015
wizard-code	0.216	0.026	0.026	0.16	0.273	0.283	0.323	0.417	0.566
wizard-math	0.218	0.193	0.149	0.187	0.221	0.452	0.322	0.141	0.161
Avg	0.180	0.092	0.101	0.100	0.170	0.157	0.225	0.079	0.268

Table 13: **Asymmetry in Results for Converse Relations:** This table reports the DISQ Score for converse relation pairs, where one relation in each converse group exhibits lower performance compared to its counterpart.

		Overall	Exp.	Cont.	Comp.	Temp.
LLaMA2-13B-Chat	w/o Context	0.588 / 0.547	0.762 / 0.536	0.381 / 0.489	0.516 / 0.718	0.323 / 0.564
		0.645 / 0.586	0.833 / 0.572	0.529 / 0.551	0.384 / 0.647	0.307 / 0.685
Wizard-Code	w/o Context	0.332 / 0.799	0.332 / 0.797	0.308 / 0.820	0.123 / 0.820	0.647 / 0.722
		0.459 / 0.672	0.456 / 0.652	0.484 / 0.695	0.228 / 0.755	0.682 / 0.610

Table 14: **Influence of Context on Targeted and Counterfactual Scores:** Each cell reports the Targeted and Counterfactual Scores as X/Y, respectively. For both LLaMA and Wizard models, we observe a significant rise in Targeted Scores accompanied by a decrease or marginal enhancement in Counterfactual Scores.

the average score for Cont.Reason is merely 0.100, in contrast to 0.170 for Cont.Result. This suggests a possible intrinsic asymmetry in Large Language Models’ (LLMs) processing of semantic relationships.

C.5 Contextual Results

To explore the effect of surrounding context on the comprehension of discourse arguments, we decompose the DISQ Score into Targeted and Counterfactual categories, as detailed in Table 14. Our analysis reveals that the overall enhancement in the DISQ Score is predominantly due to the elevation in Targeted Score. For the majority of relations, we observe a pronounced increase in Targeted Score, contrasted with a decrease or slight rise in Counterfactual Score. This indicates that context primarily benefits affirmatively answering Targeted questions.