

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 41% of the ideal performance in PDTB; (2) DiSQ is robust to domain shifts and paraphrase variations; (3) Open-source models generally lag behind their closed-source GPT counterparts, with notable exceptions being those enhanced with chat and code/math features; (4) Our analysis validates the effectiveness of explicitly signalled discourse connectives, the role of contextual information, and the benefits of using 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 through a series of questions, we introduce Discursive Socratic Questioning (DiSQ), a new method

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.*

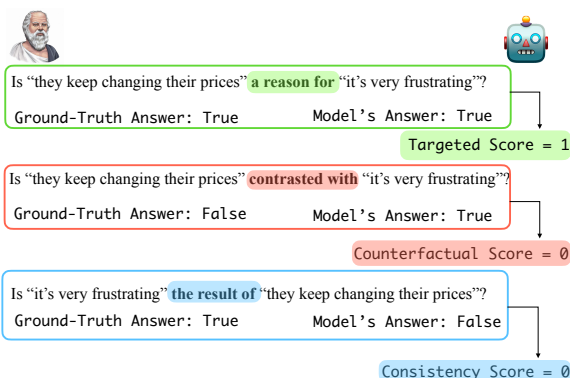


Figure 1: DiSQ combines three discourse-relevant scores: (1) Targeted Score, gauging responses to key events; (2) Counterfactual Score, assessing robustness against irrelevant queries; (3) Consistency Score, measuring logical coherence to equivalent questions.

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. While 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 generate questions centered on key spans with ground-truth 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 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). Berglund et al. (2023) find 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 test to the broader LLM research on 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. We find that while larger, closed-source models excel in responsiveness, they also struggle with the “reversal curse”, indicating a probabilistic approach to discourse semantics without full logical consistency. We further demonstrate that DISQ’s measure is robust against domain shifts (TED-MDB corpus (Zeyrek et al., 2020)) and question paraphrasing. 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

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

¹DISQ’s software and data will be released once accepted.

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. The spans are similar to elementary discourse units (EDU) in the RST/SDRT theories. They are characterized by a self-contained subject–verb–object (s–v–o) structure, identified by semantic role labeling (SRL), 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) 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. It makes its prediction on a new input X after being exposed to both a positive and a negative example (each example is structured as $X \rightarrow y$).

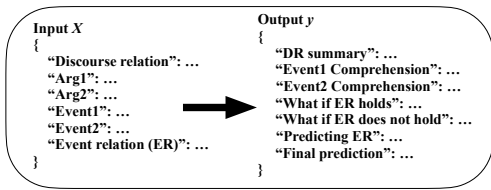


Figure 2: The input and output for in-context learning for selecting salient signals.

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).

2.3 Dataset Statistics

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

Table 2: **PDTB Dataset Statistics:** Discourse relations with their corresponding event relations, the type of questions (uni- or bi-directional), and question counts.

Table 2 outlines the 11 Level-2 relations from PDTB-3.0, with modifications in the Contingency relation to merge smaller groups (Appendix C.1). It lists each discourse relation’s corresponding event relation and whether the question type is uni- or bi-directional, aiding in generating converse questions. For bi-directional relations, the converse mirrors the original (e.g., “A happens in the same time as B” and vice versa), while for uni-directional, the converse flips the sequence (e.g., “A happens before B” becomes “B happens after A”). We annotated all implicit discourse in PDTB test set (Sections 21 and 22 in PDTB). 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.

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’s judgements (Fabbri et al., 2021a). Rather, we directly evaluate the questions’ correctness of the benchmark. Upon verification of the questions,

subsequent steps are transparent, deterministic, and verifiable — all important features for any measure.

	A1&A2	A1&ICL	A2&ICL
Agreements	85.2%	85.2%	83.7%
Cohen’s Kappa	38.5%	56.0%	52.0%
Success Rate	/	95.8%	93.8%

Table 3: Agreement rates between two annotators (A1 and A2) and ICL method, alongside the success rate.

Verification Results We invite human annotators to identify whether a relation r exists between events s_1 and s_2 in a discourse instance ($Arg1, Arg2, R$) – the same binary task that the ICL method tackles. They annotated 61 event relation instances across all Level-2 discourse relations, as detailed in Appendix B.4. Two NLP-specialized graduate students performed this task, each paid at the university’s standard rate of US\$10 per hour for 2 hours work. To ensure clarity, instances requiring extensive domain-specific knowledge, such as finance, were excluded after random sampling. The annotation began following a basic discourse semantics tutorial.

Table 3 shows strong **agreements** between ICL method’s predictions and human annotators (~85%). Furthermore, despite the majority of data samples being positive cases, ICL demonstrates a decent **Cohen’s Kappa** score with human annotators. Most importantly, the **success rate** – the proportion of positive cases confirmed by human annotation – exceeds 93%, validating the effectiveness of our ICL method in identifying 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

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

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	$Q_t = \{QG(s_2, s_1, \overleftarrow{r})\}$	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} .

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 ?*” ($\overleftarrow{r}$) corresponds to the earlier question about **reason** (r). We anticipate consistent responses from LMs. For bi-directional questions, only the entity order is reversed (e.g., “Does A happen at the same time as B?” becomes “Does B happen at the same time as A?”). For uni-directional questions, both relation and entity order are inverted (e.g., “Is A the reason for B?” to “Is B the result of A?”), detailed in Appendix C.9.

3.2 Question Answering

Algorithm 1 DISQ interrogates a language model.

```

1: Input: Discourse  $d$  and its corresponding questions  $Q$ .
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:    $\tilde{a}_i = \text{LM}(q = \tilde{q}, c = d, (q_i, a_i) \in \mathcal{H})$   $\triangleright$  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). These questions focus on events s_1 and s_2 , and reference the discourse context d ,

specifically 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 §4.4).

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. Concretely, for N questions asked:

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

$$s_{cf} = \frac{1}{N} \sum_{i=1}^N \mathbb{1}[a_i = False], q_i \in \{Q_c, \tilde{Q}_c\} \quad (2)$$

$$s_{con} = \frac{1}{N} \sum_{i=1}^N \mathbb{1}[a_i = \tilde{a}_i], q_i \in Q, \tilde{q}_i \in \tilde{Q} \quad (3)$$

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

We guide our evaluation with following research questions (RQs):

RQ1: How do models perform on DiSQ’s three scores overall?

RQ2: Are DiSQ’s scores consistent in different datasets and variations in question phrasing?

RQ3: What impact do different discourse relations have on model performance?

RQ4: What linguistic structures can help models improve their performance on DiSQ?

Datasets: Besides PDTB, we also use English sections of the TED-Multilingual Discourse Bank (TED-MDB) dataset, with PDTB-style annotations

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

from TED talks. After preprocessing it to generate a question bank similar to PDTB’s method, we obtain 448 discourse instances and 8,376 questions, which is nearly half of PDTB, see details in Appendix C.1.

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 (Appendix C.5). Despite this constraint, we noted stable performance throughout our experiment and relation distributions similar to those of the entire dataset. (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.

Implementation Details: We use a zero-shot approach in our evaluations to mirror real-world conditions. Recognizing that smaller models may falter with different instruction templates, we experimented with various templates and report the optimal performances. This method allows us to focus on evaluating the models’ ability to answer questions about discourse relation, minimizing the influence of their instruction comprehension skills. Following Zhao et al. (2021), we prompt models with concise instructions, discourse context and questions. We determine predictions based on the first token’s probability (Appendix C.4).

4.1 Overall Performance (RQ1)

Figure 3 displays the overall performance of various models in a zero-shot setting, leading to the following observations: (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. GPT-3.5 underperforms GPT-4 significantly in both the PDTB and TED-MDB datasets (almost half in s_{disq}), particularly in the Targeted Score, which shows its limitation in understanding discourse. (B) **Significant Improvements with LLaMA Enhancements** (2nd and 3rd row): Initial tests show Vanilla

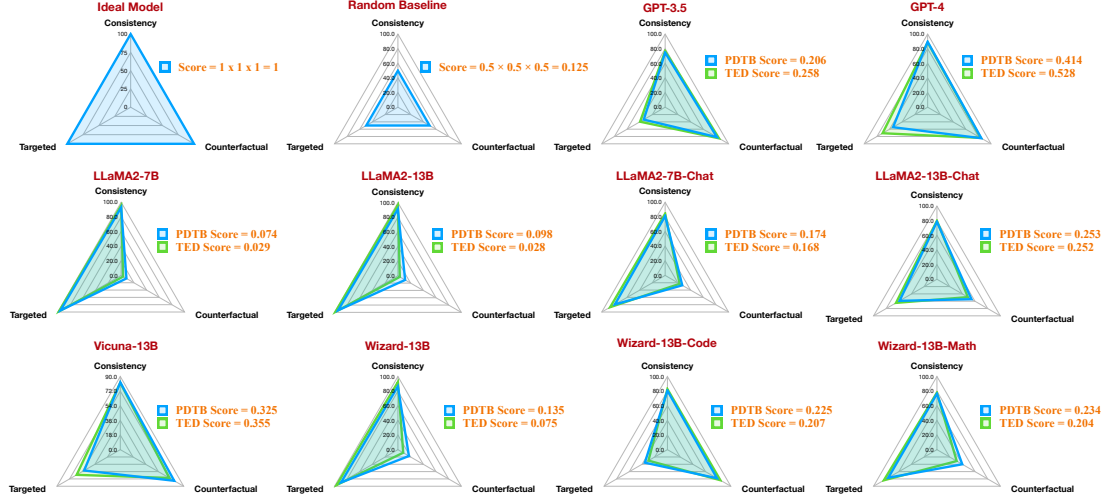


Figure 3: **Overall performance on DiSQ (RQ1):** This figure shows the overall performance (s_{disq}) of advanced GPT models (1st row), LLaMA-2 models with 7B/13B parameters and their chat variants (2nd row), and specialized Vicuna and Wizard models that are further fine-tuned from the LLaMA architecture (3rd row). Best view in color.

LLaMA-2 models (7B and 13B) perform below the random baseline. However, significant improvements are noted with their Chat variants, and further enhancements are observed with Vicuna-13B after tuning on user interaction, and upon further tuning for Code and Math based on Wizard models. Vicuna-13B notably surpasses GPT-3.5 in both datasets, suggesting open-source models can rival GPT-3.5 in discourse understanding. (C) **Consistent Performance Metrics Across Datasets:** The scores for both datasets align well (as visualized in blue and green shapes), demonstrating DiSQ is consistent in differing domains.

4.2 Consistency of DiSQ Scores (RQ2)

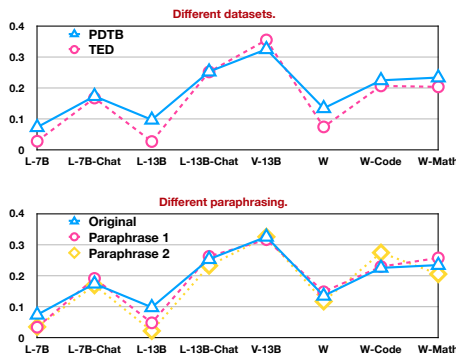


Figure 4: Models' performance under different datasets and various question paraphrasing scenarios (RQ2).

While evaluating DiSQ *intrinsically* through human assessment of question correctness, we also present *extrinsic* evidence of DiSQ Scores' robustness across domain and paraphrase variations. The first plot in Figure 4 shows DiSQ Scores for var-

ious open-source models (abbreviated by initial letters) across two datasets, with model rankings demonstrating strong consistency, evidenced by a Kendall's Tau correlation of 0.857.

The second plot, focusing on the PDTB dataset, illustrates models' resilience to question paraphrasing, involving synonym replacement and syntactic changes, and results in two paraphrase sets. This analysis shows the DiSQ Score remains stable under these variations, with a mean Spearman correlation of 93.6 across three pairs. For the TED dataset, consistency is also observed (Appendix C.8).

4.3 DiSQ Scores by Discourse Relations (RQ3)

Table 5 provides performance per discourse relation. These results yield 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 for most models, suggesting that merely increasing model and data samples do not yield comprehensive discourse understanding. (2) **Open-sourced Models Rival GPT:** The granular analysis of scores reveals that open-source models are capable of matching, and in some instances, surpassing the performance of the esteemed GPT models. For instance, within Contingency relations, models like Vicuna-13B (6A, 6B in Table 5) and Wizard-

Models	Overall	Comp.Conce	Comp.Contra	Cont.Reason	Cont.Result	Exp.Conj	Exp.Equiv	Exp.Inst	Exp.Detail	Exp.Subst	Temp.Async	Temp.Sync
I. Random Baseline	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125
PDTB												
2A. LLaMA2-7B	0.074	0.029	0.083	0.094	0.095	0.076	0.056	0.087	0.067	0.156	0.035	0.048
3A. LLaMA2-7B-Chat	0.174	0.231	0.431	0.131	0.174	0.213	0.104	0.120	0.150	0.199	0.108	0.040
4A. LLaMA2-13B	0.098	0.037	0.100	0.082	0.097	0.127	0.101	0.113	0.107	0.086	0.084	0.092
5A. LLaMA2-13B-Chat	0.253	0.193	0.477	0.129	0.172	0.288	0.157	0.326	0.373	0.291	0.195	0.028
6A. Vicuna-13B	0.325	0.087	0.513	0.200	0.353	0.369	0.000	0.334	0.462	0.195	0.511	0.069
7A. Wizard	0.135	0.221	0.256	0.067	0.107	0.170	0.072	0.167	0.128	0.108	0.097	0.082
8A. Wizard-Code	0.225	0.032	0.268	0.175	0.287	0.121	0.008	0.283	0.329	0.174	0.545	0.109
9A. Wizard-Math	0.234	0.132	0.264	0.241	0.286	0.192	0.046	0.240	0.323	0.201	0.240	0.135
10A. GPT-3.5	0.206	0.151	0.278	0.082	0.161	0.246	0.067	0.257	0.262	0.232	0.388	0.000
11A. GPT-4	0.414	0.053	0.567	0.119	0.351	0.610	0.192	0.659	0.481	0.422	0.692	0.000
TED												
2B. LLaMA2-7B	0.029	0.011	0.021	0.053	0.047	0.024	0.037	0.027	0.040	0.037	0.018	0.032
3B. LLaMA2-7B-Chat	0.168	0.182	0.315	0.139	0.160	0.184	0.105	0.089	0.124	0.099	0.167	0.063
4B. LLaMA2-13B	0.028	0.001	0.044	0.020	0.017	0.034	0.029	0.053	0.030	0.000	0.027	0.056
5B. LLaMA2-13B-Chat	0.252	0.196	0.492	0.141	0.175	0.260	0.211	0.376	0.369	0.284	0.203	0.053
6B. Vicuna-13B	0.355	0.098	0.551	0.241	0.387	0.386	0.064	0.509	0.508	0.219	0.448	0.232
7B. Wizard	0.075	0.100	0.159	0.028	0.075	0.063	0.063	0.120	0.100	0.063	0.062	0.008
8B. Wizard-Code	0.207	0.015	0.330	0.114	0.269	0.163	0.000	0.360	0.254	0.096	0.616	0.116
9B. Wizard-Math	0.204	0.126	0.240	0.228	0.258	0.165	0.062	0.302	0.303	0.208	0.224	0.220
10B. GPT-3.5	0.258	0.159	0.562	0.132	0.181	0.240	0.326	0.350	0.500	0.089	0.346	0.000
11B. GPT-4	0.528	0.061	0.688	0.238	0.481	0.652	0.593	0.652	0.403	0.314	0.812	0.592

Table 5: **Impact of Discourse Relations on DiSQ Scores (RQ3):** We highlight the top three models per discourse relation in each dataset. GPT-4 dominates, yet open-source models closely rival in several relations.

Code (8A, 8B) as well as Wizard-Math (9A, 9B) excel, even outperforming GPT models. It underscores the potential of LLaMA-based specialized training as a promising method to enhance discourse comprehension. (3) **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.6), indicating a potential asymmetry in semantic processing by language models.

4.4 Linguistic Features (RQ4)

We are driven to find what linguistic features can improve the faithfulness of discourse understanding, especially for open-source models. We examine the following features:

Feature 1: Is the Presence of Discourse Connectives Beneficial? Discourse connectives transform implicit discourse into explicit forms, enhancing comprehension (Kurfali and Östling, 2021). Does this extend to LLM? In short, yes.

Models	Overall	Exp.	Cont.	Comp.	Temp.
w/o Conn					
LLaMA2-13B-Chat	0.253	0.322	0.147	0.293	0.149
Vicuna-13B	0.325	0.374	0.262	0.220	0.357
Wizard-Code	0.225	0.220	0.223	0.101	0.382
w/ Conn					
LLaMA2-13B-Chat	0.273	0.328	0.187	0.329	0.149
Vicuna-13B	0.396	0.418	0.382	0.289	0.401
Wizard-Code	0.264	0.249	0.275	0.192	0.376

Table 6: **Feature 1:** Models’ DiSQ Scores with the help of discourse connective.

We selected LLaMA2-13B-Chat, Vicuna-13B

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 all models, increasing DiSQ by 8% to 22%. For example, it boosts Vicuna’s overall performance from 0.325 to 0.396, closely approaching GPT-4’s score of 0.414 (which lack connectives). This suggests that correctly inferring connectives significantly enhances the accuracy of LMs’ discourse comprehension.

Feature 2: Does Context Enhance Comprehension? Next, we assess the influence of surrounding context on discourse comprehension. Does it enhance LLM’s faithfulness? In short, yes.

Models	Overall	Exp.	Cont.	Comp.	Temp.
w/o Context					
LLaMA2-13B-Chat	0.253	0.322	0.147	0.293	0.149
Vicuna-13B	0.325	0.374	0.262	0.220	0.357
Wizard-Code	0.225	0.22	0.223	0.101	0.382
w/ Context					
LLaMA2-13B-Chat	0.311	0.402	0.231	0.186	0.169
Vicuna-13B	0.369	0.424	0.333	0.192	0.380
Wizard-Code	0.253	0.245	0.273	0.152	0.331

Table 7: **Feature 2:** DiSQ Scores with context’s help.

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, LLaMA’s overall performance has improved from 0.253 to 0.311. We find the improved performance is primarily due to a significant rise in Targeted Score with minimal changes in Counterfactual Score, as detailed in Appendix C.7. This indicates

that models particularly benefit from extra context when positively responding to targeted questions.

Feature 3: 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
Vicuna-13B	82.8	88.7
Wizard-Code	81.6	99.8

Table 8: **Feature 3:** Models’ Consistency Scores with the insertion of QA history.

Table 8 presents mixed outcomes: Vicuna-13B and Wizard-Code exhibit significant improvements, whereas LLaMA2-13B-Chat experiences a reduction in consistency. Further analysis into LLaMA’s Consistency Scores by question type reveals lower scores for uni-directional questions (e.g., “happen before”) and higher for bi-directional (e.g., “happen at the same time as”, which does not rephrased in converse). This pattern suggests that LLaMA might focus mainly on literal keywords, lacking in deeper reasoning abilities, while Wizard-Code’s code-based training appears to have bolstered its logical reasoning (detailed in Appendix C.10).

5 Related Work

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; Poliakov 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. There are several efforts re-formalizing discourse parsing as QA, including QADiscourse (Pyatkin et al., 2020), QA for reference/ellipsis resolution (Hou, 2020; Aralikkatte et al., 2021), and Question Under Discussion (QUD) framework (Ko et al., 2022a; Wu et al., 2023). However, their focus is on parsing rather than utilizing QA for evaluating faithfulness.

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 overlook 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

In this paper, we contribute DiSQ, the first systematic evaluation for faithful discourse comprehension. To ensure the reliability of DiSQ’s assessment, we employ both intrinsic verification via human annotation and extrinsic evaluation to demonstrate its resilience against domain shifts and paraphrase variations. Our extensive experiments reveal that even leading models like GPT-4 have their shortcomings in DiSQ, open-source models, despite trailing GPT-4, can close this gap with specialized fine-tuning. To advance LLMs’ understanding, we suggest incorporating linguistic features such as discourse connectives, contextual information, and historical QA data. In the future, we aim to extend our analysis to longer-range discourse, incorporate additional discourse annotation frameworks beyond PDTB, and distilling knowledge from larger models to benefit smaller models.

Ethical Considerations and Limitations

Our annotation of human verification is approved by institutional review board (IRB). IRB reviewed 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 respond 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 study primarily utilizes PDTB-style annotations, yet adapting DISQ to other discourse frameworks is also feasible. To the best of our knowledge, PDTB and TED-MDB are the only two compatible corpora in English, since other datasets like GUM, adhere to the RST framework, and a suitable Twitter-based PDTB corpus is not openly accessible. Consequently, we chose TED-MDB as our supplementary dataset due to its compatibility.

References

Rahul Aralikkatte, 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: Lms 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.

676	Matt Gardner, Joel Grus, Mark Neumann, Oyvind	<u>Linguistics and the 11th International Joint</u>	731
677	Tafford, Pradeep Dasigi, Nelson Liu, Matthew Pe-	<u>Conference on Natural Language Processing</u>	732
678	ters, Michael Schmitz, and Luke Zettlemoyer. 2018.	<u>(Volume 1: Long Papers)</u> , pages 1813–1827.	733
679	Allennlp: A deep semantic natural language process-		
680	ing platform. <u>arXiv preprint arXiv:1803.07640</u> .		
681	MAK Halliday. 1976. Cohesion in english. <u>Longman</u> .	Xin Liu, Jiefu Ou, Yangqiu Song, and Xin Jiang. 2020.	734
682	Yufang Hou. 2020. Bridging anaphora resolu-	On the importance of word and sentence representa-	735
683	tion as question answering. In <u>Proceedings of</u>	tion learning in implicit discourse relation classifica-	736
684	<u>the 58th Annual Meeting of the Association for</u>		737
685	<u>Computational Linguistics</u> , pages 1428–1438.		
686	Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang,	Kevin Meng, David Bau, Alex Andonian, and Yonatan	738
687	Mari Ostendorf, Ranjay Krishna, and Noah A Smith.	Belinkov. 2022. Locating and editing factual associa-	739
688	2023. Tifa: Accurate and interpretable text-to-	tions in gpt. <u>arXiv preprint arXiv:2202.05262</u> .	740
689	image faithfulness evaluation with question answer-		
690	ing. <u>arXiv preprint arXiv:2303.11897</u> .	Tasnim Mohiuddin, Shafiq Joty, and Dat Tien Nguyen.	741
691	Patrick Huber and Giuseppe Carenini. 2022. Towards	2018. Coherence modeling of asynchronous conver-	742
692	understanding large-scale discourse structures in pre-	sations: A neural entity grid approach. <u>arXiv</u>	743
693	trained and fine-tuned language models. <u>NAACL</u> .	<u>preprint arXiv:1805.02275</u> .	744
694	Feng Jiang, Yaxin Fan, Xiaomin Chu, Peifeng Li, and	Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Au-	745
695	Qiaoming Zhu. 2021. Not just classification: Rec-	tomatic sense prediction for implicit discourse rela-	746
696	ognizing implicit discourse relation on joint model-	tions in text.	747
697	ing of classification and generation. In <u>Proceedings</u>		
698	<u>of the 2021 Conference on Empirical Methods in</u>	Adam Poliak, Jason Naradowsky, Aparajita Hal-	748
699	<u>Natural Language Processing</u> , pages 2418–2431.	dar, Rachel Rudinger, and Benjamin Van Durme.	749
700	Prathyusha Jwalapuram, Shafiq Joty, and Xiang Lin.	2018. Hypothesis only baselines in natural lan-	750
701	2022. Rethinking self-supervision objectives for	guage inference. In <u>Proceedings of the Seventh</u>	751
702	generalizable coherence modeling. In <u>Proceedings</u>	<u>Joint Conference on Lexical and Computational</u>	752
703	<u>of the 60th Annual Meeting of the Association</u>	<u>Semantics</u> , pages 180–191.	753
704	<u>for Computational Linguistics (Volume 1: Long</u>		
705	<u>Papers)</u> , pages 6044–6059.	Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Milt-	754
706	Najoung Kim, Song Feng, Chulaka Gunasekara, and	sakaki, Livio Robaldo, Aravind Joshi, and Bonnie	755
707	Luis Lastras. 2020. Implicit discourse relation	Webber. 2008. The penn discourse treebank 2.0. In	756
708	classification: We need to talk about evaluation.	<u>Proceedings of the Sixth International Conference</u>	757
709	In <u>Proceedings of the 58th Annual Meeting of the</u>	<u>on Language Resources and Evaluation (LREC'08)</u> .	758
710	<u>Association for Computational Linguistics</u> , pages		
711	5404–5414.	Valentina Pyatkin, Ayal Klein, Reut Tsarfaty, and Ido	759
712	Wei-Jen Ko, Cutter Dalton, Mark Simmons, Eliza	Dagan. 2020. <u>QADiscourse - Discourse Relations as</u>	760
713	Fisher, Greg Durrett, and Junyi Jessy Li. 2022a. Dis-	<u>QA Pairs: Representation, Crowdsourcing and Base-</u>	761
714	course comprehension: A question answering frame-	<u>lines</u> . In <u>Proceedings of the 2020 Conference on</u>	762
715	work to represent sentence connections. <u>EMNLP</u> .	<u>Empirical Methods in Natural Language Processing</u>	763
716	Wei-Jen Ko, Yating Wu, Cutter Dalton, Dananjay Srin-	<u>(EMNLP)</u> , pages 2804–2819, Online. Association	764
717	ivas, Greg Durrett, and Junyi Jessy Li. 2022b. Dis-	for Computational Linguistics.	765
718	course analysis via questions and answers: Parsing		
719	dependency structures of questions under discussion.	Marco Tulio Ribeiro, Sameer Singh, and Carlos	766
720	<u>arXiv preprint arXiv:2210.05905</u> .	Guestrin. 2016. " why should i trust you?" explain-	767
721	Murathan Kurfalı and Robert Östling. 2021. Let's	ing the predictions of any classifier. In <u>Proceedings</u>	768
722	be explicit about that: Distant supervision for im-	<u>of the 22nd ACM SIGKDD international conference</u>	769
723	PLICIT discourse relation classification via connec-	<u>on knowledge discovery and data mining</u> , pages	770
724	tive prediction. In <u>Proceedings of the 1st Workshop</u>	1135–1144.	771
725	<u>on Understanding Implicit and Underspecified</u>	Rachel Rudinger, Jason Naradowsky, Brian Leonard,	772
726	<u>Language</u> , pages 1–10.	and Benjamin Van Durme. 2018. Gender bias in	773
727	Belinda Z Li, Maxwell Nye, and Jacob Andreas.	coreference resolution. In <u>NAACL-HLT (2)</u> .	774
728	2021. Implicit representations of meaning in neural		
729	language models. In <u>Proceedings of the 59th Annual</u>	Attapol Rutherford and Nianwen Xue. 2014. Dis-	775
730	<u>Meeting of the Association for Computational</u>	covering implicit discourse relations through brown	776
		cluster pair representation and coreference pat-	777
		terns. In <u>Proceedings of the 14th Conference</u>	778
		<u>of the European Chapter of the Association for</u>	779
		<u>Computational Linguistics</u> , pages 645–654.	780
		Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert	781
		rediscovers the classical nlp pipeline. In <u>Proceedings</u>	782
		<u>of the 57th Annual Meeting of the Association for</u>	783
		<u>Computational Linguistics</u> , pages 4593–4601.	784

- 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, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. arXiv preprint arXiv:2304.12244.
- Deniz Zeyrek, Amália Mendes, Yulia Grishina, Murathan Kurfalı, Samuel Gibbon, and Maciej Ogrodniczuk. 2020. Ted multilingual discourse bank (tedmdb): a parallel corpus annotated in the pdtb style. Language Resources and Evaluation, 54:587–613.
- 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

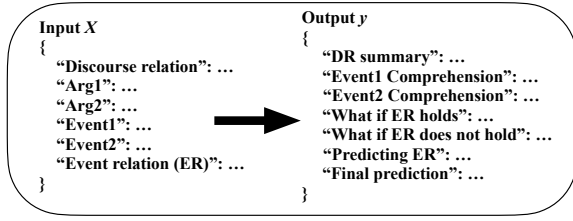


Figure 5: **ICL Template:** The input and output for in-context learning for selecting salient signals.

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. 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. Concretely, for 11 Level-2 discourse relations, we create 22 examples in total.

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 annotators to understand the task:** Tutorials and instructions 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 Event Relation (ER) cases (Figure 6). Our task instructions were straightforward, focusing on binary classification.

B.3 Annotation Interface

Our annotation interface: (Figure 7) adopts a minimalist style, straightforwardly prompting the annotators to determine whether an Event Relation (ER) is present in the given scenario.

Annotation Execution: On average, the annotators spent approximately one minute to answer each question. The 61 questions were divided into two sessions, with each session lasting about 40 minutes.

B.4 Samples for annotation

We selected these samples randomly, but excluded any that necessitate extensive domain knowledge.

Example 1 (Positive) Input: <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> Output: <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) Input: <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> Output: <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.

Discourse relation	# of event pairs
Comparison.Concession	6
Comparison.Contrast	4
Contingency.Reason	6
Contingency.Result	10
Expansion.Conjunction	8
Expansion.Equivalence	4
Expansion.Instantiation	5
Expansion.Level-of-detail	4
Expansion.Substitution	4
Temporal.Asynchronous	6
Temporal.Synchronous	4
Total	61

Table 10: **Annotation details:** Discourse relations and the number of event pairs to be judged by human annotators.

```
{
  "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.

We summarize the distribution of the number of questions (i.e., event pairs for annotators to determine) in Table 10. Samples for one discourse relation originate from one or two instances in PDTB. The presence of more pairs in some samples is attributable to the extended context in those cases.

900
901
902
903
904
905

C Experiment Details

C.1 Detailed Dataset Statistics

In Table 11, we provide detailed statistics of our dataset. Both datasets encompass 11 Level-2 discourse relations. Although the TED-MDB corpus is smaller than PDTB, it contributes 8,376 questions, aiding our examination of the cross-domain robustness of DISQ Scores.

C.2 Model Details

We list the models being evaluated in Table 12, using APIs and weights hosted on Huggingface. We also use AllenNLP (Gardner et al., 2018) for semantic role labeling toolkit.

C.3 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 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.4 Task template

We adopt the approach outlined by Zhao et al. (2021), employing a straightforward instruction template in Table 13. 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. During evaluation, we consolidate the probabilities for the “True” token (covering variations like “True,” “true,” “TRUE,” etc.), and similarly for the “False” token.

We tested three template variations and report each model’s best outcomes: (1) removing the “True or False” phrase, (2) inserting a line break at the end, and (3) placing a line break between the “question” and “answer”.

C.5 Samples for GPT Experiments

Table 14 displays the distribution of discourse relations in both the PDTB and TED-MDB datasets for GPT evaluation. We selected the first 200 samples from PDTB and randomly chose 100 samples from TED-MDB to ensure their relation distributions align closely with each dataset. This selection process was designed to match the overall distribution without needing random sampling for PDTB.

The analysis reveals that PDTB features a higher prevalence of causal discourse, whereas TED-MDB exhibits a greater number of expansions, reflecting the distinctive nature of TED Talks. This difference highlights the unique characteristics of each dataset.

C.6 Asymmetric Results

Table 15 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, the average score for Cont.Reason is merely 0.168, in contrast to 0.271 for Cont.Result. This suggests a possible intrinsic asymmetry in Large Language Models’ (LLMs) processing of semantic relationships.

C.7 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 16. 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.

C.8 Paraphrasing Performance on TED-MDB

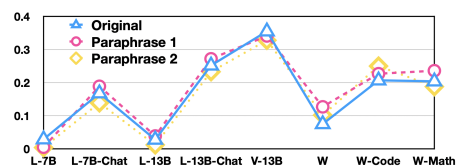


Figure 8: Models’ performance under paraphrasing in TED-MDB corpus.

Discourse Relation	Event Relation	# of TED Instance	# of Q	# of PDTB Instance	# of Q
Comparison.Concession	deny or contradict with	42	816	86	1392
Comparison.Contrast	contrast with	20	384	45	708
Contingency.Reason	reason of	38	648	162	5124
Contingency.Result	result of	45	936	113	2268
Expansion.Conjunction	contribute to the same situation	172	3084	192	4512
Expansion.Equivalence	equivalent to	11	156	26	408
Expansion.Instantiation	example of	15	432	120	2148
Expansion.Level-of-detail	provide more detail about	49	876	180	3696
Expansion.Substitution	alternative to	14	240	14	180
Temporal.Asynchronous	happen before/after	25	516	56	1236
Temporal.Synchronous	happen at the same time as	17	288	32	768
Total		448	8,376	1026	22,400

Table 11: **Comprehensive Dataset Statistics:** This summarizes the count of discourse instances within the PDTB and TED-MDB datasets, alongside the number of questions generated for each discourse relation.

Model	Resource
GPT-3.5-turbo	API GPT-3.5-turbo-0613 Version
GPT-4	API GPT-4-0613 Version
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 12: **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.

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 13: **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).

Discourse relation	PDTB	TED-MDB
Comparison.Concession	6.5%	11.0%
Comparison.Contrast	1.5%	3.0%
Contingency.Cause.Reason	20.0%	9.0%
Contingency.Cause.Result	10.5%	7.0%
Expansion.Conjunction	19.5%	38.0%
Expansion.Equivalence	4.5%	3.0%
Expansion.Instantiation	13.0%	4.0%
Expansion.Level-of-detail	18.5%	10.0%
Expansion.Substitution	1.5%	5.0%
Temporal.Asynchronous	3.5%	5.0%
Temporal.Synchronous	1.0%	5.0%

Table 14: **GPT Experiment Details:** The sample distribution for experiments used for GPT in both PDTB and TED-MDB.

	Cont.Reason	Cont.Result	Comp.Arg1-as-denier	Comp.Arg2-as-denier	Exp.Arg1-as-detail	Exp.Arg2-as-detail	Temp.Precedence	Temp.Succession
LLaMA2-13B-Chat	0.129	0.172	0.155	0.197	0.289	0.383	0.203	0.122
Vicuna-13B	0.2	0.353	0.048	0.091	0.273	0.487	0.53	0.354
Wizard-Code	0.175	0.287	0.053	0.03	0.269	0.335	0.558	0.417
Avg	0.168	0.271	0.085	0.106	0.277	0.402	0.43	0.298

Table 15: **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 16: **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.

In the main paper, we focused solely on paraphrasing within the PDTB dataset. We now extend our reporting to include model performance on the TED dataset regarding paraphrasing variability. Figure 8 demonstrates a high degree of correlation among the three sets, with a mean Spearman correlation of 94.4 across the three pairs.

To build upon our original questions, which detail the event relations in Table 11, we introduce two sets of paraphrases:

Paraphrase set 1: ‘is the consequence of’, ‘is the cause of’, ‘does occurs simultaneously as’, ‘does occurs before’, ‘does occurs after’ ‘is opposed to’, ‘is negated by’, ‘negates’ ‘serves as a substitute for’, ‘is being provided an substitute by’ ‘provide additional information about’, ‘is being provided additional information by’, ‘is equal to’, ‘are contributed to the same circumstance’ ‘is an instance of’, ‘is being instantiated by’

Paraphrase set 2: ‘is due to’, ‘leads to’, ‘takes place simultaneously as’, ‘does takes place before’, ‘does takes place after’, ‘is contrary to’, ‘is refuted by’, ‘refutes’, ‘acts as a replacement for’, ‘is replaced by’ ‘present more specifics on’, ‘is presented with more specifics by’, ‘is on par with’, ‘are contributed to the same scenario’, ‘serves as an example of’, ‘is exemplified by’.

C.9 Converse questions

A key aspect of DISQ involves the inclusion of converse questions. Table 17 outlines the discourse relations, original questions, and their converse counterparts. In bi-directional questions, we re-

verse only the order of entities (e.g., from "Does A happen at the same time as B?" to "Does B happen at the same time as A?"). For uni-directional questions, we invert both the relation and the order of entities (e.g., changing "Is A the reason for B?" to "Is B the result of A?").

C.10 Details for Experiments Using Historical QA

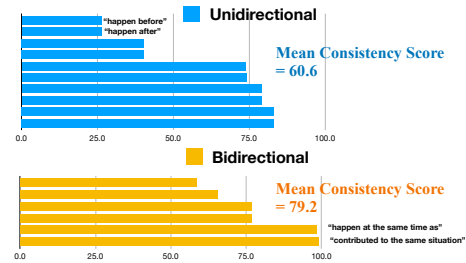


Figure 9: **Feature 3:** LLaMA’s consistency score w.r.t. relations. Bidirectional relations have higher scores.

Table 18 details consistency scores for each question type, revealing that bi-directional questions generally achieve higher Consistency Scores. For instance, "happen at the same time as" scores an impressive 98.6, while uni-directional questions, such as "happen before," score merely 26.5. Figure 9 offers a clear visual comparison, showing an average consistency score of 79.2 for bi-directional relations versus 60.6 for uni-directional ones.

This trend indicates that LLaMA-13B-Chat may predominantly rely on literal keyword matching, possibly at the expense of deeper reasoning capabilities in question answering. Conversely, the

Discourse relation	Original question	Converse question	Question type
Temporal.Synchronous	Does A happen at the same time as B?	Does B happen at the same time as A?	Bidirectional
Comparison.Contrast	Is A contrasted with B?	Is B contrasted with A?	Bidirectional
Comparison.Concession	Does A deny or contradict with B?	Is B denied or contradicted with A?	Bidirectional
Expansion.Conjunction	Does A contribute to the same situation with B?	Does B contribute to the same situation with A?	Bidirectional
Expansion.Equivalence	Is A equivalent to B?	Is B equivalent to A?	Bidirectional
Contingency.Reason	Is A the reason of B?	Is B the result of A?	Unidirectional
Contingency.Result	Is A the result of B?	Is B the reason of A?	Unidirectional
Expansion.Instantiation	Is A an example of B?	Is B exemplified by A?	Unidirectional
Expansion.Level-of-detail	Does A provide more details about B?	Is B provided more details by A?	Unidirectional
Expansion.Substitution	Is A an alternative to B?	Is B provided an alternative by A?	Unidirectional
Temporal.Asynchronous	Does A happen before B?	Does B happen after A?	Unidirectional

Table 17: **Converse Questions:** This table outlines discourse relations along with their original and converse questions, including the type of each question.

Question	Consistency Score	Question Type
happen before	26.5	Unidirectional
happen after	26.5	Unidirectional
provide more detail about	40.3	Unidirectional
being provided more detail by	40.3	Unidirectional
contrasted with	58.4	Bidirectional
equivalent to	65.4	Bidirectional
the result of	74.0	Unidirectional
the reason for	74.1	Unidirectional
denied or contradicted with	76.8	Bidirectional
deny or contradict with	76.8	Bidirectional
an example of	78.8	Unidirectional
being exemplified by	78.8	Unidirectional
an alternative to	83.3	Unidirectional
an alternative by	83.3	Unidirectional
happen at the same time as	98.6	Bidirectional
contributed to the same situation	99.2	Bidirectional

Table 18: **Historical QA Consistency:** A comparison of LLaMA2-13B-Chat’s Consistency Scores, showing bi-directional questions scoring higher in consistency than uni-directional ones.

code-based training of Wizard-Code seems to enhance its logical reasoning, leading to better overall performance.