# BeDiscovER: The Benchmark of Discourse Understanding in the Era of Reasoning Language Models

**Chuyuan Li** and **Giuseppe Carenini**
Department of Computer Science
University of British Columbia
Vancouver, BC, Canada, V6T 1Z4
chuyuan.li@ubc.ca, carenini@cs.ubc.ca

## Abstract

We introduce BeDiscovER (Benchmark of Discourse Understanding in the Era of Reasoning Language Models), an up-to-date, comprehensive suite for evaluating the discourse-level knowledge of modern LLMs. BeDiscovER compiles 5 publicly available discourse tasks across discourse lexicon, (multi-)sentential, and documental levels, with in total 52 individual datasets. It covers both extensively studied tasks such as discourse parsing and temporal relation extraction, as well as some novel challenges such as discourse particle disambiguation (e.g., "*just*"), and also aggregates a shared-task on Discourse Relation Parsing and Treebanking for multilingual and multi-framework discourse relation classification. We evaluate open-source LLMs: Qwen3 series, DeepSeek-R1, and frontier reasoning model GPT-5-mini on BeDiscovER, and find that state-of-the-art models exhibits strong performance in arithmetic aspect of temporal reasoning, but they struggle with long-dependency reasoning and some subtle semantic and discourse phenomena, such as rhetorical relation classification.

## 1 Introduction

Large Language Models (LLMs) have been tasked with solving increasingly difficult problems that require strong reasoning abilities across text understanding, common-sense, math, and coding (Shao et al., 2024; Jaech et al., 2024; Ke et al., 2025). Still, it is an open question how the discourse knowledge—that is, the ability to process and understand the interplay of multi-sentence and long-context information—of state-of-the-art reasoning models varies across various tasks and phenomena.

Discourse analysis has demonstrated impact across many downstream tasks. Its attention to both local and global levels of text understanding has also spurred advances in long-document language modeling (Prasad et al., 2023; Feng et al., 2023; Ivgi et al., 2023; Buchmann et al., 2024; Li et al., 2025). To achieve deeper understanding of natural language and induce stronger future reasoning models, it is essential for LLMs to capture and reason over discourse knowledge.

Recent studies have explored this question by evaluating LLMs' performance on a range of sentence-level tasks, such as temporal (Yuan et al., 2023; Wei et al., 2024), causal (Chan et al., 2024), and discourse relations (Yung et al., 2024; Fan et al., 2024). However, each of these studies uses a different set of models, metrics (probing, prompting, acceptability judgment), and tasks, limiting any possible bigger-picture conclusions. *Discourse understanding* is also approached from differing perspectives within the field: some work focuses on generalization ability (Braud et al., 2024; Eichin et al., 2025), while other work on the answer faithfulness (Miao et al., 2024). We, instead, use discourse understanding to refer broadly to the full range of discourse knowledge. To this end, we introduce BeDiscovER – the Benchmark of Discourse Understanding in the Era of Reasoning Language Models: the first comprehensive, practical benchmark that covers five broad areas of discourse knowledge, namely: (1) Discourse Marker Understanding, (2) Temporal Reasoning, (3) Discourse Relation Recognition, (4) Sentence Ordering, and (5) Dialogue Discourse Parsing. In total, BeDiscovER comprises around 30,000 test instances across 52 publicly available datasets (most were annotated by human experts), 16 languages, and 6 frameworks, providing a practical testbed for evaluating LLMs' reasoning skills across lexical, semantic, rhetorical, temporal, and commonsense knowledge.

We use BeDiscovER to study several reasoning-oriented LLMs: Qwen3 series (Yang et al., 2025), DeepSeek-R1 (Guo et al., 2025), DeepSeek-R1-Distll-Qwen-32B (Guo et al., 2025), and GPT-5-mini (OpenAI, 2025), since they have been shown to be more capable than models not explicitly optimized for reasoning, like GPT-4o-mini (Hurst et al.,



*Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics*
*Volume 1: Long Papers*, pages 4417–4479

2024). Empirically, we approximate *discourse understanding* ability using automated scoring based on task-specific answer correctness. This gives us indirect evidence about each model's discourse knowledge and allows us to compare models in a consistent way. Through extensive experiments, we conclude that whereas models like GPT-5-mini appear to have significant reasoning ability across many tasks, these strengths are concentrated in discourse phenomena well represented in training data, such as frequent particles (e.g., *Just*[1]) or arithmetic-style temporal reasoning.

We hope that BeDiscovER unifies research efforts across different levels of discourse knowledge by revealing phenomena where even state-of-the-art language models still exhibit notable reasoning gaps, and by drawing attention to areas that future studies (both in evaluation and in pre-/post-training) should explore in greater depth. We release the unified evaluation pipeline to support future assessment across the community[2].

## 2 Related Work

**Large Language Models.** Recent advances in computational infrastructure and the availability of large-scale corpora have driven the development of LLMs capable of remarkable generalization (OpenAI, 2025; Achiam et al., 2023; Team et al., 2023; Touvron et al., 2023). These capabilities have been further enhanced by instruction tuning (Wei et al., 2021) and reinforcement learning from human feedback (Ouyang et al., 2022), enabling models to perform complex text reasoning.

This scaling paradigm has fueled a wave of increasingly capable LLMs and pushing the boundaries across a wide spectrum of benchmarks, from graduate-level exams (Rein et al., 2024) and mathematical problem solving (Phan et al., 2025) to competitive programming and multi-agent systems (Kapoor et al., 2024; Liang et al., 2024). Despite these breakthroughs, LLMs continue to exhibit persistent limitations, particularly in modeling long-range dependencies, maintaining discourse coherence, and avoiding factual inconsistencies (Ivgi et al., 2023; Liang et al., 2023; Wu et al., 2024a). Discourse analysis, which focuses on understanding document-level structure, coherence, and information flow, offers a principled lens through which these limitations can be exposed and addressed. We pursue this line of research in this paper.

**Discourse Knowledge of LLMs.** Earlier studies evaluate LM's discourse knowledge using probing tasks (Kim et al., 2019; Li et al., 2021; Tao et al., 2024) or extracting sentence representation from PLM's attention networks (Chen et al., 2019; Huber and Carenini, 2022; Li et al., 2023). QA-based evaluation has emerged as a popular alternative, offering greater granularity and flexibility: QADiscourse (Pyatkin et al., 2020), QA for reference/ellipsis resolution (Aralikatte et al., 2021), QA for discourse faithfulness (Miao et al., 2024), and Question Under Discussion (QUD) (Wu et al., 2023, 2024b). We follow a similar approach to prompt LLMs to answer questions that designed to test different levels of discourse knowledge.

The scope of discourse knowledge evaluation varies widely. Some studies focus on specific theoretical frameworks, such as PDTB-style (Webber et al., 2019) discourse relation recognition (Miao et al., 2024; Yung et al., 2024; Mehri et al., 2025) or SDRT-style (Asher and Lascarides, 2003) dialogue parsing (Fan et al., 2024), while others examine how models generalize across frameworks (Eichin et al., 2025). Some studies target particular domains, such as stylistic discourse in literature (Wang et al., 2023), or specific relational phenomena, including event relations (Wei et al., 2024) and temporal relations (Yuan et al., 2023; Fatemi et al., 2024). Another line of research focuses on models' generation capabilities, such as DiscoScore (Zhao et al., 2023) for coherence assessment, DiSQ (Miao et al., 2024) for answer faithfulness, and LongBench (Bai et al., 2024) and LongGenBench (Wu et al., 2025) for long-text generation quality.

Closely related to our work, DiscoEval (Chen et al., 2019) proposes a sentence-level test suite for learning discourse-aware representations, but is limited to local discourse phenomena and binary classification. Disco-Bench (Wang et al., 2023) evaluates various aspects of discourse understanding and generation, but focuses on specific phenomena in Chinese and English. DiscoTrack (Bu et al., 2025) introduces new tasks that probe LLMs both explicitly (e.g., salience recognition) and implicitly (e.g., bridging inference), but primarily targets local phenomena such as entities. Unlike prior studies that concentrate on specific frameworks or narrowly defined discourse tasks, our goal is to develop a comprehensive evaluation suite that spans multiple levels of discourse understanding.

[1] *Just* is 51st most frequent word and 2nd most frequent adverb (COCA (Davies, 2009)).

[2] https://github.com/chuyuanli/BeDiscovER.

| Level | Task | Nature | Dataset | # Examples¶ | Knowledge‡ |
|---|---|---|---|---|---|
| Lexical | (1) Disc. Marker Understanding | Classification | *Just-Manual* (2025) | 90 | L&S, CS |
| | | – | *Just-Subtitle* (2025) | 149 | L&S, CS |
| | | – | *Otherwise* (2025) | 294 | L&S, CS |
| (multi-) Sentence | (2) Temporal Reasoning | Classification | TimeBank-Dense (2014) | 1515 | T, Logi, CS |
| | | – | TDDiscourse (2019) | 1,500 | T, Logi, CS |
| | | QA | ToT-Arithmetic (2024) | 1,850 | T, A, CS |
| | (3) Relation Recognition | Classification | DISRPT Shared Task* (2025) | 8,757 (43,716) | R, CS |
| Document | (4) Sentence Ordering | Seq-generation | AAN abstract (2018) | 806 (2,687) | Logi, R, CS |
| | | – | ArXiv abstract (2016) | 898 (179,691) | Logi, R, CS |
| | | – | Neurips abstract (2018) | 377 (377) | Logi, R, CS |
| | | – | NSF abstract(2018) | 814 (20,366) | Logi, R, CS |
| | | – | ROC stories (2016) | 883 (9,816) | Logi, R, CS |
| | | – | SIND (2016) | 808 (5,055) | Logi, R, CS |
| | | – | Wikipedia movie plots (2021) | 836 (3,345) | Logi, R, CS |
| | (5) Dialogue Disc. Parsing | Generation | STAC (2016) | 1,045 | R, T, L&S, CS |
| | | – | Molweni (2020) | 3,930 | R, T, L&S, CS |
| | | – | MSDC (2024b) | 1,474 | R, T, L&S, CS |

Table 1: Overview of level, task, nature of task, dataset, number of test examples, and discourse knowledge covered by BeDiscovER. The first three tasks are multi-class classification (except ToT-arithmetic), tasks (4) and (5) are generation tasks. '–' indicates same as above. # examples¶: we evaluate subsets of tasks (3) and (4) due to their extremely large size (original sizes in parentheses). Representative text examples are provided in appendix A. Discourse Knowledge‡ abbreviations: L&S, R, T, Logi, A, CS denote *Lexicon & Semantics*, *Rhetorical*, *Temporal*, *Logic*, *Arithmetic*, and *Commonsense*, respectively. DISRPT*: the latest 2025 version consists of 38 datasets.

## 3 BeDiscovER: Motivation and Method

### 3.1 Motivation

Our motivation stems from the convergence of two active research areas: discourse analysis and large language models. Traditionally, discourse analysis has centered on theoretical frameworks and phenomena such as anaphora resolution and textual coherence. More recently, it has expanded toward computational goals: modeling long-distance dependencies, generating long narratives, and supporting pragmatic reasoning (Parmar et al., 2024; Wu et al., 2024a; Cao et al., 2025). Meanwhile, LLMs have evolved from sentence-level learners to systems capable of long-document understanding, coherent generation, and advanced reasoning (Besta et al., 2025). This evolution underscores the growing need for discourse-level evaluation to diagnose weaknesses and guide further model development. Despite this convergence, research on LLMs' discourse understanding remains fragmented across tasks and frameworks (see §2), offering only a partial view of their true capabilities.

To address this gap, we introduce a unified benchmark to assess discourse knowledge across multiple dimensions, aiming to provide a big picture of the current state of discourse understanding in reasoning-oriented LLMs and their limitations.

### 3.2 Task Coverage

BeDiscovER tasks represent well-established discourse evaluation settings in three levels: **I. Lexicon**: Task (1) Discourse Marker Understanding. **II. Multi-sentence**: Task (2) Temporal Reasoning and (3) Discourse Relation Recognition. **III. Document-level structure**: Task (4) Sentence Ordering and (5) Dialogue Discourse Parsing. Examples for each task are provided in appendix A.1.

Although not exhaustive, our task selection is representative, covering diverse knowledge types. Each task is supported by multiple datasets, summarized in Table 1. When different splits exist, we only include test subset. We carefully curate datasets to capture complementary aspects of discourse phenomena. For example, within Task (2), TimeBank-Dense (Cassidy et al., 2014) targets events in adjacent sentences, while TDDiscourse (Naik et al., 2019) emphasizes long-distance relations. To further enrich this category, we include a subset of Test-of-Time (Fatemi et al., 2024) on arithmetic temporal reasoning.

To ensure diversity across languages, genres, and theoretical frameworks, we adopt the discourse relation classification subtask in DISRPT 2025[3]

[3]We do not incorporate other two tasks in DISRPT: segmentation is widely considered nearly solved ( 90% F1); connective identification is effectively a subset of relation classification (restricted to explicit cases) and thus less challenging.

rather than limiting to a single framework. The latest DISRPT version introduces a unified relation taxonomy that consolidates over 300 relation types into 17 unified relation types, facilitating cross-framework comparison (Braud et al., 2025). We further incorporate the most recent datasets: *Just* (Sheffield et al., 2025) and *Otherwise* (Liu et al., 2025) for Task (1). These corpora introduce an interesting perspective on polyfunctional discourse particles and their subtle effects on textual meaning, enabling assessment of LLMs' sensitivity to fine-grained pragmatic distinctions.

### 3.3 Model and Evaluation

We test the following reasoning-oriented LLMs: **(a) Qwen3 series** (Yang et al., 2025): We evaluate both the "thinking" mode (high reasoning effort) and "non-thinking" (low reasoning effort) mode. **(b) DeepSeek-R1** (Guo et al., 2025): DeepSeek reasoning model. **(c) DeepSeek-R1-distill-Qwen32B** (Guo et al., 2025): Qwen2.5-32B (Qwen, 2024) as the base model with direct distillation from DeepSeek-R1. **(d) GPT-5-mini** (OpenAI, 2025): We test "low" and "high" reasoning effort. In addition, we evaluate three widely used LLMs without dedicated reasoning-oriented training: GPT-4o-mini (Hurst et al., 2024), LLaMA-4-Scout (Llama, 2025), and Qwen2.5-72B-Instruct (Qwen, 2024). Model configurations are in appendix B.

We adopt QA-based prompting for its flexibility and its ability to support fine-grained interpretability. Precisely, we cast all tasks in an open-ended QA format to enable a unified evaluation pipeline. For classification tasks with a fixed label space, the label set is explicitly provided in the system prompt (appendix D). To mitigate model variation, each experiment is run three times. For clarity, we report average accuracies in the main text when applicable, while details (e.g., additional metrics and fine-grained scores per dataset) are provided in appendix E, along with error analysis in appendix C.

For open-source reasoning models such as Qwen3 and DeepSeek-R1, their reasoning process is observable, whereas in proprietary models like GPT-5, it remains implicit. Nevertheless, we can infer reasoning effort indirectly through the usage of thinking tokens. In appendix F, we visualize model performance under low (i.e., "non-thinking") versus high (i.e., "thinking") reasoning effort and present interesting observations.

## 4 Understanding on Discourse Lexicons

**Task (1) Discourse Marker Understanding** contains two newly curated datasets *Just* (with two subsets: Just-Manual and Just-Subtitle) (Sheffield et al., 2025) and *Otherwise* (Liu et al., 2025). Owing to their polyfunctional nature, discourse markers produce subtle yet diverse semantic and discourse effects (Lee, 1987; Rohde et al., 2016; Warstadt, 2020; Beltrama, 2022). While human can reliably distinguish these discourse functions in context (Rohde et al., 2018), the extent to which LLMs exhibit similar sensitivity and proficiency remains unclear.

### 4.1 Settings and Baselines

We design three QA-style prompting formats that vary in the amount of background information provided about the particle's functions: (i) BASIC: offers no explanation and directly asks the model to select a function label; (ii) DEF: adds a one-sentence definition for each label; and (iii) DEF+EXP: further includes a concrete example, resembling one-shot in-context learning. Prompt examples are provided in appendix D.1.

We compare with direct prompting results for *Just* and continuation acceptability scores for *Otherwise* in their original papers. *Just* paper adopts a prompting setup similar to our DEF+EXP, whereas *Otherwise* employs a continuation acceptability approach: given a left-hand side (LHS) and right-hand side (RHS) clause, the language model evaluates the most plausible connective linking them. These connectives—explicit markers such as *because if not*—capture the distinctive semantics associated with the adverbial *otherwise*. For evaluation, we report accuracy; other metrics (precision, recall, F1 scores) are provided in appendix E.1.

### 4.2 Results and Analysis

Table 2 presents main results of *Just-Manual*, *Just-Subtitle*, and *Otherwise* with BeDiscovER (top) and baselines (middle and bottom parts). Regarding prompting strategy, providing richer background information (DEF and DEF+EXP) generally improves performance, especially for large reasoning models (>32B), but yields only modest gains for non-reasoning models. This suggests that reasoning LLMs benefit more from in-context learning, whereas non-reasoning models rely more heavily on internal knowledge. This trend holds across both datasets. However, for smaller and medium-

| | Model | Size | *Just-Manual* (acc.) | | | *Just-Subtitle* (acc.) | | | *Otherwise* (acc.) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Basic | Def | Def+Exp | Basic | Def | Def+Exp | Basic | Def | Def+Exp |
| BeDiscovER | Qwen3-1.7B | 1.7B | $38.9_{.7}$ | $40.7_{4.2}$ | $40.2_{2.2}$ | $25.9_{0.9}$ | $26.2_{1.5}$ | $25.5_{1.3}$ | $23.9_{0.6}$ | $21.2_{1.3}$ | $16.2_{1.3}$ |
| | Qwen3-14B | 14B | $60.2_{.6}$ | $54.9_{2.0}$ | $55.6_{1.4}$ | $43.2_{2.1}$ | $32.9_{1.7}$ | $33.2_{1.3}$ | $32.1_{.7}$ | $\underline{45.2}_{2.5}$ | $51.4_{1.6}$ |
| | Qwen3-32B | 32B | $59.8_{1.8}$ | $65.6_{3.5}$ | $64.4_{2.0}$ | $48.5_{1.4}$ | $48.9_{2.0}$ | $47.1_{1.2}$ | $34.3_{.4}$ | **$49.3_{2.6}$** | $\underline{66.1}_{.6}$ |
| | DS-r1-distill-Qwen | 32B | $56.7_{1.1}$ | $63.7_{1.5}$ | $59.6_{4.0}$ | $31.8_{4.3}$ | $40.5_{2.9}$ | $35.8_{.2}$ | $33.1_{1.5}$ | $32.4_{1.8}$ | $44.3_{3.5}$ |
| | DeepSeek-r1-0528 | 37/671B | $53.7_{.4}$ | $\underline{66.3}_{.7}$ | $67.0_{1.4}$ | $48.1_{3.1}$ | $\underline{60.2}_{1.8}$ | $59.5_{4.5}$ | $31.2_{.9}$ | $38.9_{1.8}$ | $58.8_{1.7}$ |
| | GPT-5-mini (low) | Unknown | **$63.3_{2.2}$** | $65.9_{.3}$ | $65.6_{.0}$ | $\underline{56.4}_{1.4}$ | $58.4_{1.4}$ | $\underline{62.9}_{.5}$ | $\underline{34.6}_{.6}$ | $36.8_{1.1}$ | $51.9_{3.3}$ |
| | GPT-5-mini (high) | Unknown | $\underline{61.7}_{.6}$ | **$66.1_{1.7}$** | $65.6_{1.1}$ | **$57.4_{1.7}$** | **$63.1_{.0}$** | **$63.4_{.3}$** | **$35.5_{.5}$** | $43.9_{1.0}$ | **$71.8_{1.7}$** |
| | GPT-4o-mini | Unknown | $51.5_{.4}$ | $60.7_{1.8}$ | $66.8_{3.4}$ | $24.6_{1.8}$ | $29.1_{.2}$ | $28.0_{1.8}$ | $22.0_{.6}$ | $21.4_{1.0}$ | $30.6_{.7}$ |
| | Llama-4-Scout | 17/109B | $60.6_{.6}$ | $61.7_{.6}$ | $65.0_{.6}$ | $34.6_{.3}$ | $38.9_{.7}$ | $38.9_{.0}$ | $31.6_{.3}$ | $19.2_{.2}$ | $37.6_{1.2}$ |
| | Qwen2.5-72B | 72B | $61.1_{1.1}$ | $66.7_{1.1}$ | $66.1_{1.7}$ | $32.6_{1.0}$ | $31.2_{1.0}$ | $38.9_{.7}$ | $34.7_{.0}$ | $31.3_{1.0}$ | $32.0_{1.0}$ |
| Contin. | GPT-Neo-1.3B | 1.3B | | - | | | - | | | 56.0† (surprisal score) | |
| | Mistral-7B-v0.1 | 7B | | - | | | - | | | 59.0† (surprisal score) | |
| Prompting | Mistral-0.3-7B (2025) | 7B | | | $\underline{69}$* | | | 27* | | - | |
| | OLMo2-13B (2025) | 13B | | | 46* | | | 18* | | - | |
| | Llama-3.3-70B (2025) | 70B | | | **75*** | | | 35* | | - | |

Table 2: **Task (1) Discourse Marker Understanding** performance with BeDiscovER and baselines on *Just-Manual*, *Just-Subtitle*, and *Otherwise* datasets. Top: BeDiscovER scores with reasoning and non-reasoning LLMs. Middle: *Otherwise* dataset baselines (2025), † denotes surprisal scores which calculates the continuation acceptability in language models. Bottom: prompting baselines in *Just* (2025). *are *approximated* numbers from the Figure 2 in the original paper. Best score per column is in **red** and second best underlined.

sized models (small: <4B; medium: 4B–32B), additional context does not always help. On the more challenging *Just-Subtitle* and *Otherwise* datasets, excessive contextual information can even degrade performance. A relevant phenomenon is noted by Hu and Frank (2024), who show that increased task complexity can widen the demand gap for smaller models with limited parameters and training data, leading to reduced accuracy. We also observe a critical model size threshold (>2B parameters), beyond which performance improves sharply—consistent with findings by Sheffield et al. (2025). Compared to baselines, GPT-5-mini with high reasoning effort achieves the best performance, setting new state-of-the-art results on *Just–Subtitle* and *Otherwise*.

## 5 Understanding on (multi-)Sentences

### 5.1 Temporal Relation

**Task (2) Temporal Reasoning** requires EVENT-EVENT pair reasoning across sentences, both adjacent pairs as in TBD (Cassidy et al., 2014)) and distant as in TDD-Man (Naik et al., 2019)). We also include Test-of-Time (ToT) (Fatemi et al., 2024) to test arithmetic skills in temporal reasoning.

#### 5.1.1 Settings and Baselines

In TBD and TDD-Man, we use full news documents with target events annotated by <EVENT></EVENT> tags. Our prompt, adapted from the best-performing template in Yuan et al. (2023), presents relation labels as a multiple-choice QA task (see appendix D.2).

For baselines, we consider both prompting-based baselines using LLMs, such as ChatGPT (Yuan et al., 2023; Chan et al., 2024), GPT-4 (Fatemi et al., 2024), and Llama (Fan and Strube, 2025), as well as supervised approaches, including graph-based neural networks UCGraph, RSGT, and DTRE (Liu et al., 2021; Zhou et al., 2022; Wang et al., 2022), prototypical network CPTRE (Yuan et al., 2024), and a fine-tuned Llama (Fan and Strube, 2025). Most supervised baselines use much smaller pre-trained language models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) for node/label representation learning.

#### 5.1.2 Results and Analysis

Table 3 summarizes the main results across all three datasets, with details provided in appendix E.2. Overall, recent reasoning LLMs substantially outperform earlier models such as ChatGPT and non-reasoning counterparts, achieving gains of 10–20% on TBD and TDD-Man, and a notable 30-50% improvement on ToT-Arithmetic. These results indicate enhanced temporal reasoning capability, particularly in arithmetic-related temporal inference. However, compared with supervised baselines, zero-shot LLMs still underperform on TBD and TDD-Man. Both datasets involve full documents containing multiple events, requiring LLMs to reconstruct the underlying narrative to determine

| | Model | Size | TBD (f1) | TDD-Man (f1) | ToT-Ari (acc.) |
|---|---|---|---|---|---|
| BeDiscovER | Qwen3-1.7B | 1.7B | 26.2$_{0.8}$ | 21.9$_{0.6}$ | 46.5$_{0.5}$ |
| | Qwen3-14B | 14B | 39.8$_{0.1}$ | 29.5$_{0.6}$ | 80.4$_{1.2}$ |
| | Qwen3-32B | 32B | 40.8$_{0.5}$ | 31.6$_{0.3}$ | 81.6$_{0.5}$ |
| | DS-r1-distill-Qwen | 32B | 23.8$_{2.5}$ | 18.6$_{3.3}$ | 68.0$_{1.2}$ |
| | DeepSeek-r1-0528 | 671B | 33.1$_{0.0}$ | 22.2$_{0.0}$ | 63.3$_{0.5}$ |
| | GPT-5-mini (low) | Unk. | 38.7$_{0.2}$ | 34.4$_{0.2}$ | 87.2$_{0.1}$ |
| | GPT-5-mini (high) | Unk. | **43.4$_{0.3}$** | **36.7$_{0.3}$** | **88.2$_{0.0}$** |
| | GPT-4o-mini | Unk. | 27.8$_{.3}$ | 25.0$_{.2}$ | 31.4$_{.1}$ |
| | Llama-4-Scout | 109B | 29.3$_{.1}$ | 27.0$_{.2}$ | 30.4$_{.2}$ |
| | Qwen2.5-72B | 72B | 33.2$_{.7}$ | 28.4$_{.3}$ | 30.4$_{.2}$ |
| Prompting | ChatGPT (2023) | Unk. | 37.0 | 24.3 | - |
| | ChatGPT (2024) | Unk. | 27.0 | 16.8 | - |
| | GPT-4 (2024) | Unk. | - | - | 54.2 |
| | Llama3.3-70B (2025) | 70B | - | 26.8 | - |
| Supervised | UCGraph (2021) | 125B | 59.1 | 43.4 | - |
| | RSGT (2022) | 355M | 68.7 | - | - |
| | DTRE (2022) | 125M | **70.2** | 56.3 | - |
| | CPTRE (2024) | 125M | 61.1 | 56.5 | - |
| | FT Llama3.3 (2025) | 70B | - | **57.9** | - |

Table 3: **Task (2) Temporal Reasoning** performance: BeDiscovER (top), LLM-prompting baselines (middle), and supervised baselines (bottom) on TBD, TDD-Man, and ToT-arithmetic datasets. Scores are micro-F1. In the first four supervised baselines, "size" refers to the parameter count of the PLM encoders (BERT and RoBERTa). The last supervised baseline uses a fine-tuned Llama3.3-70B model. Best unsupervised scores are in **red** and best supervised scores in **blue**.

the correct event order (e.g., an *investigation* typically follows a *crime*): a task that remains challenging even for advanced reasoning models.

In an ablation study, we simplify inputs by limiting context to sentences containing the target events. This pre-selection reduces reasoning load and yields modest gains on TBD (2–3%; tables 24 and 25). Conversely, for TDD-Man, where events are farther apart, removing intermediate context leads to a 2–9% drop in performance (tables 26 and 27). These findings suggest that LLM reasoning is sensitive to both context complexity and the availability of relevant information.

## 5.2 Discourse Relation

**Task (3): Discourse Relation Recognition** adopts data from the latest version of DISRPT Shared Task, which provides a unified label set of 17 relations across 38 datasets (Braud et al., 2025). The task requires predicting a rhetorical relation (e.g., *contrast*) between two segments, Arg1 and Arg2.

### 5.2.1 Settings and Baselines

We enrich the input prompt with additional features, including language, framework, corpus name, argument direction (Arg1→Arg2 or inverse), and context in which the arguments appear, motivated

| | Model | DEP | eRST | ISO | PDTB | RST | SDRT |
|---|---|---|---|---|---|---|---|
| BeDiscovER | Qwen3-1.7B | 21.2 | 19.7 | 25.0 | 22.1 | 18.0 | 18.7 |
| | Qwen3-14B | 37.9 | 32.7 | 41.4 | 36.4 | 33.6 | 26.2 |
| | Qwen3-32B | 39.5 | 34.7 | 40.3 | 36.8 | 35.5 | 29.9 |
| | DS-r1-distill-Qwen | 32.6 | 29.9 | 38.7 | 29.8 | 32.9 | 24.6 |
| | DeepSeek-r1-0528 | 44.3 | **36.6** | 51.4 | 43.7 | 43.6 | 26.6 |
| | GPT-5-mini (low) | 46.1 | 33.9 | 47.3 | 42.9 | 42.3 | 34.2 |
| | GPT-5-mini (high) | **51.2** | 36.0 | **56.8** | **47.4** | **46.6** | **37.9** |
| | GPT-4o-mini | 28.3 | 25.9 | 28.6 | 31.2 | 24.2 | 25.3 |
| | Llama-4-Scout | 43.2 | 34.3 | 44.8 | 36.0 | 36.6 | 25.3 |
| | Qwen2.5-72B | 44.9 | 34.0 | 37.6 | 31.9 | 34.7 | 26.4 |
| Supervised | DeDisCo (2025) | **77.2** | **71.8** | **72.0** | **79.0** | **64.9** | **83.0** |
| | HITS (2025) | 74.1 | 64.5 | 72.0 | 76.3 | 61.9 | 82.1 |
| | DisCreT (2025) | 72.3 | 58.1 | 60.0 | 75.3 | 56.4 | 77.5 |
| | CLAC (2025) | 74.9 | 57.5 | 54.8 | 74.7 | 56.2 | 77.4 |
| | SeCoRel (2025) | 69.7 | 53.8 | 52.8 | 70.6 | 52.5 | 76.4 |

Table 4: **Task (3): Discourse Relation Recognition [framework-split]**. BeDiscovER results (top) vs. supervised systems (bottom; all <4B parameters). Framework abbreviations: DEP – Dependency Structure (2018); eRST – Enhanced RST (2025); ISO – ISO Framework (2016); PDTB – Penn Discourse TreeBank (2005); RST – Rhetorical Structure Theory (1988); SDRT – Segmented Discourse Representation Theory (2003).

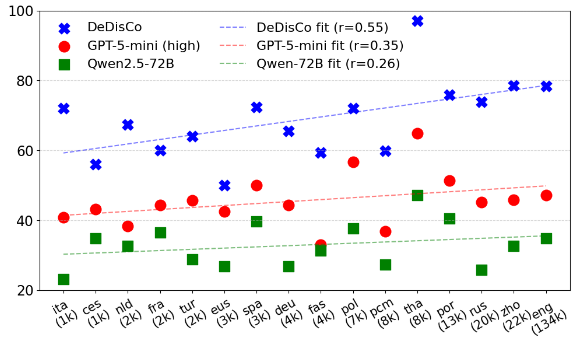


Figure 1: **Task (3): Discourse Relation Recognition [language-split]**. Best supervised (DeDisCo), reasoning and non-reasoning models (GPT-5-mini, Qwen2.5-72B), with languages ordered by increasing training size (Italian, Czech, Dutch, French, Turkish, Basque, Spanish, German, Persian, Polish, Nigerian Pidgin, Thai, Portuguese, Russian, Chinese, English).

by prior studies (Gessler et al., 2021; Metheniti et al., 2024). Our prompting setup closely follows DeDisCo (Ju et al., 2025) (see appendix D.3), the top-performing system in the DISRPT25 competition. Despite the multilingual data, we use English for all prompt instructions. All baseline systems are supervised and trained with a model size cap of 4B parameters.

### 5.2.2 Results and Analysis

Tables 4 and Figure 1 give the framework-view and language-view results, respectively. We show detailed scores for language-view and dataset-view in tables 29 to 33 in appendix E.3.

All reasoning LLMs perform worse than super-

| | Model | AAN abstract | | ArXiv abstract | | Neurips abstract | | NSF abstract | | ROC stories | | SIND | | Wiki. mv. plts. | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | PMR | Acc | PMR | Acc | PMR | Acc | PMR | Acc | PMR | Acc | PMR | Acc | PMR | Acc |
| BeDiscovER | Qwen3-1.7B | $21.1_{.5}$ | $38.0_{.5}$ | $19.0_{.5}$ | $31.4_{.3}$ | $11.0_{2.2}$ | $30.8_{.7}$ | $6.6_{0.5}$ | $14.7_{.1}$ | $17.1_{.7}$ | $44.7_{.9}$ | $14.6_{1.0}$ | $35.6_{1.0}$ | $8.9_{.3}$ | $12.7_{.1}$ |
| | Qwen3-14B | $55.5_{.9}$ | $70.6_{.3}$ | $43.4_{.3}$ | $59.1_{.6}$ | $43.0_{.8}$ | $66.1_{.1}$ | $14.3_{.0}$ | $29.2_{.6}$ | $67.3_{.4}$ | $82.5_{.4}$ | $31.2_{.9}$ | $55.1_{.4}$ | $18.2_{.1}$ | $27.1_{.2}$ |
| | Qwen3-32B | $59.6_{.4}$ | $74.0_{.3}$ | $46.1_{.4}$ | $61.9_{.4}$ | $44.2_{.4}$ | $68.6_{.2}$ | $14.3_{.3}$ | $30.2_{.4}$ | $70.6_{.7}$ | $84.4_{.3}$ | $31.2_{.5}$ | $55.3_{.3}$ | $19.1_{.6}$ | $29.1_{.5}$ |
| | DS-r1-distill-Qwen | $51.4_{.9}$ | $67.7_{.5}$ | $38.8_{.4}$ | $55.4_{.2}$ | $38.9_{.4}$ | $63.5_{.7}$ | $11.4_{.2}$ | $26.0_{.2}$ | $63.2_{2.8}$ | $79.8_{1.6}$ | $29.0_{1.8}$ | $53.3_{1.2}$ | $17.6_{.6}$ | $24.3_{.5}$ |
| | DeepSeek-r1-0528 | $64.8_{.4}$ | $75.5_{.2}$ | $46.6_{.5}$ | $61.9_{.3}$ | $56.8_{.5}$ | $74.9_{.6}$ | $15.7_{.2}$ | $28.5_{.2}$ | **$77.1_{1.2}$** | **$88.2_{.6}$** | **$36.4_{1.3}$** | **$58.4_{.9}$** | **$23.8_{.4}$** | $28.0_{.4}$ |
| | GPT-5-mini (low) | $61.1_{.6}$ | $75.4_{.0}$ | $49.3_{.1}$ | $65.6_{.4}$ | $50.5_{.4}$ | $73.1_{.1}$ | $15.8_{.1}$ | $33.2_{.4}$ | $68.3_{1.1}$ | $83.6_{.6}$ | $31.5_{1.1}$ | $54.6_{.8}$ | $20.3_{.1}$ | $32.0_{.2}$ |
| | GPT-5-mini (high) | **$66.3_{.7}$** | **$78.7_{.1}$** | **$55.0_{.1}$** | **$69.8_{.1}$** | **$58.6_{.3}$** | **$78.6_{.2}$** | **$17.8_{.2}$** | **$35.5_{.1}$** | $75.7_{1.0}$ | $87.3_{.04}$ | $35.0_{1.5}$ | $58.2_{1.0}$ | $23.0_{.5}$ | **$40.1_{.4}$** |
| | GPT-4o-mini | $37.2_{.7}$ | $54.7_{.1}$ | $30.2_{.4}$ | $43.9_{.2}$ | $22.8_{.0}$ | $49.1_{.8}$ | $10.5_{0.1}$ | $21.1_{.1}$ | $47.3_{.1}$ | $69.8_{.4}$ | $25.6_{.2}$ | $50.2_{.1}$ | $14.3_{.1}$ | $19.4_{.2}$ |
| | Llama-4-Scout | $42.9_{.0}$ | $61.9_{.1}$ | $33.6_{.0}$ | $51.2_{.1}$ | $31.4_{.7}$ | $58.1_{.1}$ | $10.5_{0.3}$ | $26.2_{.1}$ | $49.1_{.3}$ | $71.8_{.2}$ | $27.5_{.3}$ | $50.9_{.2}$ | $16.1_{.3}$ | $22.3_{.3}$ |
| | Qwen2.5-72B | $55.2_{.6}$ | $71.1_{.2}$ | $41.3_{.1}$ | $58.6_{.2}$ | $48.3_{.8}$ | $70.7_{.1}$ | $15.0_{0.5}$ | $30.7_{.2}$ | $66.4_{.5}$ | $81.9_{.1}$ | $31.9_{.2}$ | $55.8_{.2}$ | $19.7_{.0}$ | $28.8_{.3}$ |
| Supervised | BERSON (2020) | 59.8 | 78.0 | 56.1 | 75.1 | 48.0 | 73.9 | 23.1 | 50.2 | 68.2 | 82.9 | 31.7 | 58.9 | - | - |
| | Re-BART (2021) | 73.5 | 84.3 | 62.4 | 74.3 | 57.0 | 77.4 | 29.7 | 50.2 | 81.9 | 90.8 | 43.2 | 64.5 | **25.8** | **42.0** |
| | CoVER (2023) | 59.2 | 78.1 | - | - | 50.9 | 74.9 | - | - | 72.3 | 84.8 | 33.0 | 60.3 | - | - |
| | NAON-BART (2023) | **73.9** | **87.2** | **62.6** | **79.2** | **61.2** | **84.2** | **30.6** | **54.8** | **89.1** | **95.1** | **55.6** | **79.6** | - | - |

Table 5: **Task (4) Sentence Ordering** main results. We present perfect match rate (PMR) and accuracy, with additional metrics in appendix E.4. Best unsupervised scores are in **red** and best supervised scores in **blue**.

vised baselines, with an average gap of about 20 points, while non-reasoning models consistently trail reasoning models. Among the frameworks, eRST yields the lowest scores—likely because the framework is relatively new and its theoretical foundations have seen limited exposure in modern LLM—with even GPT-5-mini achieving only the low 30s. The ISO framework attains the highest LLM performance at 57%, while remains well below supervised results (72%). In contrast, all fine-tuned systems perform markedly better, with DeDisCo reaching 70–80% across frameworks.

From a language perspective (Figure 1), BeDiscovER models exhibit nearly flat (and low) performance across languages ($r = 0.2$–$0.3$, not significant), whereas supervised models show clear disparities between high-resource (e.g., Chinese (zho), English (eng)) and low-resource languages (e.g., Czech (ces), Basque (eus)). This pattern suggests that LLMs possess a certain degree of cross-lingual generalizability, but still lack robust internal representations of discourse relations, as evidenced by the low overall accuracy. Interestingly, two mid-resource languages (≈8k): Thai (tha) and Nigerian Pidgin (pcm), show completely contrasting outcomes both in all models. This difference likely reflects divergent discourse properties: the Thai corpus is annotated heavily with explicit discourse connectives (Prasertsom et al., 2024), making the task easier, whereas Nigerian Pidgin has highly flexible syntax and fewer explicit connectives (Scholman et al., 2025), increasing task difficulty.

# 6 Understanding on Document Structure

## 6.1 Sentence Ordering

**Task (4) Sentence Ordering** reorders a set of shuffled sentences into a coherent text. It has been used to assess a model's understanding of causal and temporal relations (Barzilay and Lapata, 2008) and shown to aid discourse structure extraction (Li et al., 2023). We include, in this task, 7 publicly available datasets from two domains: scientific paper abstracts (AAN, ArXiv, Neurips, NSF (Wang et al., 2018; Chen et al., 2016; Logeswaran et al., 2018)) and narratives (ROC stories, SIND, Wiki movie plots (Mostafazadeh et al., 2016; Huang et al., 2016)), with examples in appendix A.1.

### 6.1.1 Settings and Baselines

In our setup, we adopt a text-to-marker format that prompts the model to generate a sequence of reordered sentence labels (e.g., <s3>, <s1>) as output, following ReBART (Basu Roy Chowdhury et al., 2021) (prompts in appendix D.4).

Supervised baselines use PLMs to encode sentences, followed by either a Pointer Network to predict sentence positions (e.g., BERSON and CoVER (Cui et al., 2020; Jia et al., 2023)) or a sequence-to-sequence model that generates the position sequence autoregressively, such as ReBART (Basu Roy Chowdhury et al., 2021), or non-autoregressively: NAON-BART (Bin et al., 2023).

### 6.1.2 Results and Analysis

Table 5 presents the performance of BeDiscovER and supervised models across seven datasets. A clear scaling trend emerges where larger LLMs consistently outperform smaller ones. In scientific abstract domain, GPT-5-mini (high) achieves competitive results, potentially benefiting from exposure to similar text during pretraining, though there remains room for improvement on long-input dataset such as NSF. In narrative datasets, GPT-5-mini and DeepSeek-R1 perform comparably, matching supervised performance on Wiki Movie, a chal-

| | Model | Size | STAC | | Molweni | | MSDC | |
|---|---|---|---|---|---|---|---|---|
| | | | Link | Full | Link | Full | Link | Full |
| BeDiscovER | Qwen3-1.7B | 1.7B | $35.9_{1.8}$ | $5.2_{0.0}$ | $43.1_{0.3}$ | $5.3_{0.3}$ | $36.7_{2.8}$ | $3.6_{0.2}$ |
| | Qwen3-14B | 14B | $61.3_{0.8}$ | $29.8_{1.4}$ | $\underline{60.1}_{0.6}$ | $25.7_{0.7}$ | **$72.7$**$_{0.2}$ | $22.6_{0.1}$ |
| | Qwen3-32B | 32B | $62.4_{1.1}$ | $31.3_{0.3}$ | $58.3_{0.6}$ | $22.2_{0.9}$ | $\underline{70.5}_{0.1}$ | $23.0_{0.3}$ |
| | DS-r1-distill-Qwen | 32B | $61.5_{0.7}$ | $22.3_{2.3}$ | $58.6_{0.7}$ | $16.1_{0.3}$ | $69.6_{0.4}$ | $25.7_{1.1}$ |
| | DeepSeek-r1-0528 | 37/671B | **$66.3$**$_{0.1}$ | $\underline{38.7}_{0.5}$ | $56.7_{0.7}$ | $22.3_{1.3}$ | $69.6_{0.7}$ | **$34.2$**$_{0.7}$ |
| | GPT-5-mini (low) | Unk. | $61.1_{0.3}$ | $32.1_{0.8}$ | $55.6_{0.1}$ | $\underline{27.7}_{0.3}$ | $66.1_{0.1}$ | $26.8_{0.4}$ |
| | GPT-5-mini (high) | Unk. | $\underline{66.0}_{0.2}$ | **$38.8$**$_{0.7}$ | $58.4_{0.6}$ | **$30.2$**$_{0.3}$ | $66.4_{0.7}$ | $\underline{31.9}_{0.5}$ |
| | GPT-4o-mini | Unk. | $53.9_{.1}$ | $10.3_{.1}$ | $49.7_{.3}$ | $13.9_{.2}$ | $56.3_{.2}$ | $23.6_{.0}$ |
| | Llama-4-Scout | 17/109B | $57.7_{.2}$ | $13.7_{.2}$ | $58.4_{.1}$ | $11.0_{.1}$ | $70.2_{.1}$ | $17.3_{.1}$ |
| | Qwen2.5-72B | 72B | $58.8_{.6}$ | $15.4_{.0}$ | $55.7_{.1}$ | $12.9_{.4}$ | $64.6_{.1}$ | $18.1_{.2}$ |
| Prompt | ChatGPT zero-shot (2024) | Unk. | 20.0 | 4.4 | 28.3 | 5.4 | - | - |
| | ChatGPT prompt engineering (2024) | Unk. | 59.9 | 25.3 | **63.7** | 23.8 | - | - |
| Supervised | RoBERTa+pointer (2021) | 125M | 72.9 | 57.0 | $\underline{79.0}$ | 55.4 | - | - |
| | RoBERTa+CLE (2022) | 125M | $\underline{73.0}$ | $\underline{58.1}$ | 81.0 | $\underline{58.6}$ | - | - |
| | T0+transition (2024) | 3B | 72.3 | 56.6 | **83.4** | **60.0** | - | - |
| | Llama3 (2024a) | 8B | **77.5** | **60.7** | - | - | **88.3** | **79.5** |

Table 6: **Task (5) Dialogue Discourse Parsing** F1 scores on STAC, Molweni, and MDSC: BeDiscovER (top), LLM-based prompting (middle) and supervised baselines (bottom). Best unsupervised scores are in **red** and best supervised scores in **blue**.

lenging dataset with the longest document length. Overall, these results suggest that reasoning LLMs, especially larger ones, possess a good grasp of textual coherence, possibly stemming from their extensive pretraining on narrative-rich corpora.

## 6.2 Dialogue Discourse Parsing

**Task (5): Dialogue Discourse Parsing** focuses on structure construction across the entire document. We additionally target dialogue rather than monologue to introduce greater genre diversity. This task includes three widely used dialogue discourse datasets: STAC (Asher et al., 2016), Molweni (Li et al., 2020), and the recently annotated MSDC based on Minecraft Dialogue Corpus (Thompson et al., 2024b), all datasets in the SDRT framework (Asher and Lascarides, 2003).

### 6.2.1 Settings and Baselines

Directly prompting LLMs to generate SDRT-style discourse graphs from raw text remains challenging, as shown in prior studies with ChatGPT (Chan et al., 2024; Fan et al., 2024).

Following the parsing-as-generation paradigm (Li et al., 2024), we instead prompt the model to produce relation triples incrementally, where model predicts relations for each new utterance and appends the generated structure as context for subsequent predictions (prompt given in appendix D.5). This incremental approach has achieved state-of-the-art performance with both sequence-to-sequence models (e.g., T0+Transition) and decoder-only architectures such as Llamipa (Thompson et al., 2024a). For comparison, we also include earlier baselines employing smaller PLMs with task-specific decoding strategies (Liu and Chen, 2021; Chi and Rudnicky, 2022). For evaluation, we employ the micro-averaged F1 scores for link attachment (Link) and full structure (Full).

### 6.2.2 Results and Analysis

Table 6 reports BeDiscovER performance using incremental prompting. This approach establishes a new state of the art using zero-shot prompting, achieving a 7–13% improvement in full structure prediction over prior ChatGPT results (Chan et al., 2024; Fan et al., 2024). Despite these gains, all reasoning and non-reasoning LLMs still trail supervised baselines by a wide margin of 20–30%, particularly on full structure prediction, which requires accurate identification of both discourse links and their relations. The relation component appears to be the primary bottleneck, with relation recognition in the SDRT scores below 40 (discussed in §5.2.2).

In an oracle setting, where gold triples are provided at each step (see "single-turn" setup in appendix E.5), we observe only modest improvements over auto-regressive parsing, suggesting limited impact from error propagation. Notably, models with higher reasoning effort outperform their lower-effort counterparts, especially on full-structure prediction, indicating that explicit reasoning enhances discourse parsing performance.

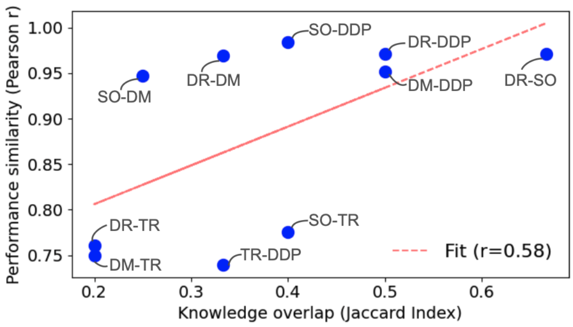


Figure 2: The correlation between sub-task performance similarity (Pearson r) vs. sub-task knowledge overlap, calculated using 7 reasoning LLMs scores). Task abbreviations: DM: discourse marker, TR: temporal reasoning, DR: discourse relation, SO: sentence ordering, DDP: dialogue discourse parsing.

## 7 Inter-task Correlation Analysis

We analyze the correlation among the five subtasks in BeDiscovER: for each task, we aggregate results (accuracies) across all datasets. This yields one task score per model. We take data points from the seven reasoning-oriented LLMs. The pairwise correlations across all subtasks exhibit strong positive relationships ($r \in [0.73, 0.98]$, $p < 0.05$).

To understand what drives the strong performance correlations among tasks, we further examine the discourse knowledge associated with each subtask (e.g., lexical, temporal, rhetorical), as shown in the last column of Table 1. Although this mapping from tasks to required knowledge is not exhaustive, it provides a useful basis for our analysis. To quantify how much knowledge two subtasks share, we compute their Jaccard overlap: Jaccard(A,B) $= \frac{|A \cup B|}{|A \cap B|}$, where the intersection indicates number of shared knowledge types, and union the number of knowledge types required by at least one of the tasks. Naturally, a higher overlap score means the two tasks rely on more of the same underlying knowledge.

Figure 2 shows the relationship between task performance similarity ($r$) and knowledge overlap (Jaccard index). We find a moderate positive correlation between knowledge overlap and performance similarity ($r' = 0.58$), suggesting that shared underlying knowledge types explain why some subtasks exhibit similar performance. For example, dialogue discourse parsing (DDP) task shares substantial knowledge with discourse relation classification (DR) and discourse marker understanding (DM), leading to high performance similarity, whereas temporal reasoning (TR) appears to rely on largely distinct knowledge and shows lower similarity. This corroborates some findings in the recent work on discursive circuits (Miao and Kan, 2025), where different frameworks show a consistent use of linguistic sub-skills in discourse relation. Additionally, we find that performance similarity is not determined solely by the knowledge taxonomy: the sentence ordering (SO) and DDP tasks exhibit the highest correlation (0.98) despite only moderate overlap (0.4). This likely reflects that both are document-level tasks requiring modeling of global structure in long context—a factor that is not captured by the knowledge types. Overall, these findings are intriguing and motivate future work to extend BeDiscovER with tasks that cover a broader range of knowledge types.

## 8 Conclusion and Discussion

We introduce BeDiscovER, a comprehensive benchmark designed to evaluate the discourse knowledge of reasoning-oriented LLMs. Covering 3 levels, 5 tasks, and 52 datasets, BeDiscovER offers a linguistically grounded and multifaceted assessment of how current LLMs understand discourse structure and coherence.

BeDiscovER also serves as a baseline and reference point for future research on discourse-aware LLMs. Our findings highlight persistent challenges, particularly in temporal reasoning (logical not arithmetic) and rhetorical relation recognition where LLMs still lag behind humans and supervised systems. Arguably, these shortcomings may stem from limited exposure to (explicit) structural phenomena during pre-/post-training and from the difficulty of transferring such abstract reasoning skills. From an interpretability perspective, the diverse tasks in BeDiscovER also offer a rich testbed for probing the internal mechanisms of LLMs, complementing recent work such as CausalGym (Arora et al., 2024) and discursive circuits (Miao and Kan, 2025).

As an initial step toward deeper insight, we visualize the relationship between thinking effort and performance (appendix F). On ToT arithmetic temporal reasoning, LLMs' performances did consistently improve with longer reasoning traces. However, across many other tasks, longer reasoning traces **do not** necessarily yield better outcomes. Models often become more verbose without producing more meaningful reasoning. Improving the quality of reasoning thus remains a promising direction for advancing discourse-aware LLMs.

## Limitations

BeDiscovER is designed to provide a broad, high-level overview of recent LLM performance across diverse discourse tasks. Although we made considerable efforts to include a broad range of datasets, it is impractical to cover all available benchmarks within each task. Also, as many underlying datasets are publicly available, it is difficult to determine what data proprietary models were trained on, and thus whether their performance reflects genuine reasoning ability or parametric memorization. To address this, we intentionally include a few recent datasets (e.g., Just, Otherwise, DISRPT 2025 version, MSDC) to reduce the chance of their usage as pretraining data. Furthermore, we release our evaluation pipeline to support evaluation of future new datasets and more advanced LLMs, aimed at making BeDiscovER a dynamic benchmark and allowing the community to continuously evaluate LLMs over time.

For evaluation, we adopt a unified open-ended QA prompting strategy to ensure simplicity and consistency across tasks. While model outputs can be sensitive to prompt phrasing, we intentionally employ minimal and standardized wording and retain default or suggested hyperparameter settings. We make all prompt templates and model hyperparameter publicly available (appendices B and D) to facilitate replication and future comparison. We acknowledge that alternative methods exist, such as sentence log-probabilities (Hu and Levy, 2023) and surprisal-based predictability scores (Giulianelli et al., 2023; Tsipidi et al., 2024) have been used to probe linguistic sensitivity. However, these approaches are less suited for structure prediction in the benchmark.

Regarding the benchmark's "ceiling performance", we do not have human evaluators providing directly comparable scores, and therefore use inter-annotation agreement (IAA) as a proxy (appendix A.2). We acknowledge that IAA is an imperfect substitute for direct human evaluation, given differences in task formulation (e.g., text prompting vs. IAA task descriptions, original label sets vs. the unified DISRPT labels). Nonetheless, IAA offers a useful indicator of how consistently human annotators perform each task, providing insights into task difficulty.

Finally, we note that, in its current form, BeDiscovER is intended as a benchmark for assessing discourse knowledge (by measuring answer correctness); it is not yet positioned as a benchmark for evaluating reasoning quality. Our initial attempt in this regard is to track the "reasoning trace" or chain-of-thought tokens understand the model's process (appendix F). To move toward assessing reasoning quality more directly, there may involve different strategies. For instance, future work can conduct human evaluation on part of the benchmark. Such evaluation could provide direct feedback on the strengths and weaknesses of model reasoning traces and, importantly, inform how high-quality reasoning traces might be leveraged in future pre-/post-training strategies.

## Acknowledgments

The authors thank the anonymous reviewers and the Area Chair for their valuable feedback and suggestions. The authors acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC). Nous remercions le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG) de son soutien.

## References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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, Online. Association for Computational Linguistics.

Aryaman Arora, Dan Jurafsky, and Christopher Potts. 2024. CausalGym: Benchmarking causal interpretability methods on linguistic tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14638–14663, Bangkok, Thailand. Association for Computational Linguistics.

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

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

Yushi Bai, Jiajie Zhang, Xin Lv, Linzhi Zheng, Siqi Zhu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. Longwriter: Unleashing 10,000+ word generation from long context llms. In *The Thirteenth International Conference on Learning Representations*.

Souvik Banerjee, Yi Fan, and Michael Strube. 2025. HITS at DISRPT 2025: Discourse segmentation, connective detection, and relation classification. In *Proceedings of the 4th Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2025)*, pages 63–78, Suzhou, China. Association for Computational Linguistics.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Somnath Basu Roy Chowdhury, Faeze Brahman, and Snigdha Chaturvedi. 2021. Is everything in order? a simple way to order sentences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10769–10779, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Andrea Beltrama. 2022. Just perfect, simply the best: an analysis of emphatic exclusion. *Linguistics and Philosophy*, 45(2):321–364.

Maciej Besta, Julia Barth, Eric Schreiber, Ales Kubicek, Afonso Catarino, Robert Gerstenberger, Piotr Nyczyk, Patrick Iff, Yueling Li, Sam Houliston, and 1 others. 2025. Reasoning language models: A blueprint. *arXiv preprint arXiv:2501.11223*.

Yi Bin, Wenhao Shi, Bin Ji, Jipeng Zhang, Yujuan Ding, and Yang Yang. 2023. Non-autoregressive sentence ordering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4198–4214.

Chloé Braud, Amir Zeldes, Chuyuan Li, Yang Janet Liu, and Philippe Muller. 2025. The DISRPT 2025 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In *Proceedings of the 4th Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2025)*, pages 1–20, Suzhou, China. Association for Computational Linguistics.

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

Lanni Bu, Lauren Levin, and Amir Zeldes. 2025. Discotrack: A multilingual llm benchmark for discourse tracking. *arXiv preprint arXiv:2510.17013*.

Jan Buchmann, Max Eichler, Jan-Micha Bodensohn, Ilia Kuznetsov, and Iryna Gurevych. 2024. Document structure in long document transformers. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1056–1073, St. Julian's, Malta. Association for Computational Linguistics.

Harry Bunt and Rashmi Prasad. 2016. Iso dr-core (iso 24617-8): Core concepts for the annotation of discourse relations. In *Proceedings 12th joint ACL-ISO workshop on interoperable semantic annotation (ISA-12)*, pages 45–54.

Zhuchen Cao, Sven Apel, Adish Singla, and Vera Demberg. 2025. Pragmatic reasoning improves llm code generation. *arXiv preprint arXiv:2502.15835*.

Taylor Cassidy, Bill McDowell, Nathanael Chambers, and Steven Bethard. 2014. An annotation framework for dense event ordering. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 501–506.

Chunkit Chan, Cheng Jiayang, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. 2024. Exploring the potential of ChatGPT on sentence level relations: A focus on temporal, causal, and discourse relations. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 684–721, St. Julian's, Malta. Association for Computational Linguistics.

Mingda Chen, Zewei Chu, and Kevin Gimpel. 2019. Evaluation benchmarks and learning criteria for discourse-aware sentence representations. 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 649–662, Hong Kong, China. Association for Computational Linguistics.

Xinchi Chen, Xipeng Qiu, and Xuanjing Huang. 2016. Neural sentence ordering. *arXiv preprint arXiv:1607.06952*.

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

Baiyun Cui, Yingming Li, and Zhongfei Zhang. 2020. BERT-enhanced relational sentence ordering network. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6310–6320, Online. Association for Computational Linguistics.

Mark Davies. 2009. The 385+ million word corpus of contemporary american english (1990–2008+): Design, architecture, and linguistic insights. *International journal of corpus linguistics*, 14(2):159–190.

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

Florian Eichin, Yang Janet Liu, Barbara Plank, and Michael A. Hedderich. 2025. Probing LLMs for multilingual discourse generalization through a unified label set. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18665–18684, Vienna, Austria. Association for Computational Linguistics.

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

Yi Fan and Michael Strube. 2025. Consistent discourse-level temporal relation extraction using large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 18605–18622, Suzhou, China. Association for Computational Linguistics.

Bahare Fatemi, Mehran Kazemi, Anton Tsitsulin, Karishma Malkan, Jinyeong Yim, John Palowitch, Sungyong Seo, Jonathan Halcrow, and Bryan Perozzi. 2024. Test of time: A benchmark for evaluating llms on temporal reasoning. *ICLR 2025*.

Shangbin Feng, Zhaoxuan Tan, Wenqian Zhang, Zhenyu Lei, and Yulia Tsvetkov. 2023. KALM: Knowledge-aware integration of local, document, and global contexts for long document understanding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2116–2138, Toronto, Canada. Association for Computational Linguistics.

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

Mario Giulianelli, Sarenne Wallbridge, and Raquel Fernández. 2023. Information value: Measuring utterance predictability as distance from plausible alternatives. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5633–5653, Singapore. Association for Computational Linguistics.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Jennifer Hu and Michael Frank. 2024. Auxiliary task demands mask the capabilities of smaller language models. In *First Conference on Language Modeling*.

Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060, Singapore. Association for Computational Linguistics.

Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. 2016. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239, San Diego, California. Association for Computational Linguistics.

Patrick Huber and Giuseppe Carenini. 2022. Towards understanding large-scale discourse structures in pre-trained and fine-tuned language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2376–2394.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Maor Ivgi, Uri Shaham, and Jonathan Berant. 2023. Efficient long-text understanding with short-text models. *Transactions of the Association for Computational Linguistics*, 11:284–299.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Sainan Jia, Wei Song, Jiefu Gong, Shijin Wang, and Ting Liu. 2023. Sentence ordering with a coherence verifier. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9301–9314.

Zhuoxuan Ju, Jingni Wu, Abhishek Purushothama, and Amir Zeldes. 2025. DeDisCo at the DISRPT 2025 shared task: A system for discourse relation classification. In *Proceedings of the 4th Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2025)*, pages 48–62, Suzhou, China. Association for Computational Linguistics.

Sayash Kapoor, Benedikt Stroebl, Zachary S Siegel, Nitya Nadgir, and Arvind Narayanan. 2024. Ai agents that matter. *arXiv preprint arXiv:2407.01502*.

Zixuan Ke, Fangkai Jiao, Yifei Ming, Xuan-Phi Nguyen, Austin Xu, Do Xuan Long, Minzhi Li, Chengwei Qin, Peifeng Wang, Silvio Savarese, and 1 others. 2025. A survey of frontiers in llm reasoning: Inference scaling, learning to reason, and agentic systems. *arXiv preprint arXiv:2504.09037*.

Najoung Kim, Roma Patel, Adam Poliak, Alex Wang, Patrick Xia, R. Thomas McCoy, Ian Tenney, Alexis Ross, Tal Linzen, Benjamin Van Durme, Samuel R. Bowman, and Ellie Pavlick. 2019. Probing what different NLP tasks teach machines about function word comprehension. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 235–249, Minneapolis, Minnesota. Association for Computational Linguistics.

Sobha Lalitha Devi, Pattabhi Rk Rao, and Vijay Sundar Ram. 2025. SeCoRel: Multilingual discourse analysis in DISRPT 2025. In *Proceedings of the 4th Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2025)*, pages 79–86, Suzhou, China. Association for Computational Linguistics.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

David Lee. 1987. The semantics of just. *Journal of pragmatics*, 11(3):377–398.

Belinda Z. Li, Maxwell Nye, and Jacob Andreas. 2021. Implicit representations of meaning in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1813–1827, Online. Association for Computational Linguistics.

Chuyuan Li, Patrick Huber, Wen Xiao, Maxime Amblard, Chloe Braud, and Giuseppe Carenini. 2023. Discourse structure extraction from pre-trained and fine-tuned language models in dialogues. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2562–2579, Dubrovnik, Croatia. Association for Computational Linguistics.

Chuyuan Li, Austin Xu, Shafiq Joty, and Giuseppe Carenini. 2025. Topic-guided reinforcement learning with LLMs for enhancing multi-document summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 12395–12412, Suzhou, China. Association for Computational Linguistics.

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

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

Xiaobo Liang, Zecheng Tang, Juntao Li, and Min Zhang. 2023. Open-ended long text generation via masked language modeling. In *Proc. ACL 2023*, pages 223–241, Toronto.

Yancheng Liang, Daphne Chen, Abhishek Gupta, Simon S Du, and Natasha Jaques. 2024. Learning to cooperate with humans using generative agents. *Advances in Neural Information Processing Systems*, 37:60061–60087.

Guifu Liu, Bonnie Webber, and Hannah Rohde. 2025. "otherwise" in context: Exploring discourse functions with language models. In *Proceedings of the 6th Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences (CODI 2025)*, pages 81–95, Suzhou, China. Association for Computational Linguistics.

Jian Liu, Jinan Xu, Yufeng Chen, and Yujie Zhang. 2021. Discourse-level event temporal ordering with uncertainty-guided graph completion. In *IJCAI*, pages 3871–3877.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zhengyuan Liu and Nancy Chen. 2021. Improving multi-party dialogue discourse parsing via domain integration. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 122–127, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

Meta Llama. 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation.

Lajanugen Logeswaran, Honglak Lee, and Dragomir Radev. 2018. Sentence ordering and coherence modeling using recurrent neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Shuhaib Mehri, Chuyuan Li, and Giuseppe Carenini. 2025. Discourse relation recognition with language models under different data availability. In *Proceedings of the 6th Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences (CODI 2025)*, pages 148–156, Suzhou, China. Association for Computational Linguistics.

Eleni Metheniti, Chloé Braud, and Philippe Muller. 2024. Feature-augmented model for multilingual discourse relation classification. In *Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI 2024)*, pages 91–104, St. Julians, Malta. Association for Computational Linguistics.

Yisong Miao and Min-Yen Kan. 2025. Discursive circuits: How do language models understand discourse relations? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32558–32577, Suzhou, China. Association for Computational Linguistics.

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

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849.

Aakanksha Naik, Luke Breitfeller, and Carolyn Rose. 2019. Tddiscourse: A dataset for discourse-level temporal ordering of events. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 239–249.

OpenAI. 2025. introducing-gpt-5.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Mihir Parmar, Hanieh Deilamsalehy, Franck Dernoncourt, Seunghyun Yoon, Ryan A. Rossi, and Trung Bui. 2024. Towards enhancing coherence in extractive summarization: Dataset and experiments with LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19810–19820, Miami, Florida, USA. Association for Computational Linguistics.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, and 1 others. 2025. Humanity's last exam. *arXiv preprint arXiv:2501.14249*.

Archiki Prasad, Trung Bui, Seunghyun Yoon, Hanieh Deilamsalehy, Franck Dernoncourt, and Mohit Bansal. 2023. MeetingQA: Extractive question-answering on meeting transcripts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15000–15025, Toronto, Canada. Association for Computational Linguistics.

Rashmi Prasad, Aravind Joshi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, and Bonnie Webber. 2005. The penn discourse treebank as a resource for natural language generation. In *Proceedings of the Corpus Linguistics Workshop on Using Corpora for Natural Language Generation*, pages 25–32.

Ponrawee Prasertsom, Apiwat Jaroonpol, and Attapol T Rutherford. 2024. The thai discourse treebank: Annotating and classifying thai discourse connectives. *Transactions of the Association for Computational Linguistics*, 12:613–629.

Robin Pujol, Firmin Rousseau, Philippe Muller, and Chloé Braud. 2025. DisCuT and DiscReT: MELODI at DISRPT 2025 multilingual discourse segmentation, connective tagging and relation classification. In *Proceedings of the 4th Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2025)*, pages 21–35, Suzhou, China. Association for Computational Linguistics.

Valentina Pyatkin, Ayal Klein, Reut Tsarfaty, and Ido Dagan. 2020. QADiscourse - Discourse Relations as QA Pairs: Representation, Crowdsourcing and Baselines. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2804–2819, Online. Association for Computational Linguistics.

Qwen. 2024. Qwen2.5: A party of foundation models.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Hannah Rohde, Anna Dickinson, Nathan Schneider, Christopher N. L. Clark, Annie Louis, and Bonnie Webber. 2016. Filling in the blanks in understanding discourse adverbials: Consistency, conflict, and context-dependence in a crowdsourced elicitation task. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 49–58, Berlin, Germany. Association for Computational Linguistics.

Hannah Rohde, Alexander Johnson, Nathan Schneider, and Bonnie Webber. 2018. Discourse coherence: Concurrent explicit and implicit relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2257–2267, Melbourne, Australia. Association for Computational Linguistics.

Merel CJ Scholman, Marian Marchal, AriaRay Brown, and Vera Demberg. 2025. Disconaija: a discourse-annotated parallel nigerian pidgin-english corpus. *Language Resources and Evaluation*, pages 1–37.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

William Berkeley Sheffield, Kanishka Misra, Valentina Pyatkin, Ashwini Deo, Kyle Mahowald, and Junyi Jessy Li. 2025. Is it JUST semantics? a case study of discourse particle understanding in LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21704–21715, Vienna, Austria. Association for Computational Linguistics.

Yunmei Shi, Haiying Zhang, Ning Li, and Teng Yang. 2024. An overview of sentence ordering task. *International Journal of Data Science and Analytics*, 18(1):1–18.

Mingxu Tao, Quzhe Huang, Kun Xu, Liwei Chen, Yansong Feng, and Dongyan Zhao. 2024. Probing multimodal large language models for global and local semantic representations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13050–13056.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Kate Thompson, Akshay Chaturvedi, Julie Hunter, and Nicholas Asher. 2024a. Llamipa: An incremental discourse parser. In *EMNLP 2024*, pages 6418–6430.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Eleftheria Tsipidi, Franz Nowak, Ryan Cotterell, Ethan Wilcox, Mario Giulianelli, and Alex Warstadt. 2024. Surprise! Uniform Information Density isn't the whole story: Predicting surprisal contours in long-form discourse. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18820–18836, Miami, Florida, USA. Association for Computational Linguistics.

Nawar Turk, Daniele Comitogianni, and Leila Kosseim. 2025. CLaC at DISRPT 2025: Hierarchical adapters for cross-framework multi-lingual discourse relation classification. In *Proceedings of the 4th Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2025)*, pages 36–47, Suzhou, China. Association for Computational Linguistics.

Liang Wang, Peifeng Li, and Sheng Xu. 2022. Dct-centered temporal relation extraction. In *Proceedings of the 29th international conference on computational linguistics*, pages 2087–2097.

Longyue Wang, Zefeng Du, Donghuai Liu, Deng Cai, Dian Yu, Haiyun Jiang, Yan Wang, Leyang Cui, Shuming Shi, and Zhaopeng Tu. 2023. Disco-bench: A discourse-aware evaluation benchmark for language modelling. *arXiv preprint arXiv:2307.08074*.

Qingyun Wang, Zhihao Zhou, Lifu Huang, Spencer Whitehead, Boliang Zhang, Heng Ji, and Kevin Knight. 2018. Paper abstract writing through editing mechanism. In *56th Annual Meeting of the Association for Computational Linguistics, ACL 2018*, pages 260–265. Association for Computational Linguistics (ACL).

Alex Warstadt. 2020. " just" don't ask: Exclusives and potential questions. In *Proceedings of Sinn und Bedeutung*, volume 24, pages 373–390.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The penn discourse treebank 3.0 annotation manual. *Philadelphia, University of Pennsylvania*, 35:108.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Kangda Wei, Aayush Gautam, and Ruihong Huang. 2024. Are LLMs good annotators for discourse-level event relation extraction? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1–19, Miami, Florida, USA. Association for Computational Linguistics.

Shengguang Wu, Shusheng Yang, Zhenglun Chen, and Qi Su. 2024a. Rethinking pragmatics in large language models: Towards open-ended evaluation and preference tuning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22583–22599.

Yating Wu, Ritika Mangla, Alex Dimakis, Greg Durrett, and Junyi Jessy Li. 2024b. Which questions should i answer? salience prediction of inquisitive questions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19969–19987.

Yating Wu, Ritika Mangla, Greg Durrett, and Junyi Jessy 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, Singapore. Association for Computational Linguistics.

Yuhao Wu, Ming Shan Hee, Zhiqiang Hu, and Roy Ka-Wei Lee. 2025. Longgenbench: Benchmarking long-form generation in long context llms. In *The Thirteenth International Conference on Learning Representations*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang and Sujian Li. 2018. SciDTB: Discourse dependency TreeBank for scientific abstracts. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 444–449, Melbourne, Australia. Association for Computational Linguistics.

Chenhan Yuan, Qianqian Xie, and Sophia Ananiadou. 2023. Zero-shot temporal relation extraction with ChatGPT. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 92–102, Toronto, Canada. Association for Computational Linguistics.

Chenhan Yuan, Qianqian Xie, and Sophia Ananiadou. 2024. Temporal relation extraction with contrastive prototypical sampling. *Knowledge-Based Systems*, 286:111410.

Frances Yung, Mansoor Ahmad, Merel Scholman, and Vera Demberg. 2024. Prompting implicit discourse relation annotation. In *Proceedings of the 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 150–165, St. Julians, Malta. Association for Computational Linguistics.

Amir Zeldes, Tatsuya Aoyama, Yang Janet Liu, Siyao Peng, Debopam Das, and Luke Gessler. 2025. eRST: A signaled graph theory of discourse relations and organization. *Computational Linguistics*, 51(1):23–72.

Wei Zhao, Michael Strube, and Steffen Eger. 2023. Discoscore: Evaluating text generation with bert and discourse coherence. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3865–3883.

Jie Zhou, Shenpo Dong, Hongkui Tu, Xiaodong Wang, and Yong Dou. 2022. Rsgt: Relational structure guided temporal relation extraction. In *Proceedings of the 29th international conference on computational linguistics*, pages 2001–2010.

# A Appendix A. Task Description

## A.1 Task Examples

Table 7.

| Task | Dataset | Example | Answer |
|---|---|---|---|
| (1) DM | Just-Manual | (i) Betsy's picky, she *just* eats chicken nuggets.<br>(ii) You are *just* in time for a demonstration. | Exclusionary<br>Temporal |
| | Just-Subtitle | (iii) I can *just* go?<br>(iv) S-same way they *just* decided. | Emphatic<br>Unexplanatory |
| | Otherwise | (v) A close-up mode of this camera will give sharp results from a subject only two feet away, *otherwise*, I thought the specifications somewhat limited for a camera of this price.<br>(vi) A respectful adaptation must consider details such as local speech and culture, translated into a more universal dimension for a global audience. *Otherwise*, the series will fall flat. | Exception<br><br><br>Argumentation |
| (2) TR | TBD | (i) ADDIS ABABA, Ethiopia (AP). The Organization of African Unity said Friday it would investigate the Hutu-organized **<EVENT e4>genocide</EVENT>** of more than 500,000 minority Tutsis in Rwanda nearly four years ago. Foreign ministers of member-states **<EVENT e5>meeting</EVENT>** in the Ethiopian capital agreed to set up a seven-member panel to investigate who shot down Rwandan President Juvenal Habyarimana's plane on April 6, 1994. The assassination touched off a murderous rampage by Hutu security forces and civilians, who slaughtered mainly Tutsis but also Hutus who favored reconciliation with the minority. It also reignited the civil war. The panel also will look at the exodus of about 2 million Rwanda Hutus to neighboring countries where they lived in U.N.-run refugee camps for 2 1/2 years. The investigation will consider the role of "internal and external forces prior to the genocide and subsequently, and the role of the United Nations and its agencies and the OAU before, during and after the genocide," the OAU said. The panel will be based in Addis Ababa, and will finish its investigation within a year, it said. It is to be funded by voluntary contributions from within and outside the continent. (aa-kjd) | BEFORE (e4, e5) |
| | TDD-Man | (ii) BUDAPEST, Hungary (AP). Tired of being sidelined, Hungarian astronaut Bertalan Farkas is **<EVENT e2>leaving</EVENT>** for the United States to start a new career, he said Saturday. "Being 48 is too early to be retired," a fit-looking Farkas said on state TV's morning talk show. With American astronaut Jon McBride, Farkas set up an American-Hungarian joint venture called Orion 1980, manufacturing space-travel related technology. Farkas will **<EVENT e11>move</EVENT>** to the company's U.S. headquarters. Farkas, an air force captain, was sent into space on board the Soyuz 36 on May 26, 1980. He spent six days aboard the Salyut 6 spacecraft with three Soviet astronauts, Valery Kubasov, Leonid Popov and Valery Riumin. McBride, 54, of Lewisburg, West Virginia, was part of a seven-member crew aboard the Orbiter Challenger in October 1984 and later served as assistant administrator for congressional relations for NASA. Farkas expressed the hope he one day follow in the footsteps of fellow astronaut John Glenn, who at 77 is about to go into space again. On May 22, 1995, Farkas was made a brigadier general, and the following year he was appointed military attache at the Hungarian embassy in Washington. However, cited by District of Columbia traffic police in December for driving under the influence of alcohol, Farkas was ordered home and retired. (ab/dc) | SIMULTANEOUS (e2, e11) |
| | ToT-arith | (iii) The war started in 360 BC and went on for 8 years. What year did the war end in? Answer with the form of: answer: <year> <era>, where year is yyyy and era is one of BC or AD. Eg: 1958 BC | 'answer': '352 BC' |

*Continued on next page*

| Task | Dataset | Example | Answer |
|---|---|---|---|
| | | (iv) In a movie, here are some activities Theodore did. He was traveling on 24-09-1912 at 15:42:07 (24hr). He also did some snowboarding and hiking on Jan 21, 1760 and in February 1836 respectively. He was also vacationing on 13 Aug, 1927 at 22:53:42 (24hr) and skateboarding on 21-01-1760 at 18:00 (24hr). Arrange the activities Theodore did in ascending order. If a date/time is at a lower granularity, assume the earliest value for the missing information. Eg: if the date is Feb, 2020, assume the complete time to be 1 Feb 2020 00:00:00. Answer with the form of an ordered list of activities: {'ordered_list': [activity1, activity2, ...]}. | {'ordered_list': ['snowboarding', 'skateboarding', 'hiking', 'traveling', 'vacationing']} |
| (3) DR | deu.rst.pcc | (i) **[ARG1]** Dagmar Ziegler sitzt in der Schuldenfalle . **[ARG2]** Auf Grund der dramatischen Kassenlage in Brandenburg hat sie jetzt eine seit mehr als einem Jahr erarbeitete Kabinettsvorlage überraschend auf Eis gelegt | causal |
| | eng.pdtb.gum | (ii) **[ARG1]** How do people look at and experience art? **[ARG2]** Which elements of specific artworks do they focus on? | conjunction |
| | fra.sdrt.annodis | (iii) **[ARG1]** Milutinovic devant le TPI . **[ARG2]** L'ancien président de Serbie Milan Milutinovic , <*> s'est rendu volontairement hier au Tribunal pénal international <*> pour l'ex - Yougoslavie de La Haye . | elaboration |
| (4) SO | ROC | (i) **<s1>** Tom let his friend borrow his phone. **<s2>** He kept draining the battery. **<s3>** The phone died shortly after. **<s4>** The friend kept using it. **<s5>** Tom got it back way later. | s1 s4 s2 s5 s3 |
| | AAN abstract | (ii) **<s1>** We represent each medical event as a time duration, with a corresponding start and stop, and learn to rank the starts/stops based on their proximity to the admission date. **<s2>** We investigate the problem of ordering medical events in unstructured clinical narratives by learning to rank them based on their time of occurrence. **<s3>** Interestingly, we observe that this methodology performs better than a classification-based approach for this domain, but worse on the relationships found in the Timebank corpus. **<s4>** Such a representation allows us to learn all of Allen's temporal relations between medical events. **<s5>** This finding has important implications for styles of data representation and resources used for temporal relation learning: clinical narratives may have different language attributes corresponding to temporal ordering relative to Timebank, implying that the field may need to look at a wider range of domains to fully understand the nature of temporal ordering. | s2 s1 s4 s3 s5 |
| | Wiki Movies | (iii) **<s1>** In 1940, after the fall of France, the fictitious Channel Island of Armorel is occupied by a small garrison of German troops under the benign command of Hauptmann Weiss (George Coulouris). **<s2>** In a race against the Germans discovering their presence, they spirit the cow onto a beach and via a special craft, onto a Royal Navy Motor Torpedo Boat which takes them to Britain, though they are pursued by German E-boat. **<s3>** Back in London, the Ministry of Agriculture realise that during the evacuation of the island, Venus, a prize pedigree cow, has been left behind. **<s4>** He finds that the hereditary ruler, the Suzerain, is away in the British army, leaving the Provost in charge. **<s5>** They petition the War Office to do something urgently due to the value of the cow's bloodline, and Major Morland (David Niven), is assigned the task of rescuing Venus. **<s6>** They contact the Provost and discover that the Hauptmann, a cattle breeder in civilian life, is about to have the cow shipped to Germany. **<s7>** When he realises that the Suzerain's sister, Nicola Fallaize (Glynis Johns) is in Wales, serving as an Auxiliary Territorial Service army cook, she is quickly posted to the War Office and the two, with a radio operator sergeant and a Channel Islander naval officer who knows the local waters, are landed on the island. | s1 s4 s3 s5 s7 s6 s2 |

*Continued on next page*

| Task | Dataset | Example | Answer |
|---|---|---|---|
| (5) DDP | STAC | (i) **0.** Gaeilgeoir: anyone have clay? :) | |
| | | **1.** Gaeilgeoir: I have wheat and wood | ELABORATION(0,1) |
| | | **2.** yiin: no sorry | QA_PAIR(0,2) |
| | | **3.** inca: not any more, | QA_PAIR(0,3) |
| | | **4.** inca: sorry | COMMENT(3,4) |
| | | **5.** nareik15: nope, | QA_PAIR(0,5) |
| | | **6.** nareik15: no clay | ELABORATION(5,6) |
| | | **7.** Gaeilgeoir: oh well | ACK(5,7) ACK(2,7) ACK(3,7) |
| | Molweni | (ii) **0.** cr1mson: apt-get i doubt my apt thing is bad though , i just installed ubuntu today | |
| | | **1.** APT-GET_INSTALL_: now you're inside your sources.list ( the file that apt uses to find servers ) | EXPLANATION(0,1) |
| | | **2.** APT-GET_INSTALL_: wait ! i found a much easier way | COMMENT(0,2) |
| | | **3.** APT-GET_INSTALL_: well , i want you to read all of that | COMMENT(0,3) |
| | | **4.** APT-GET_INSTALL_: before you start mucking around in system files | BACKGROUND(0,4) |
| | | **5.** cr1mson: most of it was rem'd out | CONTINUATION(4,5) |
| | | **6.** APT-GET_INSTALL_: you are going to learn what all of them all from the url i just pasted | COMMENT(5,6) |
| | | **7.** cr1mson: i can always use more than one terminal | COMMENT(6,7) |
| | | **8.** cr1mson: okay , so i have to add or change a 'repository' | RESULT(6,8) |
| | MSDC | (iii) **0** <Buil> Mission has started. | |
| | | **1** <Arch> We'll be making a pair of horseshoes. | CONTINUATION(0,1) |
| | | **2** <Arch> First, start by making a 3x3 square of red bricks on the ground. | ELABORATION(1,2) |
| | | **3** <Buil> place red 2 1 -1, place red 1 1 -1, place red 0 1 -1, place red 0 1 0, place red 0 1 1, place red 1 1 0, place red 2 1 0, place red 2 1 2, place red 1 1 1, pick 2 1 2, place red 2 1 1. | RESULT(2,3) |
| | | **4** <Arch> Nice! | ACK(3,4) |
| | | **5** <Arch> Now remove the middle brick, | RESULT(4,5) NARRATION(1,5) |
| | | **6** <Arch> and one brick on the edge. | CONTINUATION(5,6) |
| | | **7** <Buil> place red 1 2 0, pick 1 2 0, pick 1 1 0, pick 2 1 1. | RESULT(6,7) |
| | | **8** <Arch> Edge, | CORRECTION(7,8) |
| | | **9** <Arch> rather than corner. | CONTRAST(8,9) |
| | | **10** <Buil> place red 2 2 0, pick 2 2 0, place red 2 1 1, pick 1 1 1. | CORRECTION(7,10) RESULT(9,10) |
| | | **11** <Arch> Awesome, | ACK(10,11) |

Table 7: Representative task examples from a randomly sampled dataset. Task abbreviations: DM – discourse markers, TR – temporal reasoning, DR – discourse relation, SO – sentence ordering, DDP – dialogue discourse parsing. Dataset name in task (3) DR follows the format "language.framework.dataset".

### A.2 Inter-Annotator Agreement

Table 8. For Task (3) discourse relation classification, we adopt the labeling scheme used in the DISRPT shared task, which follows the format "language.framework.dataset". All scores are obtained from corresponding original datasets. We observe the following:

- Task (1) discourse marker disambiguation, both Just and Otherwise exhibit strong annotation consensus, with Otherwise achieving particularly high agreement (K=0.87). Despite this, current LLMs still fall short of human-level performance.
- Task (2) temporal reasoning, event-based datasets (TBD and TDD-man) also share substantial annotation agreement. Comparable results have so far only been achieved by task-specific supervised models. We find that both reasoning-oriented and general-purpose LLMs lag significantly behind.
- Task (3) discourse relation prediction: across languages and frameworks, IAA scores are generally substantial to near-perfect, with only a few challenging datasets, e.g. STAC (K=0.58) and ERT (K=0.56), showing moderate agreement. In contrast, LLM performance on BeDiscovER remains uniformly low, indicating considerable room for improvement.
- Task (4) sentence ordering, this task does not require human annotation, so we do not have a direct IAA comparison. However, as highlighted by a recent survey (Shi et al., 2024), supervised language models still underperform relative to humans, especially on multi-paragraph ordering.
- Task (5) dialogue discourse parsing, link and full structure annotation yields substantial and moderate human agreement, respectively. As noted by the STAC creators, "Bearing in mind that annotating full discourse structures is a very complex task, this is a relatively good score (Asher et al., 2016). Although supervised parsers can approach this level, general-purpose LLMs still fall noticeably behind.

| Task | Dataset | IAA | Interpretation^ |
|---|---|---|---|
| (1) Discourse marker | Just* | – | strong consensus |
| (1) Discourse marker | Otherwise | $K = 0.87$ | near perfect |
| (2) Temporal reasoning | TBD | $K = 0.56$–0.64 | moderate–substantial |
| (2) Temporal reasoning | TDD-Man | $K = 0.69$ | substantial |
| (3) Discourse relation | ces.rst.crdt | $K = 0.41$–0.66 | moderate–substantial |
| (3) Discourse relation | deu.pdtb.pcc | $K = 0.74$ | substantial |
| (3) Discourse relation | deu.rst.pcc | $K = 0.74$–0.91 | substantial–near perfect |
| (3) Discourse relation | eng.dep.covdtb** | – | high consistency |
| (3) Discourse relation | eng.dep.scidtb | $K \geqslant 0.7$ | substantial |
| (3) Discourse relation | eng.erst.gentle | $K > 0.9$ | near perfect |
| (3) Discourse relation | eng.erst.gum | $K = 0.9$ | near perfect |
| (3) Discourse relation | eng.pdtb.gentle | $K > 0.9$ | near perfect |
| (3) Discourse relation | eng.pdtb.gum | $K = 0.77$–0.83 | substantial–near perfect |
| (3) Discourse relation | eng.pdtb.tedm | $K = 0.92$ | near perfect |
| (3) Discourse relation | eng.rst.oll | – | unknown |
| (3) Discourse relation | eng.rst.rstdt | $K = 0.62$–0.82 | substantial–near perfect |
| (3) Discourse relation | eng.rst.umuc | $K = 0.31$ | fair |
| (3) Discourse relation | eng.sdrt.stac | $K = 0.58$ | moderate |
| (3) Discourse relation | eus.rst.ert | $K = 0.56$ | moderate |
| (3) Discourse relation | fas.rst.prstc | – | unknown |
| (3) Discourse relation | fra.sdrt.annodis | $K = 0.57$ | moderate |
| (3) Discourse relation | ita.pdtb.luna | – | unknown |
| (3) Discourse relation | nld.rst.nldt | $K = 0.7$ | substantial |
| (3) Discourse relation | pcm.pdtb.disconaija | $K = 0.60$–0.94 | moderate–near perfect |
| (3) Discourse relation | pol.iso.pdc | – | unknown |
| (3) Discourse relation | por.pdtb.crpc | $K = 0.71$–0.88 | substantial–near perfect |
| (3) Discourse relation | por.pdtb.tedm | $K = 0.76$ | substantial |
| (3) Discourse relation | por.rst.cstn | $K = 0.66$ | substantial |
| (3) Discourse relation | rus.rst.rrt | $K = 0.8$ | substantial |
| (3) Discourse relation | spa.rst.rststb | $K = 0.78$ | substantial |
| (3) Discourse relation | spa.rst.sctb | $K = 0.64$–1.0 | substantial–near perfect |
| (3) Discourse relation | tha.pdtb.tdtb | $K = 0.84$ | near perfect |
| (3) Discourse relation | tur.pdtb.tdb | $K = 0.73$–0.94 | substantial–near perfect |
| (3) Discourse relation | tur.pdtb.tedm | $K = 0.71$ | substantial |
| (3) Discourse relation | zho.dep.scidtb | $K = 0.72$ | substantial |
| (3) Discourse relation | zho.pdtb.cdtb | $K = 0.94$ | near perfect |
| (3) Discourse relation | zho.pdtb.ted | $K = 0.81$–0.94 | near perfect |
| (3) Discourse relation | zho.rst.gcdt | $K = 0.57$ | moderate |
| (3) Discourse relation | zho.rst.sctb | $K = 0.73$–0.84 | substantial–near perfect |
| (5) Dialogue discourse parsing | STAC | link $K = 0.72$, relation $K = 0.58$ | substantial, moderate |
| (5) Dialogue discourse parsing | Molweni | link $K = 0.91$, relation $K = 0.56$ | near perfect, moderate |
| (5) Dialogue discourse parsing | MSDC | – | unknown |

Table 8: Inter-annotation agreement (IAA) statistics across tasks and datasets. Interpretation^ follows the Standard interpretation of inner-annotator agreement (IAA) score (Landis and Koch, 1977). Just*: All sentences in both Just subsets have a strong primary reading, either by construction (in the hand-constructed corpus, i.e., Just-manual) or by annotator agreement (in the annotated corpus, i.e., Just-subtitle). eng.dep.covdtb**: The authors had several discussions with each annotator to maintain the annotation consistency at a satisfactory level.

## B Appendix B. Model Hyperparameters

Table 9.

| | | w/o reasoning | | | | w/ reasoning | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Model | Size | Max_new_tokens | Temp | Top-p | Seeds | Max_new_tokens | Temp | Top-p | Seeds | Verbosity |
| Qwen3-0.6B | 0.6B | 2048 | 0.7 | 0.8 | 0,1,42 | 4096 | 0.6 | 0.95 | 0,1,42 | - |
| Qwen3-1.7B | 1.7B | 2048 | 0.7 | 0.8 | 0,1,42 | 4096 | 0.6 | 0.95 | 0,1,42 | - |
| Qwen3-4B | 4B | 2048 | 0.7 | 0.8 | 0,1,42 | 4096 | 0.6 | 0.95 | 0,1,42 | - |
| Qwen3-8B | 8B | 2048 | 0.7 | 0.8 | 0,1,42 | 4096 | 0.6 | 0.95 | 0,1,42 | - |
| Qwen3-14B | 14B | 2048 | 0.7 | 0.8 | 0,1,42 | 4096 | 0.6 | 0.95 | 0,1,42 | - |
| Qwen3-32B | 32B | 2048 | 0.7 | 0.8 | 0,1,42 | 4096 | 0.6 | 0.95 | 0,1,42 | - |
| DS-r1-distill-Qwen | 32B | - | - | - | - | 4096 | - | - | 0,1,42 | - |
| DeepSeek-r1-0528 | 37/671B | - | - | - | - | 4096-16384 | - | - | 0,1,42 | - |
| GPT-5-mini (low) | Unk. | - | - | - | - | 4096 | - | - | 0,1,42 | low |
| GPT-5-mini (high) | Unk. | - | - | - | - | 8192-32768 | - | - | 0,1,42 | low |
| GPT-4o-mini | Unknown | 4096 | - | - | 0,1,42 | - | - | - | - | - |
| Llama-4-Scout | 17/109B | 4096 | - | - | 0,1,42 | - | - | - | - | - |
| Qwen2.5-72B | 72B | 4096 | - | - | 0,1,42 | - | - | - | - | - |

Table 9: A summary of hyperparameters used for LLMs. For Qwen3 models, we use the suggested temperature and top-p values. For the dialogue discourse parsing task, we set a higher max_new_tokens value for DeepSeek-r1-0528 and GPT-5-mini with higher reasoning effort.

# C Appendix C. Error Analysis

## C.1 Task (1) Discourse Marker Understanding

Figure 3. Confusion matrices using GPT-5-mini (high) for *Just* and *Otherwise* datasets, all in the DEF+EXP setting.

We observe a notable contrast in the results of *Just* subsets. On the cleaner, expert-curated *Just-Manual* subset, GPT-5-mini can largely distinguish the senses “Adjective”, “Emphatic”, “Exclusionary”, and “Temporal”, but fails almost entirely on “Unelaboratory” and “Unexplanatory” – two closely related readings that both deny further elaboration or explanation. On the more challenging *Just-Subtitle* subset, the model primarily identifies only “Exclusionary” and “Temporal”, with little success on the remaining senses.

For the *Otherwise* dataset, the model distinguishes “Argumentation” and “Exception” reasonably well, but tends to over-predict “Consequence” and struggles to recognize “Enumeration” which provides another option to achieve some goal – a less-used function of Otherwise.

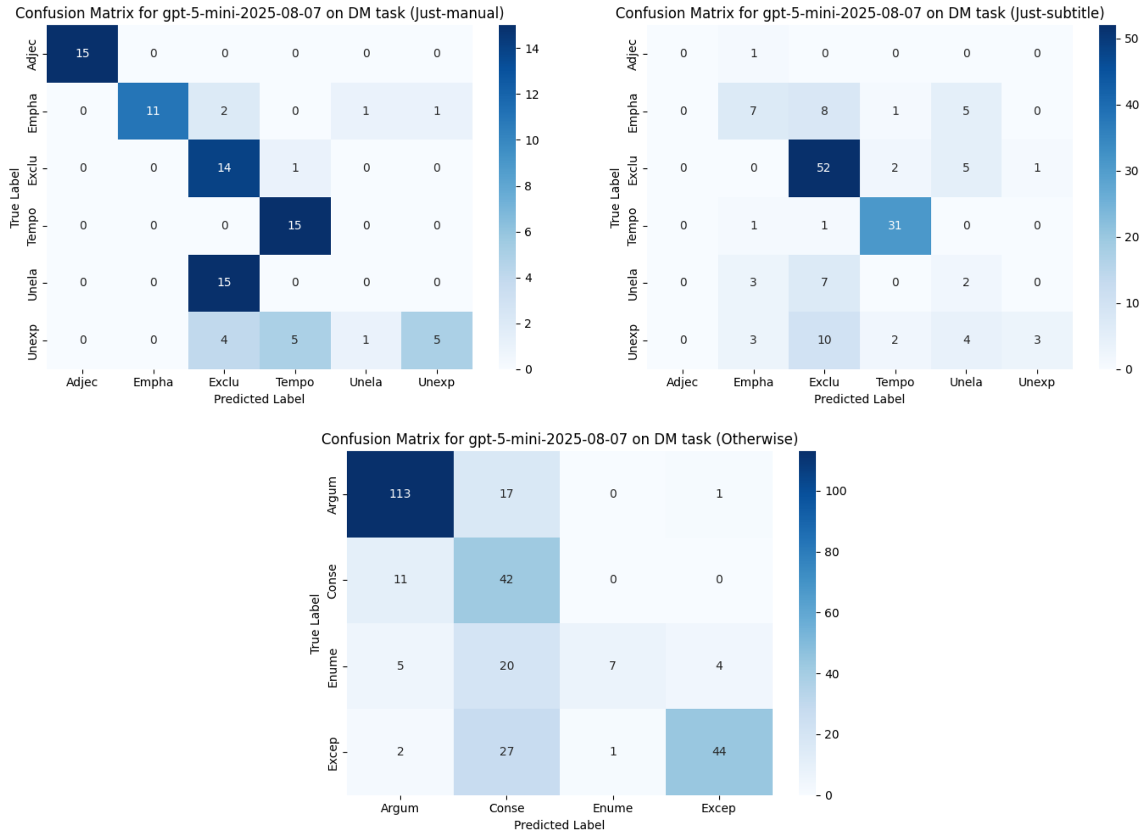


Figure 3: Confusion Matrices on **Task (1) Discourse Marker Understanding** on ***Just-Manual***, ***Just-Subtitle***, ***Otherwise*** datasets, in the DEF+EXP setup. Just relation labels are: Adjective, Emphatic, Exclusionary, Temporal, Unelaboratory, Unexplanatory. Otherwise relation labels are: Argumentation, Consequence, Exception, Enumeration.

### C.2 Task (2) Temporal Reasoning

Figure 4. GPT-5-mini can generally identify when one event occurs before or after another. However, it performs poorly when events overlap or include one another: either predicting such cases only rarely (as in TBD) or over-predicting them (as in TDD-Manual). The model also struggles to correctly identify simultaneous events. These results are extremely interesting, suggesting that while the model may possess some commonsense knowledge on straightforward temporal order (before/after), it has difficulty making finer-grained distinctions among more nuanced temporal relations.

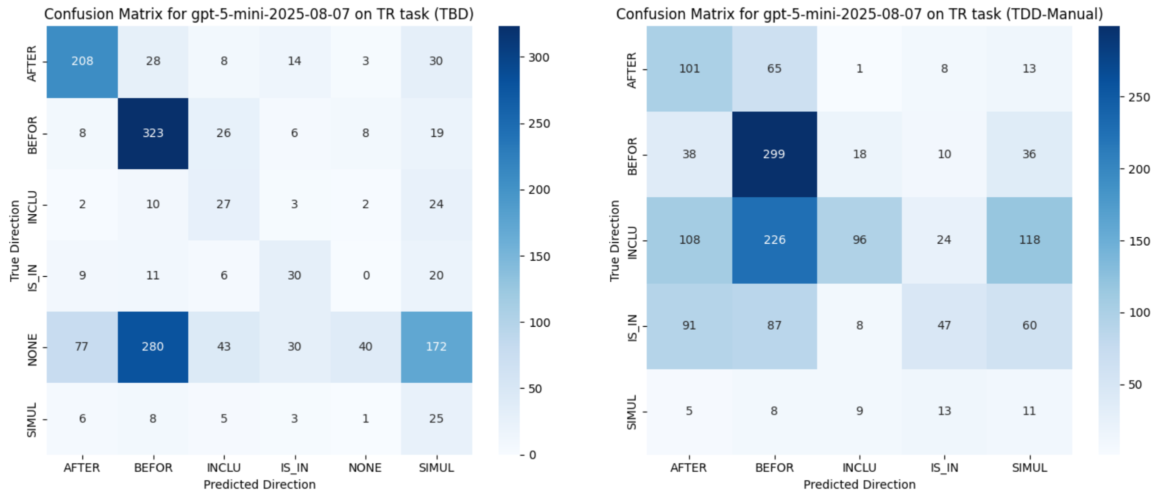


Figure 4: Confusion matrices on **Task (2) Temporal Reasoning**, **TBD** and **TDD-Man** datasets. Temporal labels are: AFTER, BEFORE, INCLUDES, IS_INCLUDED, NONE, SIMULTANEOUS.

### C.3 Task (3) Discourse Relation Recognition

Figure 5. This confusion matrix shows strong performance on several major discourse relations including “elaboration”, “conjunction”, and “temporal”. At the same time, it also reveals notable confusion between “elaboration” and “conjunction”: a pattern consistent with the findings reported in DeDisCo (Ju et al., 2025). This may stem from the semantic and structural similarity between these two relations. Also, splitting apart “cause” and “explanation” is challenging as their differences can be subtle. Finally, the model exhibits a tendency to over-predict “elaboration”, the most frequent label in the dataset.

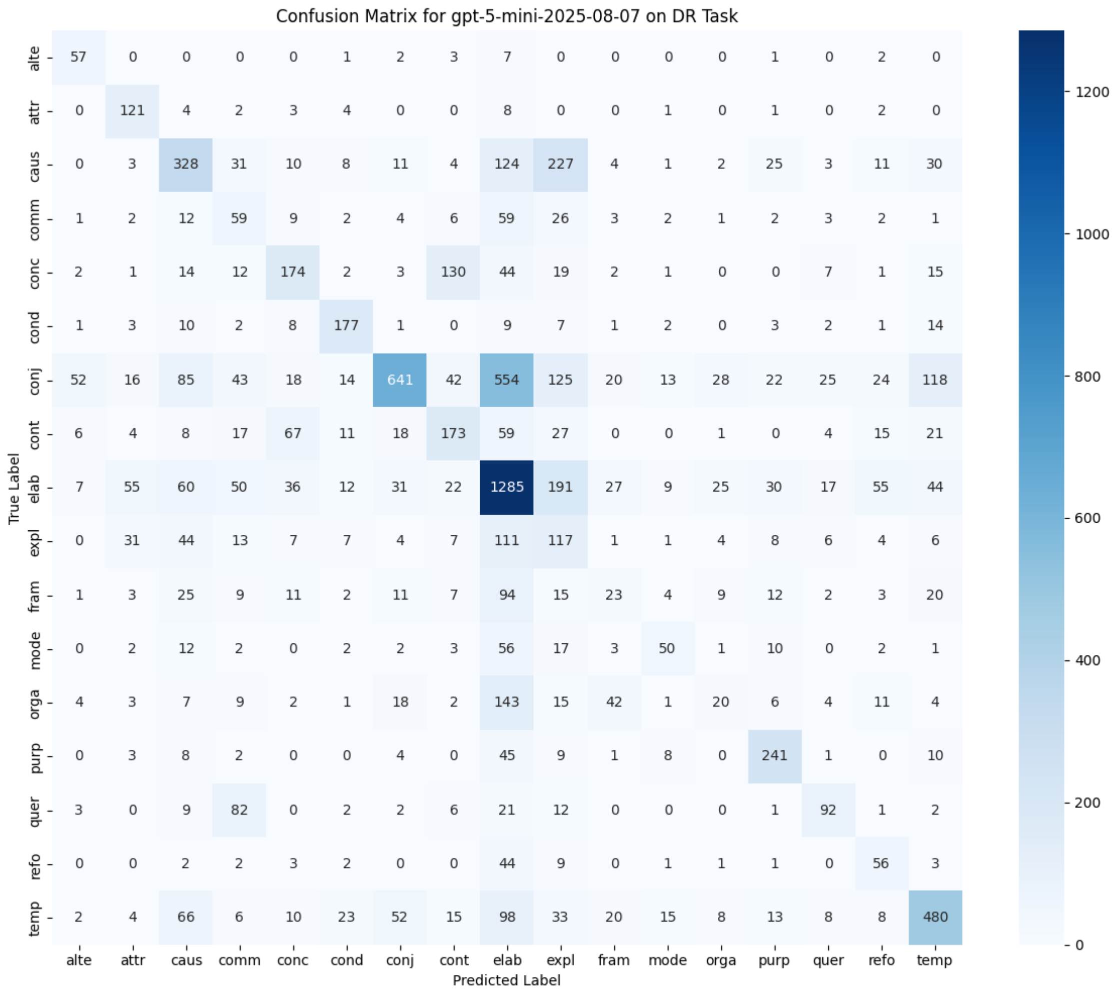


Figure 5: Confusion matrix of **Task (3) Discourse Relation Classification**, with **DISRPT 2025 Shared Task**. Relation labels are: alternation, attribution, causal, comment, concession, condition, conjunction, contrast, elaboration, explanation, frame, mode, organization, purpose, query, reformulation, temporal.

### C.4 Task (4) Sentence Ordering

Figure 6. We report results split by document-length bins and observe a consistent decline in performance as input length increases across datasets. This trend is expected, as longer documents pose greater reordering difficulty. Once a shuffled document exceeds 17 sentences, the model is unable to produce a perfect reconstruction and achieves only partial matches (around 15% accuracy). In comparison, supervised models such as reBART (Basu Roy Chowdhury et al., 2021) maintain roughly 30–40% accuracy on documents of similar length.

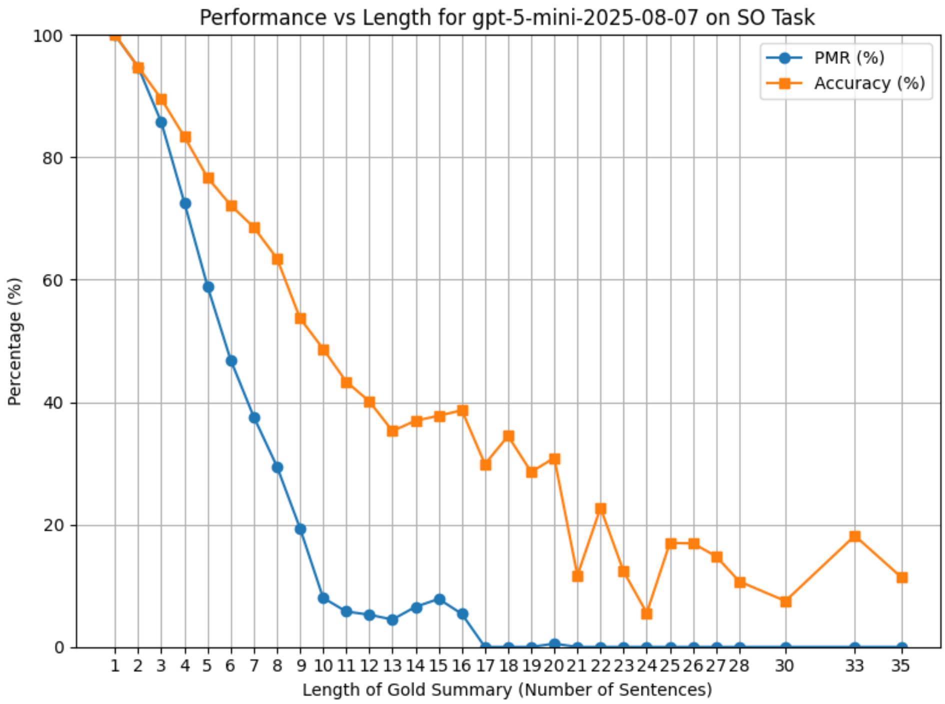


Figure 6: Performance vs. document length on **Task (4) Sentence Ordering**, all seven datasets' results aggregated. PMR: perfect match.

### C.5 Task (5) Dialogue Discourse Parsing

Figure 7. We adopt an incremental generation approach and analyze how performance varies with the amount of preceding context (including both the text and the predicted discourse structure). With a small context window of 1–4 sentences, the model achieves reasonably good performance on both link prediction (70) and full-structure prediction (38). As more context is added, we do not observe a consistent decline. From 5–16 sentences, GPT-5-mini remains relatively stable, suggesting that the model largely focuses on nearby sentences rather than the entire context. However, once the context exceeds 17 sentences, performance drops sharply.

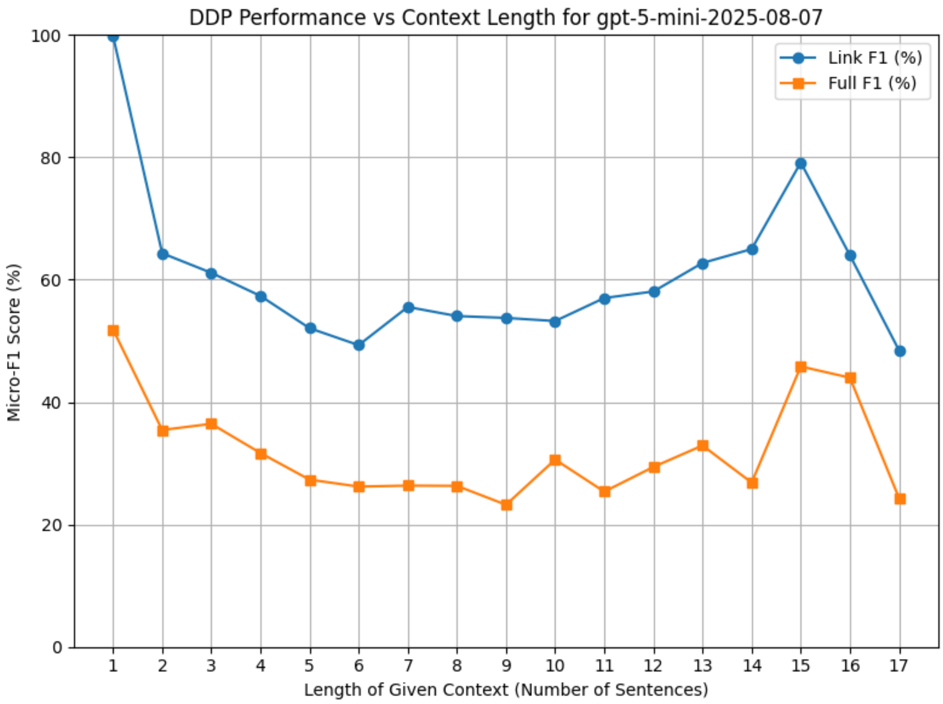


Figure 7: Performance vs. context length on **Task (5) Dialogue Discourse Parsing**, all three datasets' results aggregated.

# D Appendix D. Prompt Templates

## D.1 Task (1) Discourse Marker Understanding

Table 10.

*System prompt*:
A conversation between User and Assistant. The User provides one sentence that contains a discourse marker 'just'. The assistant identifies the **discourse function** of the marker. Choose one of the following six labels: [Exclusionary, Unelaboratory, Unexplanatory, Emphatic, Temporal, Adjective]. Choose ONLY ONE label. Keep the answer short.

(for ***Just*** dataset)

*System prompt*:
A conversation between User and Assistant. The User provides two arguments linked with a discourse marker 'otherwise': [Arg1]. Otherwise, [Arg2]. The assistant identifies the **discourse function** of the marker. Choose one of the following four labels: [Argumentation, Consequence, Exception, Enumeration]. Choose ONLY ONE label. Keep the answer short.

(for ***Otherwise*** dataset)

*User prompt*:
<TEXT>

Question: What is the function of the discourse marker 'just' in the sentence above?

(for ***Just*** dataset)

*User prompt*:
Here are the main functions of 'just' along with their definition:

Exclusionary: 'Just' is used to exclude other possibilities or options.

Unelaboratory: 'Just' is used to deny further elaboration on an event or concept.

Unexplanatory: 'Just' is used to deny that there is an explanation or to offer a weak explanation with no stronger one available.

Emphatic: 'Just' is used to add emphasis to an already strong word or phrase.

Temporal: 'Just' is used to indicate that something happened very recently, or close to another event.

Adjective: 'Just' is used used as an adjective to describe a person or idea, especially a law or policy, as fair, appropriate, or lawful.

Given these, identify the function of 'just' in the following sentence. Respond with the function label.

<TEXT>

(for ***Just*** dataset)

*User prompt*:
Here are the main functions of 'just' along with their definition and examples:

Exclusionary: 'Just' is used to exclude other possibilities or options.
Example: "I'm just looking for a job, not a career."

Unelaboratory: 'Just' is used to deny further elaboration on an event or concept.
Example: "I just don't like it."

Unexplanatory: 'Just' is used to deny that there is an explanation or to offer a weak explanation with no stronger one available.
Example: "The lights just turn on and off."

Emphatic: 'Just' is used to add emphasis to an already strong word or phrase.
Example: "It's just amazing!"

Temporal: 'Just' is used to indicate that something happened very recently, or close to another event.
Example: "I just saw him a minute ago."

Adjective: 'Just' is used used as an adjective to describe a person or idea, especially a law or policy, as fair, appropriate, or lawful.
Example: "She is a just ruler."

Given these, identify the function of 'just' in the following sentence. Respond with the function label.

<TEXT>

(for ***Just*** dataset)

---

*User prompt*:
<TEXT>

Question: What is the function of the discourse marker 'Otherwise' between [Arg1] and [Arg2]?

(for ***Otherwise*** dataset)

---

Here are the main functions of 'Otherwise' along with their definition:

Consequence: If the situation in Arg1 doesn't occur, the situation in Arg2 would arise.

Argumentation: Arg2 is undesirable and can be possibly avoided by following Arg1.

Enumeration: It doesn't take the failure of Arg1 to consider Arg2 as another option.

Exception: Arg1 is an exception to Arg2.

Given these, identify the function of 'Otherwise' between [Arg1] and [Arg2] in the following sentence. Respond with the function label.

<TEXT>

(for ***Otherwise*** dataset)

---

*User prompt*:
Here are the main functions of 'Otherwise' along with their definition and examples. Arg1 and Arg2 are placed within square brackets:

Consequence: If the situation in Arg1 doesn't occur, the situation in Arg2 would arise.
Example: [I like you too.] Otherwise, [we wouldn't be friends.]

Argumentation: Arg2 is undesirable and can be possibly avoided by following Arg1.
Example: [We have to operate immediately.] Otherwise, [she will die.]

Enumeration: It doesn't take the failure of Arg1 to consider Arg2 as another option.
Example: [You can choose to go to the beach.] Otherwise, [you can stay at home and read a book.]

Exception: Arg1 is an exception to Arg2.
Example: [He is usually very punctual.] Otherwise, [he would have arrived on time.]

| |
|---|
| Given these, identify the function of ‘Otherwise’ between [Arg1] and [Arg2] in the following sentence. Respond with the function label.<br><br><TEXT><br>(for ***Otherwise*** dataset) |

Table 10: Prompt Templates used for **Task (1) Discourse Marker Understanding**.

### D.2 Task (2) Temporal Reasoning

Table 11.

| |
|---|
| *System prompt*:<br>A conversation between User and Assistant. The User provides a document with two target EVENT, each marked with <EVENT e>...</EVENT>. The assistant identifies the **temporal relation** between the two events.<br>Choose one of the following six labels: [IS_INCLUDED, INCLUDES, SIMULTANEOUS, BEFORE, AFTER, NONE]. Answer NONE if unsure. Keep the answer short and concise.<br>(for **TBD** and **TDD-Man** datasets) |
| *System prompt*:<br>A conversation between User and Assistant. The User asks a question that requires temporal arithmetic reasoning. The Assistant must reason using the logic and semantics of time to answer. Strictly follow the instructions when answering. Keep the answer short and concise.<br>(for **ToT-Arithmetic** dataset) |
| *User prompt*:<br>Given the document:<br><br><DOCUMENT><br><br>Quesiton: What is the temporal relation of <EVENT> <EVENT ID> with respect to <EVENT> <EVENT ID>? Choose one of the following labels: [IS_INCLUDED, INCLUDES, SIMULTANEOUS, BEFORE, AFTER, NONE].<br>(for **TBD** and **TDD-Man** datasets) |
| *User prompt*:<br><QUESTION><br>(for **ToT-Arithmetic** dataset) |

Table 11: Prompt Templates used for **Task (2) Temporal Reasoning**.

### D.3 Task (3) Discourse Relation Recognition

Table 12.

| |
|---|
| *System prompt*:<br>A conversation between User and Assistant. You are a helpful assistant.<br>## Role and Goal:<br>You are an expert in discourse analysis, tasked with identifying the discourse relation between two sentence units based on the provided label. Your goal is to accurately determine the relationship between these two units.<br><br>## Guidelines:<br>1. You will receive Unit1 and Unit2. Unit1 appears before Unit2 in the original text.<br>2. You will be informed about the language of these units.<br>3. You will be informed of the corpus from which the data is drawn, which may help guide your analysis.<br>4. The framework for analysis will be provided, outlining the structure used for discourse analysis.<br>5. You will be given the context in which these two units appear.<br>6. The direction of the relationship between these two units will be given.<br>7. You will be provided with a set of labels representing possible discourse relations. Do not modify any label. Choose one label that best fits the relationship between Unit1 and Unit2, and output only the chosen label.<br><br>## Labels:<br>contrast, condition, mode, organization, frame, temporal, concession, reformulation, comment, query, attribution, alternation, purpose, explanation, elaboration, causal, conjunction. |
| *User prompt*:<br>## Language:<br><LANG><br><br>## Corpus:<br><CORPUS><br><br>## Framework:<br><FRAMEWORK><br><br>## Context:<br><CONTEXT><br><br>## Direction:<br><DIRECTION><br><br>## Unit1:<br><UNIT1><br><br>## Unit2:<br><UNIT2><br><br>Question: What is the discourse relation between Unit1 and Unit2?" |

Table 12: Prompt Templates used for **Task (3) Discourse Relation Recognition**.

### D.4 Task (4) Sentence Ordering

Table 13.

| |
|---|
| *System prompt*:<br>A conversation between User and Assistant. The User provides a list of shuffled sentences. Each sentence is enclosed within sentence tags, e.g., <s 1> Sentence one. </s 1>. The Assistant reorders the sentences to form a coherent paragraph. Return only the sequence of sentence tags in the correct order, for example: <s 2> <s 1> <s 3>. Do not include any sentence content. Do not repeat or modify any sentence tags. |
| *User prompt*:<br><s 1> <SENTENCE 1> </s 1><br><s 1> <SENTENCE 2> </s 1><br>...<br><s n> <SENTENCE n> </s n> |

Table 13: Prompt Templates used for **Task (4) Sentence Ordering**.

### D.5 Task (5) Dialogue Discourse Parsing

Table 14.

| |
|---|
| *System prompt*:<br>A conversation between User and Assistant. The User provides a boardgame dialogue history / Ubuntu chat log (context and structure), along with a new turn. The assistant identifies the **discourse structure** of the new turn in relation to the context. The structure should be represented in the format: RELATION(S_ID1, S_ID2), where RELATION is the discourse relation, S_ID2 is the ID of the new turn, and S_ID1 is the ID of the sentence in the context that it relates to. S_ID2 is bigger than S_ID1 since dialogue flow is unidirectional and there is no backward relation. Use the following discourse relations: [ACKNOWLEDGEMENT, ALTERNATION, BACKGROUND, CLARIF_Q, COMMENT, CONDITIONAL, CONTINUATION, CONTRAST, CORRECTION, ELABORATION, EXPLANATION, NARRATION, PARALLEL, QA_PAIR, Q_ELABORATION, RESULT]. If multiple relations exist, separate them with spaces. Provide only the structure without any additional text.<br>(for **STAC** and **Molweni** datasets) |
| *System prompt*:<br>A conversation between User and Assistant. The User provides a Minecraft game history (context and structure), along with a new turn. The assistant identifies the **discourse structure** of the new turn in relation to the context. The structure should be represented in the format: RELATION(S_ID1, S_ID2), where RELATION is the discourse relation, S_ID2 is the ID of the new turn, and S_ID1 is the ID of the sentence in the context that it relates to. S_ID2 is bigger than S_ID1 since dialogue flow is unidirectional and there is no backward relation. Use the following discourse relations: [ACKNOWLEDGEMENT, ALTERNATION, CLARIF_Q, COMMENT, CONDITIONAL, CONFIRM_Q, CONTINUATION, CONTRAST, CORRECTION, ELABORATION, EXPLANATION, NARRATION, QA_PAIR, Q_ELABORATION, RESULT, SEQUENCE]. If multiple relations exist, separate them with spaces. Provide only the structure without any additional text.<br>(for **MSDC** dataset) |
| *User prompt*:<br>## Context:<br><CONTEXT><br><br>## Structure:<br><STRUCTURE><br><br>## New turn:<br><NEW TURN><br><br>Question: What is the discourse structure of the new turn in relation to the context? Provide the answer in the format: RELATION(S_ID1, S_ID2). |

Table 14: Prompt Templates used for **Task (5) Dialogue Discourse Parsing**.

# E Appendix E. Detailed Results

## E.1 Task (1) Discourse Marker Understanding

Tables 15 to 23.

| | | w/o reasoning | | | | | w/ reasoning | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Size | Acc | weighted-Pre | weighted-Rec | weighted-F1 | tok. | Acc | weighted-Pre | weighted-Rec | weighted-F1 | tok. |
| Qwen3-0.6B | 0.6B | $17.3_{1.4}$ | $4.0_{2.3}$ | $17.3_{1.4}$ | $5.7_{1.8}$ | 4 | $32.0_{1.6}$ | $18.4_{0.5}$ | $32.0_{1.6}$ | $22.9_{0.9}$ | 279 |
| Qwen3-1.7B | 1.7B | $14.0_{0.4}$ | $2.2_{0.2}$ | $10.5_{0.3}$ | $3.7_{0.2}$ | 3 | $38.9_{0.7}$ | $41.9_{3.4}$ | $38.9_{0.7}$ | $31.9_{1.3}$ | 538 |
| Qwen3-4B | 4B | $34.7_{0.4}$ | $30.3_{1.7}$ | $34.7_{0.4}$ | $27.0_{0.4}$ | 3 | $55.1_{1.8}$ | $62.5_{3.2}$ | $55.1_{1.8}$ | $50.7_{1.8}$ | 543 |
| Qwen3-8B | 8B | $44.7_{0.6}$ | $44.0_{3.6}$ | $44.7_{0.6}$ | $34.5_{0.7}$ | 3 | $59.1_{2.9}$ | $67.5_{7.5}$ | $59.1_{2.9}$ | $54.0_{3.8}$ | 583 |
| Qwen3-14B | 14B | $57.8_{2.0}$ | $55.6_{3.4}$ | $57.8_{2.0}$ | $53.1_{2.0}$ | 2 | $60.2_{0.7}$ | $62.3_{5.7}$ | $60.2_{0.7}$ | $52.6_{1.4}$ | 404 |
| Qwen3-32B | 32B | $55.1_{2.0}$ | $60.6_{2.6}$ | $55.1_{2.0}$ | $51.4_{2.3}$ | 3 | $59.8_{1.8}$ | $60.0_{7.2}$ | $59.8_{1.8}$ | $53.5_{2.3}$ | 425 |
| DS-r1-distill-Qwen | 32B | - | - | - | - | - | $56.7_{0.9}$ | $60.4_{9.9}$ | $51.3_{4.6}$ | $50.3_{5.5}$ | 304 |
| DeepSeek-r1-0528 | 37/671B | - | - | - | - | - | $53.7_{0.5}$ | $20.9_{1.8}$ | $19.5_{2.0}$ | $18.5_{1.5}$ | 466 |
| GPT-5-mini (low) | Unk. | - | - | - | - | - | $63.3_{2.4}$ | $52.5_{5.0}$ | $60.4_{6.0}$ | $53.3_{5.5}$ | 94 |
| GPT-5-mini (high) | Unk. | - | - | - | - | - | $61.7_{0.6}$ | $56.6_{8.7}$ | $61.7_{0.6}$ | $53.5_{1.3}$ | 857 |

Table 15: Details on **Task (1) Discourse Marker Understanding** performance on ***Just-Manual*** dataset, using **basic** prompting strategy.

| | | w/o reasoning | | | | | w/ reasoning | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Size | Acc | weighted-Pre | weighted-Rec | weighted-F1 | tok. | Acc | weighted-Pre | weighted-Rec | weighted-F1 | tok. |
| Qwen3-0.6B | 0.6B | $16.9_{1.1}$ | $5.2_{2.1}$ | $16.9_{1.1}$ | $7.7_{2.3}$ | 3 | $28.4_{1.6}$ | $8.3_{4.2}$ | $9.1_{2.2}$ | $7.0_{1.8}$ | 1427 |
| Qwen3-1.7B | 1.7B | $34.4_{0.0}$ | $20.6_{0.7}$ | $34.4_{0.0}$ | $22.8_{0.2}$ | 3 | $40.7_{4.2}$ | $15.9_{6.2}$ | $14.1_{3.3}$ | $12.9_{3.3}$ | 1624 |
| Qwen3-4B | 4B | $28.7_{0.2}$ | $39.4_{3.3}$ | $28.7_{0.2}$ | $25.0_{0.4}$ | 2 | $43.3_{2.0}$ | $18.5_{5.8}$ | $12.7_{4.1}$ | $12.8_{3.9}$ | 1823 |
| Qwen3-8B | 8B | $49.6_{0.4}$ | $28.7_{0.9}$ | $49.6_{0.4}$ | $35.3_{0.6}$ | 2 | $50.9_{3.1}$ | $37.6_{3.3}$ | $26.1_{3.8}$ | $27.1_{4.0}$ | 1443 |
| Qwen3-14B | 14B | $47.8_{0.0}$ | $52.4_{0.1}$ | $47.8_{0.0}$ | $40.8_{0.1}$ | 3 | $54.9_{2.0}$ | $23.5_{1.3}$ | $20.3_{0.3}$ | $19.5_{0.3}$ | 1219 |
| Qwen3-32B | 32B | $68.2_{0.7}$ | $71.5_{0.9}$ | $68.2_{0.7}$ | $66.3_{0.7}$ | 3 | $65.6_{3.5}$ | $46.6_{6.4}$ | $47.5_{5.0}$ | $44.9_{4.7}$ | 608 |
| DS-r1-distill-Qwen | 32B | - | - | - | - | - | $63.7_{1.4}$ | $62.7_{6.6}$ | $57.6_{4.2}$ | $53.3_{2.8}$ | 326 |
| DeepSeek-r1-0528 | 37/671B | - | - | - | - | - | $66.3_{0.5}$ | $61.6_{7.8}$ | $60.0_{4.7}$ | $53.1_{3.6}$ | 291 |
| GPT-5-mini (low) | Unk. | - | - | - | - | - | $65.9_{0.5}$ | $69.4_{3.5}$ | $65.9_{0.5}$ | $59.8_{1.4}$ | 70 |
| GPT-5-mini (high) | Unk. | - | - | - | - | - | $66.1_{1.7}$ | $72.7_{7.5}$ | $66.1_{1.7}$ | $60.2_{3.2}$ | 531 |

Table 16: Details on **Task (1) Discourse Marker Understanding** performance on ***Just-Manual*** dataset, using **"definition"** prompting strategy.

| | | w/o reasoning | | | | | w/ reasoning | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Size | Acc | weighted-Pre | weighted-Rec | weighted-F1 | tok. | Acc | weighted-Pre | weighted-Rec | weighted-F1 | tok. |
| Qwen3-0.6B | 0.6B | $20.4_{1.2}$ | $7.2_{0.4}$ | $20.4_{1.2}$ | $10.4_{0.7}$ | 3 | $26.2_{1.8}$ | $6.0_{1.9}$ | $7.9_{1.5}$ | $6.0_{1.4}$ | 1436 |
| Qwen3-1.7B | 1.7B | $40.4_{0.5}$ | $25.4_{3.5}$ | $40.4_{0.5}$ | $28.1_{0.9}$ | 3 | $40.2_{2.2}$ | $16.9_{3.3}$ | $16.3_{1.8}$ | $14.5_{1.7}$ | 1587 |
| Qwen3-4B | 4B | $34.7_{0.4}$ | $30.6_{0.2}$ | $34.7_{0.4}$ | $28.9_{0.4}$ | 3 | $44.2_{2.7}$ | $29.5_{10.8}$ | $21.9_{8.8}$ | $22.1_{8.9}$ | 1942 |
| Qwen3-8B | 8B | $42.2_{0.0}$ | $61.6_{0.0}$ | $42.2_{0.0}$ | $33.5_{0.0}$ | 2 | $46.4_{2.9}$ | $34.1_{3.7}$ | $22.8_{2.9}$ | $24.3_{2.7}$ | 1505 |
| Qwen3-14B | 14B | $56.9_{0.4}$ | $56.2_{1.5}$ | $56.9_{0.4}$ | $49.3_{0.7}$ | 3 | $55.6_{1.3}$ | $22.5_{5.7}$ | $19.9_{5.4}$ | $19.6_{5.2}$ | 1574 |
| Qwen3-32B | 32B | $69.8_{1.5}$ | $69.1_{1.9}$ | $69.8_{1.5}$ | $67.8_{1.9}$ | 3 | $64.4_{2.0}$ | $42.1_{8.8}$ | $41.0_{8.1}$ | $39.2_{7.4}$ | 613 |
| DS-r1-distill-Qwen | 32B | - | - | - | - | - | $59.6_{2.9}$ | $33.5_{3.1}$ | $32.8_{3.7}$ | $31.2_{3.3}$ | 313 |
| DeepSeek-r1-0528 | 37/671B | - | - | - | - | - | $67.0_{1.0}$ | $52.0_{3.0}$ | $52.7_{3.9}$ | $48.0_{3.4}$ | 421 |
| GPT-5-mini (low) | Unk. | - | - | - | - | - | $65.6_{0.0}$ | $64.1_{0.4}$ | $65.6_{0.0}$ | $60.7_{0.1}$ | 74 |
| GPT-5-mini (high) | Unk. | - | - | - | - | - | $65.6_{1.1}$ | $64.0_{1.8}$ | $65.6_{1.1}$ | $60.3_{1.7}$ | 528 |

Table 17: Details on **Task (1) Discourse Marker Understanding** performance on ***Just-Manual*** dataset, using **"definition+example"** prompting strategy.

| | | w/o reasoning | | | | | w/ reasoning | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Size | Acc | weighted-Pre | weighted-Rec | weighted-F1 | tok. | Acc | weighted-Pre | weighted-Rec | weighted-F1 | tok. |
| Qwen3-0.6B | 0.6B | $14.6_{0.8}$ | $10.3_{5.0}$ | $16.7_{0.4}$ | $5.6_{1.3}$ | 4 | $25.5_{1.3}$ | $18.0_{2.6}$ | $30.4_{7.3}$ | $17.1_{1.0}$ | 260 |
| Qwen3-1.7B | 1.7B | $13.4_{0.0}$ | $16.3_{5.3}$ | $11.1_{0.4}$ | $4.3_{0.2}$ | 3 | $25.9_{0.8}$ | $14.7_{1.6}$ | $24.9_{1.0}$ | $15.7_{1.1}$ | 491 |
| Qwen3-4B | 4B | $24.8_{1.0}$ | $22.0_{1.1}$ | $21.4_{0.9}$ | $14.4_{0.7}$ | 3 | $41.5_{0.9}$ | $32.2_{2.9}$ | $33.8_{0.8}$ | $28.0_{0.5}$ | 537 |
| Qwen3-8B | 8B | $34.5_{0.5}$ | $27.1_{1.3}$ | $29.4_{0.4}$ | $21.4_{0.4}$ | 2 | $46.0_{1.1}$ | $33.9_{5.7}$ | $36.0_{0.5}$ | $30.5_{0.5}$ | 610 |
| Qwen3-14B | 14B | $33.8_{1.1}$ | $37.2_{1.6}$ | $32.2_{0.8}$ | $24.7_{0.9}$ | 2 | $43.2_{1.9}$ | $36.4_{2.2}$ | $33.8_{2.5}$ | $30.8_{2.3}$ | 463 |
| Qwen3-32B | 32B | $29.4_{0.8}$ | $29.1_{1.7}$ | $25.9_{0.3}$ | $19.3_{0.7}$ | 3 | $48.5_{1.7}$ | $35.3_{4.0}$ | $36.9_{1.3}$ | $32.7_{1.5}$ | 415 |
| DS-r1-distill-Qwen | 32B | - | - | - | - | - | $31.8_{3.3}$ | $33.4_{6.8}$ | $25.7_{4.3}$ | $21.7_{4.2}$ | 291 |
| DeepSeek-r1-0528 | 37/671B | - | - | - | - | - | $48.1_{4.0}$ | $10.1_{3.9}$ | $9.6_{4.1}$ | $9.4_{3.9}$ | 419 |
| GPT-5-mini (low) | Unk. | - | - | - | - | - | $56.4_{1.1}$ | $36.8_{5.2}$ | $38.2_{2.3}$ | $34.6_{1.5}$ | 103 |
| GPT-5-mini (high) | Unk. | - | - | - | - | - | $57.4_{1.7}$ | $37.5_{3.6}$ | $37.4_{0.4}$ | $35.2_{0.1}$ | 818 |

Table 18: Details on **Task (1) Discourse Marker Understanding** performance on ***Just-Subtitle*** dataset, using **basic** prompting strategy.

| | | w/o reasoning | | | | | w/ reasoning | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Size | Acc | weighted-Pre | weighted-Rec | weighted-F1 | tok. | Acc | weighted-Pre | weighted-Rec | weighted-F1 | tok. |
| Qwen3-0.6B | 0.6B | $13.7_{0.4}$ | $7.3_{0.6}$ | $15.1_{1.0}$ | $6.2_{0.9}$ | 4 | $19.3_{1.2}$ | $3.7_{1.2}$ | $3.4_{1.2}$ | $2.7_{0.2}$ | 1426 |
| Qwen3-1.7B | 1.7B | $19.7_{0.3}$ | $9.5_{0.1}$ | $19.8_{0.2}$ | $10.8_{0.2}$ | 3 | $26.2_{1.5}$ | $9.5_{2.0}$ | $7.6_{0.8}$ | $6.4_{0.7}$ | 1705 |
| Qwen3-4B | 4B | $29.8_{0.3}$ | $32.2_{0.2}$ | $27.8_{0.5}$ | $20.5_{0.3}$ | 3 | $30.9_{2.4}$ | $11.6_{4.2}$ | $7.6_{2.6}$ | $7.2_{2.4}$ | 1667 |
| Qwen3-8B | 8B | $31.3_{0.5}$ | $30.8_{2.9}$ | $46.1_{0.7}$ | $27.4_{1.1}$ | 3 | $32.3_{1.7}$ | $19.5_{2.6}$ | $13.6_{3.1}$ | $12.5_{1.9}$ | 1667 |
| Qwen3-14B | 14B | $31.3_{0.5}$ | $35.2_{1.6}$ | $46.1_{1.0}$ | $27.3_{0.5}$ | 3 | $32.9_{1.7}$ | $9.6_{4.4}$ | $8.3_{3.5}$ | $7.7_{3.3}$ | 1328 |
| Qwen3-32B | 32B | $32.8_{0.6}$ | $30.3_{0.4}$ | $30.5_{0.5}$ | $25.0_{0.6}$ | 3 | $48.9_{2.0}$ | $18.3_{3.2}$ | $18.3_{3.4}$ | $17.7_{3.3}$ | 569 |
| DS-r1-distill-Qwen | 32B | - | - | - | - | - | $40.5_{2.5}$ | $39.2_{5.8}$ | $33.2_{3.6}$ | $29.3_{3.6}$ | 308 |
| DeepSeek-r1-0528 | 37/671B | - | - | - | - | - | $60.2_{1.4}$ | $35.8_{6.1}$ | $34.1_{6.0}$ | $32.5_{6.1}$ | 321 |
| GPT-5-mini (low) | Unk. | - | - | - | - | - | $58.4_{1.1}$ | $43.6_{2.4}$ | $41.7_{1.2}$ | $39.3_{1.3}$ | 65 |
| GPT-5-mini (high) | Unk. | - | - | - | - | - | $63.1_{0.0}$ | $43.7_{1.7}$ | $44.2_{0.0}$ | $41.9_{1.0}$ | 509 |

Table 19: Details on **Task (1) Discourse Marker Understanding** performance on ***Just-Subtitle*** dataset, using **"definition"** prompting strategy.

| | | w/o reasoning | | | | | w/ reasoning | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Size | Acc | weighted-Pre | weighted-Rec | weighted-F1 | tok. | Acc | weighted-Pre | weighted-Rec | weighted-F1 | tok. |
| Qwen3-0.6B | 0.6B | $19.1_{0.9}$ | $7.8_{1.1}$ | $18.4_{0.7}$ | $9.6_{0.4}$ | 3 | $16.9_{1.8}$ | $3.2_{0.5}$ | $3.8_{1.1}$ | $2.4_{0.5}$ | 1439 |
| Qwen3-1.7B | 1.7B | $20.3_{0.4}$ | $9.9_{0.4}$ | $20.1_{0.5}$ | $13.0_{0.3}$ | 2 | $25.5_{1.9}$ | $7.1_{1.5}$ | $5.5_{0.5}$ | $5.0_{0.4}$ | 1743 |
| Qwen3-4B | 4B | $29.9_{0.3}$ | $24.8_{0.2}$ | $27.7_{0.4}$ | $21.7_{0.2}$ | 3 | $30.3_{1.1}$ | $21.4_{10.0}$ | $13.3_{6.6}$ | $12.8_{6.7}$ | 1944 |
| Qwen3-8B | 8B | $30.9_{0.6}$ | $28.3_{2.2}$ | $44.9_{0.6}$ | $25.5_{0.7}$ | 2 | $32.2_{1.4}$ | $22.9_{5.1}$ | $16.0_{4.7}$ | $14.0_{2.8}$ | 1593 |
| Qwen3-14B | 14B | $36.2_{0.4}$ | $38.2_{1.6}$ | $48.6_{0.5}$ | $32.3_{0.7}$ | 3 | $33.2_{1.3}$ | $14.6_{5.1}$ | $11.5_{4.8}$ | $10.8_{4.1}$ | 1427 |
| Qwen3-32B | 32B | $35.7_{1.1}$ | $34.1_{1.6}$ | $46.1_{0.5}$ | $30.3_{0.8}$ | 2 | $47.1_{1.2}$ | $15.0_{1.1}$ | $14.3_{1.1}$ | $13.9_{1.1}$ | 718 |
| DS-r1-distill-Qwen | 32B | - | - | - | - | - | $35.8_{0.3}$ | $17.9_{3.7}$ | $13.4_{3.2}$ | $12.5_{2.7}$ | 375 |
| DeepSeek-r1-0528 | 37/671B | - | - | - | - | - | $59.5_{3.6}$ | $27.0_{1.3}$ | $22.6_{2.7}$ | $21.5_{2.6}$ | 380 |
| GPT-5-mini (low) | Unk. | - | - | - | - | - | $62.9_{0.6}$ | $48.0_{1.7}$ | $44.0_{0.5}$ | $42.1_{0.3}$ | 87 |
| GPT-5-mini (high) | Unk. | - | - | - | - | - | $63.4_{0.3}$ | $46.5_{1.3}$ | $40.7_{0.0}$ | $40.1_{0.2}$ | 534 |

Table 20: Details on **Task (1) Discourse Marker Understanding** performance on ***Just-Subtitle*** dataset, using **"definition+example"** prompting strategy.

| | | w/o reasoning | | | | | w/ reasoning | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Size | Acc | weighted-Pre | weighted-Rec | weighted-F1 | tok. | Acc | weighted-Pre | weighted-Rec | weighted-F1 | tok. |
| Qwen3-0.6B | 0.6B | $44.6_{0.0}$ | $19.9_{0.0}$ | $44.6_{0.0}$ | $27.5_{0.0}$ | 3 | $22.2_{1.3}$ | $24.9_{2.8}$ | $22.2_{1.3}$ | $19.0_{1.5}$ | 277 |
| Qwen3-1.7B | 1.7B | $30.3_{0.7}$ | $20.1_{0.2}$ | $30.3_{0.7}$ | $24.1_{0.4}$ | 2 | $23.9_{0.6}$ | $26.2_{4.1}$ | $23.9_{0.6}$ | $17.5_{0.9}$ | 492 |
| Qwen3-4B | 4B | $25.7_{0.2}$ | $8.3_{0.2}$ | $25.7_{0.2}$ | $11.9_{0.2}$ | 1 | $32.3_{1.2}$ | $24.5_{17.9}$ | $32.3_{1.2}$ | $21.0_{0.7}$ | 569 |
| Qwen3-8B | 8B | $18.6_{0.1}$ | $38.2_{2.5}$ | $18.6_{0.1}$ | $6.7_{0.2}$ | 2 | $32.3_{0.7}$ | $34.1_{0.5}$ | $32.3_{0.7}$ | $24.0_{0.7}$ | 512 |
| Qwen3-14B | 14B | $34.3_{0.1}$ | $42.6_{0.6}$ | $34.3_{0.1}$ | $25.6_{0.1}$ | 2 | $32.1_{1.0}$ | $30.0_{3.4}$ | $32.1_{1.0}$ | $23.6_{0.9}$ | 509 |
| Qwen3-32B | 32B | $36.5_{1.3}$ | $50.7_{5.3}$ | $36.5_{1.3}$ | $38.1_{1.3}$ | 2 | $34.3_{0.4}$ | $19.7_{0.4}$ | $34.3_{0.4}$ | $23.8_{0.4}$ | 373 |
| DS-r1-distill-Qwen | 32B | - | - | - | - | - | $33.1_{1.1}$ | $25.9_{1.1}$ | $33.1_{1.1}$ | $24.9_{0.9}$ | 276 |
| DeepSeek-r1-0528 | 37/671B | - | - | - | - | - | $31.2_{0.8}$ | $39.1_{1.2}$ | $31.2_{0.8}$ | $25.8_{0.4}$ | 373 |
| GPT-5-mini (low) | Unk. | - | - | - | - | - | $34.6_{0.4}$ | $25.4_{0.6}$ | $34.6_{0.4}$ | $25.5_{0.4}$ | 74 |
| GPT-5-mini (high) | Unk. | - | - | - | - | - | $35.5_{0.5}$ | $31.1_{5.6}$ | $35.5_{0.5}$ | $26.5_{0.2}$ | 375 |

Table 21: Details on **Task (1) Discourse Marker Understanding** performance on ***Otherwise*** dataset, using **basic** prompting strategy.

| | | w/o reasoning | | | | | w/ reasoning | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Size | Acc | weighted-Pre | weighted-Rec | weighted-F1 | tok. | Acc | weighted-Pre | weighted-Rec | weighted-F1 | tok. |
| Qwen3-0.6B | 0.6B | $44.6_{0.0}$ | $19.9_{0.0}$ | $44.6_{0.0}$ | $27.5_{0.0}$ | 2 | $22.7_{0.9}$ | $25.0_{3.6}$ | $22.7_{0.9}$ | $18.1_{1.4}$ | 402 |
| Qwen3-1.7B | 1.7B | $24.2_{1.6}$ | $28.3_{1.7}$ | $24.2_{1.7}$ | $20.5_{1.6}$ | 1 | $21.2_{1.0}$ | $35.9_{6.0}$ | $21.2_{1.0}$ | $11.9_{1.1}$ | 515 |
| Qwen3-4B | 4B | $30.3_{0.6}$ | $56.1_{0.5}$ | $30.3_{0.6}$ | $32.3_{0.4}$ | 2 | $21.8_{0.9}$ | $61.8_{4.9}$ | $21.8_{0.9}$ | $15.1_{1.7}$ | 547 |
| Qwen3-8B | 8B | $33.6_{0.9}$ | $65.2_{1.1}$ | $33.6_{0.9}$ | $33.5_{0.9}$ | 2 | $39.0_{1.7}$ | $70.1_{1.8}$ | $39.0_{1.7}$ | $38.6_{1.9}$ | 1033 |
| Qwen3-14B | 14B | $43.5_{0.4}$ | $57.8_{0.4}$ | $43.5_{0.4}$ | $44.3_{0.5}$ | 2 | $45.2_{2.6}$ | $70.9_{1.8}$ | $45.2_{2.6}$ | $44.6_{2.1}$ | 766 |
| Qwen3-32B | 32B | $55.0_{0.2}$ | $64.0_{1.2}$ | $55.0_{0.2}$ | $54.9_{0.4}$ | 2 | $49.3_{2.3}$ | $71.3_{1.5}$ | $49.3_{2.3}$ | $50.3_{1.8}$ | 737 |
| DS-r1-distill-Qwen | 32B | - | - | - | - | - | $32.4_{2.3}$ | $60.0_{1.5}$ | $32.4_{2.3}$ | $31.4_{3.3}$ | 369 |
| DeepSeek-r1-0528 | 37/671B | - | - | - | - | - | $38.9_{1.9}$ | $70.3_{1.6}$ | $38.9_{1.9}$ | $45.2_{1.7}$ | 682 |
| GPT-5-mini (low) | Unk. | - | - | - | - | - | $36.8_{0.8}$ | $64.0_{2.8}$ | $36.8_{0.8}$ | $34.0_{1.4}$ | 138 |
| GPT-5-mini (high) | Unk. | - | - | - | - | - | $43.9_{1.0}$ | $74.5_{4.1}$ | $43.9_{1.0}$ | $43.6_{1.3}$ | 751 |

Table 22: Details on **Task (1) Discourse Marker Understanding** performance on ***Otherwise*** dataset, using **"definition"** prompting strategy.

| Model | Size | w/o reasoning | | | | | w/ reasoning | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Acc | weighted-Pre | weighted-Rec | weighted-F1 | tok. | Acc | weighted-Pre | weighted-Rec | weighted-F1 | tok. |
| Qwen3-0.6B | 0.6B | $44.6_{0.0}$ | $19.9_{0.1}$ | $44.6_{0.0}$ | $27.5_{0.1}$ | 2 | $15.7_{1.4}$ | $32.0_{14.3}$ | $15.7_{1.4}$ | $11.8_{2.1}$ | 1399 |
| Qwen3-1.7B | 1.7B | $21.0_{0.8}$ | $32.2_{2.0}$ | $21.0_{0.8}$ | $11.7_{0.8}$ | 2 | $16.2_{1.8}$ | $48.6_{9.0}$ | $16.2_{1.8}$ | $13.1_{1.4}$ | 1442 |
| Qwen3-4B | 4B | $41.8_{0.8}$ | $56.5_{0.5}$ | $41.8_{0.8}$ | $44.5_{0.7}$ | 2 | $30.7_{1.9}$ | $61.7_{1.5}$ | $30.7_{1.9}$ | $33.8_{1.9}$ | 1585 |
| Qwen3-8B | 8B | $46.7_{0.3}$ | $60.2_{1.0}$ | $46.7_{0.3}$ | $49.2_{0.4}$ | 2 | $45.6_{1.4}$ | $71.8_{2.2}$ | $45.6_{1.4}$ | $46.8_{1.3}$ | 1574 |
| Qwen3-14B | 14B | $51.8_{0.7}$ | $62.9_{0.5}$ | $51.8_{0.7}$ | $53.8_{0.6}$ | 2 | $51.4_{1.6}$ | $73.7_{0.6}$ | $51.4_{1.6}$ | $53.9_{1.1}$ | 1582 |
| Qwen3-32B | 32B | $63.8_{0.8}$ | $67.5_{0.4}$ | $63.8_{0.8}$ | $62.9_{0.6}$ | 2 | $66.1_{0.6}$ | $74.5_{0.5}$ | $66.1_{0.6}$ | $65.3_{0.6}$ | 873 |
| DS-r1-distill-Qwen | 32B | - | - | - | - | - | $44.3_{2.7}$ | $65.2_{2.7}$ | $44.3_{2.7}$ | $47.0_{2.8}$ | 453 |
| DeepSeek-r1-0528 | 37/671B | - | - | - | - | - | $58.8_{2.2}$ | $74.3_{0.7}$ | $58.8_{2.2}$ | $63.9_{1.4}$ | 656 |
| GPT-5-mini (low) | Unk. | - | - | - | - | - | $51.9_{2.4}$ | $73.9_{1.5}$ | $51.9_{2.4}$ | $54.8_{2.5}$ | 152 |
| GPT-5-mini (high) | Unk. | - | - | - | - | - | $71.8_{1.7}$ | $79.5_{0.6}$ | $71.8_{1.7}$ | $71.7_{1.8}$ | 937 |

Table 23: Details on **Task (1) Discourse Marker Understanding** performance on ***Otherwise*** dataset, using **"definition+example"** prompting strategy.

## E.2 Task (2) Temporal Reasoning

Tables 24 to 28.

| | | w/o reasoning | | | | | w/ reasoning | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Size | micro-F1 | weighted-Pre | weighted-Rec | weighted-F1 | tok. | micro-F1 | weighted-Pre | weighted-Rec | weighted-F1 | tok. |
| Qwen3-0.6B | 0.6B | $27.3_{9.3}$ | $21.9_{2.0}$ | $27.3_{9.3}$ | $16.6_{9.4}$ | 3 | $24.0_{0.8}$ | $32.5_{1.7}$ | $24.0_{0.8}$ | $21.0_{1.0}$ | 285 |
| Qwen3-1.7B | 1.7B | $28.0_{0.5}$ | $30.9_{12.1}$ | $28.0_{0.5}$ | $23.2_{0.5}$ | 2 | $31.3_{0.3}$ | $37.4_{0.9}$ | $31.3_{0.3}$ | $30.5_{0.5}$ | 568 |
| Qwen3-4B | 4B | $30.5_{0.3}$ | $41.0_{1.0}$ | $30.5_{0.3}$ | $25.8_{0.5}$ | 2 | $33.8_{0.5}$ | $44.2_{1.0}$ | $33.8_{0.5}$ | $29.6_{0.3}$ | 718 |
| Qwen3-8B | 8B | $34.5_{0.4}$ | $40.6_{2.0}$ | $34.5_{0.4}$ | $29.1_{0.7}$ | 2 | $37.6_{1.0}$ | $46.7_{1.1}$ | $37.6_{1.0}$ | $33.6_{0.8}$ | 754 |
| Qwen3-14B | 14B | $34.0_{0.2}$ | $47.9_{0.4}$ | $34.0_{0.2}$ | $30.1_{0.4}$ | 2 | $42.5_{0.1}$ | $49.7_{0.7}$ | $42.5_{0.1}$ | $42.1_{0.5}$ | 472 |
| Qwen3-32B | 32B | $30.6_{0.5}$ | $42.9_{0.5}$ | $30.6_{0.5}$ | $22.8_{0.2}$ | 2 | $42.4_{0.4}$ | $53.8_{0.9}$ | $42.4_{0.4}$ | $36.6_{0.3}$ | 464 |
| DS-r1-distill-Qwen | 32B | - | - | - | - | - | $27.7_{2.7}$ | $49.2_{1.7}$ | $27.7_{2.7}$ | $26.5_{1.3}$ | 317 |
| DeepSeek-r1-0528 | 37/671B | - | - | - | - | - | $39.9_{0.0}$ | $50.9_{0.0}$ | $39.9_{0.0}$ | $33.0_{0.0}$ | 476 |
| GPT-5-mini (low) | Unk. | - | - | - | - | - | $41.9_{0.1}$ | $55.1_{0.7}$ | $41.9_{0.1}$ | $38.8_{0.1}$ | 144 |
| GPT-5-mini (high) | Unk. | - | - | - | - | - | $45.2_{0.3}$ | $56.3_{0.0}$ | $45.2_{0.3}$ | $41.4_{0.6}$ | 731 |

Table 24: Details on **Task (2) Temporal Reasoning** performance on **TBD** dataset, using **"without context"** prompting strategy.

| | | w/o reasoning | | | | | w/ reasoning | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Size | micro-F1 | Weighted-Pre | Weighted-Rec | Weighted-F1 | tok. | micro-F1 | Weighted-Pre | Weighted-Rec | Weighted-F1 | tok. |
| Qwen3-0.6B | 0.6B | $32.0_{5.5}$ | $23.2_{0.5}$ | $32.0_{5.5}$ | $24.0_{4.7}$ | 3 | $21.0_{0.3}$ | $30.2_{0.7}$ | $21.0_{0.3}$ | $17.9_{0.2}$ | 346 |
| Qwen3-1.7B | 1.7B | $23.4_{0.4}$ | $34.7_{1.2}$ | $23.4_{0.4}$ | $15.6_{0.8}$ | 2 | $26.2_{0.8}$ | $37.6_{2.1}$ | $26.2_{0.8}$ | $22.3_{0.9}$ | 559 |
| Qwen3-4B | 4B | $28.3_{0.2}$ | $42.7_{1.5}$ | $28.3_{0.2}$ | $18.8_{0.3}$ | 2 | $31.7_{0.7}$ | $45.6_{1.6}$ | $31.7_{0.7}$ | $26.1_{0.6}$ | 778 |
| Qwen3-8B | 8B | $25.4_{0.3}$ | $46.8_{2.9}$ | $25.4_{0.3}$ | $16.8_{0.4}$ | 2 | $35.3_{0.4}$ | $50.6_{1.6}$ | $35.3_{0.4}$ | $29.6_{0.6}$ | 760 |
| Qwen3-14B | 14B | $31.1_{0.4}$ | $51.0_{0.5}$ | $31.1_{0.4}$ | $27.7_{0.7}$ | 3 | $39.8_{0.5}$ | $53.7_{0.9}$ | $39.8_{0.5}$ | $39.6_{0.5}$ | 501 |
| Qwen3-32B | 32B | $27.8_{0.7}$ | $54.4_{4.9}$ | $27.8_{0.7}$ | $21.1_{1.2}$ | 3 | $40.8_{0.2}$ | $56.6_{2.0}$ | $40.8_{0.2}$ | $34.5_{0.5}$ | 475 |
| DS-r1-distill-Qwen | 32B | - | - | - | - | - | $23.8_{2.5}$ | $52.8_{3.0}$ | $23.8_{2.5}$ | $22.5_{0.7}$ | 283 |
| DeepSeek-r1-0528 | 37/671B | - | - | - | - | - | $33.1_{0.0}$ | $57.3_{0.0}$ | $33.1_{0.0}$ | $28.6_{0.0}$ | 457 |
| GPT-5-mini (low) | Unk. | - | - | - | - | - | $38.7_{0.2}$ | $59.6_{0.2}$ | $38.7_{0.2}$ | $34.4_{0.6}$ | 166 |
| GPT-5-mini (high) | Unk. | - | - | - | - | - | $43.4_{0.3}$ | $61.9_{1.9}$ | $43.4_{0.3}$ | $38.0_{0.4}$ | 870 |

Table 25: Details on **Task (2) Temporal Reasoning** performance on **TBD** dataset, using **"with context"** prompting strategy.

| | | w/o reasoning | | | | | w/ reasoning | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Size | micro-F1 | weighted-Pre | weighted-Rec | weighted-F1 | tok. | micro-F1 | weighted-Pre | weighted-Rec | weighted-F1 | tok. |
| Qwen3-0.6B | 0.6B | $12.6_{0.1}$ | $11.7_{7.8}$ | $12.6_{0.1}$ | $2.9_{0.2}$ | 3 | $25.8_{1.1}$ | $29.1_{1.1}$ | $25.8_{1.1}$ | $25.6_{1.0}$ | 392 |
| Qwen3-1.7B | 1.7B | $12.5_{0.0}$ | $1.6_{0.0}$ | $12.5_{0.0}$ | $2.8_{0.0}$ | 2 | $21.9_{0.6}$ | $29.8_{2.5}$ | $21.9_{0.6}$ | $17.9_{0.6}$ | 594 |
| Qwen3-4B | 4B | $19.4_{0.5}$ | $11.6_{0.3}$ | $19.4_{0.5}$ | $12.5_{0.3}$ | 2 | $25.4_{0.4}$ | $37.1_{0.9}$ | $25.4_{0.4}$ | $20.1_{0.5}$ | 879 |
| Qwen3-8B | 8B | $18.2_{0.2}$ | $22.3_{0.4}$ | $18.2_{0.2}$ | $11.3_{0.1}$ | 2 | $28.8_{0.3}$ | $42.0_{1.6}$ | $28.8_{0.3}$ | $23.1_{0.4}$ | 846 |
| Qwen3-14B | 14B | $26.1_{0.2}$ | $57.2_{2.6}$ | $26.1_{0.2}$ | $13.8_{0.1}$ | 2 | $29.5_{0.5}$ | $39.2_{1.5}$ | $29.5_{0.5}$ | $22.0_{0.5}$ | 582 |
| Qwen3-32B | 32B | $27.5_{0.1}$ | $37.7_{1.5}$ | $27.5_{0.1}$ | $15.3_{0.7}$ | 2 | $31.6_{0.4}$ | $43.3_{2.5}$ | $31.6_{0.4}$ | $23.9_{0.5}$ | 509 |
| DS-r1-distill-Qwen | 32B | - | - | - | - | - | $18.6_{2.6}$ | $42.4_{1.6}$ | $18.6_{2.6}$ | $15.8_{1.4}$ | 298 |
| DeepSeek-r1-0528 | 37/671B | - | - | - | - | - | $22.2_{0.0}$ | $48.8_{0.0}$ | $22.2_{0.0}$ | $21.0_{0.0}$ | 478 |
| GPT-5-mini (low) | Unk. | - | - | - | - | - | $34.4_{0.2}$ | $49.6_{0.5}$ | $34.4_{0.2}$ | $30.5_{0.3}$ | 204 |
| GPT-5-mini (high) | Unk. | - | - | - | - | - | $36.7_{0.3}$ | $52.7_{0.5}$ | $36.7_{0.3}$ | $34.6_{0.2}$ | 1203 |

Table 26: Details on **Task (2) Temporal Reasoning** performance on **TDD-Man** dataset, using **"without context"** prompting strategy.

| | | w/o reasoning | | | | | w/ reasoning | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Size | micro-F1 | Weighted-Pre | Weighted-Rec | Weighted-F1 | tok. | micro-F1 | Weighted-Pre | Weighted-Rec | Weighted-F1 | tok. |
| Qwen3-0.6B | 0.6B | $17.4_{3.4}$ | $10.0_{3.3}$ | $17.4_{3.4}$ | $8.0_{3.5}$ | 3 | $23.1_{0.7}$ | $31.2_{1.0}$ | $23.1_{0.7}$ | $19.2_{0.5}$ | 297 |
| Qwen3-1.7B | 1.7B | $13.0_{0.3}$ | $29.5_{1.7}$ | $13.0_{0.3}$ | $5.8_{0.1}$ | 2 | $21.4_{0.1}$ | $36.7_{3.6}$ | $21.4_{0.1}$ | $18.4_{0.5}$ | 624 |
| Qwen3-4B | 4B | $19.4_{0.2}$ | $11.0_{0.4}$ | $19.4_{0.2}$ | $12.6_{0.2}$ | 2 | $23.8_{0.1}$ | $40.6_{0.2}$ | $23.8_{0.1}$ | $19.9_{0.1}$ | 833 |
| Qwen3-8B | 8B | $23.4_{0.3}$ | $20.8_{1.3}$ | $23.4_{0.3}$ | $14.3_{0.1}$ | 2 | $27.6_{0.8}$ | $44.8_{2.3}$ | $27.6_{0.8}$ | $24.8_{0.8}$ | 904 |
| Qwen3-14B | 14B | $22.8_{0.1}$ | $46.7_{8.6}$ | $22.8_{0.1}$ | $13.7_{0.1}$ | 2 | $24.0_{0.6}$ | $39.6_{2.4}$ | $24.0_{0.6}$ | $20.4_{0.7}$ | 546 |
| Qwen3-32B | 32B | $24.9_{0.5}$ | $29.6_{0.4}$ | $24.9_{0.5}$ | $15.0_{0.5}$ | 2 | $28.9_{1.4}$ | $44.1_{3.3}$ | $28.9_{1.4}$ | $24.4_{1.3}$ | 460 |
| DS-r1-distill-Qwen | 32B | - | - | - | - | - | $16.9_{2.8}$ | $42.8_{2.9}$ | $16.9_{2.8}$ | $15.4_{1.6}$ | 298 |
| DeepSeek-r1-0528 | 37/671B | - | - | - | - | - | $25.9_{0.0}$ | $47.1_{0.0}$ | $25.9_{0.0}$ | $22.7_{0.0}$ | 433 |
| GPT-5-mini (low) | Unk. | - | - | - | - | - | $27.2_{0.5}$ | $53.3_{0.8}$ | $27.2_{0.5}$ | $28.6_{0.6}$ | 171 |
| GPT-5-mini (high) | Unk. | - | - | - | - | - | $27.9_{0.2}$ | $53.5_{0.0}$ | $27.9_{0.2}$ | $30.1_{0.1}$ | 935 |

Table 27: Details on **Task (2) Temporal Reasoning** performance on **TDD-Man** dataset, using **"with context"** prompting strategy.

| Model | Size | Acc (all) | Acc (tz) | Acc (as) | Acc (scd) | Acc (mop) | Acc (dur) | Acc (cmp) | Acc (trk) | tok. |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | baselines | | | | | |
| Claude-3-Sonnet | Unk. | - | 74.00 | 58.57 | 29.60 | 26.57 | 15.00 | 39.14 | 40.40 | - |
| GPT-4 | Unk. | - | 88.00 | 76.28 | 43.60 | 46.86 | 16.00 | 63.14 | 45.60 | - |
| Gemini 1.5 Pro | Unk. | - | 90.00 | 71.14 | 40.00 | 62.57 | 13.50 | 55.43 | 41.20 | - |
| | | | | | w/o reasoning | | | | | |
| Qwen3-0.6B | 0.6B | $7.0_{0.1}$ | $17.0_{1.0}$ | $8.9_{0.2}$ | $9.7_{3.1}$ | $0.0_{0.0}$ | $0.0_{0.0}$ | $8.4_{1.3}$ | $10.8_{1.6}$ | 15 |
| Qwen3-1.7B | 1.7B | $8.9_{0.2}$ | $22.7_{1.3}$ | $14.4_{0.5}$ | $7.6_{1.2}$ | $0.3_{0.0}$ | $2.0_{0.0}$ | $11.7_{0.6}$ | $10.8_{0.4}$ | 46 |
| Qwen3-4B | 4B | $15.3_{0.3}$ | $32.0_{0.0}$ | $22.4_{0.7}$ | $7.2_{1.6}$ | $0.2_{0.1}$ | $11.2_{0.3}$ | $24.8_{0.6}$ | $17.7_{0.7}$ | 67 |
| Qwen3-8B | 8B | $18.2_{0.3}$ | $45.3_{0.7}$ | $21.7_{0.6}$ | $8.4_{0.8}$ | $0.6_{0.3}$ | $12.2_{1.8}$ | $26.7_{0.2}$ | $29.7_{0.7}$ | 50 |
| Qwen3-14B | 14B | $24.4_{0.1}$ | $53.3_{0.7}$ | $35.6_{0.7}$ | $11.7_{1.9}$ | $0.1_{0.2}$ | $18.2_{0.3}$ | $32.8_{0.3}$ | $37.3_{0.7}$ | 15 |
| Qwen3-32B | 32B | $22.3_{2.7}$ | $46.3_{0.7}$ | $35.8_{2.2}$ | $6.1_{2.3}$ | $0.0_{0.0}$ | $13.0_{8.0}$ | $33.8_{3.9}$ | $32.8_{2.4}$ | 63 |
| | | | | | w/ reasoning | | | | | |
| Qwen3-0.6B | 0.6B | $21.7_{0.5}$ | $20.7_{3.3}$ | $40.5_{1.2}$ | $6.4_{0.4}$ | $18.1_{3.6}$ | $3.3_{0.7}$ | $32.1_{0.8}$ | $16.1_{1.5}$ | 1361 |
| Qwen3-1.7B | 1.7B | $46.5_{0.7}$ | $44.7_{2.3}$ | $61.0_{1.0}$ | $16.4_{2.8}$ | $54.9_{0.8}$ | $21.8_{1.7}$ | $66.2_{1.8}$ | $37.6_{1.6}$ | 1404 |
| Qwen3-4B | 4B | $64.5_{1.1}$ | $84.7_{2.3}$ | $69.6_{2.7}$ | $49.1_{1.3}$ | $82.6_{1.4}$ | $35.0_{1.5}$ | $77.9_{0.4}$ | $44.0_{1.6}$ | 1258 |
| Qwen3-8B | 8B | $72.8_{0.9}$ | $83.7_{2.3}$ | $75.6_{1.5}$ | $58.5_{1.9}$ | $84.3_{1.7}$ | $40.7_{2.3}$ | $84.2_{0.7}$ | $72.7_{1.3}$ | 1377 |
| Qwen3-14B | 14B | $80.4_{1.3}$ | $95.7_{1.3}$ | $86.0_{1.7}$ | $78.3_{4.1}$ | $90.0_{1.4}$ | $35.8_{1.2}$ | $84.9_{0.8}$ | $84.3_{0.9}$ | 1084 |
| Qwen3-32B | 32B | $81.6_{0.4}$ | $95.3_{0.7}$ | $86.3_{1.1}$ | $83.9_{2.9}$ | $92.8_{0.3}$ | $37.5_{0.5}$ | $82.5_{0.6}$ | $85.9_{1.3}$ | 1005 |
| DS-r1-distill-Qwen | 32B | $68.0_{1.2}$ | $89.0_{1.0}$ | $71.6_{0.4}$ | $71.6_{2.0}$ | $68.4_{12.7}$ | $32.5_{2.5}$ | $80.6_{1.7}$ | $61.4_{1.8}$ | 883 |
| DeepSeek-r1-0528 | 37/671B | $63.3_{0.0}$ | $94.0_{0.0}$ | $71.4_{0.0}$ | $10.0_{0.0}$ | $80.0_{0.0}$ | $22.0_{0.0}$ | $88.6_{0.0}$ | $67.2_{0.0}$ | 2431 |
| GPT-5-mini (low) | Unk. | $87.2_{0.0}$ | $96.5_{0.5}$ | $82.1_{0.7}$ | $99.0_{0.6}$ | $97.3_{0.1}$ | $49.2_{0.2}$ | $93.1_{0.3}$ | $86.4_{0.0}$ | 286 |
| GPT-5-mini (high) | Unk. | $88.2_{0.0}$ | $97.0_{0.0}$ | $84.3_{0.0}$ | $99.8_{0.2}$ | $98.4_{0.1}$ | $49.8_{0.2}$ | $92.4_{0.4}$ | $89.4_{0.2}$ | 1028 |

Table 28: Details on **Task (2) Temporal Reasoning** performance on **ToT-Arithmetic** dataset. Question type abbreviations: all – average score, tz – Timezone, as – AddSubtract, scd – Schedule, mop – MultiOperations, dur – Duration, cmp – Compare, trk – Trick; see explanation on question types in Fatemi et al. (2024).

## E.3 Task (3) Discourse Relation Recognition

Tables 29 to 33.

| | Model | Ces | Deu | Eng | Eus | Fas | Fra | Ita | Nld | Pcm | Pol | Por | Rus | Spa | Tha | Tur | Zho |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BeDiscovER | Qwen3-1.7B | 9.3 | 10.7 | 22.1 | 16.3 | 18.7 | 25.0 | 19.5 | 16.7 | 14.6 | 25 | 23.5 | 22.5 | 19.4 | 38.7 | 18.3 | 18.0 |
| | Qwen3-14B | 25.5 | 25.0 | 36.1 | 30.4 | 29.9 | 41.8 | 30.8 | 30.4 | 26.0 | 41.4 | 38.3 | 38.1 | 36.0 | 55.8 | 34.6 | 32.0 |
| | Qwen3-32B | 28.4 | 26.5 | 37.5 | 30.7 | 32.7 | 43.6 | 32.7 | 33.0 | 24.2 | 40.3 | 41.7 | 40.7 | 37.0 | 53.7 | 34.1 | 33.3 |
| | DS-r1-distill-Qwen | 26.0 | 24.8 | 32.9 | 28.1 | 26.9 | 37.7 | 25.3 | 31.8 | 22.4 | 38.7 | 36.6 | 27.2 | 36.1 | 38.4 | 27.6 | 27.8 |
| | DeepSeek-r1-0528 | 39.8 | 38.1 | 42.1 | 38.4 | 31.7 | 43.5 | 40.4 | 36.9 | 28.0 | 51.3 | 47.6 | 38.4 | 48.1 | 61.1 | 42.2 | 42.1 |
| | GPT-5-mini (low) | 38.2 | 39.2 | 43.7 | 40.9 | 29.9 | 40.5 | 36.0 | 35.1 | 31.5 | 47.3 | 48.0 | 40.6 | 43.9 | 59.1 | 42.8 | 39.4 |
| | GPT-5-mini (high) | **43.2** | **44.3** | **47.1** | **42.5** | **33.0** | **44.3** | **40.8** | **38.4** | **36.9** | **56.7** | **51.4** | **45.1** | **50.0** | **64.9** | **45.6** | **45.8** |
| | GPT-4o-mini | 30.1 | 22.4 | 29.9 | 23.6 | 17.4 | 28.8 | 27.1 | 22.6 | 29.6 | 28.6 | 33.6 | 18.3 | 9.8 | 43.7 | 29.9 | 25.3 |
| | Llama-4-Scout | 34.1 | 32.5 | 36.6 | 31.0 | 33.5 | 36.1 | 22.9 | 32.6 | 26.1 | 44.8 | 41.3 | 32.9 | 37.1 | 60.4 | 33.2 | 35.1 |
| | Qwen2.5-72B | 34.8 | 26.9 | 34.8 | 26.9 | 31.4 | 36.5 | 23.1 | 32.6 | 27.3 | 37.6 | 40.5 | 25.8 | 39.6 | 47.2 | 28.9 | 32.7 |
| Supervised | *#Datasets* | *1* | *2* | *14* | *1* | *1* | *1* | *1* | *1* | *1* | *1* | *3* | *1* | *2* | *1* | *2* | *5* |
| | *#Train examples* | *978* | *4,072* | *133,731* | *2,533* | *4,100* | *2,177* | *944* | *1,608* | *7,834* | *7,040* | *12,942* | *20,014* | *2,679* | *8,274* | *2,444* | *22,000* |
| | DeDisCo (2025) | **56.1** | **65.5** | **78.4** | 50.1 | 59.3 | **60.1** | **72.0** | **67.4** | 59.9 | **72.0** | **75.9** | **73.9** | **72.3** | **97.1** | **64.0** | **78.6** |
| | HITS (2025) | 53.4 | 61.4 | 74.7 | 54.0 | **59.8** | 57.0 | 68.5 | 64.9 | **60.4** | 72.0 | 73.1 | 72.6 | 67.8 | 95.7 | 63.0 | 72.7 |
| | DisCreT (2025) | 48.0 | 57.3 | 70.6 | **54.4** | 57.6 | 58.0 | 66.7 | 59.7 | 57.7 | 60.0 | 74.8 | 66.7 | 63.2 | 97.0 | 63.5 | 67.8 |
| | CLAC (2025) | 48.0 | 53.7 | 70.8 | 50.9 | 55.4 | 53.8 | 60.3 | 56.3 | 56.3 | 54.8 | 72.6 | 66.4 | 60.3 | 97.0 | 60.9 | 68.1 |
| | SeCoRel (2025) | 43.9 | 49.8 | 67.5 | 52.6 | 52.2 | 55.2 | 60.0 | 54.2 | 56.1 | 52.8 | 70.2 | 62.5 | 58.8 | 96.3 | 56.6 | 62.2 |

Table 29: **Task (3) Discourse Relation Recognition [language-split]** scores across 16 languages: BeDiscovER (top) vs. supervised systems (bottom). Language abbreviations (left to right): Czech, German, English, Basque, Persian, French, Italian, Dutch, Nigerian Pidgin, Polish, Portuguese, Russian, Spanish, Thai, Turkish, and Chinese. The numbers of datasets and training examples in each language are given in supervised settings.

| | | ces.rst.crdt | | | | deu.pdtb.pcc | | | | deu.rst.pcc | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Size | w/o r | tok. | w/ r | tok. | w/o r | tok. | w/ r | tok. | w/o r | tok. | w/ r | tok. |
| Qwen3-0.6B | 0.6B | $5.1_{2.3}$ | 3 | $3.2_{1.1}$ | 2505 | $8.4_{5.1}$ | 2 | $9.1_{0.8}$ | 1979 | $5.5_{2.2}$ | 2 | $6.7_{1.2}$ | 1515 |
| Qwen3-1.7B | 1.7B | $13.1_{1.8}$ | 1107 | $9.3_{1.8}$ | 2706 | $13.4_{0.5}$ | 1504 | $10.3_{2.1}$ | 2280 | $13.9_{1.1}$ | 904 | $11.1_{0.3}$ | 1833 |
| Qwen3-4B | 4B | $16.9_{0.7}$ | 3 | $11.5_{2.0}$ | 2543 | $21.5_{0.7}$ | 2 | $12.5_{3.0}$ | 2120 | $14.3_{0.7}$ | 3 | $15.8_{0.7}$ | 1653 |
| Qwen3-8B | 8B | $22.7_{0.3}$ | 3 | $11.3_{0.9}$ | 1544 | $20.6_{0.5}$ | 5 | $16.7_{2.4}$ | 1547 | $12.9_{0.7}$ | 3 | $17.7_{4.6}$ | 1245 |
| Qwen3-14B | 14B | $28.4_{0.7}$ | 3 | $25.5_{1.5}$ | 1199 | $28.0_{0.4}$ | 3 | $23.0_{1.7}$ | 1121 | $23.3_{0.5}$ | 3 | $26.4_{1.4}$ | 950 |
| Qwen3-32B | 32B | $34.5_{0.6}$ | 2 | $28.4_{1.3}$ | 689 | $25.6_{1.2}$ | 2 | $24.2_{3.6}$ | 660 | $23.2_{1.7}$ | 2 | $28.3_{1.0}$ | 602 |
| DS-r1-distill-Qwen | 32B | - | - | $26.0_{0.3}$ | 446 | - | - | $24.2_{1.0}$ | 534 | - | - | $25.3_{0.7}$ | 498 |
| DeepSeek-r1-0528 | 37/671B | - | - | $39.9_{0.0}$ | 1108 | - | - | $35.1_{0.0}$ | 694 | - | - | $40.3_{0.0}$ | 711 |
| GPT-5-mini (low) | Unk. | - | - | $38.2_{0.3}$ | 134 | - | - | $37.1_{1.0}$ | 135 | - | - | $40.7_{0.0}$ | 134 |
| GPT-5-mini (high) | Unk. | - | - | $43.2_{0.0}$ | 954 | - | - | $41.2_{0.0}$ | 1010 | - | - | $46.5_{0.0}$ | 902 |
| DeDisCO | - | 56.08 | | | | 67.53 | | | | 64.10 | | | |
| HITS | - | 53.38 | | | | 63.92 | | | | 59.71 | | | |
| DiscReT | - | 47.97 | | | | 63.92 | | | | 52.75 | | | |

| | | eng.dep.covdtb | | | | eng.dep.scidtb | | | | eng.erst.gentle | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Size | w/o r | tok. | w/ r | tok. | w/o r | tok. | w/ r | tok. | w/o r | tok. | w/ r | tok. |
| Qwen3-0.6B | 0.6B | $5.7_{1.2}$ | 4 | $19.0_{0.3}$ | 516 | $3.9_{1.1}$ | 3 | $15.2_{0.8}$ | 553 | $8.6_{3.3}$ | 3 | $11.7_{0.4}$ | 551 |
| Qwen3-1.7B | 1.7B | $11.2_{4.1}$ | 160 | $30.8_{0.6}$ | 1045 | $6.1_{1.4}$ | 188 | $28.6_{0.6}$ | 1023 | $6.8_{0.1}$ | 155 | $16.9_{0.2}$ | 1178 |
| Qwen3-4B | 4B | $34.2_{1.1}$ | 4 | $34.9_{0.4}$ | 884 | $27.3_{0.3}$ | 2 | $33.2_{0.8}$ | 874 | $14.0_{0.3}$ | 3 | $19.5_{0.8}$ | 896 |
| Qwen3-8B | 8B | $29.9_{1.7}$ | 2 | $36.6_{1.6}$ | 1298 | $23.9_{1.4}$ | 3 | $38.8_{2.1}$ | 1256 | $13.8_{0.8}$ | 2 | $20.9_{0.9}$ | 1219 |
| Qwen3-14B | 14B | $46.9_{1.1}$ | 2 | $47.7_{0.4}$ | 804 | $44.5_{0.8}$ | 2 | $45.0_{0.3}$ | 829 | $23.1_{0.3}$ | 3 | $27.9_{0.9}$ | 756 |
| Qwen3-32B | 32B | $47.0_{1.4}$ | 2 | $48.8_{0.7}$ | 480 | $43.6_{0.8}$ | 2 | $47.2_{0.2}$ | 454 | $25.9_{0.4}$ | 2 | $30.0_{0.6}$ | 484 |
| DS-r1-distill-Qwen | 32B | - | - | $40.9_{1.0}$ | 396 | - | - | $38.0_{1.3}$ | 415 | - | - | $24.1_{2.5}$ | 382 |
| DeepSeek-r1-0528 | 37/671B | - | - | $50.4_{0.0}$ | 823 | - | - | $48.2_{0.0}$ | 799 | - | - | $25.5_{0.0}$ | 660 |
| GPT-5-mini (low) | Unk. | - | - | $56.8_{2.1}$ | 107 | - | - | $53.4_{1.0}$ | 99 | - | - | $20.8_{0.4}$ | 103 |
| GPT-5-mini (high) | Unk. | - | - | $58.5_{0.0}$ | 698 | - | - | $55.5_{0.0}$ | 592 | - | - | $25.1_{0.0}$ | 688 |
| DeDisCO | - | 71.46 | | | | 84.29 | | | | 68.30 | | | |
| HITS | - | 71.31 | | | | 81.78 | | | | 62.42 | | | |
| DiscReT | - | 69.22 | | | | 78.22 | | | | 53.53 | | | |

| | | eng.erst.gum | | | | eng.pdtb.gentle | | | | eng.pdtb.gum | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Size | w/o r | tok. | w/ r | tok. | w/o r | tok. | w/ r | tok. | w/o r | tok. | w/ r | tok. |
| Qwen3-0.6B | 0.6B | $6.9_{1.4}$ | 3 | $15.4_{0.4}$ | 484 | $13.8_{3.0}$ | 2 | $14.7_{0.4}$ | 785 | $12.2_{3.5}$ | 3 | $18.1_{0.7}$ | 564 |
| Qwen3-1.7B | 1.7B | $10.9_{1.1}$ | 119 | $23.7_{0.3}$ | 1096 | $12.4_{1.3}$ | 355 | $23.2_{1.5}$ | 1347 | $13.8_{2.6}$ | 199 | $26.0_{0.3}$ | 1139 |
| Qwen3-4B | 4B | $20.9_{0.5}$ | 3 | $28.8_{0.0}$ | 880 | $16.5_{0.2}$ | 2 | $29.6_{0.9}$ | 1176 | $18.9_{0.2}$ | 3 | $31.8_{0.4}$ | 1020 |
| Qwen3-8B | 8B | $19.0_{0.5}$ | 2 | $29.9_{1.3}$ | 1331 | $22.1_{0.0}$ | 3 | $29.7_{1.0}$ | 1458 | $23.3_{0.5}$ | 3 | $34.8_{1.0}$ | 1454 |
| Qwen3-14B | 14B | $30.3_{0.3}$ | 2 | $39.1_{0.8}$ | 743 | $31.3_{0.3}$ | 2 | $38.2_{0.7}$ | 919 | $32.3_{0.4}$ | 3 | $42.7_{0.4}$ | 826 |
| Qwen3-32B | 32B | $32.9_{0.5}$ | 2 | $41.1_{0.1}$ | 483 | $26.2_{0.8}$ | 2 | $37.9_{0.1}$ | 569 | $33.2_{1.5}$ | 2 | $44.2_{0.3}$ | 548 |
| DS-r1-distill-Qwen | 32B | - | - | $37.8_{0.6}$ | 410 | - | - | $29.1_{1.5}$ | 377 | - | - | $28.9_{1.0}$ | 368 |
| DeepSeek-r1-0528 | 37/671B | - | - | $51.6_{0.0}$ | 774 | - | - | $45.1_{0.0}$ | 758 | - | - | $46.6_{0.0}$ | 799 |
| GPT-5-mini (low) | Unk. | - | - | $51.6_{2.0}$ | 92 | - | - | $40.2_{0.6}$ | 108 | - | - | $44.6_{2.0}$ | 103 |
| GPT-5-mini (high) | Unk. | - | - | $50.8_{0.0}$ | 646 | - | - | $47.7_{0.0}$ | 780 | - | - | $47.4_{0.0}$ | 737 |
| DeDisCO | - | 76.50 | | | | 67.30 | | | | 73.48 | | | |
| HITS | - | 67.32 | | | | 64.89 | | | | 67.88 | | | |
| DiscReT | - | 64.21 | | | | 64.25 | | | | 69.31 | | | |

| | | eng.pdtb.pdtb | | | | eng.pdtb.tedm | | | | eng.rst.oll | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Size | w/o r | tok. | w/ r | tok. | w/o r | tok. | w/ r | tok. | w/o r | tok. | w/ r | tok. |
| Qwen3-0.6B | 0.6B | $16.3_{3.2}$ | 3 | $22.6_{0.6}$ | 459 | $15.4_{3.0}$ | 2 | $15.8_{1.3}$ | 1223 | $6.3_{0.6}$ | 2 | $12.4_{1.9}$ | 1530 |
| Qwen3-1.7B | 1.7B | $21.2_{2.7}$ | 107 | $27.6_{0.2}$ | 1048 | $17.9_{2.0}$ | 724 | $20.1_{0.1}$ | 1600 | $12.2_{1.1}$ | 952 | $16.6_{0.4}$ | 1845 |
| Qwen3-4B | 4B | $23.3_{0.4}$ | 6 | $35.8_{0.7}$ | 854 | $24.5_{0.0}$ | 7 | $23.7_{1.7}$ | 1487 | $25.0_{0.5}$ | 3 | $20.0_{0.3}$ | 1649 |
| Qwen3-8B | 8B | $29.6_{0.2}$ | 2 | $38.7_{1.2}$ | 1212 | $23.6_{0.6}$ | 3 | $33.1_{3.1}$ | 1509 | $24.4_{1.4}$ | 4 | $23.1_{2.4}$ | 1296 |
| Qwen3-14B | 14B | $34.7_{0.6}$ | 2 | $43.1_{0.4}$ | 715 | $30.3_{1.0}$ | 2 | $37.2_{0.7}$ | 932 | $28.9_{0.3}$ | 3 | $30.6_{1.5}$ | 961 |
| Qwen3-32B | 32B | $32.3_{0.6}$ | 2 | $43.2_{0.1}$ | 520 | $29.2_{0.7}$ | 2 | $38.0_{2.7}$ | 601 | $26.3_{1.0}$ | 2 | $31.5_{0.6}$ | 609 |
| DS-r1-distill-Qwen | 32B | - | - | $39.0_{1.6}$ | 391 | - | - | $34.7_{0.4}$ | 361 | - | - | $32.3_{0.6}$ | 350 |
| DeepSeek-r1-0528 | 37/671B | - | - | $56.1_{0.0}$ | 723 | - | - | $40.8_{0.0}$ | 745 | - | - | $38.7_{0.0}$ | 796 |
| GPT-5-mini (low) | Unk. | - | - | $57.0_{0.5}$ | 98 | - | - | $38.4_{0.0}$ | 108 | - | - | $38.9_{1.3}$ | 110 |
| GPT-5-mini (high) | Unk. | - | - | $63.6_{0.0}$ | 682 | - | - | $40.0_{0.0}$ | 849 | - | - | $41.7_{0.0}$ | 839 |
| DeDisCO | - | 83.54 | | | | 68.95 | | | | 62.73 | | | |
| HITS | - | 79.95 | | | | 64.96 | | | | 58.30 | | | |
| DiscReT | - | 75.06 | | | | 61.54 | | | | 47.23 | | | |

Table 30: Details on **Task (3) Discourse Relation Recognition [language.framework.dataset split]** scores on 38 separate datasets in DISRPT shared task (1/4).

| | | eng.rst.rstdt | | | | eng.rst.sts | | | | eng.rst.umuc | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Size | w/o r | tok. | w/ r | tok. | w/o r | tok. | w/ r | tok. | w/o r | tok. | w/ r | tok. |
| Qwen3-0.6B | 0.6B | $3.9_{0.9}$ | 2 | $15.8_{0.5}$ | 563 | $7.9_{0.9}$ | 2 | $13.2_{0.8}$ | 1290 | $3.7_{1.6}$ | 2 | $16.5_{1.1}$ | 988 |
| Qwen3-1.7B | 1.7B | $9.9_{1.4}$ | 186 | $26.3_{0.7}$ | 1190 | $13.7_{0.9}$ | 875 | $16.6_{2.3}$ | 1692 | $12.2_{2.1}$ | 450 | $22.8_{1.4}$ | 1472 |
| Qwen3-4B | 4B | $31.0_{0.6}$ | 2 | $33.7_{0.3}$ | 921 | $26.3_{1.1}$ | 11 | $21.7_{1.5}$ | 1475 | $18.6_{0.2}$ | 5 | $27.1_{0.8}$ | 1281 |
| Qwen3-8B | 8B | $25.6_{0.6}$ | 2 | $36.1_{2.0}$ | 1312 | $28.3_{1.0}$ | 3 | $21.7_{1.2}$ | 1306 | $18.5_{0.5}$ | 3 | $34.0_{3.8}$ | 1291 |
| Qwen3-14B | 14B | $38.8_{0.6}$ | 3 | $43.1_{0.6}$ | 814 | $36.0_{0.9}$ | 3 | $30.4_{1.6}$ | 924 | $30.0_{0.8}$ | 3 | $41.9_{1.5}$ | 878 |
| Qwen3-32B | 32B | $39.9_{1.1}$ | 2 | $44.5_{1.1}$ | 556 | $33.3_{1.2}$ | 2 | $29.2_{1.0}$ | 614 | $36.0_{0.4}$ | 2 | $42.9_{0.3}$ | 593 |
| DS-r1-distill-Qwen | 32B | - | - | $48.8_{0.5}$ | 389 | - | - | $32.8_{0.4}$ | 399 | - | - | $37.6_{0.4}$ | 424 |
| DeepSeek-r1-0528 | 37/671B | - | - | $63.7_{0.0}$ | 768 | - | - | $36.3_{0.0}$ | 772 | - | - | $49.6_{0.0}$ | 738 |
| GPT-5-mini (low) | Unk. | - | - | $64.0_{2.1}$ | 102 | - | - | $36.6_{0.0}$ | 114 | - | - | $50.4_{0.4}$ | 115 |
| GPT-5-mini (high) | Unk. | - | - | $69.3_{0.0}$ | 680 | - | - | $37.0_{0.0}$ | 883 | - | - | $56.6_{0.0}$ | 769 |
| DeDisCO | - | 73.09 | | | | 54.27 | | | | 65.91 | | | |
| HITS | - | 64.92 | | | | 54.27 | | | | 63.84 | | | |
| DiscReT | - | 60.93 | | | | 42.68 | | | | 59.09 | | | |

| | | eng.sdrt.msdc | | | | eng.sdrt.stac | | | | eus.rst.ert | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Size | w/o r | tok. | w/ r | tok. | w/o r | tok. | w/ r | tok. | w/o r | tok. | w/ r | tok. |
| Qwen3-0.6B | 0.6B | $3.0_{0.4}$ | 2 | $9.5_{0.2}$ | 514 | $4.5_{0.5}$ | 2 | $11.0_{0.6}$ | 651 | $4.0_{0.8}$ | 3 | $9.5_{0.5}$ | 1031 |
| Qwen3-1.7B | 1.7B | $3.8_{0.6}$ | 127 | $17.1_{0.2}$ | 1149 | $5.7_{0.2}$ | 19 | $16.0_{0.8}$ | 1330 | $5.8_{1.0}$ | 515 | $16.3_{1.6}$ | 1451 |
| Qwen3-4B | 4B | $14.2_{0.3}$ | 3 | $17.4_{0.2}$ | 808 | $7.2_{0.2}$ | 2 | $15.4_{0.4}$ | 1059 | $23.4_{0.3}$ | 3 | $17.4_{2.2}$ | 1234 |
| Qwen3-8B | 8B | $17.0_{0.6}$ | 3 | $24.9_{0.3}$ | 880 | $10.9_{0.8}$ | 2 | $18.3_{0.4}$ | 1031 | $19.9_{0.9}$ | 3 | $20.1_{3.2}$ | 1307 |
| Qwen3-14B | 14B | $16.5_{0.3}$ | 2 | $21.0_{0.6}$ | 670 | $13.8_{0.7}$ | 2 | $20.6_{0.6}$ | 873 | $30.9_{0.2}$ | 3 | $30.4_{1.4}$ | 813 |
| Qwen3-32B | 32B | $14.9_{0.4}$ | 2 | $23.8_{0.4}$ | 488 | $16.9_{1.4}$ | 2 | $27.0_{1.1}$ | 548 | $30.6_{0.5}$ | 2 | $30.7_{0.4}$ | 534 |
| DS-r1-distill-Qwen | 32B | - | - | $19.2_{1.2}$ | 370 | - | - | $21.3_{2.7}$ | 434 | - | - | $28.1_{0.4}$ | 494 |
| DeepSeek-r1-0528 | 37/671B | - | - | $17.1_{0.0}$ | 765 | - | - | $25.8_{0.0}$ | 913 | - | - | $38.4_{0.0}$ | 1006 |
| GPT-5-mini (low) | Unk. | - | - | $25.4_{0.0}$ | 95 | - | - | $41.3_{1.3}$ | 120 | - | - | $40.9_{1.2}$ | 126 |
| GPT-5-mini (high) | Unk. | - | - | $27.1_{0.0}$ | 673 | - | - | $47.6_{0.0}$ | 826 | - | - | $42.6_{0.0}$ | 838 |
| DeDisCO | - | 90.00 | | | | 77.04 | | | | 50.10 | | | |
| HITS | - | 89.60 | | | | 75.89 | | | | 54.02 | | | |
| DiscReT | - | 85.64 | | | | 69.50 | | | | 54.43 | | | |

| | | fas.rst.prstc | | | | fra.sdrt.annodis | | | | ita.pdtb.luna | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Size | w/o r | tok. | w/ r | tok. | w/o r | tok. | w/ r | tok. | w/o r | tok. | w/ r | tok. |
| Qwen3-0.6B | 0.6B | $1.9_{1.0}$ | 2 | $7.8_{0.6}$ | 932 | $5.6_{2.8}$ | 2 | $14.0_{0.3}$ | 852 | $10.3_{0.6}$ | 2 | $11.4_{0.7}$ | 1176 |
| Qwen3-1.7B | 1.7B | $5.9_{0.5}$ | 301 | $18.7_{0.2}$ | 1283 | $10.3_{2.1}$ | 356 | $25.0_{0.9}$ | 1158 | $12.4_{1.2}$ | 132 | $19.5_{1.0}$ | 1530 |
| Qwen3-4B | 4B | $19.3_{0.0}$ | 3 | $22.7_{1.1}$ | 1052 | $21.9_{0.0}$ | 5 | $31.5_{0.7}$ | 928 | $16.7_{0.4}$ | 10 | $26.2_{0.5}$ | 1356 |
| Qwen3-8B | 8B | $16.4_{1.2}$ | 2 | $27.1_{3.3}$ | 1317 | $26.3_{0.4}$ | 3 | $38.8_{1.1}$ | 737 | $22.8_{0.4}$ | 3 | $30.0_{0.9}$ | 1329 |
| Qwen3-14B | 14B | $28.0_{1.2}$ | 2 | $29.9_{2.7}$ | 865 | $35.7_{1.0}$ | 3 | $41.8_{0.4}$ | 617 | $25.9_{0.5}$ | 3 | $30.8_{1.7}$ | 856 |
| Qwen3-32B | 32B | $30.5_{0.6}$ | 2 | $32.7_{0.4}$ | 548 | $39.4_{0.5}$ | 2 | $43.6_{0.7}$ | 440 | $20.6_{2.1}$ | 2 | $32.7_{2.0}$ | 494 |
| DS-r1-distill-Qwen | 32B | - | - | $26.9_{2.3}$ | 436 | - | - | $37.7_{0.6}$ | 387 | - | - | $25.3_{2.2}$ | 390 |
| DeepSeek-r1-0528 | 37/671B | - | - | $31.8_{0.0}$ | 646 | - | - | $43.5_{0.0}$ | 1032 | - | - | $40.4_{0.0}$ | 1108 |
| GPT-5-mini (low) | Unk. | - | - | $29.9_{1.5}$ | 145 | - | - | $40.5_{1.0}$ | 97 | - | - | $36.0_{0.4}$ | 131 |
| GPT-5-mini (high) | Unk. | - | - | $33.1_{0.0}$ | 1106 | - | - | $44.4_{0.0}$ | 644 | - | - | $40.9_{0.0}$ | 912 |
| DeDisCO | - | 59.29 | | | | 60.06 | | | | 72.00 | | | |
| HITS | - | 59.80 | | | | 57.00 | | | | 68.53 | | | |
| DiscReT | - | 57.60 | | | | 57.97 | | | | 66.67 | | | |

| | | nld.rst.nldt | | | | pcm.pdtb.disconaija | | | | pol.iso.pdc | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Size | w/o r | tok. | w/ r | tok. | w/o r | tok. | w/ r | tok. | w/o r | tok. | w/ r | tok. |
| Qwen3-0.6B | 0.6B | $4.6_{2.5}$ | 2 | $11.1_{1.6}$ | 1297 | $10.4_{1.2}$ | 2 | $11.3_{0.3}$ | 702 | $8.2_{2.8}$ | 2 | $12.3_{0.8}$ | 836 |
| Qwen3-1.7B | 1.7B | $9.4_{0.8}$ | 635 | $16.7_{0.8}$ | 1645 | $12.1_{0.6}$ | 130 | $14.6_{1.0}$ | 1447 | $11.9_{2.9}$ | 410 | $25.0_{1.5}$ | 1240 |
| Qwen3-4B | 4B | $20.9_{1.3}$ | 8 | $18.8_{1.2}$ | 1477 | $16.7_{0.9}$ | 4 | $20.8_{1.6}$ | 1155 | $26.8_{0.3}$ | 4 | $30.8_{1.4}$ | 1060 |
| Qwen3-8B | 8B | $21.7_{1.7}$ | 4 | $23.3_{0.7}$ | 1235 | $17.6_{0.1}$ | 2 | $18.8_{0.7}$ | 1484 | $26.2_{0.2}$ | 3 | $34.5_{2.4}$ | 1340 |
| Qwen3-14B | 14B | $33.8_{1.3}$ | 3 | $30.4_{0.4}$ | 903 | $22.0_{0.3}$ | 2 | $26.0_{0.3}$ | 921 | $32.3_{1.1}$ | 3 | $41.4_{1.0}$ | 800 |
| Qwen3-32B | 32B | $34.2_{1.5}$ | 2 | $33.0_{0.8}$ | 578 | $17.6_{0.4}$ | 2 | $24.2_{0.7}$ | 570 | $31.9_{1.6}$ | 2 | $40.3_{1.4}$ | 526 |
| DS-r1-distill-Qwen | 32B | - | - | $34.5_{2.4}$ | 491 | - | - | $29.1_{0.8}$ | 415 | - | - | $45.6_{3.3}$ | 468 |
| DeepSeek-r1-0528 | 37/671B | - | - | $49.8_{0.0}$ | 1053 | - | - | $39.8_{0.0}$ | 1117 | - | - | $52.9_{0.0}$ | 1072 |
| GPT-5-mini (low) | Unk. | - | - | $47.4_{2.0}$ | 121 | - | - | $39.4_{0.4}$ | 124 | - | - | $56.8_{0.2}$ | 107 |
| GPT-5-mini (high) | Unk. | - | - | $50.6_{0.0}$ | 899 | - | - | $44.1_{0.0}$ | 896 | - | - | $59.2_{0.0}$ | 758 |
| DeDisCO | - | 67.38 | | | | 59.88 | | | | 72.01 | | | |
| HITS | - | 64.92 | | | | 60.37 | | | | 72.01 | | | |
| DiscReT | - | 59.69 | | | | 57.72 | | | | 60.03 | | | |

Table 31: Details on **Task (3) Discourse Relation Recognition [language.framework.dataset split]** scores on 38 separate datasets in DISRPT shared task (2/4).

| Model | Size | por.pdtb.crpc w/o r | tok. | w/ r | tok. | por.pdtb.tedm w/o r | tok. | w/ r | tok. | por.rst.cstn w/o r | tok. | w/ r | tok. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Qwen3-0.6B | 0.6B | $12.2_{5.3}$ | 3 | $17.5_{0.6}$ | 652 | $13.6_{4.2}$ | 2 | $14.3_{1.9}$ | 1207 | $6.0_{3.3}$ | 3 | $10.3_{1.2}$ | 1476 |
| Qwen3-1.7B | 1.7B | $14.7_{2.9}$ | 247 | $25.6_{0.6}$ | 1098 | $17.3_{1.1}$ | 873 | $23.5_{1.2}$ | 1530 | $10.3_{0.4}$ | 618 | $21.7_{1.8}$ | 1828 |
| Qwen3-4B | 4B | $23.7_{0.5}$ | 3 | $26.4_{1.0}$ | 963 | $25.7_{0.4}$ | 3 | $24.5_{0.5}$ | 1377 | $31.5_{0.9}$ | 18 | $27.6_{2.2}$ | 1652 |
| Qwen3-8B | 8B | $28.5_{1.1}$ | 3 | $31.6_{0.9}$ | 1353 | $27.0_{2.4}$ | 3 | $29.5_{1.3}$ | 1388 | $30.4_{2.0}$ | 3 | $34.9_{1.1}$ | 1322 |
| Qwen3-14B | 14B | $39.5_{0.9}$ | 3 | $39.7_{0.3}$ | 794 | $32.5_{0.2}$ | 3 | $35.6_{2.0}$ | 873 | $46.0_{0.3}$ | 3 | $39.6_{3.0}$ | 971 |
| Qwen3-32B | 32B | $39.9_{1.6}$ | 2 | $41.1_{0.6}$ | 545 | $33.7_{1.5}$ | 2 | $38.8_{1.9}$ | 569 | $47.1_{0.3}$ | 2 | $45.0_{1.0}$ | 624 |
| DS-r1-distill-Qwen | 32B | - | - | $34.5_{2.4}$ | 491 | - | - | $29.1_{0.8}$ | 415 | - | - | $45.6_{3.3}$ | 468 |
| DeepSeek-r1-0528 | 37/671B | - | - | $49.8_{0.0}$ | 1053 | - | - | $39.8_{0.0}$ | 1117 | - | - | $52.9_{0.0}$ | 1072 |
| GPT-5-mini (low) | Unk. | - | - | $47.4_{2.0}$ | 121 | - | - | $39.4_{0.4}$ | 124 | - | - | $56.8_{0.2}$ | 107 |
| GPT-5-mini (high) | Unk. | - | - | $50.6_{0.0}$ | 899 | - | - | $44.1_{0.0}$ | 896 | - | - | $59.2_{0.0}$ | 758 |
| DeDisCO | - | 78.61 | | | | 70.33 | | | | 71.32 | | | |
| HITS | - | 76.12 | | | | 65.11 | | | | 70.22 | | | |
| DiscReT | - | 79.09 | | | | 65.66 | | | | 68.01 | | | |

| Model | Size | rus.rst.rrt w/o r | tok. | w/ r | tok. | spa.rst.rststb w/o r | tok. | w/ r | tok. | spa.rst.sctb w/o r | tok. | w/ r | tok. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Qwen3-0.6B | 0.6B | $5.9_{1.8}$ | 2 | $13.3_{0.5}$ | 512 | $7.6_{3.4}$ | 4 | $12.2_{0.8}$ | 1063 | $9.4_{4.1}$ | 2 | $9.3_{1.0}$ | 2280 |
| Qwen3-1.7B | 1.7B | $9.4_{1.9}$ | 142 | $22.5_{0.4}$ | 1070 | $11.6_{0.8}$ | 681 | $20.9_{1.2}$ | 1368 | $3.4_{1.0}$ | 1081 | $17.4_{2.1}$ | 2562 |
| Qwen3-4B | 4B | $22.8_{0.2}$ | 3 | $29.8_{0.8}$ | 856 | $23.0_{0.9}$ | 6 | $23.2_{1.9}$ | 1262 | $29.8_{1.6}$ | 3 | $17.6_{1.9}$ | 2450 |
| Qwen3-8B | 8B | $19.7_{0.4}$ | 2 | $31.6_{1.8}$ | 1283 | $21.0_{1.1}$ | 5 | $27.3_{0.6}$ | 1247 | $35.6_{0.9}$ | 4 | $19.5_{3.1}$ | 1547 |
| Qwen3-14B | 14B | $28.3_{0.3}$ | 3 | $38.1_{0.7}$ | 783 | $28.6_{0.5}$ | 3 | $33.6_{0.9}$ | 886 | $50.3_{0.6}$ | 5 | $39.2_{1.1}$ | 1151 |
| Qwen3-32B | 32B | $35.4_{0.4}$ | 2 | $40.7_{0.4}$ | 531 | $29.8_{1.0}$ | 2 | $32.5_{0.6}$ | 577 | $46.5_{1.3}$ | 3 | $43.2_{6.5}$ | 671 |
| DS-r1-distill-Qwen | 32B | - | - | $27.2_{0.5}$ | 364 | - | - | $31.0_{1.4}$ | 478 | - | - | $43.1_{2.8}$ | 468 |
| DeepSeek-r1-0528 | 37/671B | - | - | $38.4_{0.0}$ | 1008 | - | - | $39.0_{0.0}$ | 1195 | - | - | $60.4_{0.0}$ | 1151 |
| GPT-5-mini (low) | Unk. | - | - | $40.6_{1.4}$ | 135 | - | - | $36.2_{0.5}$ | 112 | - | - | $54.4_{0.9}$ | 107 |
| GPT-5-mini (high) | Unk. | - | - | $45.2_{0.0}$ | 989 | - | - | $43.2_{0.0}$ | 833 | - | - | $59.1_{0.0}$ | 792 |
| DeDisCO | - | 73.93 | | | | 69.25 | | | | 80.50 | | | |
| HITS | - | 72.58 | | | | 65.49 | | | | 74.21 | | | |
| DiscReT | - | 66.68 | | | | 61.50 | | | | 67.92 | | | |

| Model | Size | tha.pdtb.tdtb w/o r | tok. | w/ r | tok. | tur.pdtb.tdb w/o r | tok. | w/ r | tok. | tur.pdtb.tedm w/o r | tok. | w/ r | tok. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Qwen3-0.6B | 0.6B | $9.1_{3.0}$ | 2 | $21.2_{1.1}$ | 617 | $6.4_{4.4}$ | 2 | $12.4_{0.9}$ | 1118 | $8.7_{4.1}$ | 3 | $10.9_{0.8}$ | 1252 |
| Qwen3-1.7B | 1.7B | $21.4_{3.2}$ | 248 | $38.7_{1.0}$ | 939 | $12.0_{1.3}$ | 521 | $17.7_{0.4}$ | 1501 | $13.6_{3.2}$ | 692 | $18.8_{0.8}$ | 1544 |
| Qwen3-4B | 4B | $28.8_{0.5}$ | 2 | $43.2_{1.2}$ | 685 | $19.0_{1.0}$ | 3 | $25.2_{1.2}$ | 1361 | $22.0_{0.6}$ | 3 | $22.2_{1.2}$ | 1394 |
| Qwen3-8B | 8B | $38.7_{0.9}$ | 2 | $53.2_{1.5}$ | 650 | $18.0_{1.0}$ | 3 | $26.3_{2.7}$ | 1385 | $22.8_{0.6}$ | 4 | $25.9_{0.3}$ | 1315 |
| Qwen3-14B | 14B | $43.3_{1.2}$ | 2 | $55.8_{1.7}$ | 552 | $27.8_{0.2}$ | 3 | $36.9_{2.1}$ | 875 | $27.5_{0.3}$ | 3 | $32.7_{0.4}$ | 869 |
| Qwen3-32B | 32B | $42.8_{1.7}$ | 2 | $53.7_{0.1}$ | 452 | $29.3_{0.9}$ | 2 | $35.2_{0.9}$ | 591 | $26.9_{0.9}$ | 2 | $33.3_{0.3}$ | 563 |
| DS-r1-distill-Qwen | 32B | - | - | $38.4_{0.7}$ | 342 | - | - | $28.6_{1.0}$ | 405 | - | - | $26.8_{1.2}$ | 394 |
| DeepSeek-r1-0528 | 37/671B | - | - | $61.2_{0.0}$ | 641 | - | - | $45.2_{0.0}$ | 1116 | - | - | $39.8_{0.0}$ | 1031 |
| GPT-5-mini (low) | Unk. | - | - | $59.1_{0.9}$ | 93 | - | - | $50.2_{1.2}$ | 133 | - | - | $36.8_{0.2}$ | 132 |
| GPT-5-mini (high) | Unk. | - | - | $64.9_{0.0}$ | 659 | - | - | $49.5_{0.0}$ | 875 | - | - | $42.5_{0.0}$ | 1036 |
| DeDisCO | - | 97.10 | | | | 68.65 | | | | 58.68 | | | |
| HITS | - | 95.68 | | | | 66.03 | | | | 59.50 | | | |
| DiscReT | - | 97.02 | | | | 65.80 | | | | 60.88 | | | |

| Model | Size | zho.dep.scidtb w/o r | tok. | w/ r | tok. | zho.pdtb.cdtb w/o r | tok. | w/ r | tok. | zho.pdtb.ted w/o r | tok. | w/ r | tok. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Qwen3-0.6B | 0.6B | $3.9_{2.8}$ | 3 | $6.9_{1.5}$ | 1738 | $6.9_{5.0}$ | 4 | $12.8_{1.0}$ | 807 | $11.2_{1.6}$ | 2 | $13.0_{0.4}$ | 583 |
| Qwen3-1.7B | 1.7B | $5.7_{1.3}$ | 416 | $10.9_{0.3}$ | 1995 | $8.2_{1.8}$ | 355 | $23.9_{0.4}$ | 1223 | $17.1_{0.5}$ | 250 | $20.8_{0.1}$ | 1026 |
| Qwen3-4B | 4B | $26.8_{0.6}$ | 9 | $16.7_{1.0}$ | 1954 | $19.5_{0.4}$ | 2 | $25.3_{0.4}$ | 1119 | $23.3_{0.0}$ | 2 | $25.9_{0.8}$ | 913 |
| Qwen3-8B | 8B | $24.7_{0.9}$ | 4 | $17.2_{0.5}$ | 1383 | $26.2_{0.4}$ | 3 | $29.0_{1.2}$ | 1388 | $25.2_{0.2}$ | 2 | $27.0_{1.5}$ | 1297 |
| Qwen3-14B | 14B | $34.7_{1.1}$ | 3 | $27.6_{1.2}$ | 1021 | $27.3_{0.1}$ | 2 | $35.1_{0.7}$ | 920 | $27.4_{0.4}$ | 2 | $32.8_{0.9}$ | 800 |
| Qwen3-32B | 32B | $31.2_{2.3}$ | 2 | $29.0_{0.8}$ | 530 | $31.3_{0.6}$ | 2 | $34.9_{1.0}$ | 582 | $27.7_{0.6}$ | 2 | $33.6_{0.2}$ | 503 |
| DS-r1-distill-Qwen | 32B | - | - | $24.2_{0.9}$ | 288 | - | - | $28.9_{0.7}$ | 314 | - | - | $27.4_{0.8}$ | 291 |
| DeepSeek-r1-0528 | 37/671B | - | - | $38.1_{0.0}$ | 1156 | - | - | $37.9_{0.0}$ | 968 | - | - | $45.9_{0.0}$ | 922 |
| GPT-5-mini (low) | Unk. | - | - | $35.1_{0.7}$ | 127 | - | - | $39.6_{0.9}$ | 113 | - | - | $45.1_{0.4}$ | 108 |
| GPT-5-mini (high) | Unk. | - | - | $44.2_{0.0}$ | 970 | - | - | $42.7_{0.0}$ | 920 | - | - | $52.6_{0.0}$ | 742 |
| DeDisCO | - | 75.35 | | | | 89.97 | | | | 75.64 | | | |
| HITS | - | 70.23 | | | | 81.79 | | | | 70.75 | | | |
| DiscReT | - | 69.77 | | | | 77.57 | | | | 67.74 | | | |

Table 32: Details on **Task (3) Discourse Relation Recognition [language.framework.dataset split]** scores on 38 separate datasets in DISRPT shared task (3/4).

| | | zho.rst.gcdt | | | | zho.rst.sctb | | | |
|---|---|---|---|---|---|---|---|---|---|
| Model | Size | w/o r | tok. | w/ r | tok. | w/o r | tok. | w/ r | tok. |
| Qwen3-0.6B | 0.6B | $6.3_{2.2}$ | 4 | $14.9_{0.8}$ | 703 | $4.3_{3.0}$ | 3 | $6.2_{1.1}$ | 2321 |
| Qwen3-1.7B | 1.7B | $11.8_{0.9}$ | 188 | $18.9_{0.2}$ | 1113 | $4.6_{3.6}$ | 1310 | $13.6_{1.5}$ | 2498 |
| Qwen3-4B | 4B | $19.3_{0.2}$ | 4 | $24.2_{1.7}$ | 946 | $33.5_{0.5}$ | 3 | $15.7_{1.9}$ | 2348 |
| Qwen3-8B | 8B | $20.7_{0.3}$ | 3 | $26.4_{1.9}$ | 1204 | $32.7_{1.9}$ | 4 | $17.4_{0.8}$ | 1521 |
| Qwen3-14B | 14B | $27.2_{0.4}$ | 3 | $31.1_{0.3}$ | 785 | $43.0_{0.4}$ | 4 | $33.3_{2.5}$ | 1039 |
| Qwen3-32B | 32B | $31.4_{0.4}$ | 2 | $33.2_{0.6}$ | 499 | $43.4_{1.3}$ | 2 | $37.1_{1.3}$ | 556 |
| DS-r1-distill-Qwen | 32B | - | - | $25.0_{0.8}$ | 278 | - | - | $35.2_{0.6}$ | 293 |
| DeepSeek-r1-0528 | 37/671B | - | - | $36.8_{0.0}$ | 1120 | - | - | $54.1_{0.0}$ | 1036 |
| GPT-5-mini (low) | Unk. | - | - | $31.8_{0.3}$ | 124 | - | - | $44.7_{0.6}$ | 115 |
| GPT-5-mini (high) | Unk. | - | - | $38.4_{0.0}$ | 944 | - | - | $50.3_{0.0}$ | 882 |
| DeDisCO | - | 75.35 | | | | 75.47 | | | |
| HITS | - | 71.46 | | | | 57.86 | | | |
| DiscReT | - | 61.91 | | | | 55.35 | | | |

Table 33: Details on **Task (3) Discourse Relation Recognition [language.framework.dataset split]** scores on 38 separate datasets in DISRPT shared task (4/4).

### E.4 Task (4) Sentence Ordering

Tables 34 to 40

| Model | Size | w/o reasoning | | | | | | w/ reasoning | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Acc | Lumr | Pmr | Skip | Lcs | tok. | Acc | Lumr | Pmr | Skip | Lcs | tok. |
| Qwen3-0.6B | 0.6B | $21.7_{0.3}$ | $32.4_{4.7}$ | $5.4_{0.9}$ | $50.0_{1.9}$ | $55.3_{0.5}$ | 24 | $22.3_{0.3}$ | $8.4_{0.0}$ | $6.3_{0.8}$ | $53.0_{0.4}$ | $56.3_{0.1}$ | 559 |
| Qwen3-1.7B | 1.7B | $31.2_{0.2}$ | $13.4_{0.6}$ | $13.6_{0.2}$ | $65.2_{0.3}$ | $64.3_{0.1}$ | 25 | $38.0_{0.2}$ | $8.9_{0.7}$ | $21.1_{0.4}$ | $71.5_{0.6}$ | $68.6_{0.5}$ | 720 |
| Qwen3-4B | 4B | $54.3_{0.1}$ | $1.1_{0.0}$ | $35.7_{0.2}$ | $84.8_{0.0}$ | $78.3_{0.1}$ | 25 | $58.7_{0.3}$ | $2.1_{0.5}$ | $41.0_{0.4}$ | $87.0_{0.0}$ | $81.0_{0.1}$ | 449 |
| Qwen3-8B | 8B | $60.6_{0.1}$ | $0.3_{0.1}$ | $42.1_{0.1}$ | $88.1_{0.0}$ | $82.0_{0.1}$ | 25 | $64.5_{0.2}$ | $1.0_{0.1}$ | $48.7_{0.4}$ | $89.5_{0.1}$ | $84.0_{0.2}$ | 413 |
| Qwen3-14B | 14B | $69.8_{0.0}$ | $0.9_{0.1}$ | $54.6_{0.1}$ | $91.5_{0.0}$ | $87.1_{0.0}$ | 25 | $70.6_{0.4}$ | $0.7_{0.0}$ | $55.5_{0.7}$ | $91.7_{0.1}$ | $87.2_{0.2}$ | 451 |
| Qwen3-32B | 32B | $74.6_{0.1}$ | $1.4_{0.1}$ | $60.4_{0.1}$ | $92.9_{0.2}$ | $89.3_{0.1}$ | 25 | $74.0_{0.3}$ | $1.5_{0.1}$ | $59.6_{0.4}$ | $92.8_{0.1}$ | $88.8_{0.2}$ | 445 |
| DS-r1-distill-Qwen | 32B | - | - | - | - | - | - | $67.7_{0.3}$ | $2.6_{0.9}$ | $51.4_{0.7}$ | $90.3_{0.5}$ | $85.8_{0.5}$ | 517 |
| DeepSeek-r1-0528 | 37/671B | - | - | - | - | - | - | $75.5_{0.0}$ | $0.0_{0.0}$ | $64.8_{0.0}$ | $91.1_{0.0}$ | $85.3_{0.0}$ | 1319 |
| GPT-5-mini (low) | Unk. | - | - | - | - | - | - | $75.4_{0.0}$ | $0.1_{0.1}$ | $61.1_{0.6}$ | $93.6_{0.1}$ | $89.4_{0.1}$ | 144 |
| GPT-5-mini (high) | Unk. | - | - | - | - | - | - | $78.7_{0.0}$ | $0.1_{0.0}$ | $66.3_{0.0}$ | $94.5_{0.0}$ | $91.4_{0.0}$ | 910 |

Table 34: Details on **Task (4) Sentence Ordering** on **AAN abstract** dataset. Metric abbreviations: Acc - accuracy; Lumr – length unmatch rate (proportion of outputs whose generated length differs from the original sequence); Pmr – perfect match rate; Skip – skip bigram match rate; Lcs – longest common subsequence.

| Model | Size | w/o reasoning | | | | | | w/ reasoning | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Acc | Lumr | Pmr | Skip | Lcs | tok. | Acc | Lumr | Pmr | Skip | Lcs | tok. |
| Qwen3-0.6B | 0.6B | $17.6_{0.6}$ | $29.5_{4.8}$ | $5.3_{0.8}$ | $46.4_{1.4}$ | $51.4_{0.6}$ | 27 | $19.2_{0.3}$ | $9.5_{1.5}$ | $8.3_{0.7}$ | $51.7_{0.4}$ | $52.9_{0.0}$ | 629 |
| Qwen3-1.7B | 1.7B | $26.4_{0.2}$ | $17.8_{0.1}$ | $12.0_{0.2}$ | $58.4_{0.2}$ | $59.5_{0.2}$ | 30 | $31.4_{0.6}$ | $10.4_{0.5}$ | $19.1_{0.5}$ | $67.6_{0.1}$ | $62.5_{0.5}$ | 829 |
| Qwen3-4B | 4B | $47.2_{0.1}$ | $0.4_{0.2}$ | $31.6_{0.1}$ | $82.2_{0.0}$ | $73.5_{0.1}$ | 27 | $47.7_{0.1}$ | $2.3_{0.3}$ | $32.4_{1.5}$ | $82.6_{0.2}$ | $74.5_{0.3}$ | 477 |
| Qwen3-8B | 8B | $52.3_{0.2}$ | $0.6_{0.0}$ | $35.9_{0.3}$ | $85.3_{0.0}$ | $77.2_{0.1}$ | 27 | $53.5_{0.3}$ | $1.2_{0.1}$ | $37.6_{0.5}$ | $85.5_{0.3}$ | $78.2_{0.1}$ | 458 |
| Qwen3-14B | 14B | $58.6_{0.2}$ | $0.8_{0.1}$ | $42.9_{0.2}$ | $88.2_{0.0}$ | $81.4_{0.1}$ | 27 | $59.1_{0.4}$ | $0.4_{0.4}$ | $43.4_{0.4}$ | $88.2_{0.1}$ | $81.5_{0.1}$ | 475 |
| Qwen3-32B | 32B | $63.8_{0.2}$ | $0.3_{0.0}$ | $48.2_{0.2}$ | $89.5_{0.0}$ | $84.1_{0.1}$ | 27 | $61.9_{0.2}$ | $0.2_{0.0}$ | $46.1_{0.3}$ | $89.1_{0.3}$ | $83.1_{0.2}$ | 450 |
| DS-r1-distill-Qwen | 32B | - | - | - | - | - | - | $55.4_{0.3}$ | $2.5_{0.7}$ | $38.8_{0.3}$ | $85.8_{0.2}$ | $79.2_{0.4}$ | 552 |
| DeepSeek-r1-0528 | 37/671B | - | - | - | - | - | - | $61.9_{0.0}$ | $0.0_{0.0}$ | $46.6_{0.0}$ | $86.5_{0.0}$ | $80.2_{0.0}$ | 1558 |
| GPT-5-mini (low) | Unk. | - | - | - | - | - | - | $65.6_{0.4}$ | $0.1_{0.1}$ | $49.3_{0.1}$ | $90.6_{0.1}$ | $84.6_{0.0}$ | 162 |
| GPT-5-mini (high) | Unk. | - | - | - | - | - | - | $69.8_{0.0}$ | $0.0_{0.0}$ | $55.0_{0.0}$ | $91.9_{0.0}$ | $86.9_{0.0}$ | 1568 |

Table 35: Details on **Task (4) Sentence Ordering** on **ArXiv abstract** dataset. Metric abbreviations: Acc - accuracy; Lumr – length unmatch rate (proportion of outputs whose generated length differs from the original sequence); Pmr – perfect match rate; Skip – skip bigram match rate; Lcs – longest common subsequence.

| Model | Size | w/o reasoning | | | | | | w/ reasoning | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Acc | Lumr | Pmr | Skip | Lcs | tok. | Acc | Lumr | Pmr | Skip | Lcs | tok. |
| Qwen3-0.6B | 0.6B | $15.5_{1.7}$ | $32.7_{4.7}$ | $1.7_{0.4}$ | $48.1_{2.2}$ | $51.0_{1.3}$ | 30 | $17.1_{0.4}$ | $10.4_{0.7}$ | $1.7_{0.4}$ | $52.5_{0.7}$ | $52.2_{0.2}$ | 677 |
| Qwen3-1.7B | 1.7B | $23.8_{0.2}$ | $21.0_{0.8}$ | $3.7_{0.0}$ | $61.6_{0.2}$ | $58.5_{0.3}$ | 32 | $30.8_{1.1}$ | $14.2_{0.7}$ | $11.0_{1.5}$ | $69.2_{1.3}$ | $64.6_{1.3}$ | 915 |
| Qwen3-4B | 4B | $48.5_{0.1}$ | $1.5_{0.1}$ | $21.4_{0.4}$ | $84.6_{0.0}$ | $76.2_{0.0}$ | 31 | $54.1_{0.9}$ | $2.8_{0.4}$ | $26.2_{0.9}$ | $86.5_{0.1}$ | $78.1_{0.4}$ | 508 |
| Qwen3-8B | 8B | $54.3_{0.1}$ | $0.8_{0.0}$ | $25.6_{0.1}$ | $87.1_{0.1}$ | $78.9_{0.0}$ | 31 | $60.0_{1.6}$ | $1.7_{1.2}$ | $33.5_{1.8}$ | $89.4_{0.4}$ | $81.8_{0.3}$ | 475 |
| Qwen3-14B | 14B | $66.3_{0.1}$ | $0.8_{0.0}$ | $42.2_{0.2}$ | $91.5_{0.1}$ | $85.6_{0.1}$ | 31 | $66.1_{0.1}$ | $0.3_{0.2}$ | $43.0_{1.3}$ | $91.8_{0.0}$ | $85.7_{0.4}$ | 548 |
| Qwen3-32B | 32B | $71.2_{0.0}$ | $0.7_{0.1}$ | $47.7_{0.3}$ | $92.5_{0.2}$ | $87.8_{0.1}$ | 31 | $68.6_{0.4}$ | $0.6_{0.2}$ | $44.2_{0.9}$ | $92.3_{0.1}$ | $86.9_{0.3}$ | 499 |
| DS-r1-distill-Qwen | 32B | - | - | - | - | - | - | $63.5_{1.0}$ | $3.2_{0.8}$ | $38.9_{0.4}$ | $90.0_{0.5}$ | $84.2_{0.2}$ | 599 |
| DeepSeek-r1-0528 | 37/671B | - | - | - | - | - | - | $74.9_{0.0}$ | $0.0_{0.0}$ | $56.8_{0.0}$ | $91.6_{0.0}$ | $86.2_{0.0}$ | 1494 |
| GPT-5-mini (low) | Unk. | - | - | - | - | - | - | $73.1_{0.1}$ | $0.0_{0.0}$ | $50.5_{0.4}$ | $94.1_{0.0}$ | $88.8_{0.0}$ | 180 |
| GPT-5-mini (high) | Unk. | - | - | - | - | - | - | $78.6_{0.0}$ | $0.0_{0.0}$ | $58.6_{0.0}$ | $95.2_{0.0}$ | $91.3_{0.0}$ | 1205 |

Table 36: Details on **Task (4) Sentence Ordering** on **Neurips abstract** dataset. Metric abbreviations: Acc - accuracy; Lumr – length unmatch rate (proportion of outputs whose generated length differs from the original sequence); Pmr – perfect match rate; Skip – skip bigram match rate; Lcs – longest common subsequence.

| | | w/o reasoning | | | | | | w/ reasoning | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Size | Acc | Lumr | Pmr | Skip | Lcs | tok. | Acc | Lumr | Pmr | Skip | Lcs | tok. |
| Qwen3-0.6B | 0.6B | $9.6_{0.6}$ | $41.6_{8.6}$ | $1.1_{0.1}$ | $43.5_{2.2}$ | $38.8_{0.9}$ | 62 | $10.7_{0.4}$ | $28.2_{0.2}$ | $4.3_{0.4}$ | $49.7_{0.1}$ | $41.0_{0.1}$ | 770 |
| Qwen3-1.7B | 1.7B | $12.7_{0.1}$ | $42.3_{0.9}$ | $1.9_{0.1}$ | $52.1_{0.0}$ | $44.7_{0.1}$ | 59 | $14.7_{0.2}$ | $33.2_{0.3}$ | $6.6_{0.5}$ | $57.3_{0.3}$ | $46.2_{0.7}$ | 1242 |
| Qwen3-4B | 4B | $23.3_{0.1}$ | $18.7_{0.1}$ | $11.1_{0.0}$ | $71.5_{0.1}$ | $57.5_{0.1}$ | 54 | $23.6_{0.0}$ | $19.7_{2.0}$ | $12.0_{0.4}$ | $72.0_{0.2}$ | $58.4_{0.3}$ | 610 |
| Qwen3-8B | 8B | $26.0_{0.1}$ | $11.7_{0.3}$ | $11.2_{0.1}$ | $74.8_{0.1}$ | $60.2_{0.2}$ | 65 | $26.4_{0.3}$ | $16.5_{2.4}$ | $12.0_{0.3}$ | $74.6_{0.2}$ | $61.1_{0.3}$ | 591 |
| Qwen3-14B | 14B | $30.9_{0.0}$ | $10.1_{0.6}$ | $15.6_{0.0}$ | $77.8_{0.1}$ | $65.7_{0.2}$ | 52 | $29.2_{0.5}$ | $12.1_{1.7}$ | $14.3_{0.1}$ | $76.7_{0.3}$ | $64.3_{0.1}$ | 616 |
| Qwen3-32B | 32B | $32.8_{0.3}$ | $9.3_{0.5}$ | $15.5_{0.1}$ | $78.5_{0.1}$ | $67.7_{0.1}$ | 52 | $30.2_{0.3}$ | $11.4_{0.1}$ | $14.3_{0.2}$ | $77.5_{0.2}$ | $65.6_{0.2}$ | 546 |
| DS-r1-distill-Qwen | 32B | - | - | - | - | - | - | $26.0_{0.5}$ | $23.7_{1.4}$ | $11.4_{0.4}$ | $72.9_{0.1}$ | $61.4_{0.2}$ | 729 |
| DeepSeek-r1-0528 | 37/671B | - | - | - | - | - | - | $28.5_{0.0}$ | $1.2_{0.0}$ | $15.7_{0.0}$ | $68.3_{0.0}$ | $53.8_{0.0}$ | 2488 |
| GPT-5-mini (low) | Unk. | - | - | - | - | - | - | $33.2_{0.4}$ | $3.2_{0.2}$ | $15.8_{0.1}$ | $79.7_{0.0}$ | $67.5_{0.0}$ | 310 |
| GPT-5-mini (high) | Unk. | - | - | - | - | - | - | $35.5_{0.0}$ | $0.1_{0.0}$ | $17.8_{0.0}$ | $81.2_{0.0}$ | $70.1_{0.0}$ | 3625 |

Table 37: Details on **Task (4) Sentence Ordering** on **NSF abstract** dataset. Metric abbreviations: Acc - accuracy; Lumr – length unmatch rate (proportion of outputs whose generated length differs from the original sequence); Pmr – perfect match rate; Skip – skip bigram match rate; Lcs – longest common subsequence.

| | | w/o reasoning | | | | | | w/ reasoning | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Size | Acc | Lumr | Pmr | Skip | Lcs | tok. | Acc | Lumr | Pmr | Skip | Lcs | tok. |
| Qwen3-0.6B | 0.6B | $22.8_{0.5}$ | $21.4_{8.3}$ | $1.2_{0.2}$ | $52.5_{3.4}$ | $57.9_{1.4}$ | 23 | $25.6_{0.1}$ | $7.7_{3.4}$ | $2.3_{0.4}$ | $56.6_{0.3}$ | $60.5_{0.2}$ | 487 |
| Qwen3-1.7B | 1.7B | $37.5_{0.2}$ | $0.0_{0.0}$ | $8.0_{0.3}$ | $68.8_{0.2}$ | $68.1_{0.1}$ | 24 | $44.7_{0.6}$ | $1.9_{0.1}$ | $17.1_{0.5}$ | $73.8_{0.3}$ | $74.0_{0.4}$ | 647 |
| Qwen3-4B | 4B | $63.6_{0.1}$ | $0.0_{0.0}$ | $35.1_{0.2}$ | $86.5_{0.1}$ | $83.5_{0.1}$ | 24 | $72.3_{0.6}$ | $0.1_{0.0}$ | $50.2_{1.0}$ | $90.0_{0.2}$ | $88.0_{0.4}$ | 481 |
| Qwen3-8B | 8B | $69.1_{0.0}$ | $0.0_{0.0}$ | $43.6_{0.1}$ | $89.3_{0.0}$ | $86.3_{0.0}$ | 24 | $78.4_{0.2}$ | $0.0_{0.1}$ | $59.9_{1.3}$ | $92.5_{0.1}$ | $90.6_{0.2}$ | 525 |
| Qwen3-14B | 14B | $79.5_{0.1}$ | $0.0_{0.1}$ | $61.5_{0.1}$ | $92.8_{0.1}$ | $91.1_{0.0}$ | 24 | $82.5_{0.5}$ | $0.0_{0.1}$ | $67.3_{0.7}$ | $94.0_{0.2}$ | $92.5_{0.1}$ | 367 |
| Qwen3-32B | 32B | $82.7_{0.1}$ | $0.1_{0.0}$ | $67.0_{0.3}$ | $94.2_{0.1}$ | $92.6_{0.2}$ | 24 | $84.4_{0.7}$ | $0.0_{0.1}$ | $70.6_{0.9}$ | $94.7_{0.3}$ | $93.3_{0.2}$ | 452 |
| DS-r1-distill-Qwen | 32B | - | - | - | - | - | - | $79.8_{1.0}$ | $0.1_{0.0}$ | $63.2_{1.5}$ | $92.9_{0.6}$ | $91.4_{0.5}$ | 529 |
| DeepSeek-r1-0528 | 37/671B | - | - | - | - | - | - | $88.2_{0.0}$ | $0.0_{0.0}$ | $77.1_{0.0}$ | $96.1_{0.0}$ | $94.9_{0.0}$ | 712 |
| GPT-5-mini (low) | Unk. | - | - | - | - | - | - | $83.6_{0.6}$ | $0.0_{0.0}$ | $68.3_{1.1}$ | $94.5_{0.1}$ | $92.8_{0.2}$ | 137 |
| GPT-5-mini (high) | Unk. | - | - | - | - | - | - | $87.3_{0.0}$ | $0.0_{0.0}$ | $75.7_{0.0}$ | $95.9_{0.0}$ | $94.7_{0.0}$ | 685 |

Table 38: Details on **Task (4) Sentence Ordering** on **ROC stories** dataset. Metric abbreviations: Acc - accuracy; Lumr – length unmatch rate (proportion of outputs whose generated length differs from the original sequence); Pmr – perfect match rate; Skip – skip bigram match rate; Lcs – longest common subsequence.

| | | w/o reasoning | | | | | | w/ reasoning | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Size | Acc | Lumr | Pmr | Skip | Lcs | tok. | Acc | Lumr | Pmr | Skip | Lcs | tok. |
| Qwen3-0.6B | 0.6B | $21.9_{0.8}$ | $24.6_{6.8}$ | $2.0_{0.5}$ | $47.9_{2.2}$ | $55.8_{1.1}$ | 24 | $24.6_{0.9}$ | $10.9_{1.6}$ | $5.1_{0.6}$ | $54.3_{1.4}$ | $57.8_{0.7}$ | 515 |
| Qwen3-1.7B | 1.7B | $28.3_{0.4}$ | $6.0_{0.1}$ | $3.9_{0.2}$ | $57.3_{0.2}$ | $61.5_{0.1}$ | 24 | $35.6_{1.0}$ | $2.8_{0.6}$ | $14.6_{0.7}$ | $66.4_{0.4}$ | $65.7_{0.7}$ | 699 |
| Qwen3-4B | 4B | $46.9_{0.2}$ | $0.8_{0.0}$ | $22.2_{0.2}$ | $76.4_{0.0}$ | $73.3_{0.1}$ | 24 | $49.6_{0.3}$ | $0.9_{0.2}$ | $24.8_{0.7}$ | $78.5_{0.2}$ | $75.2_{0.3}$ | 561 |
| Qwen3-8B | 8B | $50.0_{0.1}$ | $0.2_{0.1}$ | $25.9_{0.2}$ | $78.9_{0.1}$ | $75.4_{0.1}$ | 23 | $52.0_{0.7}$ | $0.3_{0.1}$ | $27.4_{1.1}$ | $80.2_{0.3}$ | $76.7_{0.3}$ | 630 |
| Qwen3-14B | 14B | $53.9_{0.1}$ | $0.0_{0.0}$ | $29.3_{0.3}$ | $81.2_{0.1}$ | $77.5_{0.1}$ | 23 | $55.1_{0.4}$ | $0.1_{0.2}$ | $31.2_{0.9}$ | $81.4_{0.6}$ | $78.1_{0.8}$ | 449 |
| Qwen3-32B | 32B | $54.0_{0.2}$ | $0.2_{0.2}$ | $29.7_{0.2}$ | $80.8_{0.0}$ | $77.6_{0.2}$ | 23 | $55.3_{0.6}$ | $0.5_{0.3}$ | $31.2_{0.9}$ | $81.8_{0.4}$ | $78.2_{0.2}$ | 491 |
| DS-r1-distill-Qwen | 32B | - | - | - | - | - | - | $53.3_{1.2}$ | $0.8_{0.8}$ | $29.0_{1.2}$ | $80.3_{0.4}$ | $76.8_{0.4}$ | 536 |
| DeepSeek-r1-0528 | 37/671B | - | - | - | - | - | - | $58.4_{0.0}$ | $0.1_{0.0}$ | $36.4_{0.0}$ | $82.4_{0.0}$ | $79.1_{0.0}$ | 1471 |
| GPT-5-mini (low) | Unk. | - | - | - | - | - | - | $54.6_{0.8}$ | $0.0_{0.0}$ | $31.5_{1.1}$ | $81.1_{0.3}$ | $77.9_{0.5}$ | 172 |
| GPT-5-mini (high) | Unk. | - | - | - | - | - | - | $58.2_{0.0}$ | $0.0_{0.0}$ | $35.0_{0.0}$ | $82.7_{0.0}$ | $79.7_{0.0}$ | 1073 |

Table 39: Details on **Task (4) Sentence Ordering** on **SIND** dataset. Metric abbreviations: Acc - accuracy; Lumr – length unmatch rate (proportion of outputs whose generated length differs from the original sequence); Pmr – perfect match rate; Skip – skip bigram match rate; Lcs – longest common subsequence.

| Model | Size | w/o reasoning | | | | | | w/ reasoning | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Acc | Lumr | Pmr | Skip | Lcs | tok. | Acc | Lumr | Pmr | Skip | Lcs | tok. |
| Qwen3-0.6B | 0.6B | $6.8_{0.3}$ | $53.7_{10.2}$ | $1.8_{0.6}$ | $38.2_{1.7}$ | $31.9_{1.5}$ | 114 | $8.5_{0.1}$ | $45.2_{1.2}$ | $4.2_{0.8}$ | $43.9_{0.8}$ | $34.4_{0.5}$ | 941 |
| Qwen3-1.7B | 1.7B | $12.1_{0.1}$ | $36.6_{1.0}$ | $5.2_{0.1}$ | $55.4_{0.1}$ | $41.8_{0.1}$ | 73 | $12.7_{0.2}$ | $47.7_{1.8}$ | $8.9_{0.3}$ | $52.9_{0.3}$ | $40.4_{0.3}$ | 1313 |
| Qwen3-4B | 4B | $20.2_{0.1}$ | $26.3_{0.7}$ | $12.8_{0.1}$ | $72.7_{0.1}$ | $53.3_{0.1}$ | 74 | $21.2_{0.3}$ | $33.7_{1.7}$ | $14.1_{0.9}$ | $71.9_{0.1}$ | $54.5_{0.4}$ | 976 |
| Qwen3-8B | 8B | $20.8_{0.1}$ | $21.1_{0.9}$ | $14.8_{0.2}$ | $72.8_{0.3}$ | $53.7_{0.3}$ | 139 | $23.2_{0.4}$ | $31.5_{0.7}$ | $15.9_{0.4}$ | $75.1_{0.1}$ | $58.2_{0.2}$ | 902 |
| Qwen3-14B | 14B | $27.2_{0.1}$ | $23.5_{1.0}$ | $17.4_{0.1}$ | $80.5_{0.1}$ | $63.1_{0.1}$ | 72 | $27.1_{0.2}$ | $29.5_{0.6}$ | $18.2_{0.2}$ | $79.1_{0.2}$ | $63.7_{0.1}$ | 608 |
| Qwen3-32B | 32B | $30.4_{0.3}$ | $15.2_{0.6}$ | $19.3_{0.3}$ | $82.7_{0.0}$ | $65.9_{0.1}$ | 73 | $29.1_{0.6}$ | $21.5_{2.4}$ | $19.1_{0.5}$ | $80.9_{0.4}$ | $65.2_{0.3}$ | 545 |
| DS-r1-distill-Qwen | 32B | - | - | - | - | - | - | $24.3_{0.6}$ | $43.1_{5.5}$ | $17.6_{0.8}$ | $74.0_{1.1}$ | $60.2_{0.7}$ | 886 |
| DeepSeek-r1-0528 | 37/671B | - | - | - | - | - | - | $28.0_{0.0}$ | $1.4_{0.0}$ | $23.8_{0.0}$ | $58.9_{0.0}$ | $42.0_{0.0}$ | 2771 |
| GPT-5-mini (low) | Unk. | - | - | - | - | - | - | $32.0_{0.2}$ | $8.9_{0.3}$ | $20.3_{0.1}$ | $84.1_{0.0}$ | $67.7_{0.0}$ | 701 |
| GPT-5-mini (high) | Unk. | - | - | - | - | - | - | $40.1_{0.0}$ | $0.7_{0.0}$ | $23.0_{0.0}$ | $87.9_{0.0}$ | $74.4_{0.0}$ | 3762 |

Table 40: Details on **Task (4) Sentence Ordering** on **Wikipedia movie plots** dataset. Metric abbreviations: Acc - accuracy; Lumr – length unmatch rate (proportion of outputs whose generated length differs from the original sequence); Pmr – perfect match rate; Skip – skip bigram match rate; Lcs – longest common subsequence.

## E.5 Task (5) Dialogue Discourse Parsing

Tables 41 to 43.

| | | single turn | | | | | | auto-regressive | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | w/o reasoning | | | w/ reasoning | | | w/o reasoning | | | w/ reasoning | | |
| Model | Size | Link | Full | tok. | Link | Full | tok. | Link | Full | tok. | Link | Full | tok. |
| Qwen3-1.7B | 1.7B | $17.5_{0.5}$ | $2.5_{0.1}$ | 12 | $39.7_{0.6}$ | $8.4_{0.9}$ | 944 | $16.0_{1.0}$ | $0.4_{0.0}$ | 9 | $35.9_{2.5}$ | $5.2_{0.2}$ | 762 |
| Qwen3-4B | 4B | $42.8_{0.2}$ | $8.0_{0.2}$ | 9 | $56.9_{1.0}$ | $22.0_{0.4}$ | 1020 | $34.8_{0.7}$ | $4.6_{0.3}$ | 9 | $58.8_{0.7}$ | $19.6_{0.8}$ | 849 |
| Qwen3-8B | 8B | $52.4_{0.2}$ | $12.4_{0.3}$ | 9 | $59.2_{0.3}$ | $21.9_{0.5}$ | 942 | $51.6_{0.9}$ | $8.9_{1.1}$ | 9 | $58.8_{0.1}$ | $17.2_{0.9}$ | 754 |
| Qwen3-14B | 14B | $55.7_{0.2}$ | $16.2_{0.2}$ | 9 | $62.5_{1.3}$ | $33.1_{0.3}$ | 750 | $55.8_{0.3}$ | $13.2_{0.9}$ | 9 | $61.3_{0.5}$ | $29.8_{1.3}$ | 703 |
| Qwen3-32B | 32B | $58.1_{0.4}$ | $21.5_{1.4}$ | 9 | $62.1_{0.2}$ | $33.3_{0.4}$ | 611 | $57.8_{0.6}$ | $16.5_{1.3}$ | 9 | $62.4_{0.5}$ | $31.3_{0.3}$ | 583 |
| DS-r1-distill-Qwen | 32B | - | - | - | $62.9_{0.4}$ | $25.5_{0.6}$ | 430 | - | - | - | $61.5_{0.7}$ | $22.3_{1.5}$ | 429 |
| DeepSeek-r1-0528 | 37/671B | - | - | - | $68.0_{0.0}$ | $42.3_{0.0}$ | 1494 | - | - | - | $66.3_{0.0}$ | $38.7_{0.0}$ | 1526 |
| GPT-5-mini (low) | Unk. | - | - | - | $62.1_{0.4}$ | $33.6_{0.8}$ | 146 | - | - | - | $61.1_{0.3}$ | $32.1_{0.8}$ | 142 |
| GPT-5-mini (high) | Unk. | - | - | - | $67.0_{0.5}$ | $39.9_{0.0}$ | 1235 | - | - | - | $66.0_{0.0}$ | $38.8_{0.0}$ | 1254 |

Table 41: Details on **Task (5) Dialogue Discourse Parsing** on **STAC** dataset.

| | | single turn | | | | | | auto-regressive | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | w/o reasoning | | | w/ reasoning | | | w/o reasoning | | | w/ reasoning | | |
| Model | Size | Link | Full | tok. | Link | Full | tok. | Link | Full | tok. | Link | Full | tok. |
| Qwen3-1.7B | 1.7B | $14.1_{0.2}$ | $3.0_{0.1}$ | 17 | $39.5_{0.3}$ | $10.6_{0.5}$ | 907 | $20.7_{0.5}$ | $0.2_{0.0}$ | 8 | $43.1_{0.5}$ | $5.3_{0.4}$ | 728 |
| Qwen3-4B | 4B | $41.1_{0.5}$ | $7.6_{0.2}$ | 9 | $55.0_{0.3}$ | $23.3_{0.7}$ | 994 | $33.7_{0.7}$ | $4.9_{0.4}$ | 9 | $57.9_{0.3}$ | $22.6_{0.2}$ | 943 |
| Qwen3-8B | 8B | $48.5_{0.2}$ | $10.4_{0.4}$ | 8 | $58.8_{0.6}$ | $20.0_{0.2}$ | 1006 | $46.4_{0.8}$ | $7.4_{0.3}$ | 8 | $60.6_{0.1}$ | $16.8_{0.1}$ | 963 |
| Qwen3-14B | 14B | $58.0_{0.3}$ | $19.9_{0.4}$ | 9 | $59.6_{0.3}$ | $27.2_{0.4}$ | 743 | $61.4_{0.1}$ | $18.0_{0.3}$ | 9 | $60.1_{0.6}$ | $25.7_{0.8}$ | 743 |
| Qwen3-32B | 32B | $59.0_{0.3}$ | $18.4_{0.9}$ | 9 | $59.0_{0.1}$ | $24.7_{0.4}$ | 611 | $58.1_{0.6}$ | $13.7_{1.7}$ | 9 | $58.3_{0.5}$ | $22.2_{0.8}$ | 629 |
| DS-r1-distill-Qwen | 32B | - | - | - | $59.3_{0.4}$ | $17.5_{0.4}$ | 445 | - | - | - | $58.6_{0.6}$ | $16.1_{0.3}$ | 446 |
| DeepSeek-r1-0528 | 37/671B | - | - | - | $57.7_{0.0}$ | $23.8_{0.0}$ | 1499 | - | - | - | $56.7_{0.0}$ | $22.3_{0.0}$ | 1511 |
| GPT-5-mini (low) | Unk. | - | - | - | $57.3_{0.6}$ | $29.0_{0.1}$ | 157 | - | - | - | $55.6_{0.1}$ | $27.7_{0.3}$ | 160 |
| GPT-5-mini (high) | Unk. | - | - | - | $59.2_{0.0}$ | $31.5_{0.0}$ | 1272 | - | - | - | $58.4_{0.0}$ | $30.2_{0.0}$ | 1344 |

Table 42: Details on **Task (5) Dialogue Discourse Parsing** on **Molweni** dataset.

| | | single turn | | | | | | auto-regressive | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | w/o reasoning | | | w/ reasoning | | | w/o reasoning | | | w/ reasoning | | |
| Model | Size | Link | Full | tok. | Link | Full | tok. | Link | Full | tok. | Link | Full | tok. |
| Qwen3-1.7B | 1.7B | $4.4_{0.0}$ | $1.4_{0.0}$ | 16 | $29.4_{0.1}$ | $6.0_{0.3}$ | 926 | $2.5_{0.2}$ | $0.3_{0.0}$ | 149 | $36.7_{2.2}$ | $3.6_{0.1}$ | 747 |
| Qwen3-4B | 4B | $35.8_{0.2}$ | $11.7_{0.2}$ | 10 | $67.6_{0.5}$ | $24.0_{0.4}$ | 893 | $19.9_{0.9}$ | $5.7_{0.1}$ | 10 | $72.2_{0.1}$ | $17.2_{0.5}$ | 764 |
| Qwen3-8B | 8B | $57.9_{0.4}$ | $20.0_{0.2}$ | 11 | $72.1_{0.1}$ | $29.4_{0.5}$ | 832 | $56.1_{0.8}$ | $8.8_{0.2}$ | 11 | $73.7_{0.2}$ | $17.9_{0.3}$ | 657 |
| Qwen3-14B | 14B | $58.5_{0.2}$ | $24.4_{0.2}$ | 12 | $67.9_{0.2}$ | $34.5_{0.2}$ | 658 | $70.4_{0.2}$ | $15.1_{0.4}$ | 10 | $72.7_{0.2}$ | $22.6_{0.1}$ | 536 |
| Qwen3-32B | 32B | $65.0_{0.3}$ | $28.1_{0.3}$ | 12 | $67.9_{0.2}$ | $36.2_{0.4}$ | 551 | $68.3_{0.8}$ | $10.4_{0.7}$ | 10 | $70.5_{0.1}$ | $23.0_{0.3}$ | 491 |
| DS-r1-distill-Qwen | 32B | - | - | - | $68.9_{0.4}$ | $36.1_{0.4}$ | 415 | - | - | - | $69.6_{0.4}$ | $25.7_{0.7}$ | 412 |
| DeepSeek-r1-0528 | 37/671B | - | - | - | $69.4_{0.0}$ | $48.6_{0.0}$ | 1702 | - | - | - | $69.6_{0.0}$ | $34.2_{0.0}$ | 1893 |
| GPT-5-mini (low) | Unk. | - | - | - | $65.8_{0.0}$ | $37.8_{0.0}$ | 141 | - | - | - | $66.1_{0.1}$ | $26.8_{0.4}$ | 146 |
| GPT-5-mini (high) | Unk. | - | - | - | $69.1_{0.0}$ | $46.3_{0.0}$ | 1321 | - | - | - | $66.4_{0.0}$ | $31.9_{0.0}$ | 1472 |

Table 43: Details on **Task (5) Dialogue Discourse Parsing** on **MSDC** dataset.

# F Appendix F. Visualizations

## F.1 Task (1) Discourse Marker Understanding

Figures 8 to 10.

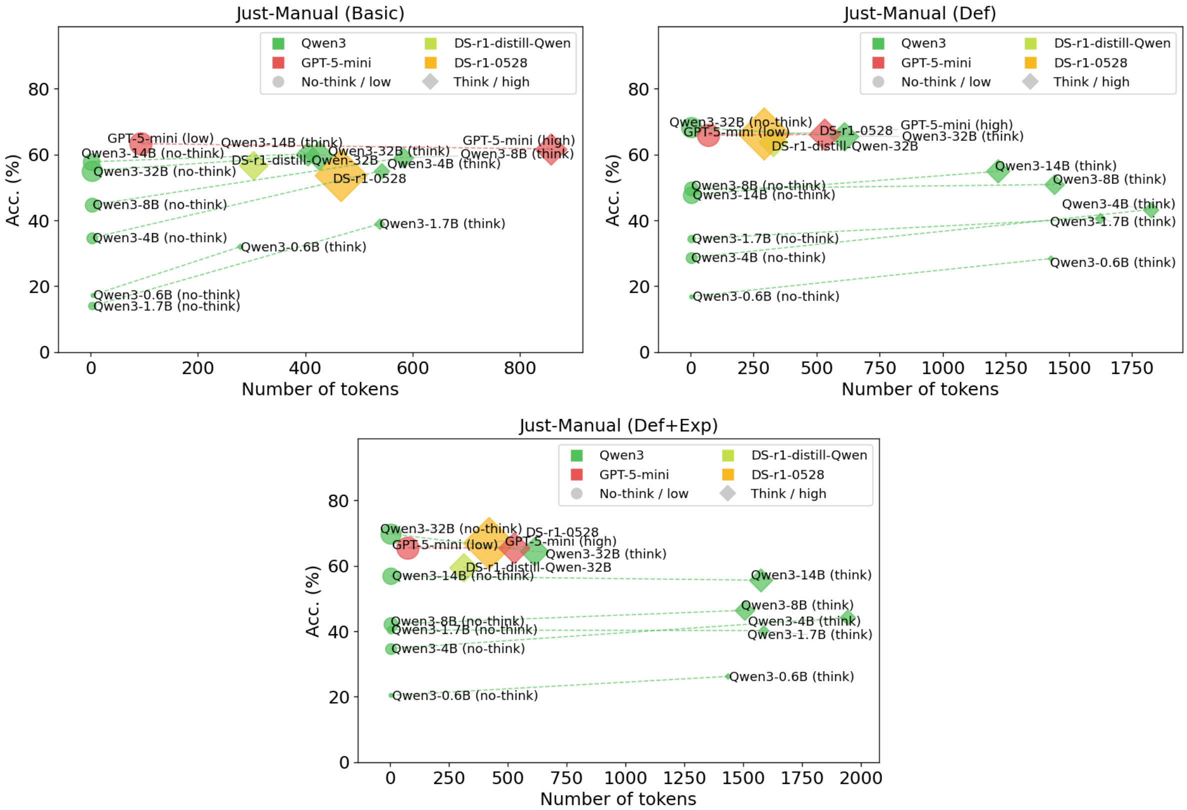


Figure 8: Performance comparison on **Task (1) Discourse Marker Understanding** on ***Just-Manual*** dataset with different **reasoning** mode or effort.

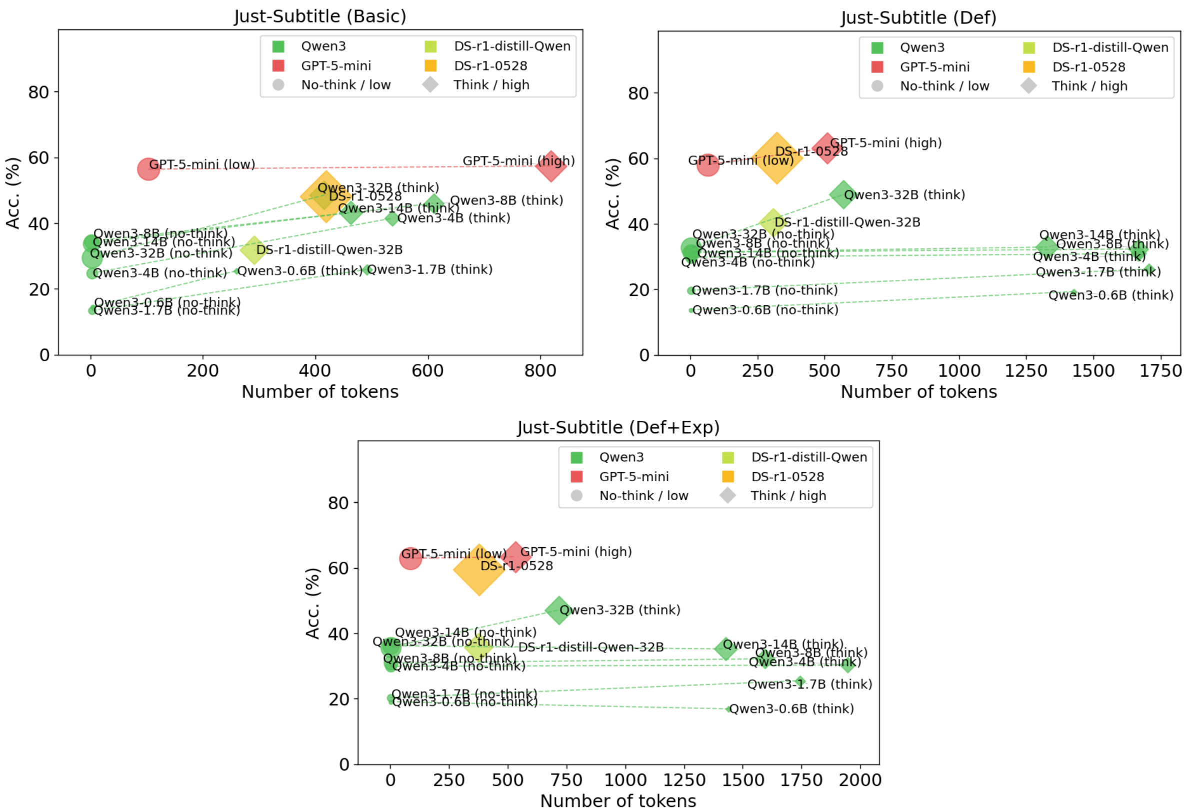


Figure 9: Performance comparison on **Task (1) Discourse Marker Understanding** on ***Just-Subtitle*** dataset with different **reasoning** mode or effort.

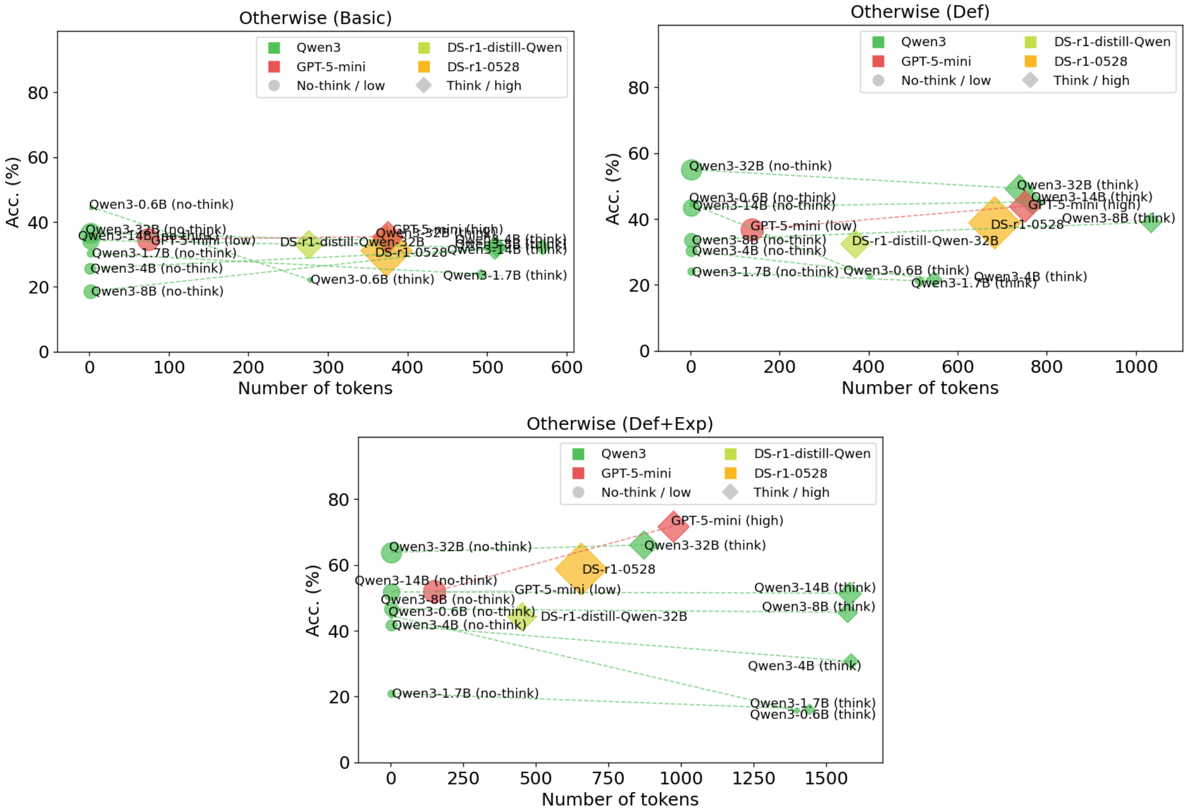


Figure 10: Performance comparison on **Task (1) Discourse Marker Understanding** on ***Otherwise*** dataset with different **reasoning** mode or effort.

### F.2 Task (2) Temporal Reasoning

Figures 11 to 13.

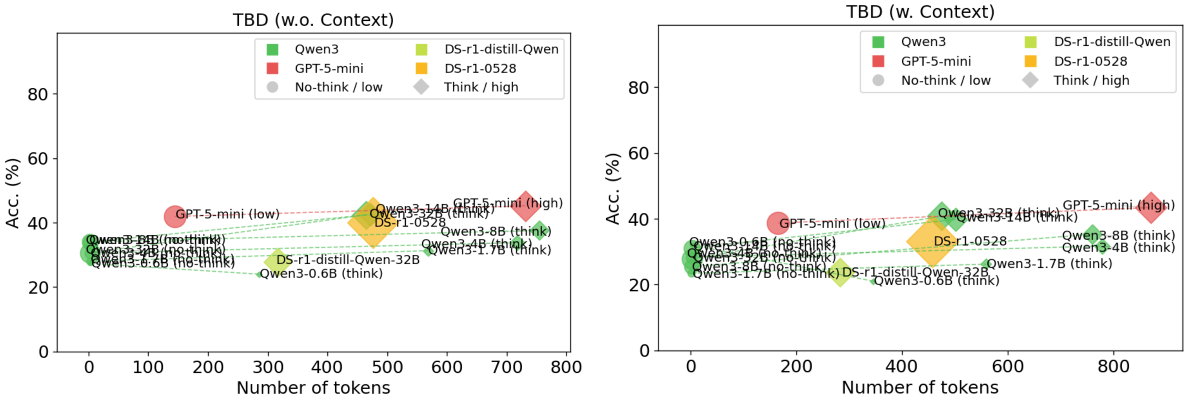


Figure 11: Performance comparison on **Task (2) Temporal Reasoning** on **TBD** dataset with different **reasoning** mode or effort.

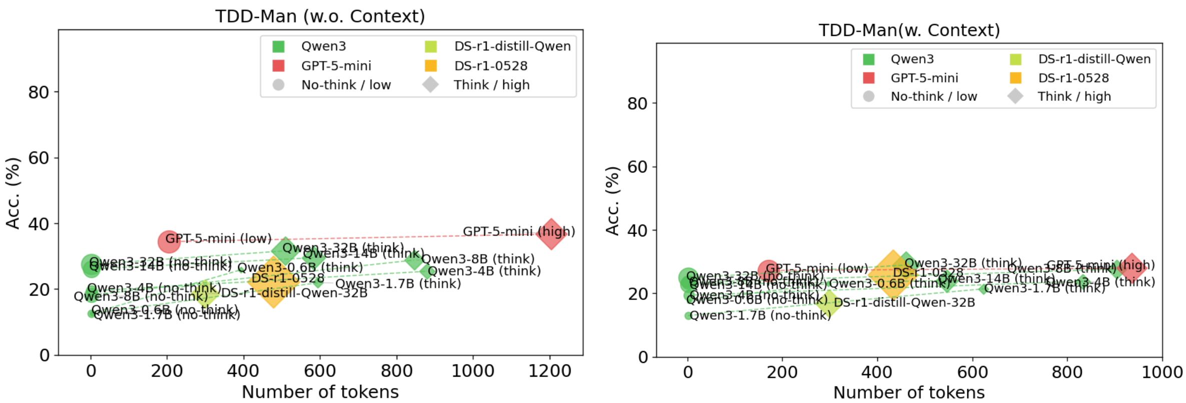


Figure 12: Performance comparison on **Task (2) Temporal Reasoning** on **TDD-Man** dataset with different **reasoning** mode or effort.

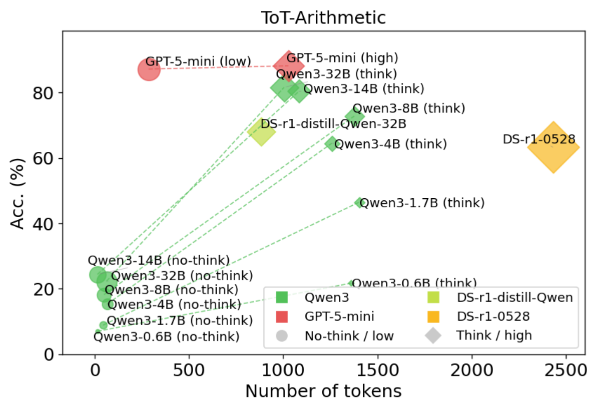


Figure 13: Performance comparison on **Task (2) Temporal Reasoning** on **ToT-Arithmetic** dataset with different **reasoning** mode or effort.

### F.3 Task (4) Sentence Ordering

Figures 14 to 17.

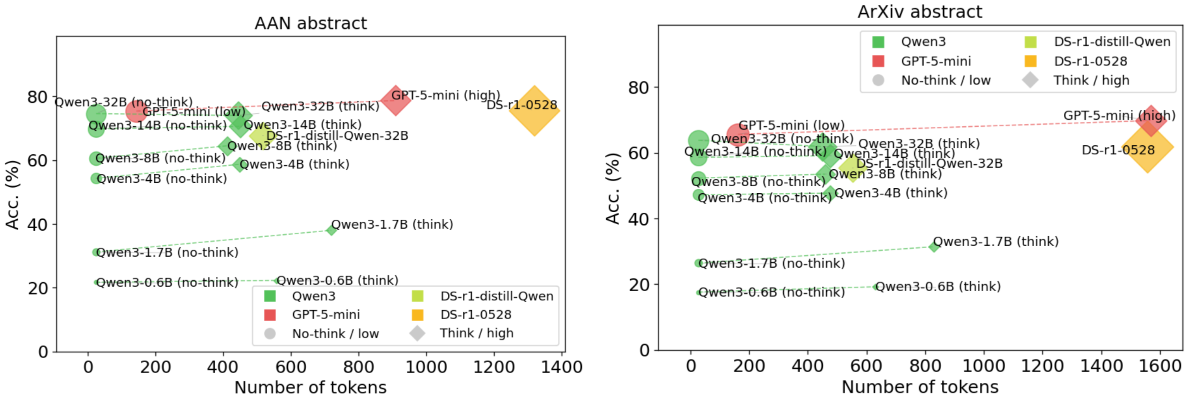


Figure 14: Performance comparison on **Task (4) Sentence Ordering** on **AAN abstract** and **ArXiv abstract** datasets with different **reasoning** mode or effort.

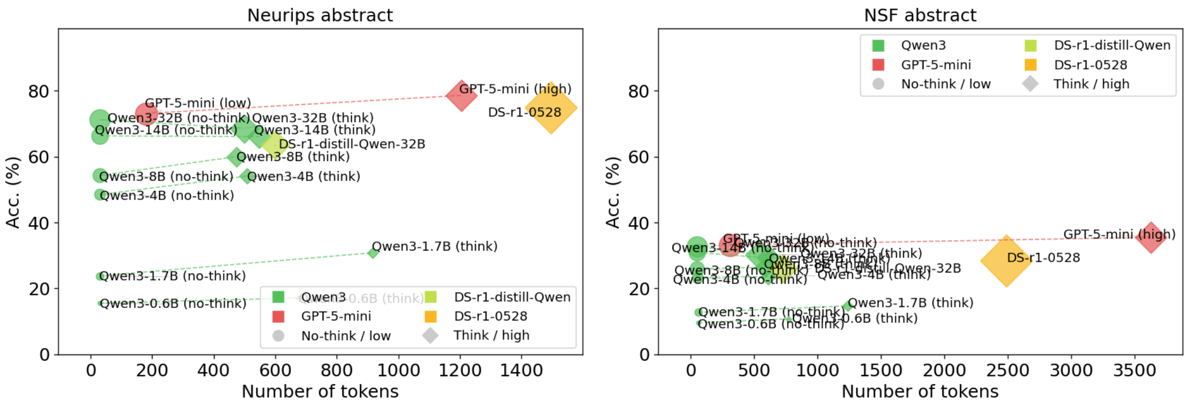


Figure 15: Performance comparison on **Task (4) Sentence Ordering** on **Neurips abstract** and **NSF abstract** datasets with different **reasoning** mode or effort.

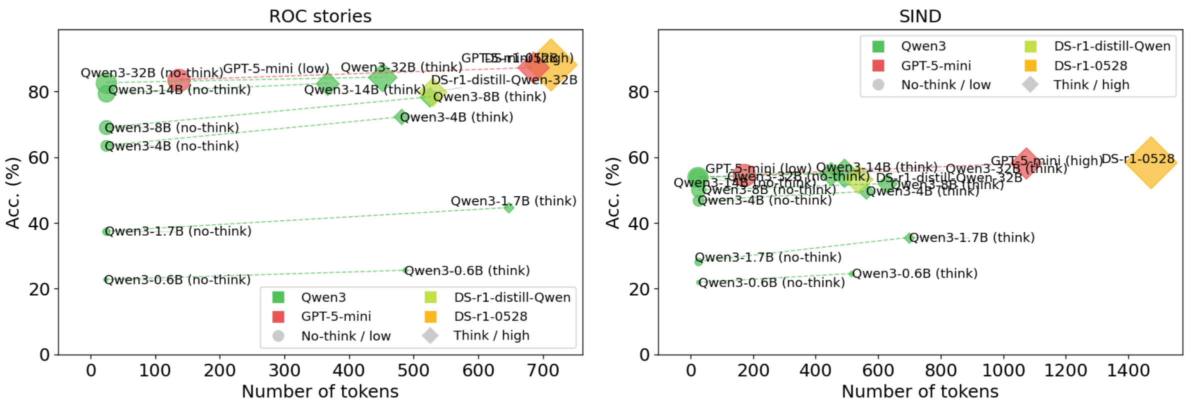


Figure 16: Performance comparison on **Task (4) Sentence Ordering** on **ROC stories** and **SIND** datasets with different **reasoning** mode or effort.

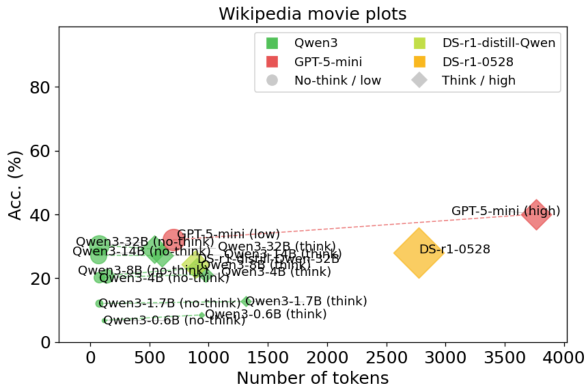


Figure 17: Performance comparison on **Task (4) Sentence Ordering** on **Wikipedia movie plots** dataset with different **reasoning** mode or effort.

### F.4 Task (5) Dialogue Discourse Parsing

Figure 18.

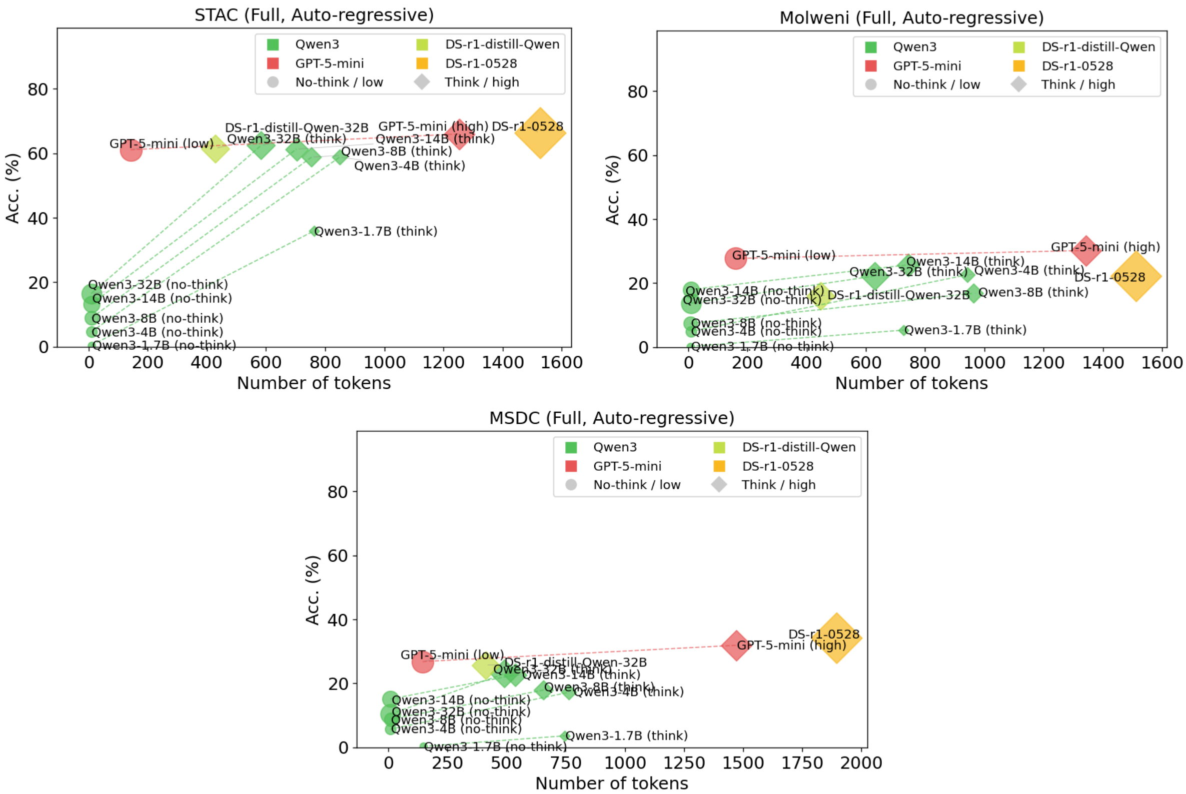


Figure 18: Performance comparison on **Task (4) Sentence Ordering** on **STAC**, **Molweni** and **MSDC** dataset with different **reasoning** mode or effort.