Exploring the Potential of ChatGPT on Sentence Level Relations: A Focus on Temporal, Causal, and Discourse Relations

Chunkit Chan[1], Cheng Jiayang[1], Weiqi Wang[1], Yuxin Jiang[2], Tianqing Fang[1], Xin Liu[1], Yangqiu Song[1]
[1]Department of Computer Science and Engineering, HKUST, Hong Kong SAR, China
[2]Information Hub, HKUST (GZ), Guangzhou, China
{ckchancc, yqsong}@cse.ust.hk

Abstract

This paper aims to quantitatively evaluate the performance of ChatGPT, an interactive large language model, on inter-sentential relations such as temporal relations, causal relations, and discourse relations. Given ChatGPT's promising performance across various tasks, we proceed to carry out thorough evaluations on the whole test sets of 11 datasets, including temporal and causal relations, PDTB2.0-based, and dialogue-based discourse relations. To ensure the reliability of our findings, we employ three tailored prompt templates for each task, including the zero-shot prompt template, zero-shot prompt engineering (**PE**) template, and in-context learning (**ICL**) prompt template, to establish the initial baseline scores for all popular sentence-pair relation classification tasks for the first time.[1] Through our study, we discover that ChatGPT exhibits exceptional proficiency in detecting and reasoning about causal relations, albeit it may not possess the same level of expertise in identifying the temporal order between two events. While it is capable of identifying the majority of discourse relations with existing explicit discourse connectives, the implicit discourse relation remains a formidable challenge. Concurrently, ChatGPT demonstrates subpar performance in the dialogue discourse parsing task that requires structural understanding in a dialogue before being aware of the discourse relation.

1 Introduction

With the proliferation of computational resources and the availability of extensive text corpora, the expeditious advancement of large language models (e.g., ChatGPT (OpenAI, 2022) and GPT-4 (OpenAI, 2023)) have prominently showcased their emergence ability resulting from the scaling up model size. Techniques such as instruction tuning (Wei et al., 2022) and reinforcement learning from human feedback (Ouyang et al., 2022) have further fortified LLM with sophisticated language understanding and logical reasoning proficiencies. Therefore, these large language models (LLMs) demonstrate remarkable few-shot, even zero-shot learning abilities in performing various tasks. Recent studies have extensively and comprehensively evaluated ChatGPT's performance on numerous language understanding and reasoning tasks, revealing that its superior performance in zero-shot scenarios when compared to other models (Bubeck et al., 2023; Bang et al., 2023; Jiao et al., 2023; Kocon et al., 2023). Besides, ChatGPT has also shown impressive powers in data annotations and has proven to be more cost-efficient than crowdworkers for several annotation tasks (Törnberg, 2023; Gilardi et al., 2023). Whilst the success of ChatGPT has been witnessed, certain obstacles persist unaddressed. Previous research has discussed the associated ethical implications and privacy concerns (Susnjak, 2022; Lukas et al., 2023; Li et al., 2023a,c). Moreover, ChatGPT's shortcomings include but are not limited to the lack of planning (Bubeck et al., 2023), the inability to perform complex mathematical reasoning (Frieder et al., 2023), and fact validation (Shahriar and Hayawi, 2023; Wang et al., 2023; Bang et al., 2023). Consequently, it is still under discussion whether large language models possess the capacity to comprehend text beyond surface forms as humans.

To comprehend the natural language text at a deeper level, it is crucial for an LLM to capture and understand the higher-level inter-sentential relations from the text, which involves mastering more complex and abstract relations beyond surface forms. These inter-sentential relations, such as temporal, causal, and discourse relations between two sentences, are widely used to form knowledge that has been proven to benefit many downstream tasks (Dai and Huang, 2019; Tang et al., 2021; Ravi et al., 2023; Su et al., 2023). In this study, we quan-

[1]The code and prompt template are available at https://github.com/HKUST-KnowComp/ChatGPT-Inter-Sentential-Relations.

Findings of the Association for Computational Linguistics: EACL 2024, pages 684–721
March 17-22, 2024

titatively evaluate the performance of ChatGPT in tasks that require an understanding of sentence-level relations, including temporal relation (Section 4), causal relation (Section 5), and discourse relation (Section 6). Under three standard prompt settings[2], we conduct extensive evaluations on the *whole* test sets of 11 datasets regarding these relations.[3] Furthermore, we conducted an in-depth study on the various intra-relations of each inter-sentential relation (e.g., *Before* and *After* relation in Temporal relations) and assessed the performance of the ChatGPT on these specific intra-relations. The detailed relation-wise performance is shown in Figure 1. The primary insights drawn from the analysis of quantitative assessments are as follows[4]:

- **Temporal relations:** ChatGPT **has difficulty** in identifying the temporal order between two events, which could be attributed to inadequate human feedback on this feature during the model's training process.
- **Causal relations:** ChatGPT **exhibits strong** performance in detecting and reasoning about causal relationships, particularly on the COPA dataset. It also outperforms fine-tuned RoBERTa on two out of three benchmarks.
- **Discourse relations:** Explicit discourse relations can **be easily recognized** by ChatGPT thanks to the explicit discourse connectives in context. However, it **struggles** with the absence of connectives for implicit discourse tasks, particularly with the link and relation prediction in dialogue discourse parsing.

We aspire to contribute to the research community through our evaluations and discoveries. By sharing the result, we intend to offer valuable insights to others in the relevant fields.

2 Related Work

Large Language Model With the increase of computational resources and available text corpora, the research community has discovered that large language models (LLMs) show an impressive ability in few-shot, even zero-shot learning with scaling up (Brown et al., 2020; Kaplan et al., 2020; Wei et al., 2022; Jiang et al., 2023). Besides, instruction tuning (Wei et al., 2022) and reinforcement learning from human feedback (Ouyang et al., 2022) also empower LLM with complicated language understanding and reasoning. Recently, ChatGPT (OpenAI, 2022) and GPT-4 (OpenAI, 2023) have achieved remarkable performance on a wide range of natural language processing benchmarks, including language modeling, machine translation, question answering, text completion, commonsense reasoning, and even human professional and academic exams. These achievements have garnered significant attention from academia and industry, and many efforts have been made to estimate the potential of artificial general intelligence (AGI) (Bang et al., 2023; Zhong et al., 2023b; Frieder et al., 2023; Davis, 2023; Yuan et al., 2023; Wang et al., 2024). It is crucial for the research community to continue exploring the capabilities of LLMs in various directions and tasks for further development of NLP.

Temporal Relation Temporal relation extraction aims to detect the temporal relation between two event triggers in the given document (Pustejovsky et al., 2003a). It is crucial for many downstream NLP tasks since reasoning over temporal relations plays an essential role in identifying the timing of events, estimating the duration of activities, and summarizing the chronological order of a series of occurrences (Ning et al., 2018b). There exists a recent work that evaluates ChatGPT's ability on zero-shot temporal relation extraction (Yuan et al., 2023). However, their manually designed prompts acquire unsatisfiable performance, and the capability of ChatGPT equipped with in-context learning has not been explored. Therefore, this work also includes the temporal relation tasks, and our results can complement and validate each other with Yuan et al. (2023).

Causal Relation Causal reasoning involves the identification of causality, which refers to the connection between a cause and its corresponding effect (Bochman, 2003). NLP models that can reason causally have the potential to improve their ability to understand language, as well as to solve complex problems in various fields, such as physical reasoning (Ates et al., 2022), event extraction (Cui et al.,

[2]Zero-shot prompting (denoted by **Prompt**), zero-shot prompt engineering (**PE**), and in-context learning (**ICL**). Prompt examples are shown in Appendix C.

[3]We exclude entailment or NLI tasks because they have already been evaluated in previous studies (Kocon et al., 2023; Zhong et al., 2023a).

[4]All evaluations were performed in April 2023 using the OpenAI API (*gpt-3.5-turbo-0301 model*), and similar performance was observed in the latest model (*"gpt-3.5-turbo-1106"*).

2022), question-answering (Zhang et al., 2022b; Sharp et al., 2016), and text classification (Choi et al., 2022). Although Tu et al. (2023) has analyzed ChatGPT's performance in a medical causality benchmark, no prior research has conducted a comprehensive study on the ability of large language models to reason upon causal relations.

Discourse Relation Discourse relation recognition is a vital task in discourse parsing, identifying the relations between two arguments (i.e., sentences or clauses) in the discourse structure. It is essential for textual coherence and is regarded as a critical step in constructing a knowledge graph (Zhang et al., 2020, 2022a) and various downstream tasks involving more context, such as text generation (Bosselut et al., 2018), text categorization (Liu et al., 2021b), and question answering (Jansen et al., 2014). Explicit discourse relation recognition (EDRR) has already shown that utilizing explicit connective information can effectively determine the types of discourse relations (Varia et al., 2019). In contrast, implicit discourse relation recognition (IDRR) remains challenging because of the absence of connectives. However, previous works have not systemically evaluated the ability of ChatGPT on these two discourse relation recognition tasks. Therefore, in this work, we assess the performance of this large language model (i.e., ChatGPT) on the PDTB-style discourse relation recognition task (Prasad et al., 2008), dialogue discourse parsing (Asher et al., 2016; Li et al., 2020), and downstream applications on discourse understanding.

3 Experimental Setting

We employ three customized prompt templates for each task: zero-shot setting, zero-shot with prompt engineering (**PE**), and the in-context learning (**ICL**) setting. The devised prompt template will serve as comprehensive and reliable baselines to exclude the variance of the prompt engineering and offer fair comparison baselines for all prevalent sentence-pair relation classification tasks. The specific template details are presented in corresponding sections and Appendix C.

- **ChatGPT$_{\text{Prompt}}$** refers to formulating the task as a multiple choice question answering problem and utilizing the prompt template in Robinson et al. (2022) as a baseline.

Method	TB-Dense	MATRES	TDDMan
Random	15.0	25.8	17.3
BERT-base	62.2	77.2	37.5
Fine-tuned SOTA	68.7	84.0	45.5
ChatGPT$_{\text{Prompt}}$	23.3	35.0	14.1
ChatGPT$_{\text{PE}}$	27.0	47.9	16.8
ChatGPT$_{\text{ICL}}$	25.0	44.9	14.7

Table 1: The Micor-F1 performance (%) of ChatGPT on temporal relation extraction.

- **ChatGPT$_{\text{Prompt Engineering}}$** refers to manually designing a more sophisticated prompt template based on the expert understanding of various tasks.

- **ChatGPT$_{\text{In-Context Learning}}$** refers to the in-context learning prompting method inspired by Brown et al. (2020). We manually select C input-output exemplars from the train split and reformulate these examples into our prompt-engineered template, where C is the number of classes. These well-selected examples for each category are distinguishable and easily understandable examples between each class.

4 Temporal Relation

Temporal relation extraction aims to determine the temporal order between two events in a text (Pustejovsky et al., 2003a), which could be formulated as a multi-label classification problem. In this section, we evaluate the temporal reasoning ability of ChatGPT on three commonly used benchmarks: TB-Dense (Cassidy et al., 2014), MATRES (Ning et al., 2018b), and TDDMan (Naik et al., 2019) (details in Appendix A). To ensure compatibility with previous research, we employ the same data split and assess ChatGPT's performance on the entire test set.

Detailed Experimental Setting. In comparison to random guess, the supervised baseline BERT-base (Mathur et al., 2021), and the supervised state-of-the-art model RSGT (Zhou et al., 2022b), we equip ChatGPT using three popular prompting strategies shown in Tables 13, 14, 15, 16, and 17 in Appendix C. For ChatGPT$_{\text{Prompt Engineering}}$, we manually design a more sophisticated prompt template to remind ChatGPT to first pay attention to the temporal order as well as the two events, which largely boosts its prediction performance.

Experimental Result. Table 1 presents the results of the experiment, where **ChatGPT lags behind fine-tuned models by more than 30% on all**

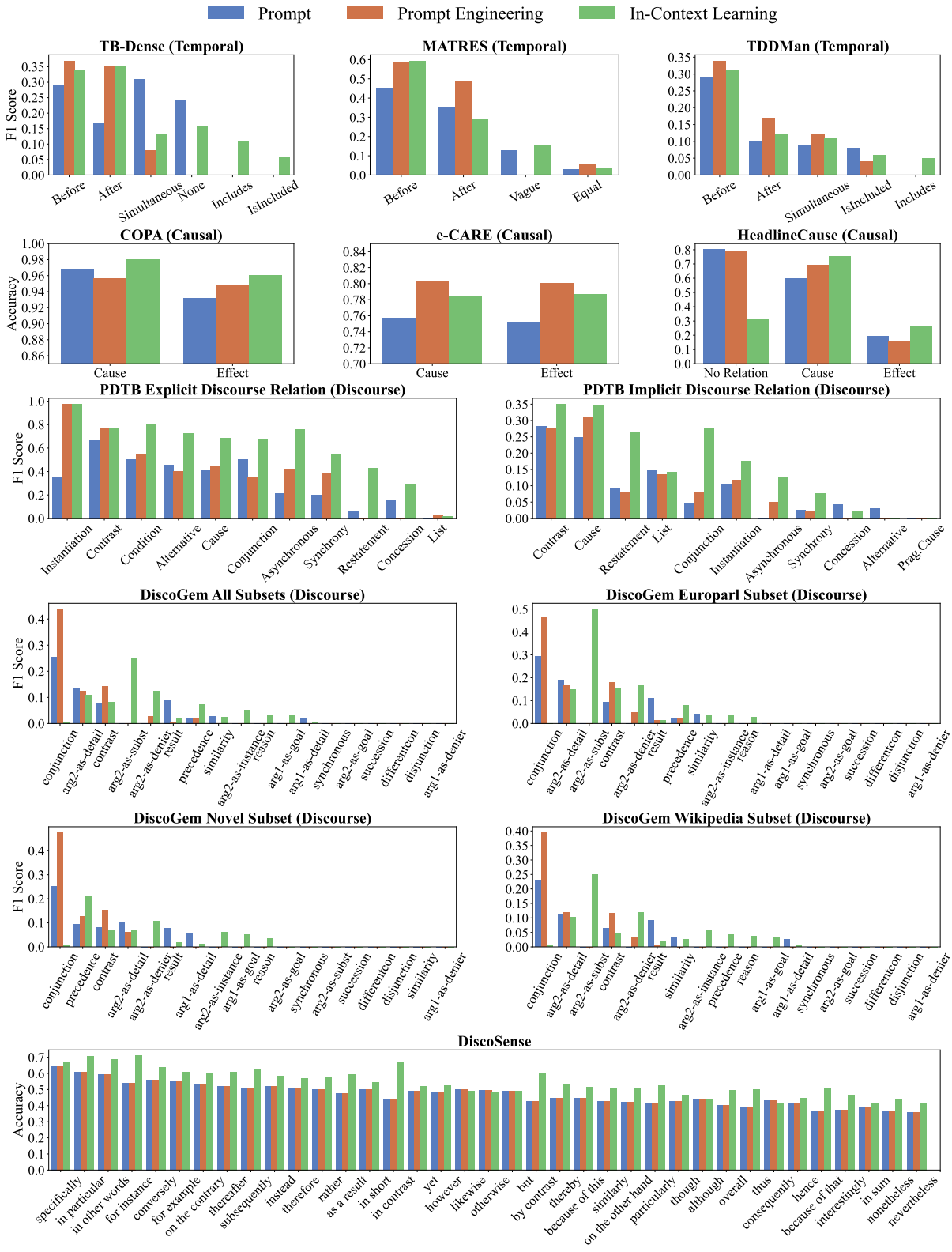

Figure 1: Relation-wise performance comparison on temporal, causal, and discourse benchmarks by ChatGPT with different prompting methods. DiscoSense is a downstream task of discourse relations.

three datasets. This suggests that ChatGPT may not be proficient in identifying the temporal order between two events, which could be attributed to inadequate human feedback on this feature during the model's training process. Additionally, our advanced prompt engineering delivers superior performance compared to the standard prompting baseline, with an improvement of 3.7%, 12.9%, and 2.7% on TB-Dense, MATRES, and TDDMan, respectively. Throughout our experiments, three significant observations emerged, which are worth noting:

(1) In temporal relation extraction tasks, ChatGPT's performance did not improve through in-context learning. The performance of in-context learning can be highly unstable across samples of examples, indicating that the process of language model acquiring information is idiosyncratic (Li and Qiu, 2023; Zhang et al., 2022c). A number of case studies are provided in Tables 13, 14, 15, 16, and 17 in Appendix C. These tables display test examples formulated into three templates using the aforementioned prompting strategies and subsequently fed to ChatGPT for response generation. The results indicate that only prompt engineering yields correct answers. We explored the underlying reasons by examining label-wise F1 performance, as illustrated in Figure 1. It appears that in-context learning enhances performance for more difficult-to-distinguish relations, such as *INCLUDES* and *IS_INCLUDED*, but negatively impacts performance for more easily distinguishable relations, like *BEFORE* and *AFTER*.

(2) ChatGPT exhibits a tendency to predict the temporal relation between $event_1$ and $event_2$ as *BEFORE*. This suggests a limited understanding of temporal order, given that the sequence of $event_1$ typically precedes $event_2$ within the text.

(3) In the context of long-dependency temporal relation extraction, ChatGPT is unsuccessful. As demonstrated in Table 1, ChatGPT, when equipped with all three prompting strategies, performs worse than random guessing on TDDMan. This dataset primarily focuses on long-document and discourse-level temporal relations, with an example provided in Tables 16 and 17 in Appendix C.

5 Causal Relation

Causal reasoning is the process of understanding and explaining the cause-and-effect relationships between events (Cao et al., 2021). It involves identifying the factors that contribute to a particular result and understanding how changes in those factors can lead to different outcomes (Ning et al., 2018a; Ponti et al., 2020). In this paper, we assess the causal reasoning ability of LLMs by benchmarking their results on three existing causal reasoning datasets (COPA (Gordon et al., 2012), e-CARE (Du et al., 2022), and HeadlineCause (Gusev and Tikhonov, 2022), details in Appendix A) and quantitatively analyzing the results. Our findings demonstrate that the LLM exhibits a robust ability to detect and reason about causal relationships, particularly those pertaining to cause and effect, without requiring advanced prompting techniques such as in-context learning.

Method	COPA	e-CARE	HeadlineCause
Random	50.0	50.0	20.0
Fine-tuned RoBERTa	90.6	70.7	73.5
Fine-tuned SOTA	100.0	74.6	83.5
$\text{ChatGPT}_{\text{Prompt}}$	94.8	74.8	71.4
$\text{ChatGPT}_{\text{PE}}$	95.2	79.6	72.7
$\text{ChatGPT}_{\text{ICL}}$	97.0	78.6	36.2

Table 2: Experiment results (Accuracy %) of fine-tuned RoBERTa and ChatGPT on causal reasoning benchmarks.

Detailed Experimental Setting. For the baseline, we report the accuracy of *random labeling* to reflect the character of each dataset and fine-tuned *RoBERTa* (Liu et al., 2019) to show the power of fine-tuned pre-trained language models. Accuracy is used as the evaluation metric to assess ChatGPT on three benchmarks using three different prompting techniques. The detailed prompts for three benchmarks are shown in Table 18, Table 19, and Table 20 in Appendix C, respectively. Table 2 presents the results of our experiments. For the $\text{ChatGPT}_{\text{Prompt Engineering}}$, we use more sophisticated prompt designs that emphasize the explanation of the question setting (what is the relationship between the given event and its options) and the causal relations.

Experimental Results. Notably, ChatGPT demonstrates exceptional performance on the COPA dataset and satisfactory performance on the other two datasets, outperforming fine-tuned RoBERTa on two out of three benchmarks and achieving comparable performance on the HeadlineCause dataset. Our engineered prompt improves performance slightly across all benchmarks, while in-context learning enhances ChatGPT's ability to excel only on the COPA dataset but has a

detrimental effect on the HeadlineCause dataset. To gain deeper insights, we conduct relation-wise comparisons of ChatGPT's performance on all three benchmarks, specifically examining its accuracy in identifying *cause* and *effect* relationships under different prompting techniques. The results are shown in Figure 1. Using the engineered prompt and in-context learning prompt tends to yield the best performance on the COPA and e-CARE datasets. However, for the HeadlineCause dataset, while in-context learning improves ChatGPT's ability to identify *cause* and *effect* relationships, it also makes it harder for the model to discriminate *no relation* entries.

In conclusion, our experiments demonstrate that **ChatGPT exhibits strong performance in detecting and reasoning about causal relationships, particularly those pertaining to cause and effect.** Our results also indicate that using engineered prompts and in-context learning can enhance ChatGPT's performance across various benchmarks, sometimes surpassing supervised baselines. However, the effectiveness of these techniques varies depending on the dataset. We hope this work can shed light on the strengths and limitations of ChatGPT in causal reasoning tasks and inform future research in this area.

6 Discourse Relation

In this section, we evaluate ChatGPT on Discourse Relation recognition tasks, including *PDTB-Style Discourse Relation Recognition*, *Multi-genre Crowd-sourced Discourse Relation Recognition*, *Dialogue Discourse Parsing*, and applications on discourse understanding. Apart from these datasets and tasks, we conduct the assessments of ChatGPT's performance on two downstream tasks which are shown in Appendix B.

6.1 PDTB-Style Discourse Relation Recognition

Detailed Experimental Setting. Explicit discourse relation recognition aims to recognize the discourse relation between two arguments, with the explicit discourse markers or connectives (e.g., "so", and "because") in between. In comparison, the implicit setting identifies the discourse relation without connectives. The labels of these two tasks for each discourse relation in the PDTB2.0 (Prasad et al., 2008) follow the hierarchical classification scheme throughout the annotation process, annotated as a hierarchy structure (shown in Figure 4 in Appendix). In this work, we evaluate ChatGPT's performance on PDTB 2.0 (Ji-setting (Ji and Eisenstein, 2015)), and the details are presented in Appendix A. The example of discourse relations in Figure 3 in Appendix A shows the *Contingency* top-level class and *Cause* second-level class. The details of three tailored prompt templates are provided in the Tables 21, 22, 23, and 24 in Appendix C.

Method	Top		Second	
	F1	Acc	F1	Acc
Random	25.12	25.70	7.30	9.19
Zhou et al. (2022a)	93.59	94.78	-	-
Varia et al. (2019)	95.48	96.20	-	-
Chan et al. (2023b)	95.64	96.73	-	-
$\text{ChatGPT}_{\text{Prompt}}$	34.94	39.38	31.92	43.26
$\text{ChatGPT}_{\text{PE}}$	69.26	70.21	39.34	50.80
$\text{ChatGPT}_{\text{ICL}}$	84.66	85.97	60.68	63.47

Table 3: The performance of ChatGPT performs on the explicit discourse relation recognition task of PDTB (*Ji*) test set.

For $\text{ChatGPT}_{\text{Prompt Engineering}}$, we manually designed a task-specified prompt as follows. Since the label of the PDTB2.0 dataset inherently forms the hierarchy, we utilized this label dependence to tailor a prompt template to predict the top-level class and second-level class simultaneously. Moreover, we select a representative connective for each discourse relation in the IDRR task, while the EDRR task already provides the explicit connectives for each instance. Therefore, we use the label dependence and the selected connectives to guide the LLM to understand the sense of each discourse relation.

6.1.1 Explicit Discourse Relation Recognition

Experimental Results. In Table 3, the performance shows that **ChatGPT can recognize each explicit discourse relation by utilizing the information from the explicit discourse connectives.** Furthermore, by utilizing the label dependence between the top-level label and the second-level label to design the prompt template, the performance of the top-level class increases significantly. With the prompt engineering template, as shown in Figure 1, ChatGPT does well on the *Contrast*, *Condition*, and *Instantiation* second-level class. Appending the input-output example from each discourse relation as the prefix part of the prompt template helps solve this task easily. Finally, the performance of ChatGPT on all second-level classes increases significantly except the *Exp.List* subclass.

Method	Top		Second	
	F1	Acc	F1	Acc
Random	24.74	25.47	6.48	8.78
Liu et al. (2020)	63.39	69.06	35.25	58.13
Jiang et al. (2022)	65.76	72.52	41.74	61.16
Long and Webber (2022)	69.60	72.18	49.66	61.69
Chan et al. (2023b)	70.84	75.65	49.03	64.58
ChatGPT$_{\text{Prompt}}$	29.85	32.89	9.27	15.59
ChatGPT$_{\text{PE}}$	33.78	34.94	10.73	20.31
ChatGPT$_{\text{ICL}}$	36.11	44.18	16.20	24.54

Table 4: The performance of ChatGPT performs on the implicit discourse relation recognition task of PDTB (*Ji*) test set.

Method	All		Europarl		Novel		Wiki.	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Random	5.5	3.2	5.5	3.2	5.8	3.1	5.6	3.2
(Liu et al., 2020)	48.7	22.3	53.3	25.9	45.3	23.1	45.6	24.0
ChatGPT$_{\text{Prompt}}$	10.8	3.5	13.7	4.2	9.9	3.7	9.4	3.1
ChatGPT$_{\text{PE}}$	20.8	4.2	21.6	5.0	25.3	4.8	17.7	3.7
ChatGPT$_{\text{ICL-1}}$	3.7	4.5	4.8	6.5	3.1	3.5	3.4	4.2
ChatGPT$_{\text{ICL-3}}$	3.3	2.8	3.1	2.4	4.3	4.2	2.9	2.5
ChatGPT$_{\text{ICL-18}}$	2.0	2.1	1.2	2.9	3.1	1.7	1.9	2.0

Table 5: Evaluation results (accuracy and Macro-averaged F1 score %) on the DiscoGeM dataset. In addition to the performance on the full test set ("All"), we also report the genre-wise performance on different sub-sets ("Europarl", "Novel", and "Wiki.").

6.1.2 Implicit Discourse Relation Recognition

Experimental Results. The performance in Table 4 demonstrates that **implicit discourse relation remains a challenging task for ChatGPT.** Even when using the information of label dependence and representative discourse connectives in the in-context learning setting, ChatGPT only achieves 24.54% test accuracy and 16.20% F1 score on the 11 second-level class of discourse relations. In particular, ChatGPT performs poorly on the second-level classes such as *Comp.Concession*, *Cont.Pragmatic Cause*, *Exp.Alternative*, and *Temp.Synchrony*. This may be because ChatGPT cannot understand the abstract sense of each discourse relation and the features from the text. When ChatGPT cannot capture the label sense and linguistic traits, it sometimes responds, "There doesn't appear to be a clear discourse relation between Argument 1 and Argument 2." or predicts as *Cont.Cause* class.

6.2 Multi-genre Crowd-sourced Discourse Relation Recognition

Detailed Experimental Setting. In this section, we evaluate the model on DiscoGeM (Scholman et al., 2022), which is a multi-genre implicit discourse relations dataset (details in Appendix A). For a fair and comprehensive evaluation, we test ChatGPT on the full test set containing 1,286 instances under the single label setting. To help ChatGPT understand the relations, we verbalize the relations in different settings[5]. In addition to the vanilla setting where the model directly predicts labels (ChatGPT$_{\text{Prompt}}$), we also replace relations that have special tokens or abbreviations with plain text, e.g. ("arg1-as-subst" is replaced with "argument 1 as substitution"). Under this setting (ChatGPT$_{\text{PE}}$), we concatenate the most typical connective[6] to ChatGPT$_{\text{Prompt}}$. We further explored in-context learning (ChatGPT$_{\text{ICL}}$): We randomly sample 1 or 3 examples from the training set as demonstrations (ChatGPT$_{\text{ICL-1}}$ and ChatGPT$_{\text{ICL-3}}$). Following the setting in Section 6.1.2, we manually curated a set of 18 typical examples from the training dataset for each relation as demonstrations (ChatGPT$_{\text{ICL-18}}$).

Experimental Results. Results are shown in Table 5. We report performance from both the random baseline and the model (Liu et al., 2020) fine-tuned on DiscoGeM (results reported in (Yung et al., 2022)). Generally, while ChatGPT slightly outperforms the random baseline, it lags behind the supervised model (Liu et al., 2020) by a significant margin (up to 30% accuracy and 20% macro-F1). Prompt engineering (ChatGPT$_{\text{PE}}$) could improve ChatGPT's performance, possibly due to the introduction of verbalization of labels that provided additional information for task understanding.

However, the introduction of different kinds of in-context learning templates (ChatGPT$_{\text{ICL}}$) did not have a positive influence on the model's ability to understand the task. In fact, the ChatGPT$_{\text{ICL}}$ model performed near-random or worse than random as the number of examples increased. This is possibly due to the fact that implicit discourse relations can express more than one meaning (Rohde et al., 2016; Scholman and Demberg, 2017), which makes it difficult to select representative and informative demonstrations. Overall, these findings suggest that it may require additional improvements or prompt engineering for ChatGPT to effectively perform tasks with complex classification requirements.

[5]We remove around 10 items with the "differentcon" relation as we do not find its explanation either in the paper or in the PDTB annotation guideline.

[6]https://github.com/merelscholman/DiscoGeM/blob/main/Appendix/DiscoGeM_ConnectiveMap.pdf

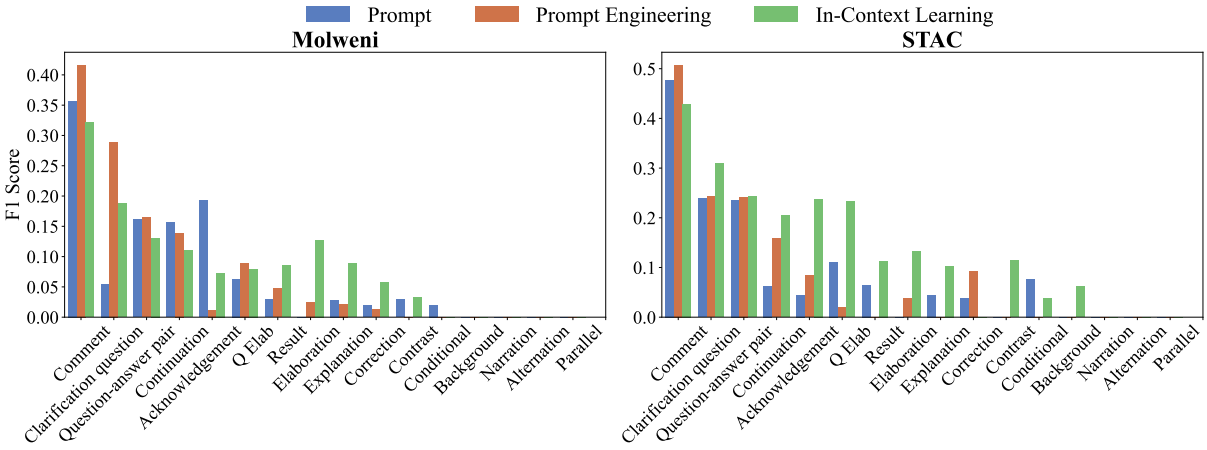

Figure 2: Relation-wise performance comparison on dialogue benchmarks by ChatGPT with different prompting methods.

Method	STAC		Molweni	
	Link	Link&Rel	Link	Link&Rel
Afantenos et al. (2015)	68.8	50.4	-	-
Perret et al. (2016)	68.6	52.1	-	-
Shi and Huang (2019)	73.2	55.7	78.1	54.8
ChatGPT$_{\text{zero w/ desc.}}$	20.5	4.3	26.7	5.0
ChatGPT$_{\text{zero w/o desc.}}$	20.0	4.4	28.3	5.4
ChatGPT$_{\text{few (n=1) w/ desc.}}$	21.0	7.1	25.7	6.0
ChatGPT$_{\text{few (n=3) w/ desc.}}$	20.7	7.3	25.1	5.7
ChatGPT$_{\text{few (n=1) w/o desc.}}$	21.2	6.2	27.2	6.8
ChatGPT$_{\text{few (n=3) w/o desc.}}$	21.3	7.4	26.5	6.9

Table 6: Evaluation results (Micro-averaged F1 score %) on the multi-party dialogue parsing datasets STAC and Molweni. Both the zero- (ChatGPT$_{\text{zero}}$) and few-shot (ChatGPT$_{\text{few}}$) baselines are tested. Under each setting, there are two variants: whether to provide a description to the labels (w/ desc.) or not (w/o desc.). The label descriptions are from Asher et al. (2016).

6.3 Dialogue Discourse Parsing

The dialogue discourse parsing task (Asher et al., 2016; Shi and Huang, 2019) is proposed to evaluate the ability to understand and respond to multi-party conversations in a coherent and context-aware manner. It focuses on extracting meaningful information from dialogues. The goal of dialogue discourse parsing is to automatically identify the structural and semantic relationships among utterances, speakers, and topics in a conversation.

Detailed Experimental Setting. The setting of discourse parsing in multi-party dialogue can be formulated as follows. Given a multi-party chat dialogue $D = \{u_1, u_2, ..., u_n\}$ with n utterances (u_1 to u_n), a system is required to predict a graph $G(V, E, R)$, where V is the vertex set containing all the utterances, E is the predicted edge set between utterances, and R is the predicted discourse relation set. According to the content of outputs, there are three evaluation settings:

Method	STAC		Molweni	
	Acc	F1	Acc	F1
Random	6.2	4.8	6.3	4.1
ChatGPT$_{\text{Prompt}}$	22.8	8.7	16.5	6.9
ChatGPT$_{\text{PE}}$	25.9	8.6	23.0	7.6
ChatGPT$_{\text{ICL}}$	24.1	13.9	14.7	8.1

Table 7: Evaluation results (Accuracy and Macro-averaged F1 (%)) on the multi-party dialogue parsing datasets STAC and Molweni. Here, the ChatGPT$_{\text{Prompt}}$, ChatGPT$_{\text{PE}}$, and ChatGPT$_{\text{ICL}}$ correspond to ChatGPT$_{\text{zero w/o desc.}}$, ChatGPT$_{\text{zero w/ desc.}}$, and ChatGPT$_{\text{few (n=1) w/ desc.}}$, respectively. The relation-wise performance is visualized in Figure 2.

- **Link prediction:** Given D, predict the links between utterances (E). Under this setting, the types of relations are ignored, and we only evaluate whether links are correctly predicted or not.

- **Link & Relation prediction:** Given D, predict the links between utterances and classify the discourse relation for the predicted links (E and R). Here, a true prediction requires both correctly predicting the link and its type of relation.

- **Relation classification:** Apart from the above two link prediction settings, we additionally evaluate ChatGPT's relation classification ability. Here, the model is given D, and the ground truth links E, and is required to predict the corresponding relations R.

In this work, we evaluate ChatGPT's performance on two multi-party dialogue discourse parsing benchmarks: STAC (Asher et al., 2016) and Molweni (Li et al., 2020). Details are presented in Appendix A.

Experimental Results. The evaluation results on the "Link prediction" and "Link & Relation prediction" settings are presented in Table 6. ChatGPT performs significantly worse than the supervised baselines (Afantenos et al., 2015; Perret et al., 2016; Shi and Huang, 2019) on both the link prediction and the link & relation prediction settings. Notably, on the link prediction setting, ChatGPT underperforms other baselines by up to 50% F1. It fails to give potential relations between utterances, indicating its poor understanding of the structure of multi-party dialogues. Adding additional examples seems to improve ChatGPT's performance under the Link & Relation prediction setting. However, these examples could have an adverse effect on link prediction (e.g., on Molweni). We also noticed that adding label descriptions does not help ChatGPT understand the task setting. We present results under the "Relation classification" setting in Table 7. ChatGPT also does not achieve very high performance under this setting, which indicates the difficulty in understanding discourse relations in dialogues. To sum up, **ChatGPT still suffers from a poor understanding of the dialogue structures** in multi-party dialogues and providing appropriate classifications.

7 Conclusion and Future Work

In conclusion, this study thoroughly examines ChatGPT's ability to handle pair-wise temporal relations, causal relations, and discourse relations by assessing its performance on the complete test sets of over 11 datasets. The result exhibits that even though ChatGPT obtains impressive zero-shot performance across other various tasks, there is still a gap for ChatGPT to achieve excellent performance on temporal and discourse relations. Though there may be numerous other capabilities of ChatGPT that go unnoticed in this paper, future work should nonetheless investigate the capability of ChatGPT on more tasks (e.g., analogy relation between two sentences (Cheng et al., 2023)).

Limitation

Evaluation Metrics In this paper, we exclusively assess the performance of ChatGPT on well-used evaluation metrics such as accuracy and F1 score. Nevertheless, these metrics are nonlinear or discontinuous metrics, and a recent study has revealed that such metrics yield conspicuous emergent capabilities, whereas linear or continuous metrics result in smooth, continuous predictable changes in model performance (Schaeffer et al., 2023). We intend to incorporate this aspect in forthcoming research endeavors.

Empirical Conclusions In this study, we give comprehensive comparisons and discussions of ChatGPT and prompts. All the conclusions are proposed based upon empirical analysis of the performance of ChatGPT to academic benchmarks. In light of the rapid evolution of the field, we will update the latest opinions timely.

Ethics Statement

In this work, we conformed to accepted privacy practices and strictly followed the data usage policy. All evaluated dataset of this paper is publicly available, and this work is in the intended use. Since we do not introduce social and ethical bias into the model or amplify any bias from the data, we can foresee no direct social consequences or ethical issues. Moreover, this study mainly formulates these sentence-level relations tasks as multi-choice tasks and requires ChatGPT to generate the English letter (e.g., "A," "B," "C," and "D"). Therefore, we do not observe or anticipate any potential toxicity, biases, or privacy in the generated context from ChatGPT. Furthermore, we also try our best to reduce these potential risks to prevent generating toxicity, biases, or privacy text by manually tailored prompt templates. These prompt templates only instruct ChatGPT to select the answer without any explanation.

Acknowledgements

The authors of this paper were supported by the NSFC Fund (U20B2053) from the NSFC of China, the RIF (R6020-19 and R6021-20) and the GRF (16211520 and 16205322) from RGC of Hong Kong. We also thank the support from the UGC Research Matching Grants (RMGS20EG01-D, RMGS20CR11, RMGS20CR12, RMGS20EG19, RMGS20EG21, RMGS23CR05, RMGS23EG08).

References

Stergos D. Afantenos, Eric Kow, Nicholas Asher, and Jérémy Perret. 2015. Discourse parsing for multi-party chat dialogues. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 928–937. The Association for Computational Linguistics.

Nicholas Asher, Julie Hunter, Mathieu Morey, Farah Benamara, and Stergos D. 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 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA).

Tayfun Ates, Muhammed Samil Atesoglu, Cagatay Yigit, Ilker Kesen, Mert Kobas, Erkut Erdem, Aykut Erdem, Tilbe Göksun, and Deniz Yuret. 2022. CRAFT: A benchmark for causal reasoning about forces and interactions. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2602–2627. Association for Computational Linguistics.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *CoRR*, abs/2302.04023.

Prajjwal Bhargava and Vincent Ng. 2022. Discosense: Commonsense reasoning with discourse connectives. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 10295–10310. Association for Computational Linguistics.

Alexander Bochman. 2003. A logic for causal reasoning. In *IJCAI-03, Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, Acapulco, Mexico, August 9-15, 2003*, pages 141–146. Morgan Kaufmann.

Antoine Bosselut, Asli Celikyilmaz, Xiaodong He, Jianfeng Gao, Po-Sen Huang, and Yejin Choi. 2018. Discourse-aware neural rewards for coherent text generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 173–184. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Túlio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *CoRR*, abs/2303.12712.

Pengfei Cao, Xinyu Zuo, Yubo Chen, Kang Liu, Jun Zhao, Yuguang Chen, and Weihua Peng. 2021. Knowledge-enriched event causality identification via latent structure induction networks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4862–4872. Association for Computational Linguistics.

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, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers*, pages 501–506. The Association for Computer Linguistics.

Chunkit Chan and Tsz Ho Chan. 2023. Discourse-aware prompt for argument impact classification. In *Proceedings of the 15th International Conference on Machine Learning and Computing, ICMLC 2023, Zhuhai, China, February 17-20, 2023*, pages 165–171. ACM.

Chunkit Chan, Xin Liu, Tsz Ho Chan, Jiayang Cheng, Yangqiu Song, Ginny Y. Wong, and Simon See. 2023a. Self-consistent narrative prompts on abductive natural language inference. *CoRR*, abs/2309.08303.

Chunkit Chan, Xin Liu, Jiayang Cheng, Zihan Li, Yangqiu Song, Ginny Y. Wong, and Simon See. 2023b. Discoprompt: Path prediction prompt tuning for implicit discourse relation recognition. *CoRR*, abs/2305.03973.

Jiayang Cheng, Lin Qiu, Tsz Ho Chan, Tianqing Fang, Weiqi Wang, Chunkit Chan, Dongyu Ru, Qipeng Guo, Hongming Zhang, Yangqiu Song, Yue Zhang, and Zheng Zhang. 2023. Storyanalogy: Deriving story-level analogies from large language models to unlock analogical understanding. *CoRR*, abs/2310.12874.

Seungtaek Choi, Myeongho Jeong, Hojae Han, and Seung-won Hwang. 2022. C2L: causally contrastive

learning for robust text classification. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 10526–10534. AAAI Press.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Shiyao Cui, Jiawei Sheng, Xin Cong, Quangang Li, Tingwen Liu, and Jinqiao Shi. 2022. Event causality extraction with event argument correlations. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 2300–2312. International Committee on Computational Linguistics.

Zeyu Dai and Ruihong Huang. 2019. A regularization approach for incorporating event knowledge and coreference relations into neural discourse parsing. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 2974–2985. Association for Computational Linguistics.

Ernest Davis. 2023. Benchmarks for automated commonsense reasoning: A survey. *CoRR*, abs/2302.04752.

Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2022. e-care: a new dataset for exploring explainable causal reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 432–446. Association for Computational Linguistics.

Tianqing Fang, Quyet V. Do, Sehyun Choi, Weiqi Wang, and Yangqiu Song. 2023. Ckbp v2: An expert-annotated evaluation set for commonsense knowledge base population. *CoRR*, abs/2304.10392.

Tianqing Fang, Quyet V. Do, Hongming Zhang, Yangqiu Song, Ginny Y. Wong, and Simon See. 2022. Pseudoreasoner: Leveraging pseudo labels for commonsense knowledge base population. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 3379–3394. Association for Computational Linguistics.

Tianqing Fang, Weiqi Wang, Sehyun Choi, Shibo Hao, Hongming Zhang, Yangqiu Song, and Bin He. 2021. Benchmarking commonsense knowledge base population with an effective evaluation dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 8949–8964. Association for Computational Linguistics.

Simon Frieder, Luca Pinchetti, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Christian Petersen, Alexis Chevalier, and Julius Berner. 2023. Mathematical capabilities of chatgpt. *CoRR*, abs/2301.13867.

Mor Geva, Eric Malmi, Idan Szpektor, and Jonathan Berant. 2019. Discofuse: A large-scale dataset for discourse-based sentence fusion. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 3443–3455. Association for Computational Linguistics.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *CoRR*, abs/2303.15056.

Andrew S. Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. Semeval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *Proceedings of the 6th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2012, Montréal, Canada, June 7-8, 2012*, pages 394–398. The Association for Computer Linguistics.

Ilya Gusev and Alexey Tikhonov. 2022. Headlinecause: A dataset of news headlines for detecting causalities. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 6153–6161. European Language Resources Association.

Xu Han, Weilin Zhao, Ning Ding, Zhiyuan Liu, and Maosong Sun. 2021. PTR: prompt tuning with rules for text classification. *CoRR*, abs/2105.11259.

Peter Jansen, Mihai Surdeanu, and Peter Clark. 2014. Discourse complements lexical semantics for non-factoid answer reranking. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 977–986. The Association for Computer Linguistics.

Yangfeng Ji and Jacob Eisenstein. 2015. One vector is not enough: Entity-augmented distributed semantics for discourse relations. *Trans. Assoc. Comput. Linguistics*, 3:329–344.

Yuxin Jiang, Chunkit Chan, Mingyang Chen, and Wei Wang. 2023. Lion: Adversarial distillation of closed-source large language model. *CoRR*, abs/2305.12870.

Yuxin Jiang, Linhan Zhang, and Wei Wang. 2022. Global and local hierarchy-aware contrastive framework for implicit discourse relation recognition. *CoRR*, abs/2211.13873.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt A good translator? A preliminary study. *CoRR*, abs/2301.08745.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *CoRR*, abs/2001.08361.

Jan Kocon, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydlo, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, Anna Kocon, Bartlomiej Koptyra, Wiktoria Mieleszczenko-Kowszewicz, Piotr Milkowski, Marcin Oleksy, Maciej Piasecki, Lukasz Radlinski, Konrad Wojtasik, Stanislaw Wozniak, and Przemyslaw Kazienko. 2023. Chatgpt: Jack of all trades, master of none. *CoRR*, abs/2302.10724.

Haoran Li, Yulin Chen, Jinglong Luo, Yan Kang, Xiaojin Zhang, Qi Hu, Chunkit Chan, and Yangqiu Song. 2023a. Privacy in large language models: Attacks, defenses and future directions. *CoRR*, abs/2310.10383.

Haoran Li, Dadi Guo, Wei Fan, Mingshi Xu, Jie Huang, Fanpu Meng, and Yangqiu Song. 2023b. Multi-step jailbreaking privacy attacks on chatgpt. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 4138–4153. Association for Computational Linguistics.

Haoran Li, Dadi Guo, Donghao Li, Wei Fan, Qi Hu, Xin Liu, Chunkit Chan, Duanyi Yao, and Yangqiu Song. 2023c. P-bench: A multi-level privacy evaluation benchmark for language models. *CoRR*, abs/2311.04044.

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, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 2642–2652. International Committee on Computational Linguistics.

Xiaonan Li and Xipeng Qiu. 2023. Finding supporting examples for in-context learning. *CoRR*, abs/2302.13539.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021a. GPT understands, too. *CoRR*, abs/2103.10385.

Xin Liu, Jiefu Ou, Yangqiu Song, and Xin Jiang. 2020. On the importance of word and sentence representation learning in implicit discourse relation classification. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3830–3836. ijcai.org.

Xin Liu, Jiefu Ou, Yangqiu Song, and Xin Jiang. 2021b. Exploring discourse structures for argument impact classification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3958–3969. Association for Computational Linguistics.

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. *CoRR*, abs/1907.11692.

Wanqiu Long and Bonnie Webber. 2022. Facilitating contrastive learning of discourse relational senses by exploiting the hierarchy of sense relations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 10704–10716. Association for Computational Linguistics.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the SIGDIAL 2015 Conference, The 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2-4 September 2015, Prague, Czech Republic*, pages 285–294. The Association for Computer Linguistics.

Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella Béguelin. 2023. Analyzing leakage of personally identifiable information in language models. *CoRR*, abs/2302.00539.

Puneet Mathur, Rajiv Jain, Franck Dernoncourt, Vlad I. Morariu, Quan Hung Tran, and Dinesh Manocha. 2021. TIMERS: document-level temporal relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021*, pages 524–533. Association for Computational Linguistics.

Nasrin Mostafazadeh, Aditya Kalyanpur, Lori Moon, David W. Buchanan, Lauren Berkowitz, Or Biran, and Jennifer Chu-Carroll. 2020. GLUCOSE: generalized and contextualized story explanations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 4569–4586. Association for Computational Linguistics.

Aakanksha Naik, Luke Breitfeller, and Carolyn P. Rosé. 2019. Tddiscourse: A dataset for discourse-level temporal ordering of events. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue, SIGdial 2019, Stockholm, Sweden, September 11-13, 2019*, pages 239–249. Association for Computational Linguistics.

Qiang Ning, Zhili Feng, Hao Wu, and Dan Roth. 2018a. Joint reasoning for temporal and causal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2278–2288. Association for Computational Linguistics.

Qiang Ning, Hao Wu, and Dan Roth. 2018b. A multi-axis annotation scheme for event temporal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1318–1328. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

TB OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. *OpenAI*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *CoRR*, abs/2203.02155.

Jérémy Perret, Stergos D. Afantenos, Nicholas Asher, and Mathieu Morey. 2016. Integer linear programming for discourse parsing. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 99–109. The Association for Computational Linguistics.

Edoardo Maria Ponti, Goran Glavas, Olga Majewska, Qianchu Liu, Ivan Vulic, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 2362–2376. Association for Computational Linguistics.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K. Joshi, and Bonnie L. Webber. 2008. The penn discourse treebank 2.0. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*. European Language Resources Association.

James Pustejovsky, José M. Castaño, Robert Ingria, Roser Saurí, Robert J. Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R. Radev. 2003a. Timeml: Robust specification of event and temporal expressions in text. In *New Directions in Question Answering, Papers from 2003 AAAI Spring Symposium, Stanford University, Stanford, CA, USA*, pages 28–34. AAAI Press.

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. 2003b. The timebank corpus. In *Corpus linguistics*, volume 2003, page 40.

Sahithya Ravi, Chris Tanner, Raymond Ng, and Vered Shwartz. 2023. What happens before and after: Multi-event commonsense in event coreference resolution. *CoRR*, abs/2302.09715.

Joshua Robinson, Christopher Michael Rytting, and David Wingate. 2022. Leveraging large language models for multiple choice question answering. *CoRR*, abs/2210.12353.

Hannah Rohde, Anna Dickinson, Nathan Schneider, Christopher N. L. Clark, Annie Louis, and Bonnie L. 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@ACL 2016, August 11, 2016, Berlin, Germany*. The Association for Computer Linguistics.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. ATOMIC: an atlas of machine commonsense for if-then reasoning. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 3027–3035. AAAI Press.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are emergent abilities of large language models a mirage? *CoRR*, abs/2304.15004.

Merel C. J. Scholman and Vera Demberg. 2017. Examples and specifications that prove a point: Identifying elaborative and argumentative discourse relations. *Dialogue Discourse*, 8(2):56–83.

Merel C. J. Scholman, Tianai Dong, Frances Yung, and Vera Demberg. 2022. Discogem: A crowdsourced corpus of genre-mixed implicit discourse relations. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 3281–3290. European Language Resources Association.

Sakib Shahriar and Kadhim Hayawi. 2023. Let's have a chat! A conversation with chatgpt: Technology, applications, and limitations. *CoRR*, abs/2302.13817.

Rebecca Sharp, Mihai Surdeanu, Peter Jansen, Peter Clark, and Michael Hammond. 2016. Creating causal embeddings for question answering with minimal supervision. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 138–148. The Association for Computational Linguistics.

Zhouxing Shi and Minlie Huang. 2019. A deep sequential model for discourse parsing on multi-party dialogues. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7007–7014. AAAI Press.

Damien Sileo, Tim Van de Cruys, Camille Pradel, and Philippe Muller. 2019. Mining discourse markers for unsupervised sentence representation learning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 3477–3486. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press.

Hung-Ting Su, Yulei Niu, Xudong Lin, Winston H. Hsu, and Shih-Fu Chang. 2023. Language models are causal knowledge extractors for zero-shot video question answering. *CoRR*, abs/2304.03754.

Teo Susnjak. 2022. Chatgpt: The end of online exam integrity? *CoRR*, abs/2212.09292.

Jialong Tang, Hongyu Lin, Meng Liao, Yaojie Lu, Xianpei Han, Le Sun, Weijian Xie, and Jin Xu. 2021. From discourse to narrative: Knowledge projection for event relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 732–742. Association for Computational Linguistics.

Petter Törnberg. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *CoRR*, abs/2304.06588.

Ruibo Tu, Chao Ma, and Cheng Zhang. 2023. Causal-discovery performance of chatgpt in the context of neuropathic pain diagnosis. *CoRR*, abs/2301.13819.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James F. Allen, Marc Verhagen, and James Pustejovsky. 2013. Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations. In *Proceedings of the 7th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2013, Atlanta, Georgia, USA, June 14-15, 2013*, pages 1–9. The Association for Computer Linguistics.

Siddharth Varia, Christopher Hidey, and Tuhin Chakrabarty. 2019. Discourse relation prediction: Revisiting word pairs with convolutional networks. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue, SIGdial 2019, Stockholm, Sweden, September 11-13, 2019*, pages 442–452. Association for Computational Linguistics.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, et al. 2023. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *arXiv preprint arXiv:2310.07521*.

Weiqi Wang, Tianqing Fang, Chunyang Li, Haochen Shi, Wenxuan Ding, Baixuan Xu, Zhaowei Wang, Jiaxin Bai, Xin Liu, Jiayang Cheng, et al. 2024. Candle: Iterative conceptualization and instantiation distillation from large language models for commonsense reasoning. *arXiv preprint arXiv:2401.07286*.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Chenhan Yuan, Qianqian Xie, and Sophia Ananiadou. 2023. Zero-shot temporal relation extraction with chatgpt. *CoRR*, abs/2304.05454.

Frances Yung, Kaveri Anuranjana, Merel Scholman, and Vera Demberg. 2022. Label distributions help implicit discourse relation classification. In *Proceedings of the 3rd Workshop on Computational Approaches to Discourse*, pages 48–53. International Conference on Computational Linguistics.

Hongming Zhang, Xin Liu, Haojie Pan, Haowen Ke, Jiefu Ou, Tianqing Fang, and Yangqiu Song. 2022a. ASER: towards large-scale commonsense knowledge acquisition via higher-order selectional preference over eventualities. *Artif. Intell.*, 309:103740.

Hongming Zhang, Xin Liu, Haojie Pan, Yangqiu Song, and Cane Wing-Ki Leung. 2020. ASER: A large-scale eventuality knowledge graph. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 201–211. ACM / IW3C2.

Minhao Zhang, Ruoyu Zhang, Yanzeng Li, and Lei Zou. 2022b. Crake: Causal-enhanced table-filler for question answering over large scale knowledge base. In *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 1787–1798. Association for Computational Linguistics.

Yiming Zhang, Shi Feng, and Chenhao Tan. 2022c. Active example selection for in-context learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 9134–9148. Association for Computational Linguistics.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2023a. Can chatgpt understand too? A comparative study on chatgpt and fine-tuned BERT. *CoRR*, abs/2302.10198.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023b. Agieval: A human-centric benchmark for evaluating foundation models. *CoRR*, abs/2304.06364.

Zexuan Zhong, Dan Friedman, and Danqi Chen. 2021. Factual probing is [MASK]: learning vs. learning to recall. In *NAACL-HLT*, pages 5017–5033.

Hao Zhou, Man Lan, Yuanbin Wu, Yuefeng Chen, and Meirong Ma. 2022a. Prompt-based connective prediction method for fine-grained implicit discourse relation recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 3848–3858. Association for Computational Linguistics.

Jie Zhou, Shenpo Dong, Hongkui Tu, Xiaodong Wang, and Yong Dou. 2022b. RSGT: relational structure guided temporal relation extraction. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 2001–2010. International Committee on Computational Linguistics.

A Experimental Setting

A.1 Evaluation Dataset

TB-Dense. TB-Dense (Cassidy et al., 2014) is a densely annotated dataset from TimeBank and TempEval (UzZaman et al., 2013) that contains six label types, including *BEFORE*, *AFTER*, *SIMULTANEOUS*, *NONE*, *INCLUDES* and *IS_INCLUDED*.

MATRES. MATRES (Ning et al., 2018b) is an annotated dataset that includes refined annotations from TimeBank (Pustejovsky et al., 2003b), AQUAINT, and Platinum documents. Four relations are annotated for the start time comparison of event pairs in 275 documents, namely *BEFORE*, *AFTER*, *EQUAL*, and *VAGUE*. Note that the two relations named *EQUAL* and *VAGUE* are equivalent to *SIMULTANEOUS* and *NONE* in TB-Dense, respectively.

TDDMan. TDDMan is a subset of the TDDiscourse corpus (Naik et al., 2019), which was created to explicitly emphasize global discourse-level temporal ordering. Five temporal relations are annotated including *BEFORE*, *AFTER*, *SIMULTANEOUS*, *INCLUDES* and *IS_INCLUDED*.

COPA. The Choice of Plausible Alternatives (COPA) (Gordon et al., 2012) dataset is a collection of questions that require causality reasoning and inferences to solve. Each question posits a commonly seen event, along with two possible options that either describe the *cause* or *effect* of the event. This requires the model to identify the relationship between a cause and its effect and then select the most likely explanation for that relationship among a set of alternatives. Such design makes COPA a very representative benchmark for evaluating causal relational reasoning. In this paper, we use the testing split of COPA, consisting of 500 questions, for evaluation.

e-CARE. The e-CARE (Du et al., 2022) dataset is a large human-annotated commonsense causal reasoning benchmark that contains over 21,000 multiple-choice questions. It is designed to provide a conceptual understanding of causality and includes free-text-formed conceptual explanations for each causal question to explain why the causation exists. Each question either focuses on the *cause* or *effect* of a given event and consists of two possible explanations. The model is still asked to select the more plausible one, given an event-and-relationship pair. Since the testing set is not publicly available, we bank on 2,132 questions in the validation set for evaluating LLMs.

HeadlineCause. HeadlineCause (Gusev and Tikhonov, 2022) is a dataset designed for detecting implicit causal relations between pairs of news headlines. It includes over 5000 headline pairs from English news and over 9000 headline pairs from Russian news, labeled through crowdsourcing. Given a pair of news, the model is first asked to determine whether a causal relationship exists between them. If yes, it needs to further determine the role of cause and effect for the two news. It serves as a very challenging and comprehensive benchmark for evaluating models' capability to detect causal relations in natural language text. We select 542 English news pairs from the testing set that are used for evaluation.

The Penn Discourse Treebank 2.0 (PDTB 2.0). PDTB 2.0 is a large-scale corpus that comprises a vast collection of 2,312 articles from the Wall Street Journal (WSJ) (Prasad et al., 2008). It utilizes a lexically grounded approach to annotate discourse relations, with three sense levels (classes, types, and sub-types) naturally forming a natural sense hierarchy. In this dataset, we assess the performance of ChatGPT on a popular setting of the PDTB 2.0 dataset, known as the Ji-setting (Ji and Eisenstein, 2015). This Ji-setting follows Ji and Eisenstein (2015) to divide sections 2-20, 0-1, and 21-22 into training, validation, and test sets, respectively. We evaluate ChatGPT on the whole test set of IDRR task and EDRR task with four top-level discourse relations (i.e., *Comparison*, *Contingency*, *Expansion*, *Temporal*) and the 11 major second-level discourse senses. The dataset statistics are displayed in Table 9 and Table 10 in Appendix.

DiscoGeM. The DiscoGeM dataset (Scholman et al., 2022) is a crowd-sourced corpus of multi-genre implicit discourse relations. Different from the expert-annotated PDTB, DiscoGeM adopts a crowd-sourcing method by asking crowd workers to provide possible *connectives* between two arguments. They curated a connective mapping from connectives to the discourse relation senses in PDTB, which is used to generate PDTB-style discourse relations from the crowd-sourced connectives. Clear differences in the distributions across three genres have been observed (Scholman et al., 2022). For instance, CONJUNCTION is more prevalent in Wikipedia text, and PRECEDENCE occurs

Dataset	Train	Validation	Test	# of labels
TB-Dense	4,032	629	1,427	6
MATRES	6,336	–	837	4
TDDMan	4,000	650	1,500	5

Table 8: Statistics of three temporal relation datasets.

more frequently in novels than in other genres. DiscoGeM includes 6,505 instances from three genres: political speech data from the Europarl corpus, texts from 20 novels, and encyclopedic texts from English Wikipedia. The data was split into 70% training, 20% testing, and 10% development sets. For a fair and comprehensive evaluation, we test ChatGPT on the full test set containing 1,286 instances under the single label setting.

STAC (Asher et al., 2016) was the first corpus of discourse parsing for multi-party dialogue. The dataset was adapted from an online multi-player game *The Settlers of Catan*, where players acquire and trade resources in order to build facilities. The STAC corpus came from the chat history in trade negotiations.

Molweni (Li et al., 2020) came from the large-scale multi-party dialogue dataset, *the Ubuntu Chat Corpus* (Lowe et al., 2015), which is a collection of chat logs between users seeking technical support on the Ubuntu operating system.Li et al. (2020) conducted additional annotations specific to dialogue discourse parsing to construct the Molweni dataset, which is larger in scale than STAC. Moreover, a preliminary study on Molweni has shown comparable baseline performance to that in STAC, which indicates the two datasets have similar quality and complexity

A.2 ChatGPT Hyperparameter

In this study, we only call the OpenAI API for conducting evaluation and do not use any GPU to train the model. For the hyperparameter for ChatGPT response generation, the temperature is 0.7, Top_p is 1, and the max_tokens is 256.

B Downstream Tasks of Discourse Relations

Discourse relations can be applied for acquiring commonsense knowledge and developing discourse-aware sophisticated commonsense reasoning benchmarks that are shown to be hard for current large language models (Bhargava and Ng, 2022). In this section, we study two NLP tasks that are applications of discourse relations, one for commonsense acquisition (Fang et al., 2021, 2023) and one for a commonsense question answering constructed with sophisticated discourse markers (Bhargava and Ng, 2022).

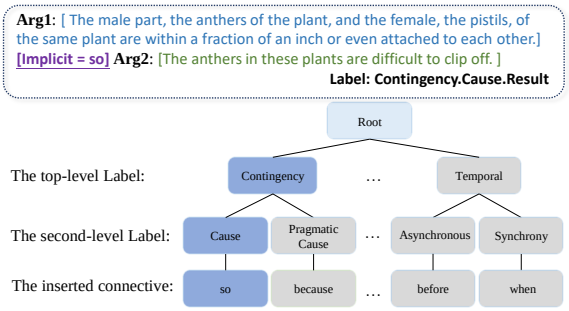

Figure 3: An example of the implicate discourse relation recognition task and the label hierarchy.

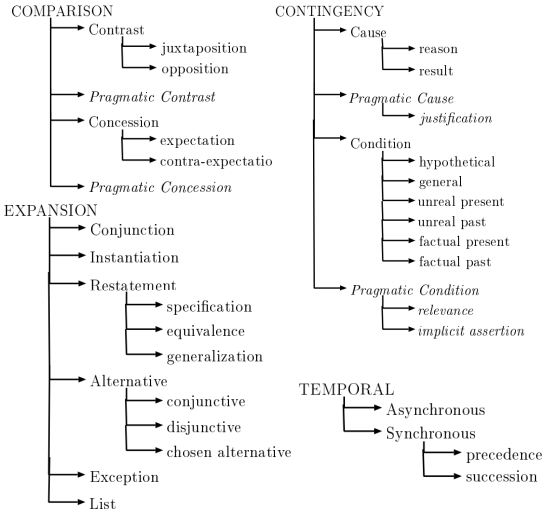

Figure 4: The sense hierarchy of implicit discourse relation in PDTB2.0 dataset

Top-level Senses	Train	Validation	Test
Comparison	1,942	197	152
Contingency	3,342	295	279
Expansion	7,004	671	574
Temporal	760	64	85
Total	12,362	1,183	1,046

Table 9: Statistics of four top-level implicit senses in PDTB 2.0.

Commonsense Knowledge Base Population. CKBP (Fang et al., 2021) is a benchmark for populating commonsense knowledge from discourse

Second-level Senses	Train	Validation	Test
Comp.Concession	180	15	17
Comp.Contrast	1566	166	128
Cont.Cause	3227	281	269
Cont.Pragmatic Cause	51	6	7
Exp.Alternative	146	10	9
Exp.Conjunction	2805	258	200
Exp.Instantiation	1061	106	118
Exp.List	330	9	12
Exp.Restatement	2376	260	211
Temp.Asynchronous	517	46	54
Temp.Synchrony	147	8	14
Total	12406	1165	1039

Table 10: The implicit discourse relation data statistics of second-level types in PDTB 2.0.

Dataset	Data source	# of dialogues/utterances/relations
STAC	Online multi-player game	111 1156 1128
Molweni	The Ubuntu chat corpus	500 4430 3911

Table 11: Statistics of the multi-party dialogue parsing datasets STAC and Molweni.

knowledge triples. For example, it requires the model to determine whether a discourse knowledge entry (*John drinks coffee*, `Succession/then`, *John feels refreshed*) represents a plausible commonsense knowledge, (*PersonX drinks coffee*, `xReact`, *refreshed*), a form of social commonsense knowledge defined in ATOMIC (Sap et al., 2019) where `xReact` studies what would *PersonX* feels after the head event. We include the latest test set of CKBP v2[7] for our experiments, which contains 4k triples converted from discourse relations to 15 commonsense relations defined in ConceptNet (Speer et al., 2017), ATOMIC (Sap et al., 2019), and GLUCOSE (Mostafazadeh et al., 2020). Prompt templates are presented in Table 31.

DISCOSENSE. DISCOSENSE is a commonsense question-answering dataset built upon discourse connectives. It's constructed from DISCOVERY (Sileo et al., 2019) and DISCOFUSE (Geva et al., 2019) where there are two sentences connected through a discourse connective and the negative options are generated through a conditional adversarial filtering process to make sure the difficulty of the dataset. The task is defined as selecting the most plausible coming sentence given the source sentence and a discourse connective such as *because*, *although*, *for example*, etc. Supervised learning models struggle on this dataset, showing a lack of subtle reasoning ability for discourse relations. We take the test set for evaluation. Prompt templates are presented in Table 32.

[7] https://github.com/HKUST-KnowComp/CSKB-Population/

Method	CKBP v2. AUC	CKBP v2. F1	DISCOSENSE Acc
Fine-tuned SOTA	73.70	46.70	65.87
$\text{ChatGPT}_{\text{PE}}$	65.77	45.93	47.25
$\text{ChatGPT}_{\text{ICL}}$	66.20	46.42	54.67

Table 12: Performance on CSKB Population and DISCOSENSE.PE and ICL indicate the prompt engineering template and in-context learning prompt template.

Experimental Results. We present the experimental results on Table 12. We compare the performance of zero-shot ChatGPT with supervised SOTA, which is PseudoReasoner-RoBERTa-large (Fang et al., 2022) for CKBP v2 and Electra-large (Clark et al., 2020) for DISCOSENSE. ChatGPT can achieve comparable F1 scores for CKBP v2. while still down performs regarding AUC. For the DISCOSENSE dataset, ChatGPT has a long way to reaching fine-tuned SOTA, letting alone human performance, indicating a lack of subtle reasoning ability to distinguish different discourse relations.

We report our experimental results summarized in Table 12 leveraging the full test sets of both CKBP and DISCOSENSE. We compare the performance of zero-shot ChatGPT with that of PseudoReasoner-RoBERTa-large (Fang et al., 2022) for CKBP v2 and ELECTRA-large (Clark et al., 2020) for DISCOSENSE, both of which are supervised state-of-the-arts. Our results show that ChatGPT achieves comparable F1 scores for CKBP v2, but it still underperforms in terms of AUC. For the DISCOSENSE dataset, ChatGPT has a long way to go to match the fine-tuned state-of-the-art performance, let alone human performance (95.40). This suggests that ChatGPT still lacks the subtle reasoning ability needed to distinguish between different discourse relations for making inferences.

C Prompt Templates

The prompting or prompt tuning method is widely applied for many downstream tasks in the Natural Language Processing (NLP) field, the sensitivity and performance variance of the prompting method has been reported in a lot of works (Han

et al., 2021; Chan et al., 2023a; Zhong et al., 2021; Liu et al., 2021a; Li et al., 2023b; Chan and Chan, 2023). Therefore, we utilized the expert knowledge on these sentence-level relation classification tasks to manually craft a prompt template that outperformed a baseline (Robinson et al., 2022) with fairly standard settings for all tasks. Our designed prompt template will be comprehensive and reliable baselines to exclude the variance of the prompt engineering and offer fair comparison baselines for further works. We list all prompt templates used in this paper as follows.

TB-Dense				
Strategies	**Template input**	**ChatGPT**	**Gold**	**T/F**
Prompt	Sentence: The Organization of African Unity said Friday it would investigate the Hutu-organized genocide of more than 500,000 minority Tutsis in Rwanda nearly four years ago. Foreign ministers of member-states meeting 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. event1: investigate event2: shot Question: What is the temporal relation between event1 and event2 in the sentence? A. AFTER B. BEFORE C. SIMULTANEOUS D. NONE E. INCLUDES F. IS_INCLUDED Answer:	NONE	AFTER	F
Prompt Engineering	Determine the temporal order from "investigate" to "shot" in the following sentence: "The Organization of African Unity said Friday it would investigate the Hutu-organized genocide of more than 500,000 minority Tutsis in Rwanda nearly four years ago. Foreign ministers of member-states meeting 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.". Only answer one word from AFTER, BEFORE, SIMULTANEOUS, NONE, INCLUDES, IS_INCLUDED. Answer:	AFTER	AFTER	T

Table 13: Prompt example for TB-Dense.

TB-Dense				
Strategies	**Template input**	**ChatGPT**	**Gold**	**T/F**
In-Context Learning	Determine the temporal order from "convictions" to "fraud" in the following sentence: "A federal appeals court has reinstated his state convictions for securities fraud.". Only answer one word from AFTER, BEFORE, SIMULTANEOUS, NONE, INCLUDES, IS_INCLUDED. Answer: AFTER Determine the temporal order from "arrested" to "said" in the following sentence: "Derek Glenn, a spokesman for the Newark Police Department, said that of nine women who had been killed last year, suspects had been arrested in only four cases.". Only answer one word from AFTER, BEFORE, SIMULTANEOUS, NONE, INCLUDES, IS_INCLUDED. Answer: BEFORE Determine the temporal order from "assassination" to "touched" in the following sentence: "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.". Only answer one word from AFTER, BEFORE, SIMULTANEOUS, NONE, INCLUDES, IS_INCLUDED. Answer: SIMULTANEOUS Determine the temporal order from "seen" to "created" in the following sentence: "Ï haven't seen a pattern yet,šaid Patricia Hurt, the Essex County prosecutor, who created the task force on Tuesday.". Only answer one word from AFTER, BEFORE, SIMULTANEOUS, NONE, INCLUDES, IS_INCLUDED. Answer: NONE Determine the temporal order from "meeting" to "agreed" in the following sentence: "Foreign ministers of memberstates meeting in the Ethiopian capital agreed to set up a sevenmember panel to investigate who shot down Rwandan President Juvenal Habyarimana's plane on April 6, 1994.". Only answer one word from AFTER, BEFORE, SIMULTANEOUS, NONE, INCLUDES, IS_INCLUDED. Answer: INCLUDES Determine the temporal order from "investigation" to "said" in the following sentence: "The panel will be based in Addis Ababa, and will finish its investigation within a year, it said.". Only answer one word from AFTER, BEFORE, SIMULTANEOUS, NONE, INCLUDES, IS_INCLUDED. Answer: IS_INCLUDED Determine the temporal order from "investigate" to "shot" in the following sentence: "The Organization of African Unity said Friday it would investigate the Hutu-organized genocide of more than 500,000 minority Tutsis in Rwanda nearly four years ago. Foreign ministers of member-states meeting 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.". Only answer one word from AFTER, BEFORE, SIMULTANEOUS, NONE, INCLUDES, IS_INCLUDED. Answer:	BEFORE	AFTER	F

Table 14: Prompt example for TB-Dense.

MATRES				
Strategies	**Template input**	**ChatGPT**	**Gold**	**T/F**
Prompt	Sentence: "It had a multiplying effect.", "We were pleased that England and New Zealand knew about it, and we thought that's where it would stop." event1: had event2: pleased Question: What is the temporal relation between event1 and event2 in the sentence? A. AFTER B. BEFORE C. EQUAL D. VAGUE Answer:	AFTER	EQUAL	F
Prompt Engineering	Determine the temporal order from "had" to "pleased" in the following sentence: ""It had a multiplying effect.", "We were pleased that England and New Zealand knew about it, and we thought that's where it would stop. "". Only answer one word from AFTER, BEFORE, EQUAL, VAGUE. Answer:	EQUAL	EQUAL	T
In-Context Learning	Determine the temporal order from "give" to "tried" in the following sentence: "It will give the rest of the world the view that Cuba is like any other nation, something the US has, of course, tried to persuade the world that it is not.". Only answer one word from AFTER, BEFORE, EQUAL, VAGUE. Answer: AFTER Determine the temporal order from "invited" to "come" in the following sentence: "Fidel Castro invited John Paul to come for a reason.". Only answer one word from AFTER, BEFORE, EQUAL, VAGUE. Answer: BEFORE Determine the temporal order from "earned" to "rose" in the following sentence: "In the nine months, EDS earned $315.8 million, or $2.62 a share, up 13 % from $280.7 million, or $2.30 a share.". Only answer one word from AFTER, BEFORE, EQUAL, VAGUE. Answer: EQUAL Determine the temporal order from "created" to "become" in the following sentence: "Ms. Atimadi says the war has created a nation of widows. Women have become the sole support of their families.". Only answer one word from AFTER, BEFORE, EQUAL, VAGUE. Answer: VAGUE Determine the temporal order from "had" to "pleased" in the following sentence: ""It had a multiplying effect.", "We were pleased that England and New Zealand knew about it, and we thought that's where it would stop. "". Only answer one word from AFTER, BEFORE, EQUAL, VAGUE. Answer:	BEFORE	EQUAL	F

Table 15: Prompt example for MATRES.

TDDMan				
Strategies	**Template input**	**ChatGPT**	**Gold**	**T/F**
Prompt	Sentence: 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. event1: rampage event2: exodus Question: What is the temporal relation between event1 and event2 in the sentence? A. AFTER B. BEFORE C. SIMULTANEOUS D. INCLUDES E. IS_INCLUDED Answer:	AFTER	BEFORE	F
Prompt Engineering	Determine the temporal order from "rampage" to "exodus" in the following sentence: "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.". Only answer one word from AFTER, BEFORE, SIMULTANEOUS, INCLUDES, IS_INCLUDED. Answer:	BEFORE	BEFORE	T

Table 16: Prompt example for TDDMan.

TDDMan				
Strategies	**Template input**	**ChatGPT**	**Gold**	**T/F**
In-Context Learning	Determine the temporal order from "thrown" to "raised" in the following sentence: "Keating's convictions were thrown out in nineteen ninety-six on a technicality. And on that basis Keating was released from prison before he was eligible for parole. Now the ninth US circuit court of appeals has ruled that the original appeal was flawed since it brought up issues that had not been raised before.". Only answer one word from AFTER, BEFORE, SIMULTANEOUS, NONE, INCLUDES, IS_INCLUDED. Answer: AFTER Determine the temporal order from "seized" to "parole" in the following sentence: "The bonds became worthless when the bankrupt thrift was seized by government regulators. Keating's convictions were thrown out in nineteen ninety-six on a technicality. And on that basis Keating was released from prison before he was eligible for parole.". Only answer one word from AFTER, BEFORE, SIMULTANEOUS, NONE, INCLUDES, IS_INCLUDED. Answer: BEFORE Determine the temporal order from "assassination" to "reignited" in the following sentence: "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.". Only answer one word from AFTER, BEFORE, SIMULTANEOUS, NONE, INCLUDES, IS_INCLUDED. Answer: SIMULTANEOUS Determine the temporal order from "war" to "genocide" in the following sentence: "It also reignited the civil war. The panel also will look at the exodus of about 2 million Rwanda Hutus to neighboring countries. The investigation will consider the role of ïnternal 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,ẗhe OAU said.". Only answer one word from AFTER, BEFORE, SIMULTANEOUS, NONE, INCLUDES, IS_INCLUDED. Answer: INCLUDES Determine the temporal order from "arrests" to "related" in the following sentence: "But over all, arrests were made in more than 60 percent of murder cases, he said. Eight of the 14 killings since 1993 were already under investigation by the Newark Police Department, Glenn said. Of the eight victims, three were stabbed, two were strangled, two were beaten to death and one was asphyxiated, he said, and these different methods of killing and other evidence seem to indicate that the eight cases are not related.". Only answer one word from AFTER, BEFORE, SIMULTANEOUS, NONE, INCLUDES, IS_INCLUDED. Answer: IS_INCLUDED Determine the temporal order from "rampage" to "exodus" in the following sentence: "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.". Only answer one word from AFTER, BEFORE, SIMULTANEOUS, INCLUDES, IS_INCLUDED. Answer:	AFTER	BEFORE	F

Table 17: Prompt example for TDDMan.

COPA				
Strategies	**Template input**	**ChatGPT**	**Gold**	**T/F**
Prompt	The cause of The cashier opened the cash register is: 1. The customer searched his wallet. 2. The customer handed her money. Only answer '1' or '2' only without any other words.	2.	2	T
Prompt Engineering	Given the event The cashier opened the cash register, which choice is more likely to be the cause of this event? 1. The customer searched his wallet. 2. The customer handed her money. Only answer '1' or '2' only without any other words.	2.	2	T
In-Context Learning	Given the event The shirt shrunk, the cause of this event is likely to be I put it in the dryer. Given the event It got dark outside, the effect of this event is likely to be The moon became visible in the sky. Given the event The cashier opened the cash register, which choice is more likely to be the cause of this event? 1. The customer searched his wallet. 2. The customer handed her money. Only answer '1' or '2' only without any other words.	2	2	T

Table 18: Prompt templates used for the COPA benchmark.

e-CARE				
Strategies	**Template input**	**ChatGPT**	**Gold**	**T/F**
Prompt	The effect of They walked along the stream is: 1. They found lots of fish in it. 2. They went to ponds. Only answer '1' or '2' only without any other words.	2.	1	F
Prompt Engineering	Given the event They walked along the stream, which choice is more likely to be the effect of this event? 1. They found lots of fish in it. 2. They went to ponds. Only answer '1' or '2' only without any other words.	1.	1	T
In-Context Learning	Given the event There is a light rain today, the effect of this event is likely to be The roots of many plants are not moistened by rain. Given the event His parents stopped him, the cause of this event is likely to be The child ran towards hippos. Given the event They walked along the stream, which choice is more likely to be the effect of this event? 1. They found lots of fish in it. 2. They went to ponds. Only answer '1' or '2' only without any other words.	1.	1	T

Table 19: Prompt templates used for the e-CARE benchmark.

HeadlineCause				
Strategies	**Template input**	**ChatGPT**	**Gold**	**T/F**
Prompt	News title A: Guv encourages creative developers during lockdown. News title B: Govt hints at lockdown extension, but promises relaxations. Is there any causal relationship between these two titles? 1. No. 2. A causes B. 3. B causes A. Only answer '1' or '2' or '3' without any other words.	1	1	T
Prompt Engineering	News title A: Guv encourages creative developers during lockdown. News title B: Govt hints at lockdown extension, but promises relaxations. Will one news cause the other one? 1. No, there is no cause-and-effect relationship between them. 2. The happening of news A will cause news B. 3. The happening of news B will cause news A. Only answer '1' or '2' or '3' without any other words.	1.	1	T
In-Context Learning	Here are three examples: News A: Why Reliance Industries share price has gained over 19% in four sessions. News B: IndusInd Bank stock rises over 6% ahead of Q4 earnings. For this pair of news titles, there is no cause-and-effect relationship between them. News A: Indian government brushes off Indian tax officers' proposal for coronavirus tax on super rich. News B: Inquiry against 50 IRS officers over suggesting tax hike for the rich: Report. For this pair of titles, the happening of news A will cause news B. News A: Insensitive or lost in translation? Twitter weighs in on Thiem's comments against a player fund. News B: Coronavirus: Why should I give money to lower-ranked players, questions Dominic Thiem. For this pair of titles, the happening of news B will cause news A. Now, answer this question. News title A: Guv encourages creative developers during lockdown. News title B: Govt hints at lockdown extension, but promises relaxations. Will one news cause the other one? 1. No, there is no cause-and-effect relationship between them. 2. The happening of news A will cause news B. 3. The happening of news B will cause news A. Only answer '1' or '2' or '3' without any other words.	2.	1	F

Table 20: Prompt templates used for the HeadlineCause benchmark.

Explicit Discourse Relation Tasks				
Strategies	**Template input**	**ChatGPT**	**Gold**	**T/F**
Top-level Prompt	Argument 1:"When used as background in this way, the music has an appropriate eeriness" Argument 2:"Served up as a solo the music lacks the resonance provided by a context within another medium" Connective between Argument 1 and Argument 2:"however" Question:What is the discourse relation between Argument 1 and Argument 2? A. Comparison B. Contingency C. Expansion D. Temporal Answer:	B. Contingency	A. Comparison	F
Second-level Prompt	Argument 1:"When used as background in this way, the music has an appropriate eeriness" Argument 2:"Served up as a solo the music lacks the resonance provided by a context within another medium" Connective between Argument 1 and Argument 2:"however" Question:What is the discourse relation between Argument 1 and Argument 2? A. Concession B. Contrast C. Cause D. Condition E. Alternative F. Conjunction G. Instantiation H. List I. Restatement J. Asynchronous K. Synchrony Answer:	B. Contrast	B. Contrast	T
Prompt Engineering	Argument 1:"When used as background in this way, the music has an appropriate eeriness" Argument 2:"Served up as a solo the music lacks the resonance provided by a context within another medium" Connective between Argument 1 and Argument 2:"however" Question:What is the discourse relation between Argument 1 and Argument 2? A. Comparison.Concession, nonetheless B. Comparison.Contrast, however C. Contingency.Cause, so D. Contingency.Condition, if E. Expansion.Alternative, instead F. Expansion.Conjunction, also G. Expansion.Instantiation, for example H. Expansion.List, and I. Expansion.Restatement, specifically J. Temporal.Asynchronous, before K. Temporal.Synchrony, when Answer:	B.Comparison. Contrast, however	B. Comparison. Contrast	T

Table 21: Prompt example for PDTB2.0 explicit discourse relation task.

Explicit Discourse Relation Tasks				
Strategies	**Template input**	**ChatGPT**	**Gold**	**T/F**
In-Context Learning	All answer select from following: A. Comparison.Concession B. Comparison.Contrast C. Contingency.Cause D. Contingency.Condition E. Expansion.Alternative F. Expansion.Conjunction G. Expansion.Instantiation H. Expansion.List I. Expansion.Restatement J. Temporal.Asynchronous K. Temporal.Synchrony Argument 1:"whose hair is thinning and gray and whose face has a perpetual pallor." Argument 2:"The prime minister continues to display an energy, a precision of thought and a willingness to say publicly what most other Asian leaders dare say only privately." Connective between Argument 1 and Argument 2:"nonetheless" Question:What is the discourse relation between Argument 1 and Argument 2? Answer:Comparison.Concession Argument 1:"they usually give current shareholders the right to buy more stock of their corporation at a large discount if certain events occur." Argument 2:"these discount purchase rights may generally be redeemed at a nominal cost by the corporation's directors if they approve of a bidder." Connective between Argument 1 and Argument 2:"however" Question:What is the discourse relation between Argument 1 and Argument 2? Answer:Comparison.Contrast ..... Argument 1:"I find it hard to ignore our environmental problems." Argument 2:"I start my commute to work with eyes tearing and head aching from the polluted air." Connective between Argument 1 and Argument 2:"when" Question:What is the discourse relation between Argument 1 and Argument 2? Answer:Temporal.Synchrony Argument 1:"When used as background in this way, the music has an appropriate eeriness" Argument 2:"Served up as a solo the music lacks the resonance provided by a context within another medium" Connective between Argument 1 and Argument 2:"however" Question:What is the discourse relation between Argument 1 and Argument 2? Answer:	B.Comparison. Contrast	B.Comparison. Contrast	T

Table 22: Prompt example for PDTB2.0 explicit discourse relation task (Continuous).

Implicit Discourse Relation Tasks				
Strategies	**Template input**	**ChatGPT**	**Gold**	**T/F**
Top-level Prompt	Argument 1:"We've been spending a lot of time in Los Angeles talking to TV production people" Argument 2:"With the competitiveness of the television market these days, everyone is looking for a way to get viewers more excited" Question:What is the discourse relation between Argument 1 and Argument 2? A. Comparison B. Contingency C. Expansion D. Temporal Answer:	C. Expansion	B. Contingency	F
Second-level Prompt	Argument 1:"We've been spending a lot of time in Los Angeles talking to TV production people" Argument 2:"With the competitiveness of the television market these days, everyone is looking for a way to get viewers more excited" Question:What is the discourse relation between Argument 1 and Argument 2? A. Concession B. Contrast C. Cause D. Pragmatic Cause E. Alternative F. Conjunction G. Instantiation H. List I. Restatement J. Asynchronous K. Synchrony Answer:	C. Cause	C. Cause	T
Prompt Engineering	Argument 1:"We've been spending a lot of time in Los Angeles talking to TV production people" Argument 2:"With the competitiveness of the television market these days, everyone is looking for a way to get viewers more excited" Question:What is the discourse relation between Argument 1 and Argument 2? A. Comparison.Concession, if B. Comparison.Contrast, however C. Contingency.Cause, so D. Contingency.Pragmatic, indeed E. Expansion.Alternative, instead F. Expansion.Conjunction, also G. Expansion.Instantiation, for example H. Expansion.List, and I. Expansion.Restatement, specifically J. Temporal.Asynchronous, before K. Temporal.Synchrony, when Answer:	C. Contingency. Cause, so	C. Contingency. Cause	T

Table 23: Prompt example for PDTB2.0 implicit discourse relation task.

Implicit Discourse Relation Tasks				
Strategies	**Template input**	**ChatGPT**	**Gold**	**T/F**
In-Context Learning	All answer select from following: A. Comparison.Concession, nonetheless B. Comparison.Contrast, however C. Contingency.Cause, so D. Contingency.Pragmatic Cause, indeed E. Expansion.Alternative, instead F. Expansion.Conjunction, also G. Expansion.Instantiation, for example H. Expansion.List, and I. Expansion.Restatement, specifically J. Temporal.Asynchronous, before K. Temporal.Synchrony, when Argument 1:"Coke could be interested in more quickly developing some of the untapped potential in those markets." Argument 2:"A Coke spokesman said he couldn't say whether that is the direction of the talks." Question:What is the discourse relation between Argument 1 and Argument 2? Answer:Comparison.Concession, nonetheless Argument 1:"Tanks currently are defined as armored vehicles weighing 25 tons or more that carry large guns." Argument 2:"The Soviets complicated the issue by offering to include light tanks, which are as light as 10 tons." Question:What is the discourse relation between Argument 1 and Argument 2? Answer:Comparison.Contrast, however ..... Argument 1:"Panamanian dictator Torrijos, he was told, had granted the shah of Iran asylum in Panama as a favor to Washington." Argument 2:"Mr.Sanford was told Mr.Noriega's friend, Mr. Wittgreen, would be handling the shah's security." Question:What is the discourse relation between Argument 1 and Argument 2? Answer: Temporal.Synchrony, when Argument 1:"We've been spending a lot of time in Los Angeles talking to TV production people" Argument 2:"With the competitiveness of the television market these days, everyone is looking for a way to get viewers more excited" Question:What is the discourse relation between Argument 1 and Argument 2? Answer:	C. Contingency. Cause, so	C. Contingency. Cause	T

Table 24: Prompt example for PDTB2.0 implicit discourse relation task (Continued).

DiscoGeM				
Strategies	**Template input**	**ChatGPT**	**Gold**	**T/F**
Prompt	Argument 1:"Allow me to make a few general comments on European solidarity, on the Solidarity Fund and on some events that may provide lessons for the future." Argument 2:"In 2002 I had the experience of leading a country that was struck by terrible floods, together with the Federal Republic of Germany and Austria. It was the scale of that disaster that provided the incentive for the creation of the Solidarity Fund." Question:What is the discourse relation between Argument 1 and Argument 2? (0) arg1-as-denier (1) arg1-as-detail (2) arg1-as-goal (3) arg2-as-denier (4) arg2-as-detail (5) arg2-as-goal (6) arg2-as-instance (7) arg2-as-subst (8) conjunction (9) contrast (10) differentcon (11) disjunction (12) precedence (13) reason (14) result (15) similarity (16) succession (17) synchronous Answer:	(2) arg1-as-goal	(4) arg2-as-detail	F

Table 25: Prompt example 1 for DiscoGeM, the multi-genre discourse classification task.

DiscoGeM				
Strategies	**Template input**	**ChatGPT**	**Gold**	**T/F**
Prompt Engineering	Argument 1:"However, the Member States are not obliged to replace fixed-term contracts with open-ended contracts assuming that there are other effective measures in place that would prevent or sanction such abuse. The European Court of Justice confirmed this interpretation in its judgment of 4 July 2006 in Case C-212/04 (Adeneler) pertaining to Greek legislation." Argument 2:"The European Court of Justice also stated that interpretation of the relevant national legislation does not fall within its competence. It is entirely for the Greek courts to provide an interpretation of relevant Greek legislation and to determine whether this legislation complies with the requirements of the Directive regarding the existence of effective measures that would prevent and sanction abuse arising from the use of successive fixed-term employment contracts." Question:What is the discourse relation between Argument 1 and Argument 2? (0) arg1-as-denier: despite the fact that (1) argument 1 as detail: in short (2) argument 1 as goal: for that purpose (3) argument 2 as denier: despite this (4) argument 2 as detail: in more detail (5) argument 2 as goal: ensuring that (6) argument 2 as instance: for instance (7) argument 2 as substitution: rather (8) conjunction: in addition (9) contrast: by comparison (10) differentcon: none (11) disjunction: or alternatively (12) precedence: subsequently (13) reason: the reasons is/are that (14) result: consequently (15) similarity: similarly (16) succession: previously (17) synchronous: at that time Answer:	(8) conjunction: in addition	(8) conjunction: in addition	T

Table 26: Prompt example 2 for DiscoGeM, the multi-genre discourse classification task.

DiscoGeM				
Strategies	**Template input**	**ChatGPT**	**Gold**	**T/F**
In-Context Learning	Candidate relations: (0) arg1-as-denier: despite the fact that (1) argument 1 as detail: in short (2) argument 1 as goal: for that purpose (3) argument 2 as denier: despite this (4) argument 2 as detail: in more detail (5) argument 2 as goal: ensuring that (6) argument 2 as instance: for instance (7) argument 2 as substitution: rather (8) conjunction: in addition (9) contrast: by comparison (10) differentcon: none (11) disjunction: or alternatively (12) precedence: subsequently (13) reason: the reasons is/are that (14) result: consequently (15) similarity: similarly (16) succession: previously (17) synchronous: at that time Argument 1:"Mr President, ladies and gentlemen, the motion for a resolution before us today is important because of its subject and the desire to protect the rule of law and press freedom. It is also very important because of the broad consensus which has finally been reached after some heated discussions behind the scenes." Argument 2:"The problem considered by the motion is a major one but, as has already been touched upon, could be regarded as minor in light of the even greater problem of the general situation in Angola which is experiencing a terrible humanitarian disaster. This situation, as in neighbouring former Zaire, is like a festering wound in which it is not clear who is infecting whom." Question:What is the discourse relation between Argument 1 and Argument 2? Answer: (14) result: consequently Argument 1:"The ship was finally able to turn around and it fled northwards as fast as possible. Then there was a terrible explosion about six hundred yards to the stern and a gigantic column of water and steam, perhaps a hundred yards high, shot out of the sea. The Oudenbourg set course for Harwich and sent out a radio warning in all directions: Ättention all shipping, attention all shipping!" Argument 2:"Severe danger on Ostende-Ramsgate lane. Underwater explosion. Cause unknown. All shipping advised avoid area!" Question:What is the discourse relation between Argument 1 and Argument 2? Answer: (4) argument 2 as detail: in more detail Argument 1:"Allow me to make a few general comments on European solidarity, on the Solidarity Fund and on some events that may provide lessons for the future. In 2002 I had the experience of leading a country that was struck by terrible floods, together with the Federal Republic of Germany and Austria." Argument 2:"It was the scale of that disaster that provided the incentive for the creation of the Solidarity Fund. The disaster occurred in August and the first payments were received by the Czech Republic the following January." Question:What is the discourse relation between Argument 1 and Argument 2? Answer:	(3) argument 2 as denier: despite this	(14) result: consequently	F

Table 27: Prompt example 3 for DiscoGeM, the multi-genre discourse classification task.

Dialogue DP-STAC			
Strategies	**Template input**	**ChatGPT**	**Gold**
Prompt (w/o desc.)	Here is a multi-party dialogue: Utterance 0: (Speaker A) sorry raef- Utterance 1: (Speaker A) at least i forgot to play it Utterance 2: (Speaker A) before that 6 was rolled Utterance 3: (Speaker B) well at least people should realize your advantage now Q: Predict all the possible discourse relations between utterances and their types line by line (e.g., 'Utterance 0 and utternace 1: (0) Utterance 0 and utterance 3: (1)') Candidate types are listed below: Choose from: (0) Comment (1) Clarification question (2) Question-answer pair (3) Continuation (4) Acknowledgement (5) Question and elaboration (6) Result (7) Elaboration (8) Explanation (9) Correction (10) Contrast (11) Conditional (12) Background (13) Narration (14) Alternation (15) Parallel	Utterance 0 and utterance 1: (2) Utterance 1 and utterance 2: (0) Utterance 2 and utterance 3: (9) Utterance 3 and utterance 4: (0) Utterance 4 and utterance 5: (5) Utterance 5 and utterance 6: (0) Utterance 6 and utterance 7: (7) Utterance 7 and utterance 8: (0) Utterance 8 and utterance 9: (3) Utterance 9 and utterance 10: (14)	Utterance 0 and utterance 1: (8) Utterance 1 and utterance 2: (13) Utterance 1 and utterance 3: (0)

Table 28: Prompt example for STAC in the multi-party dialogue discourse parsing task. Examples in Molweni are in a similar format.

Dialogue DP-STAC			
Strategies	**Template input**	**ChatGPT**	**Gold**
Prompt (w/ desc.)	Here is a multi-party dialogue: Utterance 0: (Speaker A) sorry raef- Utterance 1: (Speaker A) at least i forgot to play it Utterance 2: (Speaker A) before that 6 was rolled Utterance 3: (Speaker B) well at least people should realize your advantage now Q: Predict all the possible discourse relations between utterances and their types line by line (e.g., 'Utterance 0 and utternace 1: (0) Utterance 0 and utterance 3: (1)') Candidate types are listed below: Choose from: (0) Comment: Utterance y comments utterance x. (1) Clarification question: Utterance y clarifies utterance x. (2) Question-answer pair: Utterance x is a question and utterance y is the answer of utterance x. (3) Continuation: Utterance y is the continuation of utterance x. (4) Acknowledgement: Utterance y acknowledges utterance x. (5) Question and elaboration: Utterance x is a question and utterance y tries to elaborate utterance x. (6) Result: Utterance y is the effect brought about by the situation described in utterance x. (7) Elaboration: Utterance y elaborates utterance x. (8) Explanation: Utterance y is the explanation of utterance x. (9) Correction: Utterance y corrects utterance x. (10) Contrast: Utterance x and utterance y share a predicate or property and a difference on shared property. (11) Conditional: Utterance x is the condition of utterance y or utterance y is the condition of utterance x. (12) Background: Utterance y is the background of utterance x. (13) Narration: Utterance y is the narration of utterance x. (14) Alternation: Utterance x and utterance y denote alternative situations. (15) Parallel: Utterance y and utterance x are parallel and present almost the same meaning.	Utterance 0 and utterance 1: (2) Utterance 0 and utterance 3: (1) Utterance 1 and utterance 5: (0) Utterance 2 and utterance 3: (4) Utterance 4 and utterance 5: (0) Utterance 6 and utterance 7: (4) Utterance 8 and utterance 9: (0) Utterance 9 and utterance 10: (9)	Utterance 0 and utterance 1: (8) Utterance 1 and utterance 2: (13) Utterance 1 and utterance 3: (0)

Table 29: Prompt example for STAC in the multi-party dialogue discourse parsing task. Examples in Molweni are in the similar format.

Dialogue DP-STAC			
Strategies	**Template input**	**ChatGPT**	**Gold**
In-Context Learning	[Example 1] Here is a multi-party dialogue: Utterance 0: (Speaker A) sorry raef- Utterance 1: (Speaker A) at least i forgot to play it Utterance 2: (Speaker A) before that 6 was rolled Utterance 3: (Speaker B) well at least people should realize your advantage now Q: Predict all the possible discourse relations between utterances and their types line by line (e.g., 'Utterance 0 and utternace 1: (0) Utterance 0 and utterance 3: (1)') Candidate types are listed below: Choose from: (0) Comment (1) Clarification question (2) Question-answer pair (3) Continuation (4) Acknowledgement (5) Question and elaboration (6) Result (7) Elaboration (8) Explanation (9) Correction (10) Contrast (11) Conditional (12) Background (13) Narration (14) Alternation (15) Parallel A: Utterance 0 and utterance 1: (8) Utterance 1 and utterance 2: (13) Utterance 1 and utterance 3: (0) [Example 2] Here is a multi-party dialogue: Utterance 0: (Speaker A) I need wood, clay or ore, I can give Sheep Utterance 1: (Speaker B) i can trade wood Utterance 2: (Speaker C) just spent it all Utterance 3: (Speaker C) sorry Utterance 4: (Speaker A) 1 sheep for 1 wood? Utterance 5: (Speaker B) 2 sheep 1 wood Utterance 6: (Speaker C) sorry empty Utterance 7: (Speaker C) tough times.. Utterance 8: (Speaker B) hopefully i dont roll a 7 Utterance 9: (Speaker B) and that biotes me in the arse Utterance 10: (Speaker B) bites* Q: Predict all the possible discourse relations between utterances and their types line by line (e.g., 'Utterance 0 and utternace 1: (0) Utterance 0 and utterance 3: (1)') Candidate types are listed below: Choose from: (0) Comment (1) Clarification question (2) Question-answer pair (3) Continuation (... same as above) (14) Alternation (15) Parallel	A: Utterance 0 and utterance 1: (2) Utterance 1 and utterance 4: (2) Utterance 2 and utterance 3: (9) Utterance 4 and utterance 5: (5) Utterance 5 and utterance 6: (7) Utterance 5 and utterance 8: (3) Utterance 8 and utterance 9: (11) Utterance 9 and utterance 10: (9)	A: Utterance 0 and utterance 1: (2) Utterance 0 and utterance 2: (2) Utterance 2 and utterance 3: (0) Utterance 1 and utterance 4: (5) Utterance 4 and utterance 5: (2) Utterance 6 and utterance 7: (8) Utterance 8 and utterance 9: (3) Utterance 9 and utterance 10: (9) Utterance 2 and utterance 6: (7) Utterance 5 and utterance 8: (0)

Table 30: Prompt example for STAC in the multi-party dialogue discourse parsing task. Examples in Molweni are in the similar format.

CKBP				
Strategies	**Template input**	**ChatGPT**	**Gold**	**T/F**
Prompt Engineering	Answer whether the following statement is plausible. Answer with only Yes or No:If PersonX drinks coffee, as a result, PersonX feels, refreshed.	Yes	Yes	T
In-Context Learning	Answer whether the following statement is plausible. Answer with only Yes or No:If PersonY accept the interview, as a result, PersonY or others will, PersonX give PersonY this opportunity.A: No Answer whether the following statement is plausible. Answer with only Yes or No:If PersonX lead the line, as a result, PersonY or others feel, PersonX support PersonX family.A: No Answer whether the following statement is plausible. Answer with only Yes or No:If PersonX form PersonY conception, as a result, PersonY or others want to, PersonY want to discuss with PersonZ.A: Yes Answer whether the following statement is plausible. Answer with only Yes or No:If PersonX give, PersonX is seen as, PersonX be comunicative.A: Yes Answer whether the following statement is plausible. Answer with only Yes or No:If PersonX be nervous, as a result, PersonX will, that be important to PeopleX.A: Yes Answer whether the following statement is plausible. Answer with only Yes or No:If PersonX celebrate persony, because PersonX wanted, PersonX feel oneself.A: No Answer whether the following statement is plausible. Answer with only Yes or No:If PersonX learn to ride a bike, but before, PersonX needed, PersonX wear helmet.A: Yes Answer whether the following statement is plausible. Answer with only Yes or No:If PersonX take PersonY time, as a result, PersonX feels, PersonX feel mortified.A: No Answer whether the following statement is plausible. Answer with only Yes or No:If PersonX want to ask a tought question, as a result, PersonX wants to, PersonX want to throw out PersonX clothes.A: No Answer whether the following statement is plausible. Answer with only Yes or No:If PersonX achieve PersonX end, happens after, PersonX start a small business.A: Yes Answer whether the following statement is plausible. Answer with only Yes or No:If PersonX like the idea, happens before, PersonX call a uber.A: No Answer whether the following statement is plausible. Answer with only Yes or No:If PersonX get injure, because, PersonX feel odd.A: No Answer whether the following statement is plausible. Answer with only Yes or No:If person x be bed ridden with illness, can be hindered by, PersonX find the perfect dog.A: No Answer whether the following statement is plausible. Answer with only Yes or No:If PersonX play violin, includes the event or action, PersonX make noise.A: Yes Answer whether the following statement is plausible. Answer with only Yes or No:If PersonX could not complete something, causes, PeopleX have find it.A: No Answer whether the following statement is plausible. Answer with only Yes or No:If PersonX drinks coffee, as a result, PersonX feels, refreshed.	Yes	Yes	T

Table 31: Prompt example for CKBP.

DiscoSense				
Strategies	**Template input**	**ChatGPT**	**Gold**	**T/F**
Prompt Engineering	`Question:` Which option represents the most plausible ending of the given context? `Context:` Although it took a while to assemble, the instructions are easy to follow. *overall* `Option 1:` This tv stand is worth purchasing for. `Option 2:` The dining room set is a quality item that will last for the only thing I will complain about was the fact that there was dust in the boxes. `Option 3:` The stool works well for our needs. `Option 4:` The desk took less than 1 hour to assemble and has a contemporary look with espresso-colored legs. `Select only from ["Option 1", "Option 2", "Option 3", "Option 4"]`	Option 4	Option 1	F

In-Context Learning	`Question: Which option represents the most plausible ending of the given context?`
	`Context: Both sides have in their own way proved themselves as bad as each other. in short`
	`Option 1: The problem is not the attitude of individual men but the spirit of the times.`
	`Option 2: The us government has been taken over by and both by corporate interests and political hacks.`
	`Option 3: You have a society that has been utterly corrupted by money and power.`
	`Option 4: Blacklisting worked against labour in wales, in london, and possibly, if he tries it in scotland, it will rebound there.`
	`Select only from ["Option 1", "Option 2", "Option 3", "Option 4"]`
	`Option 4`
	
	`Question: Which option represents the most plausible ending of the given context?`
	`Context: Any trinidadian wanting to vote must prove they maintain a residence there. because of that`
	`Option 1: No one living in the streets of burlington will ever be allowed to vote.`
	`Option 2: They are not eligible to vote.`
	`Option 3: And because the official election is open to all, the town hall will remain open for voting on election day.`
	`Option 4: Most trinidadians living here wont be able to vote.`
	`Select only from ["Option 1", "Option 2", "Option 3", "Option 4"]`
	`Option 4`
	
	`...`
	
	`Question:` Which option represents the most plausible ending of the given context?
	`Context:` Although it took a while to assemble, the instructions are easy to follow. *overall*
	`Option 1:` This tv stand is worth purchasing for.
	`Option 2:` The dining room set is a quality item that will last for the only thing I will complain about was the fact that there was dust in the boxes.
	`Option 3:` The stool works well for our needs.
	`Option 4:` The desk took less than 1 hour to assemble and has a contemporary look with espresso-colored legs.
	`Select only from ["Option 1", "Option 2", "Option 3", "Option 4"]`
		Option 1	Option 1	T

Table 32: Prompt example for DiscoSense.