Consistent Discourse-level Temporal Relation Extraction Using Large Language Models

Yi Fan and **Michael Strube**
Heidelberg Institute for Theoretical Studies
{yi.fan, michael.strube}@h-its.org

Abstract

Understanding temporal relations between events in a text is essential for determining its temporal structure. Recent advancements in large language models (LLMs) have spurred research on temporal relation extraction. However, LLMs perform poorly in zero-shot and few-shot settings, often underperforming smaller fine-tuned models. Despite these limitations, little attention has been given to improving LLMs in temporal structure extraction tasks. This study systematically examines LLMs' ability to extract and infer discourse-level temporal relations, identifying factors influencing their reasoning and extraction capabilities, including input context, reasoning process and ensuring consistency. We propose a three-step framework to improve LLMs' temporal relation extraction capabilities: context selection, prompts inspired by Allen's interval algebra (Allen, 1983), and reflection-based consistency learning (Shinn et al., 2024). Our results show the effectiveness of our method in guiding LLMs towards structured processing of temporal structure in discourse.

1 Introduction

Temporal relations describe the interaction between events along the temporal dimension, forming a crucial aspect of natural language understanding. Humans can encode temporal information in language, using various linguistic expressions to convey time-related concepts. Beyond mere extraction of explicit temporal markers such as before and after, humans leverage their linguistic competence and cognitive reasoning to interpret time in language, constructing a coherent mental timeline of events. This allows them to infer implicit temporal relations, even when direct temporal cues are absent or ambiguous (Zhang and Hudson, 2018; Klein, 1994, 2009). In contrast, LLMs encode textual information in a high-dimensional latent space, capturing intricate semantic patterns without an explicitly defined mechanism for understanding time.

Extracting temporal structures is essential in a wide range of NLP applications, including document summarisation (Ng et al., 2014), storyline construction (Do et al., 2012; Minard et al., 2015), and reading comprehension (Sun et al., 2018; Ning et al., 2020).

In recent years, LLMs' impressive text generation and processing abilities have attracted significant attention, directing extensive research into their capabilities in various ranges of tasks. Relation extraction has emerged as a common task for evaluating LLMs, with studies showing that LLMs often underperform smaller fine-tuned models in zero-shot or few-shot settings (Wei et al., 2024; Gao et al., 2023, 2024). This issue is particularly pronounced in temporal relation extraction, where LLMs struggle with zero-shot and few-shot settings (Hasegawa et al., 2024; Yuan et al., 2023; Chan et al., 2024). Three critical challenges in discourse-level temporal relation tasks remain underexplored: (1) **Context Selection**: Recent studies leveraging LLMs often rely on two common strategies—feeding the entire document or using sliding windows to extract context—both of which introduce significant noise. (2) **Performance**: Even in the fine-tuning setting, LLMs still exhibit a notable performance gap compared to the methods based on smaller-scale pre-trained language models (Roccabruna et al., 2024), such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019). (3) **Consistency**: While most recent work focuses on achieving high F1 scores, limited attention has been given to evaluating and improving model consistency. This study, motivated by prior work, is dedicated to tackling the above-mentioned challenges. As noted by Naik et al. (2019), extracting complete temporal structures from discourse-level texts is complex, labour-intensive, and costly for humans. Enhanced LLM performance in this task can support corpus

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 18605–18622
November 4-9, 2025

annotation, significantly reducing time and financial costs while advancing downstream NLP applications. We propose a three-step framework to improve LLMs' temporal relation extraction capabilities: context selection, prompts inspired by Allen's interval algebra (Allen, 1983), and reflection-based consistency learning (Shinn et al., 2024). Context selection minimises noise by focusing on relevant text and addressing LLMs' challenges with long contexts and distant event relations. Prompt engineering inspired by Allen's interval algebra enables structured reasoning about temporal attributes like start and end times. Reflection-based consistency learning iteratively identifies and corrects inconsistencies, teaching LLMs temporal coherence. The ablation study demonstrates that our context selection strategy is effective and computationally efficient. Besides, our novel self-reflection strategy improves the consistency of model predictions and offers a new perspective for addressing consistency challenges within the field. To the best of our knowledge, we are the first to use self-reflection to ensure consistency in the discourse-level temporal relation extraction task. Our method enhances the temporal relation extraction ability of LLMs and achieves performance that surpasses the state-of-the-art models.

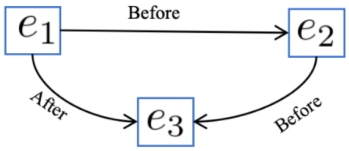
2 Related Work

Before the introduction of TDDiscourse (TDD) (Naik et al., 2019), most studies focused on local-level corpora like TimeBank (Pustejovsky et al., 2003) and TimeBank-dense (Cassidy et al., 2014). While recent advancements have improved temporal relation extraction on local corpora, progress on discourse-level datasets remains limited.

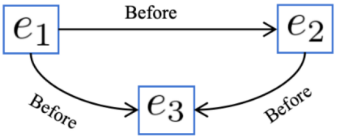
Early research relied on linguistic features such as dependency relations and grammatical cues (D'Souza and Ng, 2013; Mirza and Tonelli, 2014; Chambers et al., 2007). With neural networks, methods evolved to use BERT-R-GCN models, as demonstrated by TIMERS (Mathur et al., 2021), which set a strong foundation for subsequent studies (Liu et al., 2021; Wang et al., 2022a; Yuan et al., 2024). In contrast, SCS-EERE (Man et al., 2022) uses reinforcement learning to select relevant contexts, focusing only on those beneficial for temporal relation prediction, reducing noise compared to sliding window or all-event sentence approaches.

In discourse-level extraction, the distance between events may span ten or more sentences, but not all provide helpful temporal cues. Poor context selection introduces noise, hindering performance. Besides, LLMs are susceptible to the "Lost in the Middle" phenomenon (Liu et al., 2024), where information placed in the middle of a long document receives diminished attention. When event pairs requiring temporal relation prediction are positioned centrally in a document, providing the entire text as input causes further degraded performance.

Inconsistent temporal graph:



Consistent temporal graph:



I finished my homework. Then I went to the party. I met my friend at the party.

$$e_1 = finished, e_2 = went, e_3 = met$$

Figure 1: The above example illustrates an inconsistent temporal graph and a consistent temporal graph extracted from the same text. In the text, e_1 is before e_2 and e_2 is before e_3. Logically, if we know that e_1 is before e_2 and e_2 is before e_3, then we can easily infer that e_1 is before e_3. If the system produces an inconsistent temporal graph with e_1 after e_3, it means the system does not understand the input text well.

Consistency is a critical yet often overlooked aspect in discourse-level temporal relation extraction, which aims to construct coherent temporal graphs from text. Inconsistent model outputs undermine this goal by introducing contradictions among predicted relations, ultimately impairing document-level temporal understanding (see Figure 1). Prior work has addressed this issue via global constraints, such as formulating it as an ILP problem (Chambers and Jurafsky, 2008; Punyakanok et al., 2005; Zhao et al., 2012; Ning et al., 2017) or applying pre-training strategies like graph masking (Liu et al., 2021). Although such constraints may sometimes reduce local prediction performance, maintaining global consistency is essential for accurate temporal graph construction.

In the domain of LLMs, Yuan et al. (2023) eval-

uated ChatGPT[1] under a zero-shot setting for temporal relation extraction. However, performance was suboptimal, mainly due to an eight-sentence input limit, which constrained the model's ability to leverage broader context for reasoning. Wei et al. (2024) comprehensively investigated LLMs' performance on discourse-level event relation extraction tasks. Their study primarily focused on evaluating the models using various prompting strategies without fine-tuning. Their findings reveal two significant challenges. Firstly, the transitive rules are often violated, leading to inconsistently predicted temporal relations. Secondly, LLMs have difficulties capturing long-distance event dependencies, resulting in poor performance. Hu et al. (2025) explored fine-tuning approaches to enhance LLM performance on MAVEN-ERE (Wang et al., 2022b), a large-scale dataset encompassing temporal, causal, subevent, and coreference annotations, and MATRES (Ning et al., 2018). Despite the observed improvements, their transitivity chains did not extend to temporal relations, and thus, the issue of temporal consistency remains unresolved. While Wadhwa et al. (2023) assert that LLMs should be considered the default approach for relation extraction and serve as the benchmark for evaluating relation extraction tasks, there is currently no well-established LLM-based baseline for discourse-level event temporal relation extraction. To the best of our knowledge, existing studies have yet to give a standard method for improving LLM's performance in this task. Jain et al. (2023) conducted an in-depth investigation into LLMs' temporal reasoning capabilities, highlighting several key limitations, including difficulties in reasoning over long contexts, predicting future events, and understanding event temporal states. Chen et al. (2024) demonstrated that while LLMs struggle with maintaining logical consistency, their reasoning capabilities can be enhanced by explicitly incorporating logical constraints into the learning process. Therefore, critical gaps remain in better-guiding LLMs toward understanding implicit event temporality and improving their consistency in temporal relation extraction tasks.

3 Method

To improve LLMs' performance in temporal relation extraction, we propose a three-step framework: (1) Context selection, (2) Instruction and Chain-of-thought (CoT) (Wei et al., 2022) prompt design and (3) Consistency learning. Figure 2 shows an overview of our method.

3.1 Context Selection

Feeding the entire document into an LLM risks introducing noise, dispersing the model's attention and impairing its ability to focus on relevant temporal cues. To address this, we employ two distinct methods for selecting event-relevant contexts.

3.1.1 Entity-Based Discourse Segmentation

Our first step segments the input text into discourse segments based on entity coherence. Formally, the input document D consisting of a sentence list $S = [s_1, \ldots, s_i, \ldots, s_n]$ and an event list $E = [e_1, \ldots, e_i, \ldots, e_m]$, where n, m represent the total number of sentences and event mentions in the document D. We begin by applying neuralcoref[2] to perform coreference resolution across the document. Next, we extract each sentence's subject, objects (both direct and indirect), and all other noun and noun phrases, forming an entity list $SE_i = [se_{i1}, se_{i2}, ..., se_{im}]$, where i is the sentence number and m is the number of entities selected in sentence i. We measure semantic similarity between consecutive sentences' extracted entities to determine discourse segmentation. Assuming the current sentence is S_i, and the following sentence is S_j, we compute the cosine similarity between each pair of entities in the two lists, SE_i and SE_j, and select the maximum value Sim_{ij}. If Sim_{ij} exceeds a predefined threshold γ, then S_i and S_j are considered part of the same discourse segment; otherwise, they are assigned to different segments. The segmentation algorithm follows the procedure outlined in Appendix A. Upon completion, the document is divided into multiple discourse segments, $DS = [ds_1, ds_2, ..., ds_k]$, where k is the total number of discourse segments in document D. When constructing the input for temporal relation extraction, if event e_1 appears in ds_i and event e_2 in ds_j, we concatenate ds_i and ds_j as input to the model. If event pairs are within the same discourse segment ($ds_i = ds_j$), that segment is directly selected as context without concatenation.

3.1.2 LLM-Guided Context Selection

Each sentence is annotated in the input document with a numerical index and provides explicit instruction to the LLM. Then, the model identifies

[1] https://openai.com/blog/chatgpt

[2] https://spacy.io/universe/project/neuralcoref

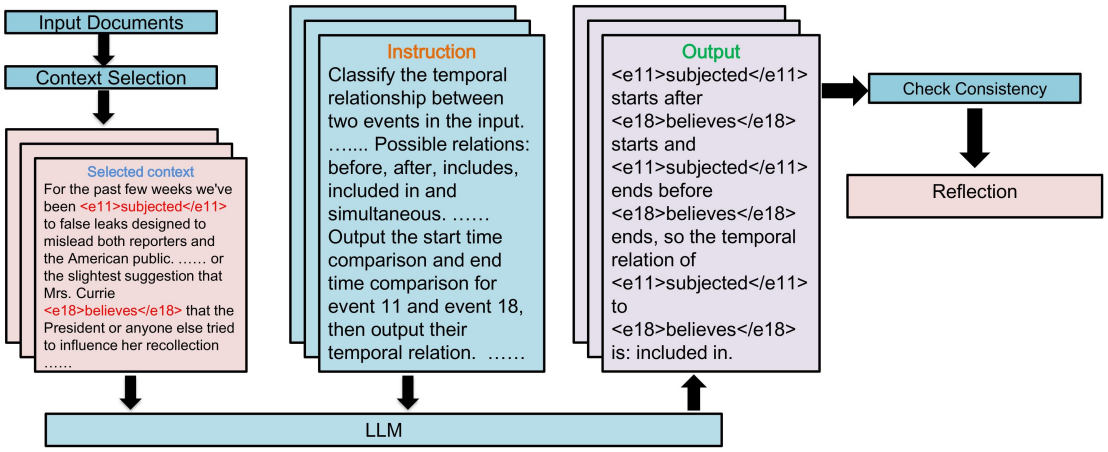


Figure 2: The input documents are split into different discourse segments in step 1. If the target event pair is *e11: subjected* and *e18: believes*, then the discourse segments, including those two events, will be concatenated and serve as the context. Then, according to Allen's interval algebra, in step 2, the model has to compare the start time and end time of two events to give the prediction. After the first round of fine-tuning is done, all the predictions are examined for consistency and form the new training data for self-reflection in step 3.

the sentences most relevant to each pair of events. This allows the model to actively select and filter the most important portions of the text for temporal relation extraction for a specific event pair. The instructions can be found in Appendix C.

3.1.3 Highlighting Target Events in Input

In order to accurately locate the target events within the extracted text, event mentions are highlighted using angle brackets to ensure the model explicitly recognises their positions within the input. For instance, the original event string *taking* and *said* in the input document are replaced by `<e1>taking</e1>` and `<e2>said</e2>`. This input design improves the model's ability to focus on temporal relations for a specific event pair rather than being distracted by irrelevant content. For instance, if the input has multiple words that are the same as our target events, the model's prediction process is affected.

3.2 Instruction and Prompt Design

Inspired by Allen's interval algebra (Allen, 1983), and insights from Cohen and Bar (2023), we propose a Chain-of-Thought (CoT) prompting framework that structures temporal reasoning into a step-by-step process. Rather than directly predicting the temporal order between two events, our approach first decomposes the task into a granular comparison of event start and end points. In this framework, the model must produce a full-sentence response articulating its reasoning process, explicitly comparing the start and end times of event pairs. For example, if the temporal relation of *Event 1* to *Event 2* is *before*, then the expected output is: *Event 1* starts before *Event 2* starts, and *Event 1* ends before *Event 2* starts, so the temporal relation of *Event 1* to *Event 2* is *before*. By structuring the task in this way, we reinforce a systematic inference process, encouraging the model to anchor events on a temporal axis and derive implicit temporal boundaries. This method enhances the model's ability to establish event sequences by reasoning over temporal intervals rather than relying solely on direct order inference. For a comprehensive breakdown of reasoning steps, refer to Appendix C.

3.3 Consistency Learning

Previous studies have highlighted the inconsistency of LLMs in temporal relation prediction tasks. A widely adopted method to address this issue is Integer Linear Programming (ILP) (Chambers and Jurafsky, 2008), which effectively enforces consistency in encoder-based models. ILP relies on the availability of probabilistic predictions for all labels to optimise outcomes. However, this approach is not directly applicable to LLMs, as their predictions are typically generated as sentences rather than discrete probabilities for each label. Inspired by Shinn et al. (2024) and Chen et al. (2024), we propose a self-reflection approach that enables LLMs to learn consistency iteratively to address this limitation, shown in Figure 3. The first fine-tuning phase in our framework involves standard temporal relation prediction, where the model pre-

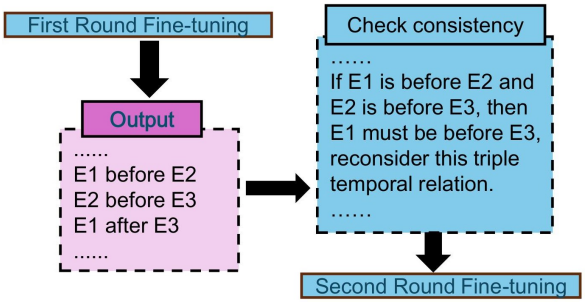


Figure 3: According to the prediction generated after the first round of fine-tuning, we check the consistency within it. If we find inconsistent triples, we will construct the new input only for the test datasets indicating the inconsistent triples and reasons. In this process, the new input does not include the true label, only pointing out the inconsistent fact. Then, our model fine-tunes again with the new input in the second round in order to improve consistency.

dicts the relation between event pairs. After this phase, we identify any inconsistent event triples in the model's predictions. These inconsistent triples form the basis for a second fine-tuning phase. In the second round, the model is provided with its predictions from the first round and explicit instructions highlighting and explaining the inconsistencies identified. The input for this phase includes the relevant event pairs and the corresponding sentences or discourse segments related to the target events. Notably, the true labels are not revealed during this process. The model must re-evaluate its inconsistent predictions and reason about them to generate updated predictions. This iterative reflection mechanism aims to teach the model the concept of consistency through self-reflection and correction. By repeatedly engaging in this reflective process, the model is expected to internalise the principle of consistency, ultimately improving its performance in maintaining logical coherence across temporal relation predictions.

4 Experiments

4.1 Dataset

We use three discourse-level corpora in this work: TDD-Man (Naik et al., 2019), MAVEN-ERE (Wang et al., 2022b) and TimeBank (Pustejovsky et al., 2003). TDD-Man is annotated by experts, aiming to provide a complete temporal structure in the document. We focus on the temporal relations for MAVEN-ERE, and our data split follows Hu et al. (2025). Since TimeBank lacks an official training/test split, we adopt the partitioning strategy from TimeBank-Dense (Cassidy et al., 2014), and simplify the original annotations to align with the TimeBank-Dense label schema.

4.2 Experimental Settings

We use the open-source model LLaMA (Grattafiori et al., 2024), specifically Llama-3.3-70B-Instruct and Llama-3.1-8B-Instruct. We use the Lora technique (Hu et al., 2022) to fine-tune LLaMA. The Lora alpha is set from [8, 16], and the Lora rank is set from [16, 32]. The learning rate we use is from [3e-5, 5e-5]. We train our model for three epochs at most. In the context selection step, we use pretrained Sentence-BERT (Reimers and Gurevych, 2019) to extract the representation of extracted entities and then compute the cosine similarity. We use the grid search method for the threshold to select the optimal γ from [0.7, 0.75, 0.8, 0.85, 0.9]. All the above-mentioned models can be accessed freely through Huggingface[3]. For comparison, we also employ the topic modelling tool BERTopic (Grootendorst, 2022) to retrieve sentences that share the same topic as those containing the target events. These sentences are used as input to compare against our proposed context selection strategy in the ablation study. We chose the micro F1-Score for the evaluation metric, following previous work. We evaluate the consistency within the output of our model by using the method in Naik et al. (2019). Specifically, we examine each triple of events (e_1, e_2, e_3) to determine whether temporal relations exist between e_1 and e_2, e_2 and e_3, and e_1 and e_3. If such relations exist, we apply transitivity rules to check for consistency among the three events. For instance, if e_1 occurs before e_2 and e_2 occurs before e_3, then e_1 must occur before e_3. If this condition is met, we identify the temporal relations among the three events as consistent. Then, the consistency rate is computed by using the number of consistent triples to divide the number of triples within the prediction. We reproduced CP-TRE (Yuan et al., 2024) for comparison purposes. And all experimental results are averaged over five runs.

4.3 Experimental Results

In Table 1, we select five models based on BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2019), etc., as reference points. We also compare the performance of two zero-shot and few-shot LLMs and

[3] https://huggingface.co/

Method	P	R	F1	Cons.
BERT (Liu et al., 2021)	39.9	39.9	39.9	51.7
BERT+ILP (Liu et al., 2021)	39.2	39.2	39.2	53.8
UCGraph (Liu et al., 2021)	44.5	42.3	43.4	55.9
SCS-EERE (Man et al., 2022)	-	-	51.1	-
CPTRE (Yuan et al., 2024)	56.5	56.5	56.5	65.9
(Yuan et al., 2023)	26.8	22.3	24.3	-
(Chan et al., 2024)	-	-	16.8	-
(Zhang et al., 2024)	-	-	52.6	-
Ours$_{Llama-3.1-8B-Instruct}$	55.3	55.3	55.3	91.3
Ours$_{Llama-3.3-70B-Instruct}$	57.9	57.9	57.9*	93.6

Table 1: Performance comparison of different methods on the TDD-Man dataset. 'Ours' indicates using entity-based context selection, prompt engineering inspired by Allen's interval algebra and self-reflection. The last five results are based on large language models, while the first five results are based on pre-trained language models such as BERT or RoBERTa. * indicates statistical significance tested at a p-value of 0.05 compared with CPTRE (Yuan et al., 2024). All the results above, except our models and CPTRE, are copied from the original paper. Cons. means the consistency rate of the model. If the original paper reported this performance, we also show it above.

Method	F1 Score	Cons.
Majority	84.8	100
Llama-3.1-8B-Instruct$_{zs}$	10.2	36.9
Llama-3.3-70B-Instruct$_{zs}$	76.3	87.1
Llama-3.1-8B-Instruct$_{ft}$	88.8	90.1
Llama-3.3-70B-Instruct$_{ft}$	89.2	90.3
Ours$_{Llama-3.1-8B-Instruct}$	91.5	91.8
Ours$_{Llama-3.3-70B-Instruct}$	91.9*	92.7

Table 2: Results on the MAVEN-ERE dataset. The subscript *zs* means zero-shot setting, and *ft* means fine-tuning setting without our proposed strategies. Majority means the majority baseline. * indicates statistical significance tested at a p-value of 0.05 compared with the fine-tuning setting.

Method	F1 Score	Cons.
CPTRE (Yuan et al., 2024)	61.1	51.7
Llama-3.1-8B-Instruct$_{zs}$	11.3	50
Llama-3.3-70B-Instruct$_{zs}$	27.9	55.5
Llama-3.1-8B-Instruct$_{ft}$	51.7	42.0
Llama-3.3-70B-Instruct$_{ft}$	62.2	46.6
Ours$_{Llama-3.1-8B-Instruct}$	59.6	82.7
Ours$_{Llama-3.3-70B-Instruct}$	66.0*	87.8

Table 3: Results on TimeBank. * indicates statistical significance tested at a p-value of 0.05 compared with CPTRE. The CPTRE result above is based on our reproduction.

one fine-tuned LLM. Our fine-tuned model, built upon Llama-3.3-70B-Instruct, surpasses the current state-of-the-art (Yuan et al., 2024), achieving both a high F1 score and strong global consistency. The version based on Llama-3.1-8B-Instruct also demonstrates competitive performance. Note that a high F1 score enables high consistency, but high consistency does not necessarily guarantee a high F1 score. Besides, despite utilising a large-scale model, the F1 score gain over previous state-of-the-art models, which rely on smaller pre-trained language models, remains modest. However, the gain in consistency is substantial, indicating that our model exhibits stronger logical consistency. Further analysis is provided in the Discussion.

For MAVEN-ERE, our setup provides the event pairs directly as input, and the model performs only relation classification. Table 2 provides a majority baseline since the label imbalance is severe in MAVEN-ERE. Since most data is labelled as `before`, fine-tuning settings can have relatively high consistency. We offer more details in Appendix D. We mainly compare our models with the fine-tuned LLaMA baseline we provide. We can observe that our model performs better than the fine-tuned baseline, showing the effectiveness of our strategy.

In Table 3, our method outperforms CPTRE and the fine-tuned baseline in TimeBank, especially in consistency. These results state our proposed strategy's advantages in improving classification accuracy and producing more logically coherent temporal relation predictions.

5 Discussion

5.1 Ablation Studies

In this subsection, we discuss the findings from ablation studies. For simplicity, we focus on reporting the experimental results of our model using Llama-3.1-8B-Instruct on TDD-Man, as the model's performance based on Llama-3.1-8B-Instruct and Llama-3.3-70B-Instruct settings on the tested corpora is similar. Table 4 shows the ablation experiments' results.

Model setting	F1 Score	Cons.
ECS+PE+SR	55.3	91.3
MS+PE+SR	47.8	88.8
BT+PE+SR	53.3	89.4
ECS+PE+PA	55.2	84.5
PE+SR	47.5	88.4
ECS+SR	52.9	88.3
ECS+PE	55.8	82.6
MS+PE	50.3	81.9

Table 4: Ablation study results with Llama-3.1-8B-Instruct. ECS: entity-based context selection, PE: prompt engineering inspired by Allen's interval algebra, SR: self-reflection, MS: model selected context, BT: BERTopic context selection, PA: predict again for inconsistent predictions.

The first three rows of Table 4 demonstrate that our proposed entity-based context selection method significantly outperforms the context selected by the LLM and BERTopic. To better understand this improvement, we analysed the contexts selected by the model and BERTopic. The analysis reveals that in some cases, the model includes low-relevance sentences, resulting in excessively long inputs. Such input length negatively impacts the model's ability to predict temporal relations, as evidenced by the performance drop under full-document input (Table 4, Row 5), where excessive noise leads to degraded performance. Our method mitigates this issue by selectively including only sentences from the discourse segments containing the target events, limiting the input length to 20 sentences. Moreover, while the contexts selected by the model and BERTopic tend to be sparse and non-contiguous, our approach ensures continuity, providing more coherent and complete information surrounding the event pair. Our experimental results emphasise the critical role of context selection in discourse-level temporal relation extraction, aligning with previous work (Jain et al., 2023; Wei et al., 2024; Liu et al., 2024). While our method partially mitigates this issue, future work should develop more effective context selection strategies or explore stronger topic-aware models, such as TopicGPT (Pham et al., 2024). Row 4 reports results obtained by re-prompting the LLM with inconsistent event predictions, without additional fine-tuning. This step aims to disentangle the contribution of inconsistency detection and self-reflection to consistency improvement. While inconsistency detection alone leads to a slight improvement in consistency, the most critical factor is to learn what constitutes consistency (Rows 1 and 4 in Table 4). This finding aligns with Chen et al. (2024).

Besides, when our CoT prompt is removed, the model is required to predict a label directly from five temporal relations. Row 6 shows that guiding the model using Allen's Interval Algebra improved its ability to extract temporal relations. However, whether the model genuinely understands temporal information behind the event—specifically, the start and end times of events—remains an open question, which we further explore in the Label Imbalance subsection.

The results in the last two rows reveal that leaving out self-reflection fine-tuning improves the model's F1 score at the expense of a decline in consistency. This aligns with (Liu et al., 2021), who observed that enhancing consistency through ILP or other constraints often reduces F1 performance. Additionally, after fine-tuning without self-reflection, our model still yields moderate consistency. Future work should aim to jointly improve both the F1 score and temporal consistency, striking a balance between predictive accuracy and structural coherence.

5.2 Label Imbalance

Figure 4 shows the confusion matrix of our best-performing model. We observe that the model performs better on labels with a higher number of instances, which aligns with our expectations. We can observe that the most significant confusion occurs between the labels "before" and "includes". The model predicts 114 instances of "includes" instead of true "before" labels, indicating difficulty distinguishing between these temporal relations. In this case, the model correctly compared the starting time of both events but wrongly estimated the events' end time. Also, the model is confused when the duration of Event 1 includes Event 2 and usu-

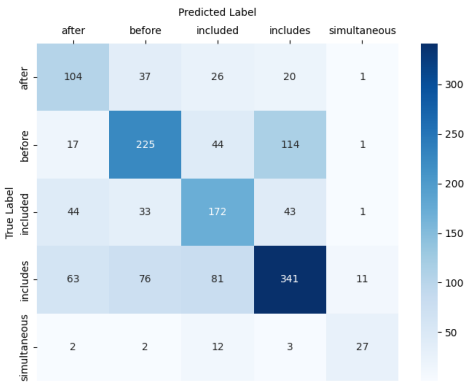


Figure 4: The confusion matrix of our model in TDD-Man's test set.

ally predicts it as "is_included". That erroneous prediction indicates that the model misunderstands the duration of both events. One reason is that LLMs do not acquire the real-world knowledge of one event's duration (Qiu et al., 2024). This problem can be alleviated if LLMs are fine-tuned properly (Xiong et al., 2024). Our results show that LLMs struggle to extract the temporal information conveyed by stative verbs relating to opinions, feelings, perceptions, etc., such as believe, relate, and seem. Since some stative verbs inherently lack clear temporal boundaries, it is difficult to determine their precise duration. Future work can combine temporal relation extraction with temporal reasoning tasks to fine-tune the model. Given that existing corpora are annotated across entire documents rather than single sentences, achieving a balance among labels is nearly impossible.

5.3 Sentence Distance

Sentence distance	Number of data	Accuracy
0-2	245	57.02
3-5	472	53.39
6-8	381	61.89
9-11	231	62.34
12-14	101	66.34
15-17	43	62.79
18-20	21	38.1
21-23	6	100

Table 5: Statistics of intervals between sentences containing events in TDD-Man's test set.

In Table 5, sentence distance refers to the number of sentences between the two events to be predicted. Our model performs poorly in the 0–5 range. Initially, we suspected this might be due to our context selection strategy. However, further analysis reveals that even when using the full context as input, performance in this range remained suboptimal and is, in fact, worse than with our selection-based input. This suggests that LLMs when trained on discourse-level temporal relation corpora, struggle to learn short-distance temporal dependencies. These findings indicate that future work focusing on long-range event pairs should not overlook the challenges of short-distance temporal relation extraction. We offer more details in Appendix E.

6 Conclusions

Our three-step strategy effectively addresses key limitations of LLMs in temporal relation extraction. Experimental results demonstrate that our model achieves a high F1 score while maintaining a strong level of temporal consistency, which is often overlooked in previous research.

Our ablation study highlights that context selection is essential for discourse-level tasks in LLMs. Future research on LLMs handling long-text inputs should prioritise context selection techniques to improve performance and mitigate noise. Additionally, while our prompt inspired by Allen's Interval Algebra encourages the model to reason about implicit temporal information, our model still struggles to accurately predict event duration in many cases. Future work should explore the integration of temporal reasoning tasks with temporal relation extraction, enabling models to better capture implicit temporal information encoded in the events. Besides, our reflection-based learning strategy significantly enhances prediction consistency. In tasks requiring temporal coherence, researchers should not solely focus on F1 scores and accuracy but also emphasise ensuring logical consistency in model predictions.

Finally, although our model achieves high F1 scores and strong consistency, its performance does not meet our expectations. Given that our model utilises significantly larger parameters than previous works, the limited performance improvement suggests LLMs remain relatively weak at capturing temporal structures in text. This indicates that increasing the model scale does not necessarily yield substantial gains in temporal relation extraction, emphasising the need for more effective strategies to enhance LLMs' temporal understanding.

Limitations

Since TDD-Man, MAVEN-ERE, and TimeBank predominantly consist of news articles, the label distribution inherently reflects the characteristics of this genre. However, this presents a limitation: the label distribution observed in news texts may not generalise well to other text genres, such as narrative or procedural texts. Consequently, models trained on such corpora lack the ability to generalise across different types of texts. Moreover, we have run experiments on all existing discourse-level corpora that we can use, except TIMELINE (Alsayyahi and Batista-Navarro, 2023). We cannot access its original text, which requires a specific membership to download it. Although extensive empirical results demonstrate that our context selection strategy outperforms full-document input, we observe that some irrelevant sentences are still included in the selected context. Precisely identifying sentences that contain cues essential for predicting the temporal relation of each event pair remains challenging. Future work should address this limitation by exploring more effective context selection strategies.

Ethics Statement

This study uses the Llama model from Meta exclusively for this specific research task, adhering to Meta's Acceptable Use Policy. The TDD-Man corpus and MAVEN-ERE used in our experiments are publicly available and intended solely for research purposes. However, as the corpus consists of news reports, it may contain inaccuracies or potentially harmful content. The presence of such content does not reflect the opinions of the authors.

Acknowledgements

The authors would like to thank the anonymous reviewers for their comments. This work has been funded by the Klaus Tschira Foundation, Heidelberg, Germany.

References

James F Allen. 1983. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843.

Sarah Alsayyahi and Riza Batista-Navarro. 2023. TIMELINE: Exhaustive annotation of temporal relations supporting the automatic ordering of events in news articles. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16336–16348, Singapore. Association for Computational Linguistics.

Regina Barzilay and Mirella Lapata. 2005. Modeling local coherence: An entity-based approach. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 141–148, Ann Arbor, Michigan. 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 (Volume 2: Short Papers)*, pages 501–506, Baltimore, Maryland. Association for Computational Linguistics.

Nathanael Chambers and Daniel Jurafsky. 2008. Jointly combining implicit constraints improves temporal ordering. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 698–706, Honolulu, Hawaii. Association for Computational Linguistics.

Nathanael Chambers, Shan Wang, and Dan Jurafsky. 2007. Classifying temporal relations between events. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 173–176, Prague, Czech Republic. Association for Computational Linguistics.

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

Meiqi Chen, Yubo Ma, Kaitao Song, Yixin Cao, Yan Zhang, and Dongsheng Li. 2024. Improving large language models in event relation logical prediction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9451–9478, Bangkok, Thailand. Association for Computational Linguistics.

Omer Cohen and Kfir Bar. 2023. Temporal relation classification using Boolean question answering. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1843–1852, Toronto, Canada. Association for Computational Linguistics.

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

Quang Do, Wei Lu, and Dan Roth. 2012. Joint inference for event timeline construction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 677–687, Jeju Island, Korea. Association for Computational Linguistics.

Jennifer D'Souza and Vincent Ng. 2013. Classifying temporal relations with rich linguistic knowledge. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 918–927, Atlanta, Georgia. Association for Computational Linguistics.

Chufan Gao, Xulin Fan, Jimeng Sun, and Xuan Wang. 2024. PromptRE: Weakly-supervised document-level relation extraction via prompting-based data programming. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 132–145, Bangkok, Thailand. Association for Computational Linguistics.

Jinglong Gao, Xiao Ding, Bing Qin, and Ting Liu. 2023. Is ChatGPT a good causal reasoner? a comprehensive evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11111–11126, Singapore. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil,

Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Kimihiro Hasegawa, Nikhil Kandukuri, Susan Holm, Yukari Yamakawa, and Teruko Mitamura. 2024. Formulation Comparison for Timeline Construction using LLMs. *ArXiv*, abs/2403.00990.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Zhilei Hu, Zixuan Li, Xiaolong Jin, Long Bai, Jiafeng Guo, and Xueqi Cheng. 2025. Large language model-based event relation extraction with rationales. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7484–7496, Abu Dhabi, UAE. Association for Computational Linguistics.

Raghav Jain, Daivik Sojitra, Arkadeep Acharya, Sriparna Saha, Adam Jatowt, and Sandipan Dandapat. 2023. Do language models have a common sense regarding time? revisiting temporal commonsense reasoning in the era of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6750–6774, Singapore. Association for Computational Linguistics.

W. Klein. 1994. *Time in Language*. Germanic linguistics. Routledge.

Wolfgang Klein. 2009. How time is encoded. In Wolfgang Klein and Ping Li, editors, *The Expression of Time*, pages 39–82. Mouton de Gruyter, Berlin, New York.

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

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy

Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv*, abs/1907.11692.

Hieu Man, Nghia Trung Ngo, Linh Ngo Van, and Thien Huu Nguyen. 2022. Selecting optimal context sentences for event-event relation extraction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11058–11066.

Puneet Mathur, Rajiv Jain, Franck Dernoncourt, Vlad 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 (Volume 2: Short Papers)*, pages 524–533, Online. Association for Computational Linguistics.

Anne-Lyse Minard, Manuela Speranza, Eneko Agirre, Itziar Aldabe, Marieke van Erp, Bernardo Magnini, German Rigau, and Rubén Urizar. 2015. SemEval-2015 task 4: TimeLine: Cross-document event ordering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 778–786, Denver, Colorado. Association for Computational Linguistics.

Paramita Mirza and Sara Tonelli. 2014. Classifying temporal relations with simple features. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 308–317, Gothenburg, Sweden. Association for Computational Linguistics.

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

Jun-Ping Ng, Yan Chen, Min-Yen Kan, and Zhoujun Li. 2014. Exploiting timelines to enhance multi-document summarization. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 923–933, Baltimore, Maryland. Association for Computational Linguistics.

Qiang Ning, Zhili Feng, and Dan Roth. 2017. A structured learning approach to temporal relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1027–1037, Copenhagen, Denmark. Association for Computational Linguistics.

Qiang Ning, Hao Wu, Rujun Han, Nanyun Peng, Matt Gardner, and Dan Roth. 2020. TORQUE: A reading comprehension dataset of temporal ordering questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1158–1172, Online. Association for Computational Linguistics.

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

Chau Minh Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2024. TopicGPT: A prompt-based topic modeling framework. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2956–2984, Mexico City, Mexico. Association for Computational Linguistics.

Vasin Punyakanok, Dan Roth, Wen-tau Yih, and Dav Zimak. 2005. Learning and inference over constrained output. In *IJCAI*, volume 5, pages 1124–9.

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

Yifu Qiu, Zheng Zhao, Yftah Ziser, Anna Korhonen, Edoardo Ponti, and Shay Cohen. 2024. Are large language model temporally grounded? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7064–7083, Mexico City, Mexico. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Conference on Empirical Methods in Natural Language Processing*.

Gabriel Roccabruna, Massimo Rizzoli, and Giuseppe Riccardi. 2024. Will LLMs replace the encoder-only models in temporal relation classification? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20402–20415, Miami, Florida, USA. Association for Computational Linguistics.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2024. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36.

Yawei Sun, Gong Cheng, and Yuzhong Qu. 2018. Reading comprehension with graph-based temporal-casual reasoning. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 806–817, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Somin Wadhwa, Silvio Amir, and Byron Wallace. 2023. Revisiting relation extraction in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15566–15589, Toronto, Canada. Association for Computational Linguistics.

Liang Wang, Peifeng Li, and Sheng Xu. 2022a. DCT-centered temporal relation extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2087–2097, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Xiaozhi Wang, Yulin Chen, Ning Ding, Hao Peng, Zimu Wang, Yankai Lin, Xu Han, Lei Hou, Juanzi Li, Zhiyuan Liu, Peng Li, and Jie Zhou. 2022b. MAVEN-ERE: A unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 926–941, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

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

Siheng Xiong, Ali Payani, Ramana Kompella, and Faramarz Fekri. 2024. Large language models can learn temporal reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10452–10470, Bangkok, Thailand. Association for Computational Linguistics.

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

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

Meng Zhang and Judith A Hudson. 2018. The development of temporal concepts: Linguistic factors and cognitive processes. *Frontiers in Psychology*, 9:2451.

Xiaobin Zhang, Liangjun Zang, Qianwen Liu, Shuchong Wei, and Songlin Hu. 2024. Event temporal relation extraction based on retrieval-augmented on llms. *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Ran Zhao, Quang Do, and Dan Roth. 2012. A robust shallow temporal reasoning system. In *Proceedings of the Demonstration Session at the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 29–32, Montréal, Canada. Association for Computational Linguistics.

Appendix

A Entity-based Discourse Segmentation Algorithm

The algorithm we use for entity-based discourse segmentation is shown in Algorithm 1.

Algorithm 1 Entity-Based Discourse Segmentation

Require: Input document D as a sentence list $S = \{s_1, s_2, \ldots, s_n\}$, predefined similarity threshold γ
Ensure: Discourse segments DS
Initialize $DS \leftarrow \emptyset$
Apply coreference resolution on D using neuralcoref
for each sentence $s_i \in S$ **do**
    Extract the subject, direct objects, indirect objects, and all noun and noun phrases from s_i
    Create an entity list $SE_i = \{se_{i1}, se_{i2}, ..., se_{im}\}$ for sentence s_i
end for
Initialize a new discourse segment $current\_segment \leftarrow \{s_1\}$
for $i = 2$ to n **do**
    Compute cosine similarity $sim(SE_{i-1}, SE_i)$ between entities in entity sets of s_{i-1} and s_i
    if $max(sim(SE_{i-1}, SE_i)) \geq \gamma$ **then**
        Add s_i to $current\_segment$
    else
        Append $current\_segment$ to DS
        Start a new segment $current\_segment \leftarrow \{s_i\}$
    end if
end for
Append the final $current\_segment$ to DS
return DS

In temporal relation extraction, most events are verbs, as verbs are closely tied to the subject and object within a sentence. We hypothesise that each

discourse unit centres around a coherent topic involving its associated participants, typically, interactions or developments concerning a specific subject-object pair, such as Person A and Person B. Therefore, when the algorithm transitions between discourse units, it typically signals a shift in topic or participating entities. Based on this observation, we argue that providing only the discourse units containing Event A and Event B is sufficient for relation classification. Other discourse units are likely to discuss unrelated or only weakly related subjects, introducing noise rather than useful context.

To support this, we analyse MAVEN-ERE training and validation sets and find 56,377 instances where the sentence distance between two events exceeds 15, with an average of 668 tokens per instance. The longest sentence distance is 74, and the maximum token count reaches 2,126. Similarly, in TDD-Man, 1,192 such instances are observed, with an average of 672 tokens, a maximum sentence distance of 63, and up to 1,679 tokens in a single instance. In TimeBank, the maximum sentence distance is 76, with 2394 tokens between two events. Prior discourse-level temporal relation extraction work has typically relied on full-text input or sliding window strategies. However, our ablation studies suggest that such approaches introduce substantial noise, often degrading model performance. Moreover, studies on entity-based coherence (Grosz et al., 1995; Barzilay and Lapata, 2005) suggest that excessive shifts in sentence focus often lead to incoherent discourse. However, the input documents in our setting are formal news articles written by professionally trained authors, whose coherence can reasonably be assumed. Consequently, the key information required to determine the temporal relation between two events is unlikely to be scattered across the document, but rather concentrated around the events themselves. While our method may not represent the optimal solution, we hope it offers insights for future research. As in the ablation studies, future work could use more powerful topic-aware models to explore, since the BERTopic (Grootendorst, 2022) also brings a performance gain in this task. Besides, in our work, we could not try with TopicGPT (Pham et al., 2024) due to the research funding limit.

Additionally, the coreference resolution component replaces pronouns with their corresponding entities to facilitate downstream semantic similarity computations. For example, multiple pronouns such as he or she may appear as subjects within the same document; resolving these references to specific named entities enables more accurate and meaningful semantic comparisons.

B Model Selection Explanation

As illustrated in Figure 6, we first index all sentences in the input document corresponding to a target event pair. The indexed document is then provided to the model, which is prompted to identify the informative sentences for predicting the temporal relation between the two events. Then, we reconstruct the input containing only the relevant context based on the selected sentence indices by the model and the event containing sentences. Figure 7 shows one prediction generated by the model. In Figure 7, the model selects 0, 1, 4, 9 as the context, and the event containing sentences are 0 and 7. Therefore, the input sentence indices are 0, 1, 4, 7 and 9.

C Our Prompts and Instructions

We show an example of using LLM to extract the context based on the input event pair in Figure 6, and an example of input text is shown in Figure 8. The instruction we use to fine-tune our model is shown in Figure 9. The instruction of self-reflection for the second round of fine-tuning is shown in Figure 10. In the setting where we do not use our prompts inspired by Allen's interval algebra, our instruction is shown in Figure 11. The true labels we adapted from our prompts are shown in Table 6. Figure 5 illustrates the comparison and definition of the start and end points of the event pairs corresponding to the five labels in TDD-Man on the time axis.

D More Details About The Dataset

We follow the same data split strategy for MAVEN-ERE in Hu et al. (2025). Since the test set of MAVEN-ERE is not publicly available, we randomly split the original training data into training and validation sets with an 80/20 ratio, and repeat this partition five times. Thus, we have five different training and validation sets. Then we use the original validation set as our test set. All reported results are averaged over the five runs on different training sets. As distinguished from our setup, the experimental setting in Hu et al. (2025) requires the model to take the entire document as input and jointly identify event pairs with temporal relations

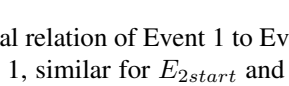
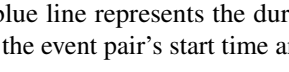
Label	Definition	Time Axis
Before	$E_{1start} > E_{2start} \wedge E_{1end} > E_{2start}$	
After	$E_{1start} < E_{2start} \wedge E_{1start} < E_{2end}$	
Includes	$E_{1start} > E_{2start} \wedge E_{1end} > E_{2end}$	
Is_included	$E_{1start} < E_{2start} \wedge E_{1end} < E_{2end}$	
Simultaneous	$E_{1start} = E_{2start} \wedge E_{1end} = E_{2end}$	

Figure 5: Five labels in TDDiscourse indicate the temporal relation of Event 1 to Event 2.E_{1start} means the start time of Event 1 and E_{1end} means the end time of Event 1, similar for E_{2start} and E_{2end}. In the time axis, the orange line represents the duration of Event 1, and the blue line represents the duration of Event 2. Following Allen's interval algebra, the model is required to compare the event pair's start time and end time, according to the definition above, to make the prediction.

Original label	Expected Output
Before	Event 1 starts before Event 2 starts and Event 1 ends before Event 2 starts, so the temporal relation of Event 1 to Event 2 is: before.
After	Event 1 starts after Event 2 and Event 1 starts after Event 2 ends, so the temporal relation of Event 1 to Event 2 is: after.
Includes	Event 1 starts before Event 2 starts and Event 1 ends after Event 2 ends, so the temporal relation of Event 1 to Event 2 is: included.
Is_included	Event 1 starts after Event 2 starts and Event 1 ends before Event 2 ends, so the temporal relation of Event 1 to Event 2 is: included in.
Contains	Event 1 starts before Event 2 starts and Event 1 ends after Event 2 ends, so the temporal relation of Event 1 to Event 2 is: contains.
Overlap	Event 1 starts before Event 2 starts, Event 1 ends after Event 2 starts and Event 1 ends before Event 2 ends, so the temporal relation of Event 1 to Event 2 is: overlap.
Begins-on	Event 1 starts at the same time as Event 2 starts and Event 1 ends before Event 2 ends, so the temporal relation of Event 1 to Event 2 is: begins-on.
Ends-on	Event 1 starts after Event 2 starts and Event 1 ends at the same time as Event 2 ends, so the temporal relation of Event 1 to Event 2 is: ends-on.
Simultaneous	Event 1 starts at the same time as Event 2 and Event 1 ends at the same time as Event 2, so the temporal relation of Event 1 to Event 2 is: simultaneous.

Table 6: Table of labels for prompts inspired by Allen's interval algebra

before classifying their relation types. Besides, their model is trained to predict not only temporal relations but also causal and subevent relations, etc., which differs from the primary focus of our work. In our setup, the model is provided with the complete document and all candidate event pairs, and is tasked solely with predicting each pair's temporal relation. Although MAVEN-ERE is a large-scale dataset, it suffers from severe label imbalance—specifically, the before label accounts for 683,581 out of 792,445 instances in the training set. As a result, a majority baseline achieves as high as 84.8 F1 on the test set.

E More Details About The Results

We reproduce the CPTRE results on TDD-Man and TimeBank using the same five-run evaluation protocol. We didn't reproduce CPTRE in MAVEN-ERE because MAVEN-ERE does not contain the features required by the model, such as linguistic features, document creation time, etc. Following the methodology in TIMERS (Mathur et al., 2021), we employ the Wilcoxon signed-rank test to assess statistical significance, and apply this test consistently across all comparisons.

Results labelled as fine-tuning are obtained by fine-tuning the model with a simple prompt (Figure 11), where the entire document is provided as input and the model is trained to predict the gold labels directly. Results labelled as zero-shot are derived using a pure in-context learning setting, without any parameter updates. These two settings clarify the baseline setup and demonstrate the effectiveness of our proposed strategies, while also providing insight into the models' pre-trained knowledge. Additional results under both zero-shot and fine-tuning settings on the TDD-Man dataset are reported in Table 8.

As shown in Table 7, the model performs poorly in the 0–5 sentence distance range under the zero-shot setting and full-context input setting, worse than when using our context selection strategy (in Table 5). This result clearly demonstrates the advantage of our method over full-context input, highlighting its effectiveness in identifying relevant discourse segments to support temporal relation extraction. It also reinforces the observation that providing the full context may introduce excessive noise, ultimately degrading the performance of LLMs in this task.

Sentence distance	Acc_{zs}	Acc_{full}
0-2	18.72	44.68
3-5	15.47	45.13
6-8	17.39	51.66
9-11	16.45	49.35
12-14	15.84	51.49
15-17	20.93	67.44
18-20	14.29	33.33
21-23	16.67	83.33

Table 7: Statistics of intervals between sentences containing events in TDD-Man's test set. The subscript zs means zero-shot setting, and full means fine-tuning setting with full context. All results are obtained by using Llama-3.1-8B-Instruct.

Method	F1 Score	Cons.
Llama-3.1-8B-Instruct_{zs}	16.8	43.8
Llama-3.3-70B-Instruct_{zs}	26.8	76.0
Llama-3.1-8B-Instruct_{ft}	48.5	83.2
Llama-3.3-70B-Instruct_{ft}	51.1	84.6

Table 8: Llama-3.1-8B-Instruct and Llama-3.3-70B-Instruct performances using zero-shot setting and fine-tuning setting on TDD-Man.

F Consistency Explanation

In the temporal relation extraction task, a high F1 score typically means a high level of consistency. However, a high level of consistency does not guarantee the F1 score. For example, TDD-Man's test dataset contains 1500 data, and 46 data are labelled as simultaneous. If a system predicts all the data as simultaneous, the consistency rate is 100%, while the F1 score is extremely low.

When we fine-tune the model, we specify the need to maintain consistency in the prediction, shown in 9. We input the selected context, event pair, and file name to inform the model of other event pairs in the same document. That's why, without our self-reflection step, our model can perform a moderate level of consistency. For previous works using BERT or RoBERTa, the setting is typically to input the full document or sliding window cut sentences and event pairs. Thus, their models do not know whether other event pairs are also in the same document when predicting the temporal relation between two events, since the models treat the task as a single classification problem for each event pair. This is a shortcoming of BERT-based

models in maintaining consistency.

Our consistency strategy is not based on repeatedly querying the model until F1 and consistency metrics improve. Instead, we adopt a two-stage fine-tuning approach. In the first stage, the model is fine-tuned using training data where each input consists of an event pair, and the model learns to predict the temporal relation between them. Based on the model's predictions on the training set, we identify inconsistent event triplets and construct a second round of fine-tuning data to explicitly teach the model consistency constraints. The model can learn from supervised feedback since this stage still uses gold labels. During inference, the model first predicts temporal relations for all event pairs. We then extract inconsistent triplets from the initial predictions and re-query the model using these triplets as input. No gold labels are provided at this stage, and the model is queried only twice—we do not iterate the process. In contrast, Chen et al. (2024) propose a fine-tuning approach that relies on constructing a high-order event relation dataset specifically designed to teach logical consistency. Their method requires an additional dataset, whereas ours is entirely data-driven: all consistency prompts are constructed directly from the predictions on the training and test data, without introducing new data or manual labelling.

G Computational Resources

The GPU we use for this study is NVIDIA H200.

H Best Hyper Parameters

The best γ value for our context selection method is 0.7. The best Lora alpha and Lora rank are 16 and 32, and 32 and 64 for Llama-3.3-70B-Instruct and Llama-3.1-8B-Instruct, for both fine-tuned models. The best learning rate for Llama-3.3-70B-Instruct is $3e-5$, and for Llama-3.1-8B-Instruct is $5e-5$. In the first round, we fine-tuned both of our models for three epochs at most, then we only fine-tuned the self-reflection for one epoch. During the first round, we also set the warm-up step to 20. When we evaluate our models, the top-p sampling value is set to 0.7, and the temperature is set to 0.9. For our context selection and BERTopic, we use all-MiniLM-L6-v2[4] as the Sentence-Transformers model.

[4] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

Instruction for selecting context

Task:
You are given a document consisting of multiple numbered sentences. Your task is to identify the most relevant sentences that provide useful information for determining the temporal relation between <e50>aware</e50> and <e24>do</e24>.
Guidelines:
1. Only select sentences that:
- Contain direct or indirect references to <e50>aware</e50> or <e24>do</e24>.
- Provide background, causal, or sequential clues that help infer the temporal relation.
- Contribute to understanding when <e50>aware</e50> happened relative to <e24>do</e24>.
2. Ignore sentences that are completely unrelated to <e50>aware</e50> and <e24>do</e24>.
3. Output the sentence numbers in a standardized format. For example, if sentences 0, 2, 3 and 5 are related, then you should output: [0, 2, 3, 5]"
4. Explain briefly why selected sentences are relevant to <e50>aware</e50> and <e24>do</e24> after giving the related sentences list.

Figure 6: An instruction example for model selection from source file: NYT19980206.0466

Prediction for the Context

After analyzing the given document, I have identified the most relevant sentences that provide useful information for determining the temporal relation between <e50>aware</e50> and <e24>do</e24>. Here are the selected sentences with their corresponding numbers:
[0, 1, 4, 9]
Here's why these sentences are relevant: Sentence 0 is relevant because it introduces the context of the discussion about the White House secretary, Betty Currie, and her alleged involvement with President Clinton. Sentence 1 is relevant because it directly mentions <e50>aware</e50> and clarifies that there is no implication that Mrs. Currie was aware of any legal or ethical impropriety. Sentence 4 is relevant because it mentions <e24>do</e24> in the context of telling the truth, which implies a temporal relation with <e50>aware</e50> as it suggests that Mrs. Currie was not aware of any unethical conduct before being asked to tell the truth. Sentence 9 is relevant because it mentions <e50>aware</e50> again and provides a temporal clue by stating that Ms. Currie's lawyers stated unambiguously that she's not aware of any unethical conduct, implying that this awareness or lack thereof occurred before the events described in the document. These sentences provide useful information for determining the temporal relation between <e50>aware</e50> and <e24>do</e24> by clarifying when Mrs. Currie became aware of certain information and how that relates to her being asked to do something (tell the truth).

Figure 7: The prediction of the context generated by the model for a specific event pair from source file: NYT19980206.0466

Input for selecting context

0: Following are statements made Friday and Thursday by Lawrence Wechsler, a lawyer for the White House secretary, Betty Currie; the White House; White House spokesman Mike McCurry, and President Clinton in response to an article in The New York Times on Friday about her statements regarding a meeting with the president: Wechsler on Thursday "Without commenting on the allegations raised in this article, to the extent that there is any implication or suggestion that Mrs. Currie was <e50>aware</e50> of any legal or ethical impropriety by anyone, that implication or suggestion is entirely inaccurate.
1: " White House on Thursday"For the past few weeks we've been subjected to false leaks designed to mislead both reporters and the American public.
2: We're not going to dignify the latest false leak with a response.
……
7: " Once again, I never asked anybody to <e24>do</e24> anything but tell the truth.
8: I know about the stories today.
……

Figure 8: An input example for model selection from source file: NYT19980206.0466

Instruction

Classify the temporal relation between two events in the input. You must consider chain reasoning, temporal information, whether the two events in the pair will occur in the future, assumptions or negations, event coreference, causal or prerequisite connections, and world knowledge to make predictions. You will be given the source file name, you should make sure your predictions in the same source file are consistent. Possible relations: before, after, includes, included in and simultaneous.

During the prediction, the consistency should be obtained. For example, if e1 is before e2 and e2 is before e3, then e1 must be before e3.
Output the start time comparison and end time comparison for event 1 and event 2, then output their temporal relation. If event 1 is before event 2, then the output should be like: Event 1 starts before Event 2 starts and Event 1 ends before Event 2 starts, so the temporal relation of Event 1 to Event 2 is: before.
File name: NYT19980206.0466
Event 1: <e50>aware</e50> and Event 2: <e24>do</e24>\

Figure 9: The instruction

Self-reflection instruction

Reconsider your prediction and classify the temporal relation between two entities in the input. From your previous response, the temporal relations you predicted is not consistent. Now, you will be given the inconsistent predictions you gave, check it and return the consistent predictions. You must consider chain reasoning, temporal information, whether the two events in the pair will occur in the future, assumptions or negations, event coreference, causal or prerequisite connections, and world knowledge to make predictions.Possible relations: before, after, includes, included in and simultaneous.

Output the start time comparison and end time comparison, then output their temporal relation. If event 1 is before event 2, then the output should be like: Event 1 starts before Event 2 starts and Event 1 ends before Event 2 starts, so the temporal relation of Event 1 to Event 2 is: before. Output this format for all event pairs.

The inconsistent information about your last time prediction: If Event 1 is before Event 2 and Event 2 is before Event 1, then the temporal relation of event Event 1 to Event 3 must be: before.",

Figure 10: The self-reflection instruction

Instruction without Cot

Classify the temporal relation between two events in the input. You must consider chain reasoning, temporal information, whether the two events in the pair will occur in the future, assumptions or negations, event coreference, causal or prerequisite connections, and world knowledge to make predictions. You will be given the source file name, you should make sure your predictions in the same source file are consistent. Possible relations: before, after, includes, included in and simultaneous. Only output one of those relations, with no more words.

During the prediction, the consistency should be obtained. For example, if e1 is before e2 and e2 is before e3, then e1 must be before e3.

Figure 11: The instruction without CoT