Getting Sick After Seeing a Doctor? Diagnosing and Mitigating Knowledge Conflicts in Event Temporal Reasoning

Anonymous ACL submission

Abstract

Event temporal reasoning aims at identifying the temporal relations between two or more events from narratives. However, *knowledge conflicts* arise when there is a mismatch between the actual temporal relations of events in the context and the prior knowledge or biases learned by the model. In this paper, we propose to detect knowledge-conflict examples in event temporal reasoning using bias indicators, which include *event relation prior* bias, *tense* bias, *narrative* bias, and *dependency* bias. We define conflict examples as those where event relations are opposite to biased or prior relations. To mitigate event-related knowledge conflicts, we introduce a Counterfactual Data Augmentation (CDA) based method that can be applied to both Pre-trained Language Models (PLMs) and Large Language Models (LLMs) either as additional training data or demonstrations for In-Context Learning. Experiments suggest both PLMs and LLMs suffer from knowledge conflicts in event temporal reasoning, and CDA has the potential for reducing *hallucination* and improving model performance.

1 Introduction

An important goal of event understanding is to identify the temporal relations (TEMPRELS) among events described in natural language text (Chambers et al., 2007). This task aligns with human's cognitive ability (Zacks and Tversky, 2001; Zacks et al., 2007), which often involves routinely reasoning about how events happening around us are temporally sequenced, planned, and lead to consequences and decisions (Schank and Abelson, 1977). From the intelligent system perspective, it also benefits many NLP applications for narrative understanding (Li et al., 2018; Cai et al., 2022), schema induction (Li et al., 2021), and question answering (Zhu et al., 2017; Stricker, 2021).

In event temporal reasoning, the input includes two parts, the event mentions and the context. The

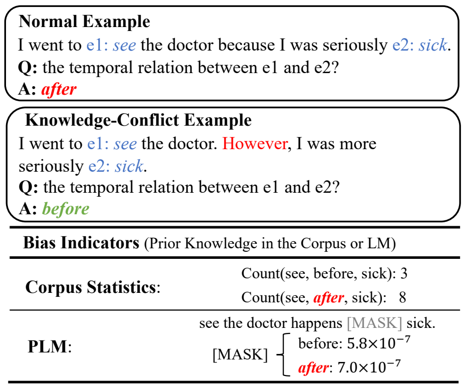

Figure 1: An example of a knowledge-conflict instance. The actual TEMPREL in the context differs from the biased or prior TEMPREL in the corpus and the language model, leading to the emergence of *knowledge conflicts*.

TEMPREL a model seeks to infer should be based on the context, rather than only revealed by the event mentions themselves. For example, in Fig. 1, without a context, the event mention *see* (the doctor) and *sick* have certain temporal prior where *see the doctor* statistically happen more often after *sick*, either by corpus statistics or probing a masked PLM. However, under the context of "I went to see the doctor, *However*, I was more seriously sick," we can infer that *see* happens before *sick* instead of after due to the presence of the connective *However*. This is known as the phenomenon of *knowledge conflicts* (Longpre et al., 2021), where the contextual information contradicts the knowledge memorized by the language model. Hence, the essential requirement for accountable temporal reasoning is *context-faithfulness* (Wang et al., 2023; Zhou et al., 2023), where models are expected to perform reasoning based on the context instead of *guessing* using only the prior knowledge about the events encoded in their parameters.

However, most current language models, including both Pre-trained Language Models (PLMs) and Large Language Models (LLMs) [1], rely on short-

[1] PLMs, or smaller models, are used in a pre-train and fine-

cuts from the mentions without being faithful to the context (Xu et al., 2022; Bender et al., 2021) to varying degrees, leading to *hallucination*. This issue is particularly severe in contexts where event or entity mentions have a different relation prior than what is presented in the context. Though entity-related knowledge conflicts (Longpre et al., 2021; Wang et al., 2022; Li et al., 2022) have recently attracted much attention, questions about event-related knowledge conflicts remained intact.

First, it is necessary to understand the conflicts regarding *relations* of events, which is more complicated than that of a single event. Second, the substitution-based paradigm defined in entity knowledge conflicts or spurious correlation detection (Longpre et al., 2021) cannot be directly applied to events. Entity mentions can often be replaced randomly with other entities with the same typing to study the faithfulness towards the context other than the entity mention, which remains unchanged after the replacement. For example, in open-domain QA, a possible question can be "Who is the CEO of Twitter?" based on the context "Yaccarino succeeded Elon Musk as the CEO of Twitter". To check whether models faithfully rely on the context instead of hallucinating, *Yaccarino* in the context can be changed to a random name to see if the model can still output the "correct" CEO instead of *Yaccarino* as they have learned in pre-training. However, events are usually denoted by predicates in the context (Bethard et al., 2007), and directly substituting the predicate (e.g., from *see* in Fig. 1 to another random verb such as *play*) will alter the semantic meaning of the whole context, including both the predicate and its dependency with the arguments, making it infeasible to analyze the *faithfulness* towards the *original* context. Thus, instead of resorting to a substitution, in this paper, we study the effect of knowledge conflicts in event temporal reasoning by *selecting* conflict examples from the original dataset based on corpus statistics, and evaluate models on the conflict subsets.

We outline the contributions of this paper as follows. First, we define four types of bias that can lead to knowledge conflicts, including *event-relation* bias, *narrative* bias, *tense* bias, and *dependency* bias. The data instances where the actual TEMPREL contradicts with the prior TEMPREL are referred to as knowledge-conflict instances (§3), as they conflict with the prior knowledge provided to language models. Second, to mitigate the effect of knowledge conflicts, we propose a Counterfactual Data Augmentation (CDA) technique that explicitly generates contexts with knowledge-conflict elements, thereby reducing the overall bias in the data distribution. CDA can be applied to both fine-tuned PLMs and LLMs with (test-time) in-context learning (§3.3). Third, we study the effect of various kinds of knowledge conflicts and our proposed bias mitigation method on two popular event temporal reasoning benchmarks, TORQUE (Ning et al., 2020) and MATRES (Ning et al., 2018). We show that models suffer from performance drop on knowledge-conflict subsets, and our bias-aware data augmentation method outperforms baselines by a remarkable margin on both bias mitigation and overall performance (§4).

2 Related Works

Event Temporal Reasoning. Event temporal reasoning aims at identifying the temporal relations (TEMPREL) of events in narratives. There are two common ways of formulating this problem. The first formulation is the TEMPREL extraction task, which involves determining the TEMPREL between two annotated event triggers from a pre-defined relation set (Bethard et al., 2007, 2017; Ning et al., 2018; Naik et al., 2019). Meanwhile, another formulation is a reading comprehension task, which involves determining more complicated TEMPRELS expressed in natural language questions (Ning et al., 2020; Han et al., 2021). To conduct event temporal reasoning, literature has leveraged various approaches, including graph neural networks (Zhang et al., 2022; Zhou et al., 2022), rhetorical discourse features and temporal arguments from semantic role labels (Mathur et al., 2021), and distant supervision (Zhou et al., 2021; Zhao et al., 2021). In addition, Wang et al. (2023) study the effect of counterfactual inference as well as Dirichlet parameterization to improve uncertainty calibration of the model. LLMs such as GPT3 (Brown et al., 2020) and ChatGPT are also leveraged for event temporal reasoning (Chan et al., 2023) with carefully designed prompts and In-Context Learning. Our work differs from previous studies in that we study the knowledge conflicts in event temporal reasoning and how to mitigate them.

Knowledge Conflict in Language Models. Knowledge conflicts have been widely studied for

tune paradigm, while LLMs, larger and more powerful models with over 10B parameters, are commonly employed through in-context learning (Sun, 2023).

entity-centric NLU tasks (Schuster et al., 2021). For example, Longpre et al. (2021) studied the knowledge conflict in open-domain question answering using entity substitution. Li et al. (2022) also adopted this strategy to study the enhancement of a PLM's robustness against context noise with a knowledge-aware working memory. Xu et al. (2022) systematically formulate six types of biases in entity typing to study spurious correlations. Certain types of biases, such as Mention-Context and Named Entity bias, can reflect knowledge conflicts in entities. Zhou et al. (2023) use opinion-based prompting and counterfactual demonstration to enhance the context-faithfulness of test-time-only LLMs against knowledge conflicts. Feng et al. (2022) proposed a dataset studying the differential effects of TEMPREL reasoning given additional contexts, while their focus is on annotating additional out-of-distribution data instead of exploring existing knowledge conflicts within the dataset. Our work systematically defines and detects knowledge conflicts in event temporal reasoning and proposes a data-augmentation-based method to mitigate those conflicts based on the detected bias.

3 Event Knowledge Conflict

In this section, we introduce the problem definition (§3.1) and formally define four types of bias and how to select knowledge-conflict data (§3.2). We then introduce our proposed Counterfactual Data Augmentation (§3.3).

3.1 Problem Definition

In event temporal reasoning, the primary objective is to determine the TEMPREL between two or more events, which previous studies (Ning et al., 2018; Naik et al., 2019) typically classify as *before*, *after*, *equal* (indicating two events occurring simultaneously), and *vague*. Without the loss of generality, our study is based on pairwise event relations: the relation r of an event pair (e_1, e_2) based on the context c. More complex cases can be easily addressed by breaking down the relations involving multiple events into pairwise relations. The case where evaluating the temporal status of a single event (*happened*, *happening*, *will happen*, etc.) can also be easily adapted in this framework by replacing the features of event pairs to a single event. Detailed adaptations to different datasets will be introduced in §4.2.

To study event-related knowledge conflict, we create an automated framework to use corpus co-occurrence statistics to select conflict subsets. Similar to the co-occurrence statistics in *reporting bias* (Gordon and Durme, 2013), to obtain knowledge-conflict data, we first define bias, as the *opposite* side of the conflict. We identify four types of bias in event temporal reasoning and defined corresponding bias statistics. We then selected a subset of the original dataset where feature-relation pairs were rare (i.e., *knowledge-conflict*) based on the bias scores. As the (reporting) bias in the training corpus is usually learned and amplified by the language models (Shwartz and Choi, 2020), our selected subsets, which represent the opposite side of the bias, conflict with the knowledge encoded in the language models.

3.2 Knowledge Conflict Diagnosis

We first define a bias score $b(P_1, P_2, r)$ with regard to certain patterns (P_1 and P_2) against a specific relation $r \in R$, where R is a subset of all relations defined in a certain dataset. Patterns P_i can be the event lemmas themselves, tense, dependency patterns, and narrative orders of either event. Sometimes (P_1, P_2) is represented by one feature only, for example, the dependency relation and narrative orders between two events. Denote $c(P_1, P_2, r)$ as the number of occurrences of (P_1, P_2) under relation r in a corpus, and the bias score is defined as:

$$b(P_1, P_2, r) = \frac{c(P_1, P_2, r)}{\sum_{r' \in R} c(P_1, P_2, r')} \quad (1)$$

For example, in the tense bias defined below, the bias score of the tense pattern (VBD, VBZ) (*past tense* and *third person singular present tense*) when we only consider two relations $R = \{\textit{before, after}\}$ is defined as:

$$b(\mathsf{VBD}, \mathsf{VBZ}, \textit{before}) = \frac{c(\mathsf{VBD}, \mathsf{VBZ}, \textit{before})}{c(\mathsf{VBD}, \mathsf{VBZ}, \textit{before}) + c(\mathsf{VBD}, \mathsf{VBZ}, \textit{after})} \quad (2)$$

Knowledge Conflict Detection. In a set of relations, those with higher bias scores indicate higher degrees of bias towards certain relations, and others with lower bias scores indicate higher degrees of knowledge conflict. We select instances whose patterns do *not* follow the majority distribution in the training set as knowledge-conflict instances. A new instance in the test set with a pattern-relation pair (P_1, P_2, r) is considered knowledge conflict if the bias score is less than the context-free frequency of relations $b(P_1, P_2, r) < \frac{c(r)}{\sum_{r' \in R} c(r')}$. Moreover, to ensure a significant degree of conflicts,

Type	Context & Label	Bias Scores
Relation Prior (relation)	(TORQUE) Chidambaram e_1:drew up the previous United Front government's Indian budget for 1997-98 which is to be e_2: approved by parliament this week . Gujral has e_3: adopted the same budget. **Question:** What will happen after e_1: drew? **True label:** e_2: approve. **Biased Prediction:** e_3: adopted	$b(\textit{draw}, \textit{adopt}, \textit{before}) = 1.0$ $b(\textit{draw}, \textit{approve}, \textit{before}) = 0$
Relation Prior (warm-up)	**Question:** What will happen in the future? **True label:** e_2: approve. **Biased Prediction:** e_3: adopt	$b(\textit{approve}, \textit{happened}) = 0.9$ $b(\textit{approve}, \textit{future}) = 0.05$
Tense (relation)	(MATRES) Albright e_1: told (VBD) ambassadors of 30 African countries in Washington, who came to the State Department to e_2: offer (VB) condolences. **True label**: e_1 happens *after* e_2; **Biased Prediction:** *before*	$b(\texttt{VBD}, \texttt{VB}, \textit{before}) = 0.70$ $b(\texttt{VBD}, \texttt{VB}, \textit{after}) = 0.27$ $b(\texttt{VBD}, \texttt{VB}, \textit{equal}) = 0.03$
Tense (warm-up)	(TORQUE) That's what will e_1: keep computer makers e_2: coming (VBG) in spite of the e_3: irritation of e_4: bugs. **Question:** What will happen in the future? **True Label:** e_1, e_2: coming; **Biased Prediction**: e_1	$b(\texttt{VBG}, \textit{happened}) = 0.42$ $b(\texttt{VBG}, \textit{future}) = 0.13$ $b(\texttt{VBG}, \textit{happening}) = 0.45$
Narrative	(MATRES) Now events are e_1: doing the work for Schumer. Slepian's death was among the first topics e_2: raised in Saturday night's debate between the two men, ... ; **True label**: e_1 happens *after* e_2; **Biased Prediction:** *before*	$b(p_1 < p_2, \textit{before}) = 0.59$ $b(p_1 < p_2, \textit{after}) = 0.37$ $b(p_1 < p_2, \textit{equal}) = 0.04$
Dependency	(MATRES) Castro e_1: said Gonzalez would e_2: travel with his current wife and their son (Dependency: *says* $\rightarrow$ ccomp $\rightarrow$ *travel*) **True label**: e_1 happens *before* e_2; **Biased Prediction:** *after*	$b(\texttt{ccomp}, \textit{before}) = 0.66$ $b(\texttt{ccomp}, \textit{after}) = 0.32$ $b(\texttt{ccomp}, \textit{equal}) = 0.02$

Table 1: Examples of different forms of knowledge conflicts.

we set a threshold $\mathcal{T}_r$ such that $b(P_1, P_2, r) < \mathcal{T}_r < \frac{c(r)}{\sum_{r' \in R} c(r')}$, to ensure that the conflict is large enough. For example, a test instance where the event with a past tense happens *after* the event with a present tense may be selected as a knowledge-conflict instance, as the context makes the actual TEMPREL different from the biased relation *before*.

Next, we introduce the definitions of different forms of bias in detail. Data instances that counteract the biased distribution are selected as corresponding knowledge-conflict subsets.

Relation Prior Bias. Bias toward certain TEMPRELS exists because there are natural *selectional preference* (Wilks, 1975) between the specific events. For example, in the TORQUE dataset, *arresting* dominantly happen after *killing*, and *voting* more often happens before *winning*. These findings suggest that the occurrence of certain events may be more likely to follow or precede other events, which can however, lead to bias when the context describes the TEMPREL differently from the most frequent cases. Our definition of the bias scoring function is based on the frequency of the co-occurrence of event e_1 and e_2 under relation r:

$$b(e_1, e_2, r) = \frac{c(e_1, e_2, r)}{\sum_{r' \in R} c(e_1, e_2, r')} \tag{3}$$

Narrative Bias. Narrative bias in event temporal reasoning is the tendency for the model to interpret the chronological order of the events to be the same as their narrative order. However, these two orders, though more often accord with each other, do not always necessarily follow the same (Zwaan et al., 1995). In this sense, we only study *before*, *after*, and *equal* relations for narrative bias. Denote $p = P(e, c)$ as the position of event e in context c, where the earlier position of e indicates that this event is described earlier in the narrative. The bias scoring function is defined as follows for the case where the positions of the two events follow the order of $p_1 < p_2$:

$$b(p_1 < p_2, \textit{before}) = \frac{c(p_1 < p_2, \textit{before})}{\sum_{r' \in R} c(p_1 < p_2, r')} \tag{4}$$

We select the event pairs where $p_1 < p_2$ while the actual relation is (e_1, *after/equal*, e_2) or $p_1 > p_2$ while the actual relation is (e_1, before/equal, e_2) as the knowledge-conflict examples.

Tense Bias. Tense bias is the tendency to rely on the grammatical tense of verbs as evidence for the temporal order of events. For example, past tense is typically used to describe events that occurred *before* the present moment, while present tense is typically used for events that are happening now or in the future. However, this grammatical convention does not always correspond to the actual temporal order of events. Denote t_1 and t_2 as the tense (POS-tags parsed by Spacy [2] as more fine-grained tense information) of event e_1 and e_2 under context c, then the bias score is defined as:

$$b(t_1, t_2, r) = \frac{c(t_1, t_2, r)}{\sum_{r' \in R} c(t_1, t_2, r')} \tag{5}$$

[2] https://spacy.io/

Dependency Bias Dependency bias is the tendency to rely on syntactic dependency patterns in language as evidence for the temporal order of events. For example, if two events e_1 and e_2 are directly connected in the dependency tree, the dependency pattern $(e_1, \texttt{dobj}, e_2)$ (where e_1 is the subject of the sentence, e_2 is the direct object, and `dobj` is the dependency between them) often indicates that e_1 is the entity performing an action on e_2. This pattern may suggest that e_1 must occur before e_2 in time, but this is not always the case. Denote d as the dependency relation between e_1 and e_2 in context c (d is null if e_1 and e_2 are not directly linked in their dependency tree).

$$b(d, r) = \frac{c(d, r)}{\sum_{r' \in R} c(d, r')} \tag{6}$$

We summarize the core features of each defined bias associated with examples in Tab. 1. Our focus is particularly on two datasets, namely TORQUE and MATRES, which will be presented in §4.1. Prior to that, we introduce our proposed conflict-mitigating method.

3.3 Counterfactual Data Augmentation

In this sub-section, we introduce our proposed Counterfactual Data Augmentation (CDA) method for mitigating knowledge conflicts (Fig. 2). We discuss the usage of CDA on both PLM and LLM separately, as they differ in their applications.

Pre-trained Language Models. PLMs are usually fine-tuned on a training corpus, which naturally contains event-relation biases that tend to be amplified after fine-tuning (Hall et al., 2022). To mitigate bias, our proposed method automatically generates context that contains event pairs whose actual temporal relation is different from the biased relation. Such knowledge-conflict (counterfactual) counterparts are trained together with the original training corpus to mitigate the biased training distribution. To be more specific, for each event pair (e_1, e_2) that is identified as biased, we ask an Instruction-finetuned Language Models (Chung et al., 2022) to generate context where (e_1, e_2) is associated with a TEMPREL that leads to a low bias score of a certain bias type, entitled augmented knowledge-conflict data. The intuition is that, even though language models may suffer from bias and cannot directly solve the task, they can be well applied to generate synthetic data under structured instructions (Josifoski et al., 2023).

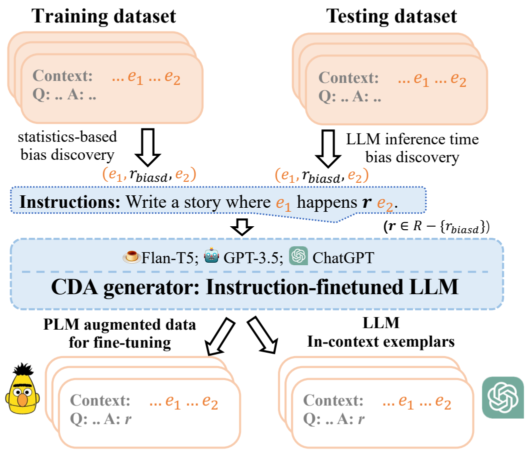

Figure 2: An overview of the CDA pipeline.

Large Language Models. The *de facto* way of leveraging LLMs for downstream tasks is test-time In-Context Learning, as further fine-tuning of the LLM is typically impractical or unviable. In this case, we extend the idea of Counterfactual Data Augmentation to automatically generate counterfactual examples for in-context learning. Unlike the data augmentation in PLMs, we generate counterfactual counterparts for every event pair to be studied instead of only for the biased ones. For a new event pair (e_1, e_2) to be studied, we acquire the predicted relation r_{LLM} by the LLM, which is regarded as a "*factual*" prediction as it is what the LLM itself *hallucinates*. We leverage the LLM to generate context examples where (e_1, e_2) are associated with relations that belong to $R - \{r_{LLM}\}$ as counterfactual examples to showcase the LLM the alternative cases when (e_1, e_2) happens following a different TEMPREL. Note that this method is still considered a zero-shot as no training examples are seen during inference.

4 Experiments

In this section, we introduce the datasets (§4.1), the settings of knowledge conflict diagnosis (§4.2), and conflict mitigation (§4.3), the primary experimental results and analysis (§4.4).

4.1 Datasets

We select two event temporal reasoning datasets:

TORQUE: TORQUE (Ning et al., 2020) is a reading comprehension benchmark with a focus on event temporal reasoning questions. In TORQUE, each passage is associated with around 10 human-annotated questions regarding the TEMPREL between certain events, and the task objective is to select the correct answers from the pre-defined set of annotated event triggers. We evaluate the model performance using exact-match (EM) and Macro F1. TORQUE is more flexible than simple relation extraction benchmarks as the reading comprehension framework allows more complicated TEMPRELS including uncertain relations (e.g., *might before*), hypothetical relations (e.g., *what will happen if ...*), and negated relations (e.g., *not after*).

MATRES: MATRES (Ning et al., 2018) is a TEMPREL extraction dataset that includes refined annotations from documents in TimeBank (Pustejovsky et al., 2003), AQUAINT (Louis and Nenkova, 2012), and Platinum (UzZaman et al., 2013). The task in MATRES is defined as identifying the TEMPREL between two events in the context, where $R = \{before, after, equal, vague\}$. We use the pre-processing by Wang et al. (2020) to acquire the training and development set from the raw annotations in MATRES, where the context includes the sentences containing the two events e_1 and e_2, together with a precedent sentence to give more contextual information. We randomly sample 1,000 entries (out of ∼6k) from the development set to perform evaluations for LLMs[3].

4.2 Knowledge Conflict Diagnosis

We apply the bias statistics introduced in §3 on the training set to select knowledge-conflict subsets from both TORQUE and MATRES development sets. In MATRES, we directly make use of the TEMPREL information (e_1, e_2, r) provided in each data entry to count the occurrence and calculate bias. However, in TORQUE, the problem is formulated as reading comprehension, which requires further pre-processing to acquire pairwise TEMPRELS. Specifically, we parse each question to acquire the temporal predicate and arguments. For example, for the question "*What happened after Bush gave four key speeches?*" and answers "{*called*, *elect*, *vote*}" under a certain context, we can acquire three event relation triples (*gave*, *before*, *called*), (*gave*, *before*, *elect*), and (*gave*, *before*, *vote*). We use those triples for calculating and detecting bias. In addition, TORQUE includes *warm-up* questions that analyze whether a single event has *happened*, *will happen*, or *is happening*. Our study calculates bias statistics based on a single event and its temporal status (*happened*, *will happen*, or *is happening*) relative to a time expression in the context. The bias in warm-up questions is labeled with *warm-up*, while the other questions studying event-pair relations are labeled with *relation*. In addition, Tab. 7 in the appendix lists the most biased features selected for both datasets. We can find some intuitive bias, for example, a past tense is more often predicted as *before* a present tense.

For each type of bias, we empirically set thresholds to select knowledge-conflict subsets. For a feature-relation pair f (e.g., f represents dependency) and r, it is knowledge-conflict if $b(f, r) < \frac{c(r)}{\sum_{r' \in R} c(r')}$, indicating that it does not conform to the dominant distribution of relation r. Such selection criteria can be further enhanced by setting a threshold $\mathcal{T}_r < \frac{c(r)}{\sum_{r' \in R} c(r')}$, which increases the level of conflicts by further restricting $b(f, r)$ to be less than $\mathcal{T}_r$. The hyperparameters we used are listed in Appx. §A. The statistics of the knowledge-conflict subsets we acquired are presented in Tab. 6.

4.3 Setup for Conflict Mitigation

Counterfactual Data Augmentation. We introduce the details of conducting Counterfactual Data Augmentation here. In augmentations for PLM, we choose Flan-T5 (11B) (Chung et al., 2022) as the generator. For each event pairs (e_1, e_2, r) identified as being biased according to *Relation Prior Bias*, we generate context with the prompt *Write a story where e_1 happens r' e_2:*, where $r' \in R - \{r\}$ (e.g., r'=*before*). In TORQUE, we thus construct a question Q=*"What happened r' e_2"*, and the corresponding answer is e_1. In MATRES, we require the model to directly predict the relation r given the generated context. Based on the above coarse data, we apply additional filters to only retain the generated data that are not biased in terms of tense and narrative.

For LLMs, we ask the model itself to predict the labels of the test data first. Take MATRES as an example, denote r_{LLM} as the *factual* prediction by the LLM, and then we ask the LLM itself to *Generate a paragraph where event e_1 happens r' e_2*, where $r' \in R - \{r_{LLM}\}$. More detailed prompts are presented in Appx. §C.2.

Model Configuration. We perform experiments using both PLMs and LLMs[4]. To use LLMs, we use the following prompt template for

[3]A common practice when doing GPT3-related experiments to reduce the overall cost (Bian et al., 2023).

[4]We refer readers to Appx. §C for more details experimental setups. We also present the effects of different prompt templates and the number of few-shot exemplars.

	all		Rel.Prior (relation)		Rel.Prior (warm-up)		Narrative (relation)		Tense (relation)		Tense (warm-up)		Dep. (relation)		Confl.Avg.	
	EM	F1*	EM	F1*	EM	F1*	EM	F1*	EM	F1*	EM	F1*	EM	F1*	EM	F1*
PLM																
RoBERTa-L	50.4	75.7	29.5	73.3	50.0	75.1	31.4	69.0	33.5	72.9	48.4	72.4	41.7	78.6	39.1	73.6
PoE	33.3	65.8	21.6	**76.1**	22.7	59.8	23.5	67.1	27.5	71.1	22.5	57.0	32.3	**79.2**	25.0	68.4
L.-mixin	46.8	74.8	27.2	75.2	50.0	72.1	27.8	68.4	30.8	72.6	**49.3**	69.8	33.8	76.8	36.5	72.5
L.-mixin+H	37.6	70.6	20.4	73.4	40.9	71.6	28.5	**69.6**	28.8	71.6	38.0	67.7	32.3	76.0	31.5	71.7
Cont. Inf.	**53.1**	75.9	28.4	75.3	50.0	72.5	**35.7**	68.9	35.4	**73.1**	**49.3**	70.2	**44.1**	78.9	40.5	73.2
AFLite	50.5	75.8	**34.1**	73.5	48.5	72.1	26.4	68.2	34.6	72.7	47.9	69.8	39.7	77.3	38.5	72.3
CDA (Ours)	51.0	**76.1**	33.7	75.4	**50.0**	**75.9**	30.7	68.6	**35.5**	**73.1**	48.8	**73.2**	**44.1**	79.1	**40.5**	**74.2**
LLM																
GPT-3.5	**8.36**	**45.5**	4.82	59.9	4.62	47.0	2.13	50.7	4.46	53.5	5.71	45.9	2.94	57.7	4.12	52.5
+ ICL	7.22	44.9	**9.09**	60.2	**9.09**	**55.6**	2.14	51.3	**5.35**	55.5	**8.45**	**52.6**	**4.41**	58.8	**6.42**	**55.7**
+ GDA	4.85	44.0	5.68	60.0	1.54	49.4	**3.19**	54.6	3.18	**56.1**	1.43	48.3	2.94	58.6	3.00	54.5
+ CDA	5.53	45.1	5.68	**60.6**	1.52	48.0	2.14	**56.5**	4.53	54.1	1.41	50.1	2.94	**61.2**	3.04	55.1
ChatGPT	**17.7**	40.7	**9.09**	40.3	**4.55**	38.3	**6.43**	42.3	**10.3**	41.4	**4.23**	35.8	**7.35**	42.2	**6.99**	40.0
+ ICL	3.92	43.9	4.55	**58.3**	**4.55**	50.1	1.43	48.9	3.70	52.8	**4.23**	47.9	1.47	54.8	3.32	52.1
+ GDA	4.38	44.2	3.41	56.2	1.52	50.6	1.43	50.0	3.29	52.9	1.41	48.3	2.94	57.4	2.33	52.6
+ CDA	6.72	**45.2**	3.41	55.6	1.52	**50.9**	1.43	**51.4**	2.06	**53.3**	2.82	**50.0**	**4.41**	**59.1**	2.60	**53.3**

Table 2: Experimental results on the TORQUE dataset. Exact-Match (EM) rate and Macro-F1 (F1, regarded as the primary metric * since EM can be susceptible to manipulation by simply predicting 'none') scores are reported. Best-performed results are **bold-faced** and the second-best are underlined.

TORQUE: *Q: [question], select none or several from [all events] [context] \n A:*. We use GPT-3.5 (`text-davinci-003`) and ChatGPT (`gpt-3.5-turbo`) as the backbone LLM. For MATRES, we formalize the problem as a multi-choice question-answering format[5].

Baselines. We compare our proposed methods with other representative bias mitigation approaches, including Product-of-Experts (PoE; Hinton 2002; He et al. 2019), Learned-mixin (Clark et al., 2019), Counterfactual Inference (Wang et al., 2022, 2023), and AFLite (Le Bras et al., 2020). These baselines are typical bias-agnostic debiasing baselines that address known or unknown bias with statistical approaches. For LLMs, we use the vanilla In-Context Learning (ICL) by randomly retrieving one set of exemplars from the training set as demonstrations. Note that ICL is considered few-shot learning while our method is purely *zero-shot*. In addition, to study the effect of the strategy for generating *counterfactual* exemplars, we add an additional baseline named *Generative Data Augmentation* (GDA) that performs exemplar generation without counterfactual guidance[6].

4.4 Results and Analysis

We present the main experimental results for TORQUE in Tab. 2 and for MATRES in Tab. 3. The *all* row indicates the performance on the whole evaluation set. The *Confl.Avg.* column is an average of all knowledge-conflict subsets, measuring models' ability on knowledge conflicts.

Impact of Knowledge Conflicts. Models on both TORQUE and MATRES show a decrease in performance when evaluated on knowledge-conflict subsets. Tab. 4 shows a comparison of baseline model performance on the conflict and non-conflict partitions of MATRES. The comparison on TORQUE is presented in Tab. 5 in the Appendix, showing a similar trend. This finding indicates that the selected conflict subsets are indeed more confusing for language models, proving the effectiveness of our conflict detection framework.

For LLMs, the overall performance is not satisfactory compared with fully-supervised models, which is in line with the findings in several evaluation works on LLMs (Chan et al., 2023; Zhou et al., 2023; Yuan et al., 2023), due to the fact that such tasks focusing on specific types of contextualized reasoning, when not trained with instruction fine-tuning, often lead to poor performance (Zhang et al., 2023). Nonetheless, since LLMs are not fine-tuned on the biased training set, their performance on knowledge-conflict subsets does not drop as significantly in comparison to that on the entire evaluation set, while even being better in some cases. This suggests that zero-shot predictions using LLM can be more generalizable when not trained on smaller and biased data.

[5] Details of prompts are listed in Appx. §C.2.

[6] Details of all baselines are in Appx. §C.1

	all		Rel. Prior		Narrative		Tense		Dependency		Confl.Avg.	
	Micro	Macro*	Micro	Macro*	Micro	Macro*	Micro	Macro*	Micro	Macro*	Micro	Macro*
PLM												
RoBERTa-large	70.8	44.9	59.7	28.5	59.2	27.1	54.8	33.2	58.5	38.3	58.0	31.8
PoE	69.4	45.3	60.0	30.7	52.6	32.8	61.1	29.0	53.1	36.7	56.7	32.3
Learned-mixin	71.0	45.0	60.4	29.5	55.7	**34.6**	60.9	27.5	60.0	**40.1**	59.2	32.9
Learned-mixin+H	70.5	44.8	59.6	29.2	54.3	34.0	62.2	27.7	58.5	39.8	58.6	32.6
Cont. Inf.	67.6	45.0	60.3	**31.4**	**60.7**	27.3	48.8	32.5	55.3	38.9	56.3	32.5
AFLite	64.3	43.4	52.4	28.8	50.3	32.8	**62.5**	30.0	55.0	39.3	55.1	32.7
CDA (Ours)	**72.2**	**45.5**	**61.5**	29.3	58.8	27.3	57.2	**35.1**	**62.2**	39.9	**59.9**	**32.9**
LLM												
GPT-3.5	**53.3**	19.7	54.7	25.3	2.57	3.98	36.7	17.2	28.6	13.0	30.6	14.9
+ ICL	51.6	18.4	**56.1**	20.9	1.52	2.31	35.7	16.4	26.2	10.6	29.9	12.6
+ GDA	45.6	27.6	52.0	32.4	15.1	14.9	37.6	24.0	**33.3**	18.9	34.5	22.6
+ CDA	51.3	**30.0**	53.4	**36.0**	**16.6**	**26.8**	**38.1**	**27.2**	**33.3**	**21.5**	**35.4**	**27.9**
ChatGPT	39.8	25.9	31.1	22.3	**37.6**	**32.5**	27.0	17.6	21.4	13.8	29.3	21.6
+ ICL	43.1	23.8	**53.4**	23.5	34.8	22.2	11.3	12.7	28.6	11.1	32.0	17.4
+ GDA	45.7	30.8	36.5	**25.1**	29.5	26.2	**32.5**	**20.7**	**40.5**	**24.4**	34.7	**24.1**
+ CDA	**49.3**	**32.0**	42.6	24.3	37.1	31.0	31.2	**20.7**	33.3	19.3	**36.1**	23.8

Table 3: Experimental results on MATRES. We use two evaluation metrics, Micro-F1 (denoted as Micro) and Macro F1 (denoted as Macro; regarded as the primary metric * due to the significant class imbalance). Best-performed results are **bold-faced** and the second-best is underlined.

	Conflict		Non-Conflict	
	Micro	Macro	Micro	Macro
RoBERTa-large				
Relation Prior	59.7$\downarrow$	28.5$\downarrow$	75.7	40.9
Narrative	59.2$\downarrow$	27.1$\uparrow$	76.8	21.7
Tense	54.8$\downarrow$	33.2$\downarrow$	72.8	47.2
Dependency	58.5$\downarrow$	38.3$\downarrow$	70.0	45.7
GPT-3.5				
Relation Prior	54.7$\downarrow$	25.3$\downarrow$	56.8	28.6
Narrative	2.57$\downarrow$	3.98$\downarrow$	85.8	26.3
Tense	36.7$\downarrow$	17.2$\downarrow$	60.3	27.2
Dependency	28.6$\downarrow$	13.0$\downarrow$	57.7	28.9

Table 4: Experimental results on the model performance on knowledge conflict and non-conflict data in MATRES. The RoBERTa-Large model suffers from a performance drop when tested on the conflict subsets. $\downarrow$ indicates a performance drop in the conflict subsets.

Knowledge Conflicts Mitigation. CDA significantly improves the performance of the vanilla PLM RoBERTa-large both on the entire evaluation set and on each of the knowledge-conflict subsets. Bias-agnostic baselines adopt a model trained only with event arguments and without context, which performs debiasing by countering event-relation bias. This yields competent results related to the relationship prior bias. The counterfactual inference is more effective than other fine-tuned-based methods, as also reported by previous work (Wang et al., 2022). However, bias-aware data augmentation methods are generally more effective, as they explicitly address different forms of bias and have a more focused performance on biased datasets. In the appendix, we show that more CDA data better help the model training (Fig. 3), and compare CDA with several plain data augmentation techniques in Tab. 8 and Tab. 9.

As for LLMs, on MATRES, CDA-based demonstrators can improve the performance on both the whole evaluation set and all the knowledge conflict datasets, with the exception of a minor setback compared to ChatGPT-GDA in terms of *Confl.Avg.* Macro-F1. On TORQUE, CDA on ChatGPT outperforms all baselines in terms of overall performance and *Confl.Avg.* on the main metric F1. For GPT-3.5, the zero-shot setting surprisingly achieves the best overall performance. However, CDA can outperform GDA, indicating that adding a counterfactual prior can better help LLMs to understand event temporal reasoning. Another noteworthy point is that our CDA method is purely zero-shot compared with ICL, showing the superiority of applying counterfactual guidance to LLMs.

5 Conclusion

In this paper, we investigate knowledge conflicts in event temporal reasoning by formally defining four types of biases to identify a knowledge conflict diagnoses evaluation set. We observe that both PLMs and LLMs are susceptible to knowledge conflicts in this task, resulting in decreased performance on knowledge-conflict datasets. To address this issue, we propose a CDA method that is suitable for both PLMs through pre-training and LLMs through In-Context Learning. Our experiments demonstrate the effectiveness of our proposed method in mitigating knowledge conflicts.

Limitations

This paper only discussed bias calculated based on statistics in the training set. However, there are various other ways of characterizing bias, such as using predictions of zero-shot pre-trained language models (Xu et al., 2022) and context masking, are not discussed, which can be left as a future work.

Ethics Statement

There are no direct societal implications of this work. The datasets we use, TORQUE and MATRES, are publicly available and shared via open-access licenses for research purposes. Even though we are detecting bias and conflicts in the original datasets, we focus on bias toward temporal relations of events and do not involve any bias toward certain gender or ethnics groups. The context where event relations are derived from TimeBank[7], AQUAINT[8], and Platinum[9], which has not shown to contain any obvious social biases that would raise concerns within the community. The counterfactual data augmentation technique we propose can effectively mitigate bias in event relation extraction. In conclusion, to the best of our knowledge, this paper does not raise ethical concerns.

References

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, pages 610–623. ACM.

Steven Bethard, James H. Martin, and Sara Klingenstein. 2007. Timelines from text: Identification of syntactic temporal relations. In *Proceedings of the First IEEE International Conference on Semantic Computing (ICSC 2007), September 17-19, 2007, Irvine, California, USA*, pages 11–18. IEEE Computer Society.

Steven Bethard, Guergana Savova, Martha Palmer, and James Pustejovsky. 2017. Semeval-2017 task 12: Clinical tempeval. In *Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval@ACL 2017, Vancouver, Canada, August 3-4, 2017*, pages 565–572. Association for Computational Linguistics.

Ning Bian, Xianpei Han, Le Sun, Hongyu Lin, Yaojie Lu, and Ben He. 2023. Chatgpt is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models. *CoRR*, abs/2303.16421.

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*.

Bibo Cai, Xiao Ding, Bowen Chen, Li Du, and Ting Liu. 2022. Mitigating reporting bias in semi-supervised temporal commonsense inference with probabilistic soft logic. 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 10454–10462. AAAI Press.

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.

Chunkit Chan, Jiayang Cheng, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. 2023. Chatgpt evaluation on sentence level relations: A focus on temporal, causal, and discourse relations. *CoRR*, abs/2304.14827.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. 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 4067–4080. Association for Computational Linguistics.

[7]https://catalog.ldc.upenn.edu/LDC2006T08
[8]https://catalog.ldc.upenn.edu/LDC2002T31
[9]https://bitbucket.org/leondz/te3-platinum

Yu Feng, Ben Zhou, Haoyu Wang, Helen Jin, and Dan Roth. 2022. Generic temporal reasoning with differential analysis and explanation. *CoRR*, abs/2212.10467.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction, AKBC@CIKM 13, San Francisco, California, USA, October 27-28, 2013*, pages 25–30. ACM.

Melissa Hall, Laurens van der Maaten, Laura Gustafson, and Aaron Adcock. 2022. A systematic study of bias amplification. *CoRR*, abs/2201.11706.

Rujun Han, I-Hung Hsu, Jiao Sun, Julia Baylon, Qiang Ning, Dan Roth, and Nanyun Peng. 2021. ESTER: A machine reading comprehension dataset for reasoning about event semantic relations. 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 7543–7559. Association for Computational Linguistics.

He He, Sheng Zha, and Haohan Wang. 2019. Unlearn dataset bias in natural language inference by fitting the residual. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 132–142, Hong Kong, China. Association for Computational Linguistics.

Geoffrey E. Hinton. 2002. Training products of experts by minimizing contrastive divergence. *Neural Comput.*, 14(8):1771–1800.

Martin Josifoski, Marija Sakota, Maxime Peyrard, and Robert West. 2023. Exploiting asymmetry for synthetic training data generation: Synthie and the case of information extraction. *CoRR*, abs/2303.04132.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases. In *International Conference on Machine Learning*, pages 1078–1088. PMLR.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. 2022. Large language models with controllable working memory. *arXiv preprint arXiv:2211.05110*.

Manling Li, Sha Li, Zhenhailong Wang, Lifu Huang, Kyunghyun Cho, Heng Ji, Jiawei Han, and Clare Voss. 2021. The future is not one-dimensional: Complex event schema induction by graph modeling for event prediction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5203–5215, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhongyang Li, Xiao Ding, and Ting Liu. 2018. Constructing narrative event evolutionary graph for script event prediction. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4201–4207.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Annie Louis and Ani Nenkova. 2012. A corpus of general and specific sentences from news. In *LREC*, volume 1818, page 10. Citeseer.

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.

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, 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 2020, Online, November 16-20, 2020*, pages 1158–1172. 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, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1318–1328. Association for Computational Linguistics.

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.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Roger C Schank and Robert P Abelson. 1977. Scripts, plans, goals and understanding: An inquiry into human knowledge structures.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics.

Vered Shwartz and Yejin Choi. 2020. Do neural language models overcome reporting bias? In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 6863–6870. International Committee on Computational Linguistics.

Armand Stricker. 2021. Question answering in natural language: the special case of temporal expressions. In *Proceedings of the Student Research Workshop Associated with RANLP 2021*, pages 184–192, Online. INCOMA Ltd.

Zhongxiang Sun. 2023. A short survey of viewing large language models in legal aspect. *CoRR*, abs/2303.09136.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9.

Haoyu Wang, Muhao Chen, Hongming Zhang, and Dan Roth. 2020. Joint constrained learning for event-event relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 696–706. Association for Computational Linguistics.

Haoyu Wang, Hongming Zhang, Yuqian Deng, Jacob R. Gardner, Dan Roth, and Muhao Chen. 2023. Extracting or guessing? improving faithfulness of event temporal relation extraction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 541–553. Association for Computational Linguistics.

Yiwei Wang, Muhao Chen, Wenxuan Zhou, Yujun Cai, Yuxuan Liang, Dayiheng Liu, Baosong Yang, Juncheng Liu, and Bryan Hooi. 2022. Should we rely on entity mentions for relation extraction? debiasing relation extraction with counterfactual analysis. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 3071–3081. Association for Computational Linguistics.

Yorick Wilks. 1975. A preferential, pattern-seeking, semantics for natural language inference. *Artif. Intell.*, 6(1):53–74.

Nan Xu, Fei Wang, Bangzheng Li, Mingtao Dong, and Muhao Chen. 2022. Does your model classify entities reasonably? diagnosing and mitigating spurious correlations in entity typing. 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 8642–8658. Association for Computational Linguistics.

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

Jeffrey M Zacks, Nicole K Speer, Khena M Swallow, Todd S Braver, and Jeremy R Reynolds. 2007. Event perception: a mind-brain perspective. *Psychological bulletin*, 133(2):273.

Jeffrey M Zacks and Barbara Tversky. 2001. Event structure in perception and conception. *Psychological bulletin*, 127(1):3.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Kai Zhang, Bernal Jiménez Gutiérrez, and Yu Su. 2023. Aligning instruction tasks unlocks large language models as zero-shot relation extractors.

Shuaicheng Zhang, Qiang Ning, and Lifu Huang. 2022. Extracting temporal event relation with syntax-guided graph transformer. In *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 379–390. Association for Computational Linguistics.

Xinyu Zhao, Shih-Ting Lin, and Greg Durrett. 2021. Effective distant supervision for temporal relation extraction. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 195–203, Kyiv, Ukraine. Association for Computational Linguistics.

Ben Zhou, Kyle Richardson, Qiang Ning, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2021. Temporal reasoning on implicit events from distant supervision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1361–1371, Online. Association for Computational Linguistics.

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

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models. *arXiv preprint arXiv:2303.11315*.

Linchao Zhu, Zhongwen Xu, Yi Yang, and Alexander G. Hauptmann. 2017. Uncovering the temporal context for video question answering. *Int. J. Comput. Vis.*, 124(3):409–421.

Rolf A Zwaan, Joseph P Magliano, and Arthur C Graesser. 1995. Dimensions of situation model construction in narrative comprehension. *Journal of experimental psychology: Learning, memory, and cognition*, 21(2):386.

Appendices

A Knowledge-conflict Selection Hyperparameters

In TORQUE, we set an empirical $\mathcal{T}^{\text{Relation Prior}}_{before} = \mathcal{T}^{\text{Relation Prior}}_{after} = \mathcal{T}^{\text{Relation Prior}}_{equal} = 0.25$ by investigating the distribution of *before*, *after*, and *equal* relations. For tense bias, we set $\mathcal{T}^{\text{tense}}_{before} = \mathcal{T}^{\text{tense}}_{after} = 0.25$, and $\mathcal{T}^{\text{tense}}_{equal} = 0.2$ for the relations indicating two events happening simultaneously. For narrative and dependency bias, the threshold is simply set as 0.5. In MATRES, we set $\mathcal{T}_{before} = \mathcal{T}_{after} = 0.3$ and $\mathcal{T}_{equal} = 0.1$.

B Impact of Knowledge Conflict

We compare the model performance on knowledge conflict subsets and the non-conflict subsets to show the impact of knowledge conflicts on model performance in Tab. 5. In general, models perform more poorly on the conflict subsets, compared with those without conflicts. This discovery suggests that the chosen conflict subsets pose greater challenges for PLMs and LLMs, thus validating the efficacy of our conflict detection framework.

C Additional Details of the Models

C.1 Baselines

For TORQUE, the model consists of a one-layered perceptron built on top of RoBERTa. The transformers' output corresponding to the token being analyzed serves as input to the perceptron layer as a sequence tagging task, where the expected output is either 0 or 1, indicating whether this event argument is a correct answer or not. Following the original paper of TORQUE, we fine-tuned RoBERTa-large on the training set of TORQUE, using a batch size of 6 (each input is a concatenation of one passage and one question, and the output is a vector measuring the probability of each event argument token). The learning rate is 1e-5, total epoch is 10, and three random seeds were selected. The experiments are conducted on NVIDIA A5000 GPUs, which takes around 30 minutes for training one epoch.

In MATRES, each data entry is composed of a passage and the corresponding positions of the two event triggers. The model consists of a one-layer perceptron to aggregate the embeddings of the two event triggers provided by the transformers. We use pre-trained Big Bird (Zaheer et al.,

	Conflict		Non-Conflict	
	EM	F1	EM	F1
RoBERTa-large				
Rel.Prior	29.5↓	73.3↓	40.7	74.5
Rel.Prior (warm-up)	50.0↓	75.1↓	75.0	76.2
Narrative	31.4↑	69.0↓	48.4	75.2
Tense	33.5↓	72.9↓	50.7	75.0
Tense (warm-up)	48.4↓	72.4↓	77.3	78.6
Dependency	41.7↑	78.6↓	37.5	81.2
GPT-3.5				
Rel.Prior	4.82↓	59.9↑	4.87	51.1
Rel.Prior (warm-up)	4.62↓	47.0↑	25.0	30.4
Narrative	2.13↓	50.7↑	7.21	44.4
Tense	4.46↓	53.5↑	7.27	42.6
Tense (warm-up)	5.71↓	45.9↑	25.3	30.0
Dependency	2.94↑	57.7↑	2.72	56.7

Table 5: Experimental results on the model performance on knowledge conflict and non-conflict data in TORQUE. The RoBERTa-Large model suffers from performance drop when tested on the conflict subsets. On the contrary, GPT-3.5, when not fine-tuned on the biased training set, suffer less from the knowledge conflict in general. However, there is still a large performance gap on warm-up questions for GPT-3.5, dropping from an EM of around 25% to 5%.

	TORQUE	**MATRES**
Whole Dev Set	1,483	1,000
Rel. Prior (relation)	88	148
Rel. Prior (warm-up)	66	-
Narrative	140	477
Tense (relation)	243	210
Tense (warm-up)	71	-
Dependency	68	42

Table 6: Statistics of each knowledge-conflict subset in TORQUE and MATRES.

2020), a RoBERTa variation that deals with longer documents, following Wang et al. (2023). The experiments are conducted on NVIDIA A5000 GPUs, which takes around 2 minutes for training one epoch.

We then introduce the bias-agnostic baselines that we adopt.

PoE (Hinton, 2002) and Learned-mixin (Clark et al., 2019). In this line of approaches, a biased model is trained to specifically target biased features in the data. The output of the biased model is then combined with the output of the robust model using product of predicted probabilities. This enables the robust model to focus less on the biased features and improve its overall performance. Denote the probabilities predicted by the biased model for element i as b_i, and the probabilities by the robust model as p_i, the ensemble to predict the final label by PoE is:

$$\hat{p}_i = softmax(\log(p_i) + \log(b_i))$$

As PoE assumes conditional independence between the bias in the data and all the features except for bias in the data, which may be too strong, learned-mixin is thus proposed to make the relations between p_i and b_i learnable. A function $g(x)$ of the input x is learned to dynamically adjust how much to trust the biased model, leading to the final estimation as:

$$\hat{p}_i = softmax(\log(p_i) + g(x_i)\log(b_i))$$

However, a model could learn to set $g(x_i)$ to 0 to ignore the effect of biased model, learned-mixin + H is thus proposed by adding an entropy penalty:

$$R = wH(softmax(g(x_i)\log(b_i))$$

Here the entropy function takes the form $H(z) = -\sum_j z_j \log(z_j)$. The entropy term can help encourage the biased term to be non-uniform, providing more biased information.

To train the biased model for all these three baselines, we mask all context except for the event triggers. Other hyperparameters are the same as training a RoBERTa baseline.

Counterfactual Inference (Wang et al., 2022, 2023). Counterfactual inference focus on event trigger bias and frequent label bias that leads to spurious correlations. A causal graph is established to analyze the causal relations between the effect of event triggers, the whole context, and the final prediction. To mitigate event trigger bias and label bias, element-wise subtraction operation is conducted to get the final prediction:

$$y = y_x - \lambda_1 y_{\bar{x},e} - \lambda_2 y_{\bar{x}}$$

where y_x is the prediction given by the model trained on the original data without any masking, $y_{\bar{x},e}$ is the prediction of the model trained on the data where context except for event triggers are masked, and $y_{\bar{x}}$ is the prediction where the model sees nothing as input, which reflects label bias. λ_1 and λ_2 are tuned by conducting 5-fold cross-validations on the training set. The parameters that yield the best cross validation are selected. The search space is $[-1, 1]$ with an interval of 0.1. For TORQUE, $\lambda_1 = -0.8, \lambda_2 = -0.1$. For MATRES, $\lambda_1 = -0.1, \lambda_2 = 0.3$.

AFLite (Sakaguchi et al., 2021; Le Bras et al., 2020). AFLITE, which stands for Lightweight Adversarial Filtering, is an alternative bottom-up approach to algorithmic bias reduction proposed by (Sakaguchi et al., 2021). AFLITE trains an ensemble of linear classifiers on random subsets of the training data and filters other instances in the training data that linear classifiers can correctly classify. The rationale of this baseline is that instances that can be classified correctly by a shallow linear model wound contain artifacts.

In this paper, we use logistic regression as the linear classifier. We repeat training the logistic regression model 20 times on randomly sampled subsets of the training data. Then, we used the trained logistic regression model to predict the labels of the rest of the training instances. We compute a score for every instance e based on the following equation:

$$score(e) = \frac{\textit{the times of } e \textit{ is predicted correctly}}{\textit{the times of } e \textit{ is predicted}}.$$

After repeating, we filter instances that owns a score higher than 0.8. Following previouse work (Sakaguchi et al., 2021), we use dense representations produced by frozen `robert-large` and `bigbird-roberta-large`, instead of manually identified lexical features, to train logistic regression classifiers on TORQUE and MATRES, respectively.

C.2 Large Language Models

Prompts for the Tasks. For TORQUE, the prompt template we use is "Q: {question}, select none or several from {all_events} \n {context} \n A:". Here, question, context are provided in each data entry in TORQUE. all_events indicates all the annotated event triggers in the context. GPT3 is expected to generate none or several events that are the answers to the question given the context. We also check another prompt as an additional analysis, which is "Given the context {context}, {question}, select none or several from all_events} \n A:". The performance analysis are introduced in Tab. 10.

For MATRES, we formulate the problem as a multi-choice question answering (MCQA) task format, as it's inherently a four-way classification task. The prompt takes the form "Given the context:\n {context} \n\n Q: What's the temporal relation between the event {e1} and {e2}? \n Choice A: {e1} happens before {e2}. \n Choice B: {e1} happens after {e2}. \n Choice C: {e1} happens during {e2}. \n Choice D: unknown. \n Answer only with A, B, C, or D. \n\n A: Choice". Here, e1 and e2 are the target event triggers to be studied. The expected output is either A, B, C, or D. In addition, we compare our MCQA template with other templates that have been used in previous works, denoted as template 2 (Chan et al., 2023) and template 3 (Yuan et al., 2023). A comparison of different templates are presented in Tab. 11. We also present the effect of the three prompt templates in Tab. 13, and find that our MCQA template achieves the best performance.

TORQUE	
Rel.Prior	b(kill, arrest, *before*)=0.69, b(bombing, condemn, *before*)=0.67
	b(incident, *happened*)=1,b(host, *future*)=0.91, b(progress, *happening*)=1
Tense	b(`VBN`, `VB`, *before*)=0.64,b(`VBN`, `VBD`, *before*)=0.48, b(`VBD`, `VB`, *before*)=0.55
	b(`VBD`, *happened*)=0.95,b(`VB`, *future*)=0.60, b(`VBZ`, *happening*)=0.62
Narrative	b($p1$<$p2$, *before*)=0.50,b($p1$<$p2$, *after*)=0.32, b($p1$<$p2$, *equal*)=0.03,b($p1$<$p2$, *vague*)=0.13
Dependency	b(`xcomp`, *before*)=0.81,b(`ccomp`, *after*)=0.70
MATRES	
Rel.Prior	b(say, have, *after*)=1,b(rise, close, *before*)=1, b(have, close, *before*)=0.83
Tense	b(`VBN`, `VB`, *before*)=0.80,b(`VBN`, `VBP`, *before*)=0.78, b(`VBD`, `VB`, *before*)=0.70
Narrative	b($p1$<$p2$, *before*)=0.50,b($p1$<$p2$, *after*)=0.32, b($p1$<$p2$, *equal*)=0.03,b($p1$<$p2$, *vague*)=0.13
Dependency	b(`xcomp`, *before*)=0.61,b(`ccomp`, *after*)=0.60

Table 7: Selected top biased event features in TORQUE and MATRES.

Baselines We use In-Context Learning (ICL) and Generative Data Augmentation (GDA) as two intuitive baseline that can be directly comparable to our CDA method. For ICL, specifically, we retrieve one passage-question pair in TORQUE, and retrieve one example per relation from *before*, *after*, *equal*, and *unknown* as as set of exemplars for MATRES (denoted as 1-shot), to form the ICL demonstration. Note that ICL is considered few-shot learning while our method is purely zero-shot. We study the variability of different sets of exemplars as well as the effect of 1-shot and 3-shot ICL in Tab. 13. We can find that the performance of ICL is quite stable across different sets of random exemplars, and 3-shot exemplars help on template 1 but not the other two templates.

In addition, we add an additional baseline named *Generative Data Augmentation* (GDA) that performs exemplar generation without a counterfactual guidance. That is to say, we ask LLMs to generate exemplars under all relations from R, instead of only the counterfactual relations.

Counterfactual Data Augmentation We introduce how to do Counterfactual Data Augmentation (CDA) for both PLMs and LLMs.

In CDA for PLM, we generate augmented data at scale. For TORQUE, we first retrieve all event pairs that are identified as biased in the training set. For an event-relation triple (e_1, e_2, r), where r is identified as knowledge-conflict, which appears less frequently in the training set, we ask Flan-T5 to generate some context where e_1, e_2 happens under relation r, to augment the undervalued distribution of these two events under the conflict relation r. The prompt is: "Write a story where e_1 happens r' e_2:". We set temperature as 1 and use greedy decoding to get the results. After generating the context, the question associated with the context is thus *Q=What happened* r' e_2 and the corresponding answer is e_1. We do similar generations for warm-up questions that asks what events have happened / is happening / will happen. We first acquire events that are knowledge-conflict with regard to a relation $r \in \{$*happened, will happen, happening*$\}$, and randomly sample two or events that are conflict with regard to r. We ask Flan-T5 "Write a story where e_1 and e_2 r". The corresponding question associate with the generated context is then *Q=What have happened/will happen in the future/is happening?*, based on what r is. After such augmentations, we conduct an additional filtering step by selecting only knowledge-conflict augmented data. We keep a proportion of augmented data that is scored with low loss by a fine-tuned PLM on TORQUE to boost the initial learning process when trained on augmented data. For MATRES, the prompt given to Flan-T5 is "Write a story where e_1 happens r' e_2". Then r is used as the final label.

In CDA for LLM, we generate demonstrations to perform in-context learning. In MATRES, for an example (c, e_1, e_2, r), we first ask the LLM to predict the temporal relation r_{LLM}. Then we use the same prompt as in CDA for PLM to generate counterfactual examples dedicated to the event pair (e_1, e_2), under relations other than r_{LLM}. The generated examples are thus served as exemplars. In TORQUE, the pipeline is more complicated. An entry is composed of context c, the set of event triggers E in c, the question q, and the answers a, which is a subset of E. We first ask an LLM to predict the answers a_{LLM}, which is also expected to be a subset of E. We then ask the LLM itself to generate some context where the ground answers are sampled from $E - a_{LLM}$, using the same prompt as in CDA for PLM. Examples on MATRES are presented in Tab. 12.

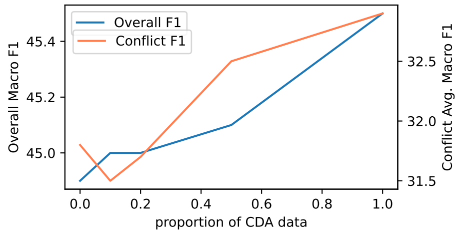

Figure 3: Effect of varying proportions of Counterfactual Data Augmentation (CDA) on MATRES. Models benefit from increased amounts of CDA data.

D Additional Ablations

In this section, we compare our Counterfactual Data Augmentation method with other popular data augmentation methods to show the effectiveness of CDA with regard to knowledge conflict mitigation. Specifically, we adopt EDA and Synonym replacement as representative text-editing-based data augmentation baselines, and we use a Generative Data Augmentation (GDA) baseline to automatically generate task data using the same backbone language model, Flan-T5-11B, to generate training data without counterfactual constraints. The only difference between GDA and CDA is that GDA does not use counterfactual constraints, and GDA can serve as an ablation to study the effect of counterfactual constraints. The results for TORQUE are presented in Tab. 8 and the results for MATRES are presented in Tab. 9

	all		Rel.Prior (relation)		Rel.Prior (warm-up)		Narrative (relation)		Tense (relation)		Tense (warm-up)		Dep. (relation)		Confl.Avg.	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
RoBERTa-L	50.4	75.7	29.5	73.3	50.0	75.1	31.4	69.0	33.5	72.9	48.4	72.4	41.7	78.6	39.1	73.6
+EDA	50.2	75.5	33.5	74.2	**50.7**	71.7	30.7	67.9	33.9	71.8	**50.0**	69.4	41.1	**79.6**	40.0	72.4
+Synonym	49.7	**76.1**	28.0	71.8	49.5	72.3	29.5	68.7	33.5	72.0	47.4	69.7	35.8	75.9	37.3	71.7
+GDA	49.9	75.8	30.3	73.8	50.5	74.0	**31.7**	**69.1**	34.4	72.6	34.4	72.6	**49.3**	71.5	38.4	72.3
+CDA	**51.0**	**76.1**	**33.7**	**75.4**	50.0	**75.9**	30.7	68.6	**35.5**	**73.1**	48.8	**73.2**	44.1	79.1	**40.5**	**74.2**

Table 8: Experimental results on the TORQUE dataset using different data augmentation techniques. Exact-Match (EM) rate and Macro-F1 (F1) scores are reported. Best-performed results are **bold-faced** and the second-best are underlined.

	all		Rel. Prior		Narrative		Tense		Dependency		Confl.Avg.	
	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro
RoBERTa-large	70.8	44.9	59.7	28.5	59.2	27.1	54.8	33.2	58.5	38.3	58.0	31.8
+EDA	70.5	**46.0**	60.9	29.8	58.7	27.4	55.1	33.8	60.0	38.4	58.7	32.4
+Synonym	70.4	45.0	59.6	28.3	59.5	26.9	55.5	33.7	61.9	41.3	57.8	32.5
+GDA	**72.2**	43.6	62.0	27.2	57.5	25.3	54.0	31.4	58.1	36.0	57.9	30.0
+CDA (Ours)	**72.2**	45.5	**61.5**	29.3	58.8	27.3	57.2	**35.1**	**62.2**	39.9	**59.9**	**32.9**

Table 9: Experimental results on MATRES using different data augmentation techniques. We use two evaluation metrics, Micro-F1 (denoted as Micro) and Macro F1 (denoted as Macro). Best-performed results are **bold-faced** and the second-best are underlined.

	EM	F1
CDA (1-shot)	5.16	44.6
CDA (3-shot)	14.5	50.1
template 1 (zero-shot)	8.36	45.5
template 2 (zero-shot)	8.16	45.9
template 1 (1-shot)-1	4.52	43.4
template 1 (1-shot)-2	6.00	44.7
template 1 (1-shot)-3	13.1	46.9
template 1 (1-shot)-avg	7.87	45.0
template 2 (1-shot)-1	9.51	50.5
template 2 (1-shot)-2	12.6	51.2
template 2 (1-shot)-3	10.5	48.8
template 2 (1-shot)-avg	10.9	50.2
template 1 (3-shot)-1	13.0	46.7
template 1 (3-shot)-2	16.4	48.5
template 1 (3-shot)-3	11.2	48.2
template 1 (3-shot)-avg	13.5	47.8
template 2 (3-shot)-1	19.3	56.1
template 2 (3-shot)-2	18.6	55.4
template 2 (3-shot)-3	23.3	54.0
template 2 (3-shot)-avg	20.4	55.2

Table 10: Experimental results on TORQUE using different prompt templates.

MATRES				
Strategies	**Template input**	**GPT3.5**	**Gold**	**T/F**
Prompt 1 (MCQA)	`Given the context:\n` Jim Unruh, Unisys's president, said he is approaching next year with caution. He said the strength of the world-wide economy is suspect, and doesn't see much revenue growth in the cards. He also said that the price wars flaring up in parts of the computer industry will continue through next year. He said the move toward standard operating systems means customers aren't locked into buying from their traditional computer supplier and can force prices down. `\n\nQ: What's the temporal relation between the event "suspect" and "flaring"? \n Choice A: suspect happens before flaring. \n Choice B: suspect happens after flaring. \n Choice C: suspect happens during flaring. \n Choice D: unknown. \Answer only with A, B, C, or D. \n\nA: Choice`	A	A	T
Prompt 2 (Chan et al., 2023)	`Determine the temporal order from "suspect" to "flaring" in the following sentence: ""`Jim Unruh, Unisys's president, said he is approaching next year with caution. He said the strength of the world-wide economy is suspect, and doesn't see much revenue growth in the cards. He also said that the price wars flaring up in parts of the computer industry will continue through next year. He said the move toward standard operating systems means customers aren't locked into buying from their traditional computer supplier and can force prices down. `"". Only answer one word from AFTER, BEFORE, EQUAL, VAGUE. Answer:`	BEFORE	BEFORE	T
Prompt 3 (Yuan et al., 2023)	`Given the document` Jim Unruh, Unisys's president, said he is approaching next year with caution. He said the strength of the world-wide economy is suspect, and doesn't see much revenue growth in the cards. He also said that the price wars flaring up in parts of the computer industry will continue through next year. He said the move toward standard operating systems means customers aren't locked into buying from their traditional computer supplier and can force prices down. `and a list of temporal relations [before, after, vague, equal] and event triggers` suspect and flaring. `what is the temporal relation between` suspect and flaring? `Answer vague if unsure. Keep the answer short and concise.`	before	before	T

Table 11: Prompt templates for MATRES.

MATRES				
Strategies	**Template input**	**GPT3.5**	**Gold**	**T/F**
Zero-shot	`Given the context:\n` [Context] `\n\nQ: What's the temporal relation between the event "`e_1`" and "`e_2`"? \n Choice A: `e_1` happens before `e_2`. \n Choice B: `e_1` happens after `e_2`. \n Choice C: `e_1` happens during `e_2`. \n Choice D: unknown. \Answer only with A, B, C, or D. \n\nA: Choice`	A	B	F

Counterfactual generation	`Generate a paragraph where event `e_1` happens before `e_2`:`
	`Generate a paragraph where event `e_1` happens after `e_2`:`
	`Generate a paragraph where event `e_1` happens in the same time as `e_2`:`
	`Generate a paragraph where the temporal relation of `e_1` and `e_2` cannot be determined based on the context:`
		c_A, c_B, c_C, c_D	/	/

CDA prompting	`Given the context:\n `c_B` \n\nQ: What's the temporal relation between the event "`$\cdots$` A: Choice B`
	`Given the context:\n `c_C` \n\nQ: What's the temporal relation between the event "`$\cdots$` A: Choice C`
	`Given the context:\n `c_D` \n\nQ: What's the temporal relation between the event "`$\cdots$` A: Choice D`
	`Given the context:\n` [Context] `\n\nQ: What's the temporal relation between the event "`e_1`" and "`e_2`"? \n Choice A: `e_1` happens before `e_2`. \n Choice B: `e_1` happens after `e_2`. \n Choice C: `e_1` happens during `e_2`. \n Choice D: unknown. \Answer only with A, B, C, or D. \n\nA: Choice`
		B	B	T

Table 12: A running example of CDA in MATRES. The LLM itself first predict the label of the example, where the prediction is denoted as r_{LLM}. Then, the LLM is asked to generate four context given e_1 and e_2 under four different temporal relations, using the prompts in the second columns, where the corresponding generated context are then c_A, c_B, c_C, c_D. Then, the generated contexts other than under the predicted relation r_{LLM} are used as demonstrations for in-context learning.

	Micro F1	Macro F1
CDA (1-shot)	51.3	**30.0**
CDA (3-shot)*	51.5	26.3
template 1 zero-shot (MCQA)	53.3	19.7
template 2 (Chan et al., 2023)	52.1	17.1
template 3 (Yuan et al., 2023)	13.4	13.0
template 1 (1-shot)-1	52.3	18.5
template 1 (1-shot)-2	53.1	20.4
template 1 (1-shot)-3	51.6	18.4
template 1 (1-shot)-avg	52.3	19.1
template 1 (1-shot)-MV	52.1	19.0
template 2 (1-shot)-1	49.9	22.0
template 2 (1-shot)-2	49.3	22.1
template 2 (1-shot)-3	50.1	19.8
template 2 (1-shot)-avg	49.8	21.3
template 2 (1-shot)-MV	50.0	20.6
template 3 (1-shot)-1	32.7	18.6
template 3 (1-shot)-2	34.4	20.7
template 3 (1-shot)-3	28.8	17.8
template 2 (1-shot)-avg	32.0	19.0
template 3 (1-shot)-MV	31.9	18.5
template 1 (3-shot)-1*	57.5	24.1
template 1 (3-shot)-2*	**57.0**	28.0
template 1 (3-shot)-3*	50.0	23.4
template 1 (3-shot)-avg*	54.8	25.2
template 1 (3-shot)-MV*	57.0	24.4
template 2 (3-shot)-1*	46.5	18.2
template 2 (3-shot)-2*	47.0	18.1
template 2 (3-shot)-3*	47.5	24.9
template 2 (3-shot)-avg*	47.0	20.4
template 2 (3-shot)-MV*	48.0	19.2
template 3 (3-shot)-1*	35.5	21.3
template 3 (3-shot)-2*	29.0	15.7
template 3 (3-shot)-3*	34.0	20.2
template 2 (3-shot)-avg*	32.8	19.1
template 3 (3-shot)-MV*	33.0	19.2

Table 13: Experimental results on MATRES using different prompt templates. $*$ indicates we test the performance on the same 200 randomly down-sampled examples from MATRES. We run 3 different random seeds per few-shot in-context learning experiments. 'avg' indicates the average between the three runs, and 'MV' indicates the majority voting across the three runs.