A Rationale-centric Counterfactual Data Augmentation Method for Cross-Document Event Coreference Resolution

Anonymous ACL submission

Abstract

Based on Pre-trained Language Models (PLMs), event coreference resolution (ECR) systems have demonstrated outstanding performance in clustering coreferential events across documents. However, the state-of-the-art system exhibits an excessive reliance on the 'trigger lexical matching' spurious pattern in the input mention pair text. We formalize the decision-making process of the baseline ECR system using a Structural Causal Model (SCM), aiming to identify spurious and causal associations (i.e., rationales) within the ECR task. Leveraging the debiasing capability of counterfactual data augmentation, we developed a rationale-centric counterfactual data augmentation method with LLM-in-the-loop. This method is specialized for pairwise input in the ECR system, where we conduct direct interventions on triggers and context to mitigate spurious associations while emphasizing causation. Our approach achieves state-of-the-art performance on three popular cross-document ECR benchmarks and demonstrates robustness in out-of-domain scenarios [1].

1 Introduction

The goal of cross-document event coreference resolution (ECR) is to group event mentions referring to the same real-world event together across documents. It is an essential task in NLP and has provided valuable prior event-related knowledge for many downstream tasks, e.g., topic detection and tracking (Allan et al., 1998), multi-hop question answering (Yang et al., 2018) and information extraction (Humphreys et al., 1997). In real life, event coreference systems commonly assist decision-makers in important fields such as intelligence analysis and security event warnings (Palantir, 2023).

To resolve the task, existing state-of-the-art ECR systems perform binary classification to pairwise

[1] Our code is available at XXX

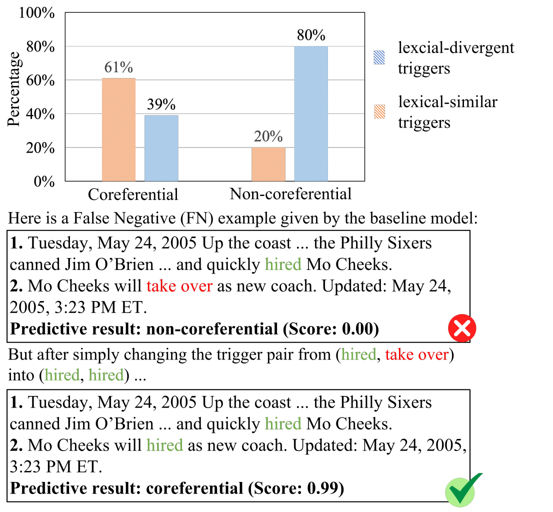

Figure 1: The distribution of 'lexical triggers matching' in mention pairs from the ECB+ training set, along with a false negative example from Held et al.'s system which shows that forcing the event trigger in the second mention to lexically match the first one causes a significant change in the predicted coreference score.

compare event mentions (Barhom et al., 2019; Yu et al., 2022; Caciularu et al., 2021; Held et al., 2021). In their pipelines, a fraction of coreferential and non-coreferential mention pairs are retrieved from the corpus to fine-tune a cross-encoder, which is used as a coreference scorer to gauge the likelihood of pairwise events being coreferential. Finally, coreferential mentions are merged into clusters based on the predicted coreference score.

However, most coreference scorers are troubled by the curse of 'lexical triggers matching' (Ravi et al., 2023; Ahmed et al., 2023). Figure 1 demonstrates that when constructing event-mention pairs, it is natural that coreferential mentions frequently share lexical similar event triggers, whereas non-coreferential mentions typically have lexical divergent ones. This skewed feature distribution results in trigger-centric ECR systems that use event representation in Appx. A.2.2 (Yu et al., 2022; Held et al., 2021) excessively relying on 'lexical triggers matching', which is a spurious association. Essentially, what truly determines the ECR outcome of event mentions is the coreference of event-relevant arguments, which include (non-)human participants, times, locations, and actions (i.e., event triggers) (Cybulska and Vossen, 2015). In other words, these deeper semantic features constitute the rationales of the ECR task, as they demonstrate the task's corresponding causal associations. Unfortunately, some state-of-the-art systems only learn the surface feature of trigger term similarity (Ravi et al., 2023; Ahmed et al., 2023) (Figure 1 also provides an example), and the underlying reason could be that lexical-similar trigger words in the data often correspond to coreference. Through data governance, we can adjust the distribution of key features in the training data to resolve this issue.

Counterfactual data augmentation (DA) is a promising way for debiasing the classification system (Garg et al., 2019; Madaan et al., 2021), which enhances the robust causal thinking ability of models with a human-like logic: '*What the output label would be if certain phrases within the input text were altered?*'. In practice, we can intervene with rationales in the original example input text to ensure minimal editing to flip the output label, thus generating counterfactual augmented data (CAD). The minimal editing constraint is to prevent the introduction of unnecessary noise into the augmented data, which allows the model trained with CAD to focus directly on the causal associations from rationales, rather than on other parts of the input text (Keane and Smyth, 2020).

Given this, we propose LLM-RCDA, a **r**ationale-centric **c**ounterfactual **DA** method with a **l**arge **l**anguage **m**odel (LLM) in the loop, aiming to enhance the model to think causally and understand deeper semantics in the pairwise context. As shown in Figure 2, our method focuses on intervening triggers and rationales in the event-mention sentence. In the phase of trigger intervention, lexical divergent synonyms of the original trigger are generated to force the system to capture the coreferential meaning between triggers, while in the phase of context intervention, we use the LLM to merely change the rationales of the target event-mention based on prompts, and keep the discourse of the event-mention unchanged.

To evaluate the efficacy of our method, we evaluate our method on three popular cross-document ECR benchmarks: ECB+ (Cybulska and Vossen, 2014), Football Coreference Corpus (FCC) (Bugert et al., 2021) and Gun Violence Corpus (GVC) (Vossen et al., 2018). Our enhanced system achieves state-of-the-art performance on all of them, with improvements varying from 1.8 to 2.3 CoNLL F1 over baselines. On ECB+, our approach significantly surpasses the performance of directly employing LLMs, showcasing its superiority to the current LLM-QA paradigm in the task. Additionally, the cross-corpus experiment on the out-of-the-domain data shows a robustness improvement of our method, with a 7.2 CoNLL F1 gain over the baseline.

To the best of our knowledge, we are the first to evaluate and analyze the performance of popular LLMs on the cross-document ECR benchmark, and the first to formalize the decision process of the mainstream ECR model from a causal view. Moreover, we are the first to utilize rationale-centric CAD generated by the LLM to causally enhance the ECR system.

2 Related Work

Event Coreference Resolution Currently, the pretrained language models (Devlin et al., 2019; Liu et al., 2019) have significantly enhanced the contextual semantics of text data. In recent ECR systems (Kenyon-Dean et al., 2018; Caciularu et al., 2021; Held et al., 2021; Chen et al., 2023), the pairwise representation of events (known as cross-encoding) becomes mainstream. Such representation combines contextual embeddings with pairwise token-level trigger embeddings to represent the event-mention pair and expects the embeddings to encode the event-relevant argument information implicitly. Some other works enhanced the pairwise events representation by explicitly fusing the encoding of event-relevant arguments which are extracted by semantic Role Labeling (SRL) systems (Barhom et al., 2019; Zeng et al., 2020; Yu et al., 2022), achieving success in performance improvements. In our approach, we also emphasize the crucial features that influence event coreference, such as argument features. However, we do not alter the model structure or existing representation methods. Instead, we induce the model to learn these key features through rationale-centric counterfactual data augmentation, thereby enhancing the causal reasoning capability of the ECR system.

Counterfactual Data Augmentation Counterfactual data augmentation is widely used in NLP tasks

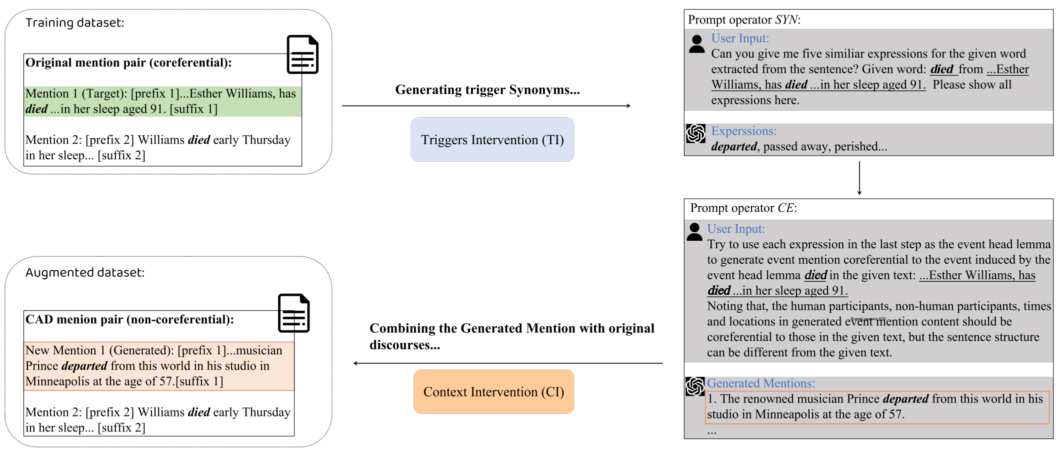

Figure 2: The procedure of our rationale-centric counterfactual DA with LLM-in-the-loop (LLM-RCDA).

to improve the system's performance and robustness. The methods for generating counterfactual augmented data (CAD) vary across tasks, such as SA (Yang et al., 2022a), NLI (Pope and Fern, 2021; Robeer et al., 2021) and NMT (Liu et al., 2021). In early works, CAD generation either relies on a human-in-the-loop system (Kaushik et al., 2020; Srivastava et al., 2020) or relies on PLMs (Tucker et al., 2021; Wu et al., 2021) or external knowledge bases automatically (Wang and Culotta, 2020; Yang et al., 2022a). Recently, Li et al. explored and confirmed the feasibility of using LLMs to generate CAD on SA, NLI, NER and RE tasks, demonstrating good efficiency. Our work is the first one specifically designed for the ECR task, which also involves the LLM-in-the-loop to automatically and efficiently construct required CAD that meets the task's causal requirements.

3 Baselines

LLM Currently, little work has been done to evaluate LLMs' performance on cross-document ECR. To achieve this task, LLMs must possess the ability to comprehend and process a long context for understanding and comparing event mentions across multiple documents. Therefore, we utilize Claude-2(100K maximum input length) (Anthropic, 2023) and GPT-4 (8K maximum input length) (OpenAI, 2023), two LLMs with strong long context comprehensions, to perform the evaluation. We compare the zero-shot results of LLMs with a rule-based system which employs the same head lemma matching technique (Barhom et al., 2019), an end-to-end neural system (Cattan et al., 2020), the state-of-the-art pipeline system (Held et al., 2021) and our causally enhanced system.

Fully fine-tuned Baseline Our method is built upon the state-of-the-art ECR system (Held et al., 2021), which serves as our main baseline. Held et al. applies the discourse coherence theory to create event mention pairs for training and inference. For each event mention, they retrieve the K nearest mentions in a trained event representation space to establish matches. These event-mention pairs are encoded by RoBERTa-large (Liu et al., 2019) (Appx. A.2.2), and then fed to fine-tune the coreference scorer. During inference, the system prunes non-coreferential mention pairs and merges the remaining greedily to construct the coreferential cluster. On top of that, we also compare with an ELMo-based system (Barhom et al., 2019), where the entity and event coreference are jointly modelled; an end-to-end cross-document coreference resolution system for both event and entity (Cattan et al., 2020); a robust feature-based system (Bugert et al., 2021); a CDLM-based system (Caciularu et al., 2021), which uses a larger longformer (Beltagy et al., 2020) model for document-level representation; a system with pairwise triggers and arguments representation (Yu et al., 2022) and a system trained with pruned mention pairs (Ahmed et al., 2023).

4 Method

We analyze the decision process of the ECR system on event coreference with Structural Causal Model (SCM) (Pearl, 2000). Formally, the event coreference process of the baseline ECR system into the following equation:

$$Y = f\left(T\left(X\right), A\left(X\right), U\right) \quad (1)$$

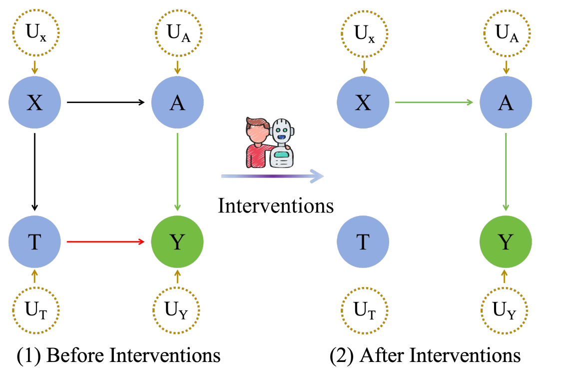

Figure 3: SCM illustration. (1) stimulates the decision process of the baseline ECR system; (2) shows the decision process of the causally enhanced system after interventions.

where X and Y represent the input pairwise data and the output label, A represents the coreference of counterpart event-relevant augments in the context of input, T denotes the scenario of triggers matching lexically and U refers to the unobserved variable. Equation 1 demonstrates the fact that 'lexical triggers matching' influences the prediction heavily in our baseline ECR system. At the same time, the coreferential counterparts of event-relevant arguments are rationales for ECR according to the definition of event coreference (Cybulska and Vossen, 2015).

4.1 Causality Analysis and Interventions

Equation 1 can be represented by the causal graph in Figure 3 (1), where the input pairwise data X serves as a confounder of triggers matching (T) and the coreference result (Y), indicating a backdoor path $T \leftarrow X \rightarrow Y$, where the ECR system does not capture the causality by recognizing the semantic rationales for ECR provided in the context completely but relies on the linguistic surface feature of the trigger pair, which is a spurious association to the coreference prediction.

To address this problem, we perform the Trigger Intervention (TI) on the path $X \rightarrow T$, as well as the Context Intervention (CI) on the path $X \rightarrow A \rightarrow Y$. TI aims to decompose the spurious association from the lexical matching of trigger terms. We use the prompt operator *SYN* (Appx. Table 8) to generate synonyms of existing trigger terms but lexical divergence from them, which expands the limited expressions of triggers in the corpus. This allows us to adjust the distribution of trigger-matching features and ultimately block the confounding influence. As for CI, it aims to emphasize that rationales are a key factor influencing the output label and enhance the ECR understanding in a causal manner. We leverage CAD to achieve the purpose. Therefore, we developed a counterfactual generation algorithm specifically for ECR, ensuring that the generated counterfactual data not only have diverse trigger expressions but also emphasize causal features. We will introduce this algorithm in Section 4.2.

4.2 Counterfactual Generation

We design an LLM-in-the-loop counterfactual generation algorithm to generate CAD candidates automatically for a given original event-mention pair *MP*. All prompt operators in Algorithm 1 are presented in the appendix section A.3.2. Since we follow the discourse setup in the system of Held et al., each event-mention text is associated with a maximum of $2w + 1$ sentences totally in a discourse context. Therefore, *MP* can be symbolized as $(S^{(1)}_{i-w}...S^{(1)}_{i}...S^{(1)}_{i+w};\ S^{(2)}_{j-w}...S^{(2)}_{j}...S^{(2)}_{j+w})$, where $S^{(1)}_{\cdot}$ represents the sentence associated with the first event-mention in the pair, and $S^{(2)}_{\cdot}$ represents the sentence associated with the second one. For simplicity, we divide the text of an event-mention into prefix sentences $(S^{(1)}_{i-w},...,S^{(1)}_{i-1})$, mention sentence $S^{(1)}_{i}$ and suffix sentences $(S^{(1)}_{i+1},...,S^{(1)}_{i+w})$.

When generating CAD, we only consider making adjustments to the text related to the first mention. As shown in Algorithm 1, a mention sentence within the *MP* will serve as the target for intervention with its selection depending on the original example's label (lines 2-3&9-10). It will then undergo Trigger Intervention (lines 4&11) followed by the generation of several non-coreferential (or coreferential) mention sentence candidates using the LLM (lines 5&12). If the original *MP* is coreferential, generating CAD becomes somewhat simpler, where we only need to sequentially replace the target mention sentence $S^{(1)}_{i}$ with each of the generated candidates s_g (lines 20-21). In this way, the counterfactual dataset D_{cf} is constructed while adhering to the constraint of minimal edits. However, additional operations are required to generate coreferential CAD from the original non-coreferential example. This is necessary to ensure that all event-relevant arguments co-refer with their counterparts within the pairwise context, as the definition of event coreference (Cybulska and Vossen, 2015). Therefore, a new event-mention with discourse context, which co-refers to the second event-mention in

MP needs to be constructed. We begin with utilizing a paraphraser to generate prefix and suffix sentences based on those of the second event-mention (line 23-24) and then combine them with the mention sentence s_g to construct the required event-mention text $\tilde{m}_1$ in line 25. In line 26, we pair $\tilde{m}_1$ with the original text of the second event-mention in the original *MP*. Thus, the desired commonsense reasonable CAD is constructed with relatively minor text changes.

The plausible counterfactual should ensure minimal changes compared with the original data, otherwise, it may hurt the model's performance and robustness (Keane and Smyth, 2020). Inspired by Yang et al. (2022a), we use MoverScore, an edit-distance scoring metric (Zhao et al., 2019), to evaluate the plausibility of our generated counterfactual data. The average MoverScore on sampled counterfactual data is 0.7314, demonstrating the minor changes in our CAD and validating the plausibility of these generated instances.

Algorithm 1 LLM-in-the-loop Counterfactual Generation

Input: Original data *MP* with label *Y*; Large language model LLM; trigger terms of two mentions ($T^{(1)}$,$T^{(2)}$).
Prompt operators: Synonyms generator SYN; Coref events generator CE; Non-coref events generator NCE; Paraphraser $PARA$.
Output: Counterfactual dataset D_{cf}

1: **while** sentence s in MP **do**
2:   **if** $Y == coref$ **then**
3:     **if** $s == S_i^{(1)}$ **then**
4:       $T_{syns}^{(1)} = LLM\left(SYN, T^{(1)}\right)$
5:       $S_{gens} = LLM\left(NCE, T_{syns}^{(1)}, S_i^{(1)}\right)$
6:     **else**
7:       continue
8:     **end if**
9:   **else if** $Y == not\ coref$ **then**
10:     **if** $s == S_i^{(2)}$ **then**
11:       $T_{syns}^{(2)} = LLM\left(SYN, T^{(2)}\right)$
12:       $S_{gens} = LLM\left(CE, T_{syns}^{(2)}, S_j^{(2)}\right)$
13:     **else**
14:       continue
15:     **end if**
16:   **end if**
17: **end while**
18: **while** sentence s_g in S_{gens} **do**
19:   **if** $Y == coref$ **then**
20:     $\tilde{m}_1 = (S_{i-w}^{(1)}, ..., s_g, ..., S_{i+w}^{(1)})$
21:     $MP_{cf} = concat\left\{\tilde{m}_1, \left(S_{j-w}^{(2)}, ..., S_{j+w}^{(2)}\right)\right\}$
22:   **else if** $Y == not\ coref$ **then**
23:     $pre = LLM\left(PARA,\ (S_{j-w}^{(2)}, ..., S_{j-1}^{(2)})\right)$
24:     $suf = LLM\left(PARA,\ (S_{j+1}^{(2)}, ..., S_{j+w}^{(2)})\right)$
25:     $\tilde{m}_1 = concat\{pre, s_g, suf\}$
26:     $MP_{cf} = concat\left\{\tilde{m}_1, \left(S_{j-w}^{(2)}, ..., S_{j+w}^{(2)}\right)\right\}$
27:   **end if**
28: **end while**
29: Add MP_{cf} to the set D_{cf}
30: **return** D_{cf}

5 Experimental Settings

5.1 Evaluation Metrics and Datasets

Evaluation Metrics Since we do not conduct identification of the mention, we use B^3 F1 proposed by Bagga and Baldwin (1998) to select the best model during training because Moosavi and Strube (2016) to identify that it has the fewest relevant drawbacks under the condition (Held et al., 2021). For a comprehensive comparison with recent works, we also report MUC (Vilain et al., 1995), $CEAF_e$ (Luo, 2005), LEA (Moosavi and Strube, 2016) and CoNLL F1 (the arithmetic average of the value of B^3, MUC and $CEAF_e$).

Datasets Our experiments are performed on three benchmarks: Event Coreference Bank Plus (ECB+), Football Coreference Corpus (FCC) and Gun Violence Corpus (GVC). For ECB+, we follow the data split by Cybulska and Vossen (2015), while following the data split by Bugert et al. (2021) for FCC and GVC. The data details are presented in Appx. Table 5.

5.2 Implementation Details

LLM To evaluate the ECR performance of LLMs, we employ the document template prompt (Appx. Table 7) as suggested by Le and Ritter (2023). This prompt has shown considerably superior performance compared to the standard QA prompt and competes well with existing unsupervised entity coreference resolution systems. The evaluation is performed on the test set of ECB+. In practice, we begin by clustering the documents into golden topics, and subsequently, we evaluate the event coreference within each topic individually. ECB+ does not include cross-topic coreference links, so this operation will overlook incorrect coreference links across topics, thus simplifying the task. We do this to ensure that each prompt does not exceed the maximum acceptable length of GPT-4.

Fully Fine-tuned Experiments To compare with the main baseline (Held et al., 2021) fairly, we follow their setup. For main experiments on three benchmarks, we retrieve the nearest 15 (K=15) and 5 (K=5) mention pairs for training and inference in main experiments on three benchmarks. For the ablation study and generalization test, we retrieve 5

Methods	CoNLL F1
Lemma Matching (Barhom et al., 2019)	76.5
E2E Neural System (Cattan et al., 2020)	81
Pipeline System (Held et al., 2021)	85.7
Causally Enhanced System (Ours)	**86**
Claude-2	56.9↓
GPT-4	<u>70↓</u>

Table 1: LLMs performance compared with other systems. The best overall result is highlighted in bold, while the best result among LLMs is underlined.

(K=5) mention pairs for both training and inference. Considering a trade-off between the training time and the increasing amount of augmented data, we only add two CAD for each original data from the top 5 nearest pairwise data in the training set, and keep the others unchanged. After data augmentation, we receive 68.2K, 35.8K and 97.3K mention pairs to train the cross-encoder on ECB+, FCC and GVC respectively. All of our models are trained and evaluated on a single Nvidia Tesla V100 GPU. All our augmented data originates from GPT-3.5-turbo (OpenAI, 2023).

6 Experimental Results

LLMs Table 1 shows the cross-document ECR results by LLMs. Claude-2 lags significantly behind GPT-4 by 13.1 CoNLL F1 points. After checking the answers, we found that the low performance of Claude-2 is attributed to its neglect of a significant number of golden event mentions that should have been classified. Claude-2 missed 15% of the total golden mentions of ECB+, including all golden mentions within Topic 37. GPT-4 predicted more completely, with only 16 out of a total of 1780 golden event mentions being missed. Although GPT-4 performs better than Claude-2, it still falls short compared to other baselines. GPT-4's performance decreases by 6.5 CoNLL F1 points compared to the simple baseline (Barhom et al., 2019), which relies solely on event head lemma matching for coreference. Also, it falls significantly behind the current state-of-the-art pipeline method (Held et al., 2021). In particular, our method outperforms GPT-4 with 16.0 CoNLL F1 points. The experimental results provide direct evidence that LLMs are not enough to solve the cross-document ECR problem and also demonstrate the effectiveness of our causally enhanced system based on LLM-RCDA.

Causally Enhanced System As shown in Table 2, our causally enhanced ECR system has achieved state-of-the-art performance across multiple evaluation metrics. In terms of CoNLL F1, the system surpasses the baseline by 1.8, 2.6, and 2.3 points on ECB+, FCC, and GVC, respectively.

In the case of ECB+, we observe a significant improvement in Recall for our enhanced system compared to the baseline system, as measured by MUC, B^3, and LEA, with an average improvement of 3.5 points. This improvement can be attributed to the trigger intervention in Algo. 1. The introduction of diverse trigger expressions enhances the model's comprehension of event semantics, thereby rectifying false negatives caused solely by literal differences in trigger terms.

FCC and GVC represent single-domain datasets focused on football and gun violence news. They include a substantial volume of coreferential event mentions across various topics, resulting in a considerable number of challenging negatives (i.e., non-coreferential event-mention pairs with very similar contexts). Nevertheless, several metrics exhibit notable enhancements in Precision, such as a 13.1-point increase for B^3 on FCC and a 5.7-point increase for $CEAF_e$ on GVC. These results suggest that our LLM-RCDA method is well-suited for such scenarios, as it guides the model to make decisions based on fine-grained causal terms within the context.

7 Analysis

Ablation Study Our LLM-RCDA algorithm assists the cross-document ECR system in disentangling spurious patterns via Trigger Intervention (TI) while emphasizing causal associations through Context Intervention (CI) when understanding the pairwise context. To investigate the efficacy of each intervention, we modified the augmented data generation algorithm (Algo. 1), conducting ablations for both TI and CI. During the TI ablation (Algo. 2 in Appx.), we no longer diversify the expressions of the target mention's trigger, resulting in trigger pairs in the augmented data that remain consistent with those in the original mention pair. In the CI ablation (Appx. Algo. 3), we intentionally introduced more substantial modifications to the text of the original mention pair. This deliberate approach leads to the generation of relatively implausible counterfactuals, aiming to reduce the emphasis on rationales within the context (Keane and Smyth, 2020; Yang et al., 2022a). The augmented data from TI and CI ablation are short for TIA and CIA.

Methods	MUC			B^3			$CEAF_e$			LEA			CoNLL
	R	P	F1	R	P	F1	R	P	F1	R	P	F1	F1
ECB+													
Barhom et al. (2019)	77.6	84.5	80.9	76.1	85.1	80.3	81	73.8	77.3	-	-	-	79.5
Cattan et al. (2020)	85.1	81.9	83.5	82.1	82.7	82.4	75.2	78.9	77	-	-	-	81
Bugert et al. (2021)	76	76.1	76.1	71.8	81.2	76.2	72.2	72.1	72.2	55.1	67.9	60.8	74.8
Caciularu et al. (2021)	87.1	**89.2**	88.1	84.9	87.9	86.4	83.3	81.2	82.2	76.7	77.2	76.9	85.6
Held et al. (2021)	87	88.1	87.5	85.6	87.7	86.6	80.3	85.8	82.9	74.9	73.2	74	85.7
Yu et al. (2022)	**88.1**	85.1	86.6	**86.1**	84.7	85.4	79.6	**83.1**	81.3	-	-	-	84.4
Ahmed et al. (2023)	80	87.3	83.5	79.6	85.4	82.4	83.1	75.5	79.1	70.5	73.3	71.9	81.7
Baseline System$_{-DA}$	82.5	88.6	85.4	82.6	**88.6**	85.5	**85.1**	78.5	81.7	74	77.4	75.6	84.2
Enhanced System$_{+DA}$	86.4	88.6	**87.5**	85.7	88.4	**87**	84.7	82.2	**83.4**	**77.4**	**79.6**	**78.5**	**86**
FCC													
Barhom et al. (2019)	-	-	-	36	83	50.2	-	-	-	-	-	-	-
Bugert et al. (2021)	82.7	78.3	80.4	**70.8**	38.3	49.2	28.2	40.4	33.2	**60.4**	30.4	39.8	54.3
Held et al. (2021)	**86.4**	75.7	80.7	61.6	65.4	63.5	39.1	65.3	48.9	47.2	57	51.6	64.4
Baseline System$_{-DA}$	79.2	**88.9**	**83.7**	64.4	61.6	63	**73.3**	46	56.5	58.1	47.2	52.1	67.7
Enhanced System$_{+DA}$	79.2	88.2	83.4	66.8	**74.7**	**70.5**	72.7	**46.7**	**56.9**	60.1	**60.1**	**60.1**	**70.3**
GVC													
Barhom et al. (2019)	-	-	-	81	66	72.7	-	-	-	-	-	-	-
Bugert et al. (2021)	66.3	78.1	71.7	49.9	73.6	59.5	60.9	38.2	47	38.2	56.5	45.6	59.4
Held et al. (2021)	**91.8**	91.2	**91.5**	82.2	83.8	83	75.5	**77.9**	**76.7**	79	**82.3**	**80.6**	83.7
Ahmed et al. (2023)	84	91.1	87.4	79	76.4	77.7	69.6	52.5	59.9	74.1	63.9	68.6	75
Baseline System$_{-DA}$	89.3	**92.3**	90.8	82.1	85.7	83.8	76.6	67.5	71.7	76.9	78.8	77.8	82.1
Enhanced System$_{+DA}$	90.4	92.1	91.3	**84.8**	**86.8**	**85.8**	**78.9**	73.2	76	**79.8**	80.7	80.2	**84.4**

Table 2: Performance comparison of different cross-document ECR systems on ECB+, FCC and GVC. Baseline System$_{-DA}$ results are obtained by reproducing the work of Held et al. without data augmentation. Enhanced System$_{+DA}$ is trained by the original data combined with CAD from LLM-RCDA. Bold values represent the overall best results, while underlined values indicate results that beat the Baseline System.

Training Data (data volume)	Percentage	MoverScore	MUC			B^3			$CEAF_e$			LEA			CoNLL
			R	P	F1	R	P	F1	R	P	F1	R	P	F1	F1
ORI$_{-TI/-CI}$ (14.3K)	70.0%	-	83	85.9	84.4	84.2	85.7	85	81.9	78.7	80.2	74.6	74.2	74.4	83.2
ORI&TAD$_{+TCDA}$(42.8K)	19.9%	-	79.8	88	83.7	82	87.9	84.9	85.1	76.5	80.6	73.2	75.1	74.2	83.1
ORI&TIA$_{-TI/+CI}$ (42.8K)	91.7%	0.7324	82.1	85.7	83.9	83.3	85.2	84.2	**82.7**	78.6	80.6	73.9	73.7	73.8	82.9
ORI&CIA$_{+TI/-CI}$ (42.8K)	19.9%	0.6971	84.1	86.2	85.2	84.5	**86.5**	85.5	82.4	80	81.2	74.7	75.5	75.1	84
ORI&CAD$_{+TI/+CI}$ (42.8K)	19.9%	0.7314	**87.2**	**86.5**	**86.8**	**86.7**	85.4	**86.1**	81.8	**82.7**	**82.2**	**77.2**	**76.2**	**76.7**	**85**

Table 3: Results of the system trained with different data combinations on ECB+. Percentage denotes the proportion of examples with lexical similar triggers in coreferential pairs. $\cdot_{+/-TI(CI)}$ indicates whether the Trigger (Context) Intervention is included or excluded. $\cdot_{+TCDA}$ means that the augmented data is from Ravi et al.'s TCDA method.

In table 3, we compare the results obtained by training on the original data with those achieved using various data combinations. As observed from the Percentage column, ORI&TIA shows a notably higher percentage of coreferential data involving similar trigger pairs, approximately 92% v.s. 70% in ORI. This extremely imbalanced distribution exacerbates the model's reliance on 'lexical triggers matching', which is a spurious association. Consequently, the model trained with ORI&TIA shows evident performance degradation across various coreference metrics when compared with ORI (e.g., 0.3 CoNLL F1 points decrease), despite the larger dataset size and the high-quality counterfactuals with its MoverScore 0.7324. Despite CIA being a relatively implausible counterfactual (MoverScore 0.6971), it still brings a gain of 0.8 CoNLL F1 points compared to the ORI baseline, attributed to TI. However, due to the ablation of CI, the performance of ORI&CIA lags behind ORI&CAD by 1.0 CoNLL F1 points. Overall, ORI&CAD, which combines both interventions, outperforms other data combinations in multiple metrics, whether focusing on Recall or Precision. The ablation study highlights the importance of TI and validates our analysis of spurious associations and causal associations. It also underscores that the most efficient way to enhance cross-document ECR performance is by fully utilizing our LLM-RCDA algorithm.

Comparison with Temporal commonsense DA Ravi et al. enriched the event context by introducing possible preceding or succeeding scenarios related to event mentions based on the Temporal Commonsense Event Coreference Data Augmentation (TCDA) method, thereby increasing the distinction between events. They also designed an inference-enhanced pairwise scorer specifically to capture such temporal information. We reproduced

Methods	MUC	B^3	CEAFe	LEA	CoNLL
Bugert et al. (2021)	52.4	33.2	27	14.1	33.2
Baseline System (Held et al., 2021)	57.5	38.4	31.5	23.9	42.5
Enhanced System (ours)	**68.6**	**46.0**	**35.0**	**31.2**	**49.9**

Table 4: Performance comparison of our enhanced system with baselines on the OOD dataset FCC.

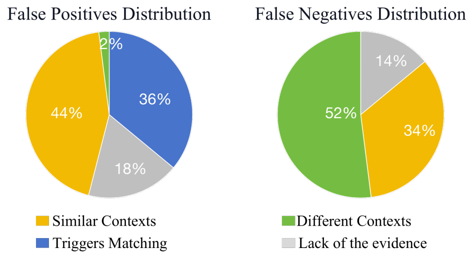

Figure 4: Error distribution.

their method (Appx. Algo. 4) for the temporal commonsense augmented data (TAD) and incorporated it into the original data (ORI) to train the cross-document ECR system (Held et al., 2021). From results in Table 3, we observe that ORI&CAD outperforms ORI&TAD across various metrics, improving by 1.9 CoNLL F1, while ORI&TAD's performance is worse than that of the system trained solely with ORI. This observation indicates that TCDA heavily relies on their well-designed scorer. In contrast, LLM-RCDA improves performance without requiring system modifications, with more convenience and scalability.

Robustness in the Generalization Test We train the system on ECB+ but test it on FCC to evaluate its out-of-the-domain (OOD) robustness. For comparison, we take the cross-corpus results from Bugert et al. (2021) and the reproduced results from Held et al. (2021) as our baselines. The enhanced system is trained with ORI and CAD from LLM-RCDA.

As shown in Table 4, our enhanced system shows the best performance in multiple metrics. It surpasses the baseline system by 7.4 CoNLL F1 points, proving the stronger robustness of LLM-RCDA. To better demonstrate how our enhanced system performs more robustly in OOD scenarios, we perform an error analysis. We randomly sample 100 errors made by the baseline system but correctly predicted by our enhanced ECR system. The 100 error samples consist of 50 false positives (FPs) and 50 false negatives (FNs). FPs refer to the non-coreferential examples being wrongly predicted as coreferential, while FNs refer to coreferential examples being incorrectly predicted as non-coreferential. According to the context of these mention pairs, we manually categorize them into four error types: 'Triggers matching', 'Similar contexts', 'Different contexts' and 'Lack of the evidence', and analyze the error distribution of FPs and FNs.

From Figure 4, we can observe that 44% of FPs fall under the category of 'Similar contexts,' where the pairwise contexts show a notable overlap in lexical similar terms. These examples pose a significant challenge for the model, as the presence of similar features strongly implies coreference, making it more prone to making incorrect predictions. Conversely, 'Different contexts' examples account for the majority of FNs, constituting 52% of the errors. Such examples have dissimilar contextual content, making it challenging for the model to assign a coreferential label. Furthermore, we notice that the issue of 'Triggers matching' is serious in FPs, accounting for 36% of the errors. In these examples, only the triggers exhibit high lexical similarity. If the model pays attention to the rationales in the context, it would avoid making such errors. This highlights that the baseline system, especially in the OOD context, heavily relies on spurious association but neglects the rationales in the context when making predictions.

The error analysis reveals a deficiency in the baseline system's comprehension of complete semantics, leading to instances where similar or dissimilar context features can significantly impact prediction outcomes. In contrast, our improved ECR system effectively addresses these errors. This can be attributed to LLM-RCDA, which not only eliminates spurious associations but also strengthens the system's ability to capture additional information related to event-relevant argument relations within the context, which is the key to prediction.

8 Conclusion

We proposed a novel rationale-centric counterfactual data augmentation method specialized for the pairwise text input of cross-document ECR systems, which leverages interventions of the LLM to mitigate the spurious association and enhance causal associations for event coreference decisions. Experimental results verify the significant performance and robustness improvement of the enhanced ECR system with our method.

Limitations

The LLM used in our LLM-RCDA method is GPT-3.5-turbo (OpenAI, 2023), and it is not an open-source model. In the future, we plan to attempt to implement our method based on some open-source large models, such as LLaMA (Touvron et al., 2023). Additionally, we aim to apply it to other cross-document tasks, not limited to event coreference resolution. We are also interested in adapting our method to other pairwise input text tasks, such as natural language inference, stance detection, and entity coreference resolution.

Ethical Statement

We honour the Code of Ethics of ACL.

References

Shafiuddin Rehan Ahmed, Abhijnan Nath, James H. Martin, and Nikhil Krishnaswamy. 2023. $2*n$ is better than n^2: Decomposing event coreference resolution into two tractable problems.

James Allan, Jaime G. Carbonell, George R. Doddington, Jonathan Yamron, and Yiming Yang. 1998. Topic detection and tracking pilot study final report.

Anthropic. 2023. Ai research and products that put safety at the frontier.

Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 79–85, Montreal, Quebec, Canada. Association for Computational Linguistics.

Shany Barhom, Vered Shwartz, Alon Eirew, Michael Bugert, Nils Reimers, and Ido Dagan. 2019. Revisiting joint modeling of cross-document entity and event coreference resolution.

Cosmin Bejan and Sanda Harabagiu. 2010. Unsupervised event coreference resolution with rich linguistic features. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1412–1422, Uppsala, Sweden. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer.

Michael Bugert, Nils Reimers, and Iryna Gurevych. 2021. Generalizing cross-document event coreference resolution across multiple corpora. *Computational Linguistics*, 47(3):575–614.

Avi Caciularu, Arman Cohan, Iz Beltagy, Matthew E. Peters, Arie Cattan, and Ido Dagan. 2021. Cdlm: Cross-document language modeling.

Arie Cattan, Alon Eirew, Gabriel Stanovsky, Mandar Joshi, and Ido Dagan. 2020. Streamlining cross-document coreference resolution: Evaluation and modeling.

Xinyu Chen, Sheng Xu, Peifeng Li, and Qiaoming Zhu. 2023. Cross-document event coreference resolution on discourse structure. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4833–4843, Singapore. Association for Computational Linguistics.

Agata Cybulska and Piek Vossen. 2014. Using a sledgehammer to crack a nut? lexical diversity and event coreference resolution. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4545–4552, Reykjavik, Iceland. European Language Resources Association (ELRA).

Agata Cybulska and Piek Vossen. 2015. " bag of events" approach to event coreference resolution. supervised classification of event templates. *Int. J. Comput. Linguistics Appl.*, 6(2):11–27.

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.

Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H. Chi, and Alex Beutel. 2019. Counterfactual fairness in text classification through robustness.

William Held, Dan Iter, and Dan Jurafsky. 2021. Focus on what matters: Applying discourse coherence theory to cross document coreference. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1406–1417, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kevin Humphreys, Robert Gaizauskas, and Saliha Azzam. 1997. Event coreference for information extraction. In *Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data.

Mark T. Keane and Barry Smyth. 2020. Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable ai (xai).

Kian Kenyon-Dean, Jackie Chi Kit Cheung, and Doina Precup. 2018. Resolving event coreference with supervised representation learning and clustering-oriented regularization. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 1–10, New Orleans, Louisiana. Association for Computational Linguistics.

Nghia T. Le and Alan Ritter. 2023. Are large language models robust zero-shot coreference resolvers?

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Yongqi Li, Mayi Xu, Xin Miao, Shen Zhou, and Tieyun Qian. 2023. Large language models as counterfactual generator: Strengths and weaknesses.

Qi Liu, Matt Kusner, and Phil Blunsom. 2021. Counterfactual data augmentation for neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 187–197, Online. Association for Computational Linguistics.

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

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, page 25–32, USA. Association for Computational Linguistics.

Nishtha Madaan, Inkit Padhi, Naveen Panwar, and Diptikalyan Saha. 2021. Generate your counterfactuals: Towards controlled counterfactual generation for text.

Nafise Sadat Moosavi and Michael Strube. 2016. Which coreference evaluation metric do you trust? a proposal for a link-based entity aware metric. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 632–642, Berlin, Germany. Association for Computational Linguistics.

OpenAI. 2023. Creating safe agi that benefits all of humanity.

Palantir. 2023. Ai-powered operations, for every decision.

Judea Pearl. 2000. *Causality: Models, Reasoning and Inference*. New York: Cambridge University Press.

Quintin Pope and Xiaoli Z. Fern. 2021. Text counterfactuals via latent optimization and shapley-guided search.

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

Marcel Robeer, Floris Bex, and Ad Feelders. 2021. Generating realistic natural language counterfactuals. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3611–3625, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Megha Srivastava, Tatsunori Hashimoto, and Percy Liang. 2020. Robustness to spurious correlations via human annotations.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Mycal Tucker, Peng Qian, and Roger Levy. 2021. What if this modified that? syntactic interventions via counterfactual embeddings.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*.

Piek Vossen, Filip Ilievski, Marten Postma, and Roxane Segers. 2018. Don't annotate, but validate: a data-to-text method for capturing event data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Zhao Wang and Aron Culotta. 2020. Robustness to spurious correlations in text classification via automatically generated counterfactuals.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S. Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models.

Linyi Yang, Jiazheng Li, Pádraig Cunningham, Yue Zhang, Barry Smyth, and Ruihai Dong. 2022a. Exploring the efficacy of automatically generated counterfactuals for sentiment analysis.

Linyi Yang, Lifan Yuan, Leyang Cui, Wenyang Gao, and Yue Zhang. 2022b. Factmix: Using a few labeled in-domain examples to generalize to cross-domain named entity recognition.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Xiaodong Yu, Wenpeng Yin, and Dan Roth. 2022. Pairwise representation learning for event coreference. In *The 11th Joint Conference on Lexical and Computational Semantics*, page 69.

Yutao Zeng, Xiaolong Jin, Saiping Guan, Jiafeng Guo, and Xueqi Cheng. 2020. Event coreference resolution with their paraphrases and argument-aware embeddings. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3084–3094.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

A Appendix

ECB+	Train	Dev	Test
Topics	25 (50)	8 (16)	10 (20)
Documents	574	196	206
Sentences	9366	2837	3505
Event Mentions	3808	1245	1780
FCC	Train	Dev	Test
Topics	3	1	1
Documents	207	117	127
Sentences	7018	3648	4274
Event Mentions	1604	680	1074
GVC	Train	Dev	Test
Topics	1 (170)	1 (37)	1 (34)
Documents	358	78	74
Sentences	7607	1325	1360
Event Mentions	5313	977	1008

Table 5: Statistics for ECB+, FCC and GVC. For Topics rows, values outside the parentheses indicate the number of topics, while values inside the parentheses represent the number of subtopics of the data split (e.g., 25 (50) means that 25 topics including 50 subtopics are in the data split).

A.1 Dataset Details

Event Coreference Bank Plus (ECB+) The ECB+ corpus is the most popular benchmark for the cross-document ECR task (Cybulska and Vossen, 2014). It is an extension of the Event Coref Bank corpus (ECB) annotated by Bejan and Harabagiu (2010). ECB+ expands on the original topics by incorporating various seminal events as subtopics and annotates the coreference relationships between events within each topic. In terms of statistics, the ECB+ corpus consists of 982 documents, covering 43 topics, and includes 26,712 coreference links among 6,833 event mentions.

Football Coreference Corpus (FCC) The Football Coreference Corpus (FCC) serves as a benchmark for cross-document Event Coreference Resolution (ECR) specifically in the domain of football tournaments (Bugert et al., 2021). This dataset is unique as it includes a significant number of cross-subtopic event coreference links, which is uncommon but highly valuable for research purposes. Overall, the FCC comprises 451 documents and contains a total of 145,272 links between 3,563 event mentions.

Gun Violence Corpus (GVC) The GVC (Vossen et al., 2018) is a challenging cross-document ECR benchmark. It consists of 510 documents that are lexically similar, posing a challenge for document clustering. The dataset comprises 29,398 links between 7,298 event mentions, with all the links being within subtopics.

A.2 Experimental Details

A.2.1 LLM Evaluation

We utilize MUC, B^3, $CEAF_e$, LEA and CoNLL, five metrics to evaluate the performance of Claude-2 and GPT-4 on cross-document ECR. Claude-2 accepts 100K input, but we have no access to adjust the parameters currently, so we interact with it directly on its official website [2]. The GPT-4 we used accepts 8K tokens shared between the prompt and output in maximum [3]. We set the temperature parameter as zero for reproducibility, and adjust the maximum length as 4500. Other parameters are set as default.

A.2.2 Representation

Following Held et al. (2021), we arrange a text snippet for each mention by including sentences from a context window preceding and following the mention sentence. Subsequently, a fine-tuned bi-encoder initialized with pre-trained weights from RoBERTA-large (326M) (Liu et al., 2019) encodes the token-level boundary representation used by (Lee et al., 2017). The mention is then represented by concatenating these token-level representations. As a result, the pairwise representation for the mention pair can be constructed as follows:

$$[m_1, m_2, m_1 \odot m_2]$$

where m_1 and m_2 refer to the representation of the first and second mention in the pair, and $m_1 \odot m_2$ denotes the element-wise multiplication between two mention.

The original data, along with the augmented data, is encoded into the pairwise representation and then passed through a cross-encoder for training a coreference classifier.

A.2.3 Hyperparameters

Hyperparameters for training the bi-encoder and cross-encoder are presented in Table 6.

A.3 Prompts

A.3.1 The prompt for LLM evaluation

To evaluate the performance of LLM on the cross-document ECR, we follow and modify the document template prompt (Table 7) proposed by Le

[2] https://www.anthropic.com
[3] https://platform.openai.com

	bi-encoder	cross-encoder
Batch Size	16	40 (8)
Learning Rate	0.00001	0.00001
Maximum Epochs	50	40
Optimizer	Adam	Adam
Warmup Proportion	0.1	0.1
Early Stop Patience	10	5
Max Grad Norm	1	1
Input Turncation	512	512

Table 6: Hyperparameters settings. We set the batch size as 40, 40, 8 for cross-encoder when performing corpus-tailored study on ECB+, FCC and GVC respectively, while setting it as 8 for ablation and generalization study.

and Ritter (2023), which outperforms several existing unsupervised coreference systems in the entity coreference resolution task on the OntoNotes dataset (Le and Ritter, 2023).

A.3.2 The prompt for data generation

Algorithm 1 contains the following prompt-based operations: generating the synonyms SYN and non-coreferential(coreferential) event mention sentence $NCE(CE)$, paraphrasing the given text $PARA$. From Table 8, prompt Step 1 illustrates the operation *SYN* in line 4, while prompt Step 2 corresponds to the operation *NCE* in line 12. Meanwhile, the prompt Step 1 and Step 2 in Table 9 refer to the operation SYN and CE in lines 11 and 12 respectively. Table 10 shows the $PARA$ prompt about paraphrasing the discourse context, corresponding to lines 22 and 23. Table 11 displays the prompts used to generate TAD, which is utilized in Algorithm 4 in the "Comparison with Temporal Commonsense DA" subsection of Section 7.

A.4 Data Examples

Table 12 and 13 show examples of Counterfactual Augmented Data (CAD), Trigger Intervention Ablation (TIA) Data, Context Intervention Ablation (CIA) Data as well as the Temporal commonsense Argumented Data(TAD), where the special token <s> and <\s> indicate the start and end of the sentence. The average MoverScore for CAD and TIA are both approximately 0.73, whereas for CIA it is around 0.69, which indicates CAD and TIA are more plausible counterfactual data. Yang et al. (2022b) have demonstrated that the performance and robustness of the model cannot be infinitely improved by adding more counterfactuals. Therefore, in our data augmentation process, we randomly select two CAD from counterfactual candidates generated by Algorithm 1 for each specified original data. To ensure fair comparison, we also two TIA and CIA instances for each original data in the ablation study. The generation process of TIA, CIA and TAD are presented in algorithm 2, 3 and 4 respectively.

A.5 Case study for generalization test

"Triggers matching"[4] refers to samples that have a pair of trigger terms that are highly similar either in their lexical or semantic attributes. When the contexts in two sentences, excluding the trigger terms, are noticeably dissimilar, the ECR system should have no trouble predicting them as non-coreferential samples, given that it effectively captures the underlying meaning of the sentences. As an illustration, we consider a false positive sample that falls under the category of "Triggers matching" as follows:

> 1. France were really made to work for this, in a way they haven't been throughout this ***World Cup***, and in a way that made them look so mortal and often so much less than world champions for the first time.
> 2. Brazil has too much showtime for Mexico at ***World Cup***, regardless of cries of bad acting

When we manually replace the term '***World Cup***' in the first sentence with '***World championship of soccer***', which is synonymous with the trigger term used in the initial mention, the coreference score predicted by the original ECR system significantly decreases by over 99%. This highlights the original system's heavy reliance on lexical matching of triggers when making coreference predictions. However, even in such a modified scenario, the enhanced system is able to generate the correct prediction consistently, suggesting that the enhanced system mitigates spurious impact of the confounder on cross-encoder prediction. **"Similar contexts"** pertains to samples where the sentence pair contains a notable number of lexically similar terms in their respective contexts. This can potentially lead the model to incorrectly predict them as coreferential. To illustrate, we continue using an example from the false positives:

> 1. England had 11 shots in this match but failed to score. They also had 11 shots in their previous ***match*** against Panama , scoring six times.
> 2. Gareth Southgate made eight changes from England's ***win*** against Panama.

Such examples pose difficult challenges within false positives. Successfully predicting them ne-

[4] In this paper, the method used to evaluate lexical matching of trigger terms involves comparing the fuzz ratio of the etyma forms of two terms. A fuzz ratio greater than 80 is considered indicative of a lexical match.

Document Template prompt:
Annotate all event mentions in the following text with coreference clusters. Use Markdown tags to indicate clusters in the output, with the following format [mention](#cluster_name):
(...) Lindsay Lohan [checks into](#) Betty Ford Center May 03, 2013. After skipping out on [entering](#) a Newport Beach rehabilitation facility and [facing](#) the prospect of [arrest](#) for [violating](#) her [probation](#), Lindsay Lohan has [checked into](#) the Betty Ford Center to [begin](#) a 90-day court-mandated [stay](#) in her reckless [driving](#) [conviction](#). (...)
LLM Response:
(...) Lindsay Lohan [checks into](#cluster1) Betty Ford Center May 03, 2013. After skipping out on [entering](#cluster2) a Newport Beach rehabilitation facility and [facing](#cluster3) the prospect of [arrest](#cluster4) for [violating](#cluster5) her [probation](#cluster6), Lindsay Lohan has [checked into](#cluster1) the Betty Ford Center to [begin](#cluster7) a 90-day court-mandated [stay](#cluster8) in her reckless [driving](#cluster9) [conviction](#cluster10). (...)

Table 7: An example of Document Template prompt for event coreference resolution, where mention head lemmas are marked with [mention](#), allowing the LLM to provide the coreference cluster IDs for based on the semantics of the entire document (Le and Ritter, 2023).

cessitates a comprehensive comprehension of the underlying semantics and cannot solely rely on distinguishing between similar and dissimilar features within a sentence. In other words, the model must go beyond surface-level analysis to accurately identify coreference relationships in such cases. Considering the example above, even though two event mentions '***match***' and '***win***' share the same subject '*England*' and object '*Panama*', they refer to two different events '*match between England and Panama*' and '*England's victory*' respectively.

"Different contexts" refers to examples where two sentences are described in a distinct manner, despite potentially sharing some similar entity features. Unlike "Similar contexts" samples, these examples pose challenges within false negatives. For demonstration purposes, we take an example from false negatives:

1. In 1998 when Croatia also reached the ***last four*** of the World Cup they were coached by Miroslav Blazovic, who was also born in Bosnia.
2. This is the furthest Croatia has advanced since 1998 when it made a miraculous run to the ***semifinals***.

In this example, both the two sentences and the event triggers are described in different ways. However, through a comprehensive understanding of the entire context, we can ascertain that the two mentions are indeed referring to the same event, specifically the *'1998 World Cup semi-final'*. Fortunately, our enhanced ECR system adeptly handles such cases, and we firmly believe that our rationales-based data enhancement approach enhances the model's capability to capture the semantic nuances within the paired texts as a whole.

"Lack of evidence" denotes the samples in which it is difficult to find evidence in the sentence pair to explain the ground truth. Perhaps we need to consider a longer context to find corresponding information. Considering an example from the false positives:

1. Defender Samuel Umtiti scored the winning goal for 1998 ***champions*** in the second half with a towering header from Antoine Griezmann's corner.
2. Why omens are good for France to repeat 1998 ***glory*** Best defender in the tournament?

There is no direct evidence to explain why two mentions above are non-coreferential. We consider that through data augmentation, our enhanced system guesses these examples correctly based on limited semantics.

Demonstration:
Please perform a two-step task based on commonsense inference.
Step1: Can you give me five similar expressions for the given word extracted from a sentence? Given word: ’***fire***’ from ’A man has been charged on suspicion of arson following a ***fire*** that devastated a Somerset supermarket.’ Please show all expressions here.
Step2: Try to use each expression in Step1 as the event head lemma to generate event mention not coreferential to the event induced by the event head lemma ’***fire***’ in the given text: ’A man has been charged on suspicion of arson following a ***fire*** that devastated a Somerset supermarket.’ Noting that, the human participants, non-human participants, times and locations in generated event mention content should be not conferential to and different from those in the given text, keeping the sentence structure same as the given text as possible.
Expressions: ***blaze, inferno, conflagration, flames, combustion***
Event mentions:
1. ’A woman has been charged with arson after a ***blaze*** at the local library in Bristol.’
2. ’A teenager has been charged with arson after an ***inferno*** at the shopping mall in Birmingham.’
3. ’A basketball athlete has been charged with arson after a ***conflagration*** at the historic museum in York.’
...
Input:
Please perform a two-step task based on commonsense inference.
Step1: Can you give me five similar expressions for the given word extracted from a sentence? Given word: ’***died***’ from ’Golden girl of screen and pool, Esther Williams, has ***died*** peacefully in her sleep aged 91.’ Please show all expressions here.
Step2: Try to use each expression in Step1 as the event head lemma to generate event mention not coreferential to the event induced by the event head lemma ’***died***’ in the given text: ’Golden girl of screen and pool, Esther Williams, has ***died*** peacefully in her sleep aged 91.’ Noting that, the human participants, non-human participants, times and locations in generated event mention content should be not conferential to and different from those in the given text, keeping the sentence structure same as the given text as possible.
LLM Response:
Expressions: ***departed, expired, perished, left us, passed away***
1. ’The renowned musician Prince ***departed*** from this world in his studio in Minneapolis at the age of 57.’
2. ’The legendary actor Marlon Brando ***expired*** in his mansion in Los Angeles at the age of 80.’
3. ’The famous singer Whitney Houston ***perished*** in her hotel room in New York at the age of 48.’
...

Table 8: The prompt about producing non-coreferential mention candidates for the input mention, where step 1 indicates the prompt operator SYN and step 2 indicates the prompt operator NCE.

Demonstration:
Please perform a two-step task based on commonsense inference.
Step1: Can you give me five similiar expressions for the given word extracted from the sentence? Given word:’***free throw***’ from ’McDermott broke Rodney Buford’ s school scoring record of 2,116 points with a ***free throw*** with 4:43 to play in the first half.’ Please show all expressions here.
Step2: Try to use each expression in Step1 as the event head lemma to generate event mention coreferential to the event induced by the event head lemma ***’free throw’*** in the given text: ’McDermott broke Rodney Buford’ s school scoring record of 2,116 points with a ***free throw*** with 4:43 to play in the first half.’ Noting that, the human participants, non-human participants, times and locations in generated event mention content should be coreferential to those in the given text, but the sentence structure can be different from the given text.
Expressions: ***basketball shot, scoring toss, uncontested shot, charity toss, foul shot***
Event mentions:
1. ’McDermott broke Rodney Buford’s school scoring record of 2,116 points with a ***basketball shot*** at 4:43 remaining in the first half.’
2. ’McDermott surpassed Rodney Buford’s school scoring record of 2,116 points with a ***scoring toss*** in the first half with 4:43 left on the clock.’
3. ’McDermott set a new school scoring record of 2,116 points with an ***uncontested shot*** during the first half with 4:43 remaining.’
...
Input:
Please perform a two-step task based on commonsense inference.
Step1: Can you give me five similar expressions for the given word extracted from a sentence? Given word: ***’protect’*** from ’"Microsoft has released a security update for Internet Explorer that will help ***protect*** its customers from malicious attacks”, said the company in a statement.’ Please show all expressions here.
Step2: Try to use each expression in Step1 as the event head lemma to generate event mention not coreferential to the event induced by the event head lemma ***’protect’*** in the given text:’"Microsoft has released a security update for Internet Explorer that will help ***protect*** its customers from malicious attacks”, said the company in a statement.’ Noting that, the human participants, non-human participants, times and locations in generated event mention content should be coreferential to those in the given text, but the sentence structure can be different from the given text.
LLM Response:
Expressions: ***secured, fortifying, defend, safeguard, shield***
1. ’A statement from Microsoft confirms that the company has ***secured*** its customers from malicious attacks by releasing a security update for Internet Explorer.’
2. ’Microsoft has released a security update for Internet Explorer, ***fortifying*** its customers against malicious attacks, according to a recent statement.’
3. ’Microsoft has released a security update for Internet Explorer in order to ***defend*** its customers from malicious attacks, according to a statement.’
...

Table 9: The prompt about producing coreferential mention candidates for the input mention, where step 1 indicates the prompt operator SYN and step 2 indicates the prompt operator CE.

Demonstration:
We have a snippet of text: 'Indianapolis Colts clinch playoff berth with win over Kansas City Chiefs December 23, 2012. Going into week 16, the Indianapolis Colts controlled their own destiny of making it to post-season play. The Colts could clinch a playoff berth with a win over the Kansas City Chiefs or a Pittsburgh Steelers loss. As they have done all season, the Colts refused to let their fate be decided by anyone other than themselves. The young team fought hard to defeat the Chiefs in another fourth quarter victory, 20-13. Although they started the game with a three-and-out, the Colts were able to light up the scoreboard first.'
The text can be divided into prefix, mention and suffix as following:
Prefix: 'Indianapolis Colts clinch playoff berth with win over Kansas City Chiefs December 23, 2012. Going into week 16, the Indianapolis Colts controlled their own destiny of making it to post-season play.'
Mention: 'The Colts could clinch a playoff berth with a win over the Kansas City Chiefs or a Pittsburgh Steelers loss.'
Suffix: 'As they have done all season, the Colts refused to let their fate be decided by anyone other than themselves. The young team fought hard to defeat the Chiefs in another fourth quarter victory, 20-13. Although they started the game with a three-and-out, the Colts were able to light up the scoreboard first.'
Can you paraphrase Prefix and Suffix in five different ways, where human participants, non-human participants, times, locations and actions in generated examples are conferential to those in the original one?
Prefix:
1. The Indianapolis Colts secured a spot in the playoffs by defeating the Kansas City Chiefs on December 23, 2012, in week 16 of the season. The Colts had control of their own destiny and could have also clinched a playoff berth with a Pittsburgh Steelers loss.
2. On December 23, 2012, the Indianapolis Colts earned a playoff spot by winning against the Kansas City Chiefs in week 16. The Colts had the power to determine their own fate and could have also secured a playoff berth if the Pittsburgh Steelers lost.
...
Suffix:
1. Throughout the season, the Indianapolis Colts refused to let anyone else decide their fate. In another fourth-quarter victory, the young team fought hard to defeat the Kansas City Chiefs with a score of 20-13, despite starting the game with a three-and-out.
2. The Indianapolis Colts demonstrated their determination to control their own destiny throughout the season. They fought hard to secure another fourth-quarter victory against the Kansas City Chiefs, winning 20-13, despite starting the game with a three-and-out.
...
Input:
We have a snippet of text: 'Microsoft has said that an emergency security update has fixed a flaw in Internet Explorer that left millions of computers vulnerable to hacking and hijack. The software patch, which was released last night, has closed a loophole that has seen the computers of at least two million users hacked by cyber criminals. The problem related to a "zero day " flaw that tricked people into visiting an infected website, enabling hackers to gain access to online banking passwords and e-shopping logon details. "Microsoft has released a security update for Internet Explorer that will help protect its customers from malicious attacks", said the company in a statement." Like a vaccine developed to fight a virus, this "security update" will protect computers only if it is installed." Computers that are set to automatically update and install software will already be protected by the patch. Those users who manage their own updates are advised to download the fix as soon as possible from the Microsoft website.'
The text can be divided into prefix, mention and suffix as following:
Prefix: 'Microsoft has said that an emergency security update has fixed a flaw in Internet Explorer that left millions of computers vulnerable to hacking and hijack. The software patch, which was released last night, has closed a loophole that has seen the computers of at least two million users hacked by cyber criminals. The problem related to a "zero day" flaw that tricked people into visiting an infected website, enabling hackers to gain access to online banking passwords and e-shopping logon details.'
Mention: "'Microsoft has released a security update for Internet Explorer that will help protect its customers from malicious attacks', said the company in a statement."
Suffix: 'Like a vaccine developed to fight a virus, this "security update" will protect computers only if it is installed." Computers that are set to automatically update and install software will already be protected by the patch. Those users who manage their own updates are advised to download the fix as soon as possible from the Microsoft website.'
Can you paraphrase Prefix and Suffix in five different ways, where human participants, non-human participants, times, locations and actions in generated examples are conferential to those in the original one?
LLM Response:
Prefix:
1. 'Microsoft has addressed a flaw in Internet Explorer that left millions of computers vulnerable to hacking and hijack by releasing an emergency security update. The software patch, which was released last night, has closed a loophole that allowed cyber criminals to hack into the computers of at least two million users by exploiting a "zero day" flaw that tricked people into visiting an infected website.'
2.'An emergency security update from Microsoft has fixed a flaw in Internet Explorer that left millions of computers vulnerable to hacking and hijack. The software patch, which was released last night, has closed a loophole that allowed cyber criminals to hack into the computers of at least two million users by exploiting a "zero day" flaw that enabled them to gain access to online banking passwords and e-shopping logon details.'
...
Suffix:
1. 'Microsoft has likened the security update for Internet Explorer to a vaccine that helps protect computers from malicious attacks. The update will only be effective if it is installed, and computers set to automatically update and install software will already be protected by the patch. Users who manage their own updates are advised to download the fix from the Microsoft website as soon as possible.'
2. 'The security update for Internet Explorer released by Microsoft is similar to a vaccine that helps protect computers from malicious attacks. However, the update will only be effective if it is installed. Computers set to automatically update and install software will already be protected by the patch, but those who manage their own updates should download the fix from the Microsoft website as soon as possible.'

Table 10: The prompt about paraphrasing the discourse context (i.e., prompt operator $PARA$).

Demonstration:
We have a template sentence: ‘A publicist says Tara Reid has checked herself into rehab.’ Please generate three Prefixes and Suffixes for the template sentence. Prefix content should be about what typically happens before the event head lemma ‘checked herself’ in the given template sentence, while Suffix content should be about what typically happens after the event head lemma ‘checked herself’ in the given template sentence. Note: Each generated Prefix or Suffix contains three sentences.
Prefixes:
1. ‘After a series of public appearances where she appeared to be under the influence, rumors began to circulate that Tara Reid was struggling with addiction. Friends and family members reportedly urged her to seek help and get treatment before things got worse.’
2. ‘Tara Reid often faces mounting pressure from friends, family, and inner demons. The weight of her addiction or mental health challenges becomes increasingly burdensome. Seeking relief and stability, she reaches a breaking point where seeking professional help is no longer just an option, but a necessity.’
...
Suffixes:
1. ‘Reid’s representatives have confirmed that she is taking her recovery seriously and is committed to staying in rehab for as long as necessary. She has also expressed gratitude for the support she has received from fans and loved ones during this difficult time. It is hoped that with the help of professionals, she will be able to overcome her addiction and move forward in a positive direction.’
2. ‘Tara Reid begins a transformative journey towards healing after checking herself into rehab. She commits herself to a comprehensive treatment plan tailored to her specific needs. With dedication and the support of professionals, she embarks on a path of self-discovery, growth, and sobriety.’
...
Input:
We have a template sentence: ‘After stealing a valuable painting from the museum, the thief managed to escape through the back door.’ Please generate three Prefixes and Suffixes for the template sentence. Prefix content should be about what typically happens before the event head lemma ‘escape’ in the given template sentence, while Suffix content should be about what typically happens after the event head lemma ‘escape’ in the given template sentence. Note: Each generated Prefix or Suffix contains three sentences.
LLM Response:
Prefixes:
1. ‘The thief carefully scoped out the museum for weeks, studying the security protocols and identifying the perfect time to strike. They communicated with a network of fellow criminals, planning the heist down to the smallest detail. Once the valuable painting caught their eye, they knew it was time to put their plan into action.’
2. ‘Feeling the thrill of the adrenaline pumping through their veins, the thief meticulously studied the layout of the museum, looking for weaknesses in the security system. They spent days plotting their escape route, ensuring that every step was calculated and foolproof. The moment they laid eyes on the valuable painting, they knew this was their chance to execute the perfect crime.’
...
Suffixes:
1. ‘The thief disappeared into the night, leaving authorities baffled and determined to bring them to justice. They carefully weaved through a labyrinth of narrow alleyways, evading capture at every turn. With the stolen painting securely hidden, they planned their next move, aware that the chase was far from over.’
2. ‘After successfully evading the museum security, the thief meticulously covered their tracks, leaving no evidence behind. They vanished into the underworld, becoming a legend among thieves. With the valuable painting in their possession, they now had the difficult task of finding a buyer who would be discreet and willing to pay a handsome sum.’
...

Table 11: The prompt about producing temporal commonsense prefixes and suffixes for a given event-mention sentence with its trigger. (i.e., prompt operator TC). This prompt is inspired by Ravi et al.’s work.

ORI

<s>Esther Williams-Esther Williams: "Million Dollar Mermaid" Swimmer And Actress Dies Aged 91. 07 June 2013. Golden girl of screen and pool, Esther Williams, has ***died*** peacefully in her sleep aged 91. It has been confirmed by publicist Harlan Boll that the 1940s Hollywood actress and record-setting swimmer, Esther Williams, died on 6th June in her sleep. Stunning Williams quickly achieved movie success and pin-up status, due to her swimmer's physique being regularly snapped in bathing suits.<\s>
<s>Esther Williams, Olympic swimmer turned actress and pinup girl, dies at 91. Esther Williams, the swimming champion turned actress who starred in glittering and aquatic Technicolor musicals of the 1940s and 1950s, has died. She was 91. Williams ***died*** early Thursday in her sleep, according to her longtime publicist Harlan Boll. Williams in a bathing suit became a favorite pinup of GI's in World War II , and her popularity continued afterward .<\s>

CAD

<s>Esther Williams-Esther Williams: "Million Dollar Mermaid" Swimmer And Actress Dies Aged 91. 07 June 2013. The renowned musician Prince ***departed*** from this world in his studio in Minneapolis at the age of 57. It has been confirmed by publicist Harlan Boll that the 1940s Hollywood actress and record-setting swimmer, Esther Williams, died on 6th June in her sleep. Stunning Williams quickly achieved movie success and pin-up status, due to her swimmer's physique being regularly snapped in bathing suits.<\s>
<s>Esther Williams, Olympic swimmer turned actress and pinup girl, dies at 91. Esther Williams, the swimming champion turned actress who starred in glittering and aquatic Technicolor musicals of the 1940s and 1950s, has died. She was 91. Williams ***died*** early Thursday in her sleep, according to her longtime publicist Harlan Boll. Williams in a bathing suit became a favorite pinup of GI's in World War II , and her popularity continued afterward .<\s>

TIA

<s>Esther Williams-Esther Williams: "Million Dollar Mermaid" Swimmer And Actress Dies Aged 91. 07 June 2013. The renowned musician Prince ***died*** from this world in his studio in Minneapolis at the age of 57. It has been confirmed by publicist Harlan Boll that the 1940s Hollywood actress and record-setting swimmer, Esther Williams, died on 6th June in her sleep. Stunning Williams quickly achieved movie success and pin-up status, due to her swimmer's physique being regularly snapped in bathing suits.<\s>
<s>Esther Williams, Olympic swimmer turned actress and pinup girl, dies at 91. Esther Williams, the swimming champion turned actress who starred in glittering and aquatic Technicolor musicals of the 1940s and 1950s, has died. She was 91. Williams ***died*** early Thursday in her sleep, according to her longtime publicist Harlan Boll. Williams in a bathing suit became a favorite pinup of GI's in World War II , and her popularity continued afterward .<\s>

CIA

<s>On June 6, 2013, Esther Williams, the iconic Hollywood actress and swimmer, passed away peacefully in her sleep at the age of 91. The renowned musician Prince ***departed*** from this world in his studio in Minneapolis at the age of 57. Esther Williams, the Hollywood actress and record-breaking swimmer, passed away on June 6, 2013, at the age of 91. Williams' beauty and athleticism made her a beloved figure in both the film industry and the world of swimming.<\s>
<s>Esther Williams, the Olympic swimmer who later became an actress and pinup girl, passed away at the age of 91. She was known for her dazzling performances in Technicolor musicals during the 1940s and 1950s. Williams ***died*** early Thursday in her sleep, according to her longtime publicist Harlan Boll. Esther Williams' beauty and talent made her a favorite pinup of GI's during World War II, and her legacy continued to captivate audiences for years to come.<\s>

TAD

<s>In the weeks leading up to his tragic departure, rumors began to circulate about Prince's declining health. Concerned fans and loved ones expressed their worries, hoping he would seek proper medical attention and take care of himself. The renowned musician Prince ***departed*** from this world in his studio in Minneapolis at the age of 57. Tributes poured in, celebrating his iconic career and the impact he had on the world of music. His immense talent and legacy continue to inspire new generations of musicians and fans alike.<\s>
<s>Esther Williams, like many people in the public eye, faced her fair share of personal struggles and demons. She had battled with addictions and mental health issues throughout her life, which had impacted her overall health. Despite attempts to seek treatment and find stability, her health continued to deteriorate, ultimately leading to her untimely death. Williams ***died*** early Thursday in her sleep, according to her longtime publicist Harlan Boll. Following her passing, Williams' publicist Harlan Boll released a statement expressing his condolences and sharing the sadness of her death. Boll highlighted the impact Williams had on the entertainment industry and the void her absence would leave. He also mentioned plans for a memorial service to honor her memory and celebrate her life<\s>

Table 12: A coreferential original data (ORI) with its Counterfactual Augmented Data (CAD), Trigger Intervention Ablation (TIA) Data, Context Intervention Ablation (CIA) Data and Temporal commonsense Augmented Data (TAD). Mention sentences are underlined, with bold trigger terms.

ORI

<s>Microsoft Rushes Emergency Fix To Address Internet Explorer Attacks. September 17, 2013 4:16 PM ET. Microsoft has rushed out a temporary fix to ***address*** ongoing attacks targeting an Internet Explorer zero-day vulnerability. The software giant said the Fix-It temporary workaround should be effective in preventing a successful attack. The company said the vulnerability impacts all currently supported versions of the browser, but attacks have been limited to users of Internet Explorer 8 and Internet Explorer 9. "On completion of this investigation, Microsoft will take the appropriate action to protect our customers, which may include providing a solution through our monthly security update release process , or an out-of-cycle security update, depending on customer needs, " the company said in a security advisory issued Tuesday.<\s>

<s>Microsoft has said that an emergency security update has fixed a flaw in Internet Explorer that left millions of computers vulnerable to hacking and hijack. The software patch, which was released last night, has closed a loophole that has seen the computers of at least two million users hacked by cyber criminals. The problem related to a ``zero day" flaw that tricked people into visiting an infected website, enabling hackers to gain access to online banking passwords and e-shopping logon details. ``Microsoft has released a security update for Internet Explorer that will help ***protect*** its customers from malicious attacks," said the company in a statement. ``Like a vaccine developed to fight a virus , this 'security update' will protect computers only if it is installed. "Computers that are set to automatically update and install software will already be protected by the patch. Those users who manage their own updates are advised to download the fix as soon as possible from the Microsoft website.<\s>

CAD

<s>An emergency security update from Microsoft has fixed a flaw in Internet Explorer that left millions of computers vulnerable to hacking and hijack. The software patch, which was released last night, has closed a loophole that allowed cyber criminals to hack into the computers of at least two million users by exploiting a "zero day" flaw. A statement from Microsoft confirms that the company has ***secured*** its customers from malicious attacks by releasing a security update for Internet Explorer. ' The security update for Internet Explorer released by Microsoft is like a vaccine that protects computers from malicious attacks. However, the update will only be effective if it is installed. Computers set to automatically update and install software will already be protected by the patch, but those who manage their own updates should download the fix from the Microsoft website without delay.<\s>

<s>Microsoft has said that an emergency security update has fixed a flaw in Internet Explorer that left millions of computers vulnerable to hacking and hijack. The software patch, which was released last night, has closed a loophole that has seen the computers of at least two million users hacked by cyber criminals. The problem related to a ``zero day" flaw that tricked people into visiting an infected website, enabling hackers to gain access to online banking passwords and e-shopping logon details. ``Microsoft has released a security update for Internet Explorer that will help ***protect*** its customers from malicious attacks," said the company in a statement. ``Like a vaccine developed to fight a virus , this 'security update' will protect computers only if it is installed. "Computers that are set to automatically update and install software will already be protected by the patch. Those users who manage their own updates are advised to download the fix as soon as possible from the Microsoft website.<\s>

TIA

<s>An emergency security update from Microsoft has fixed a flaw in Internet Explorer that left millions of computers vulnerable to hacking and hijack. The software patch, which was released last night, has closed a loophole that allowed cyber criminals to hack into the computers of at least two million users by exploiting a "zero day" flaw. A statement from Microsoft confirms that the company has ***protected*** its customers from malicious attacks by releasing a security update for Internet Explorer. ' The security update for Internet Explorer released by Microsoft is like a vaccine that protects computers from malicious attacks. However, the update will only be effective if it is installed. Computers set to automatically update and install software will already be protected by the patch, but those who manage their own updates should download the fix from the Microsoft website without delay.<\s>

<s>Microsoft has said that an emergency security update has fixed a flaw in Internet Explorer that left millions of computers vulnerable to hacking and hijack. The software patch, which was released last night, has closed a loophole that has seen the computers of at least two million users hacked by cyber criminals. The problem related to a ``zero day" flaw that tricked people into visiting an infected website, enabling hackers to gain access to online banking passwords and e-shopping logon details. ``Microsoft has released a security update for Internet Explorer that will help ***protect*** its customers from malicious attacks," said the company in a statement. ``Like a vaccine developed to fight a virus , this 'security update' will protect computers only if it is installed. "Computers that are set to automatically update and install software will already be protected by the patch. Those users who manage their own updates are advised to download the fix as soon as possible from the Microsoft website.<\s>

CIA

<s>In response to the Internet Explorer attacks, Microsoft has rapidly provided an emergency fix on September 17, 2013, at 4:16 PM ET. A statement from Microsoft confirms that the company has ***secured*** its customers from malicious attacks by releasing a security update for Internet Explorer. ' According to Microsoft, the temporary Fix-It workaround is capable of preventing a successful attack on all currently supported versions of the browser. However, the attacks have only affected Internet Explorer 8 and 9 users.<\s>

<s>An emergency security update from Microsoft has fixed a flaw in Internet Explorer that left millions of computers vulnerable to hacking and hijack. The software patch, which was released last night, has closed a loophole that allowed cyber criminals to hack into the computers of at least two million users by exploiting a "zero day" flaw ``Microsoft has released a security update for Internet Explorer that will help ***protect*** its customers from malicious attacks," said the company in a statement. The security update for Internet Explorer released by Microsoft is like a vaccine that protects computers from malicious attacks. However, the update will only be effective if it is installed. Computers set to automatically update and install software will already be protected by the patch, but those who manage their own updates should download the fix from the Microsoft website without delay.<\s>

TAD

<s>Leading up to the release of the new security update, Microsoft has been monitoring user reports and conducting extensive testing to identify and understand the nature of the glitch in Internet Explorer. They have been working diligently behind the scenes to develop an effective solution. A statement from Microsoft confirms that the company has ***secured*** its customers from malicious attacks by releasing a security update for Internet Explorer. ' Now that the security update has been released, Microsoft urges all users of Internet Explorer, including those on Windows 7 with IE11, to promptly install the patch to fix the glitch. They emphasize the importance of keeping browsers up to date to ensure optimal security and protection against potential cyber threats. <\s>

<s>As part of their proactive approach to cybersecurity, Microsoft regularly conducts audits and vulnerability assessments of Internet Explorer. During one such routine assessment, the company discovered a glitch that could potentially exploit users' systems. This prompted the swift development of a security update to mend the vulnerability. ``Microsoft has released a security update forInternet Explorer that will help ***protect*** its customers from malicious attacks," said the company in a statement. With the successful release of the security update, Microsoft assures users that their browsing experience with Internet Explorer will now be more secure and free from the glitch. They encourage users to reach out to their support team if they encounter any further issues or require additional assistance.<\s>

Table 13: A non-coreferential original data (ORI) with its Counterfactual Augmented Data (CAD), Trigger Intervention Ablation (TIA) Data, Context Intervention Ablation (CIA) Data and Temporal commonsense Augmented Data(TAD).

Algorithm 2 Generating Trigger Intervention Ablation (TIA) Data for the original mention pair

Input:
Original data MP=($S_{i-w}^{(1)}...S_{i}^{(1)}...S_{i+w}^{(1)}, S_{j-w}^{(2)}...S_{j}^{(2)}...S_{j+w}^{(2)}$) with label Y; Large language model LLM; Trigger terms of two mentions ($T^{(1)}$,$T^{(2)}$).
Prompt operators: Synonyms generator SYN; Coref events generator CE; Non-coref events generator NCE; Paraphraser $PARA$.
Output: Generated dataset: D_{TIA}

while sentence s in MP **do**
    if $Y == coref$ **then**
        if $s == S_i^{(1)}$ **then**
            $S_{gens} = LLM\left(NCE, T^{(1)}, S_i^{(1)}\right)$
        else
            continue
        end if
    else if $Y == not\ coref$ **then**
        if $s == S_i^{(2)}$ **then**
            $S_{gens} = LLM\left(CE, T^{(2)}, S_j^{(2)}\right)$
        else
            continue
        end if
    end if
end while
while sentence s_g in S_{gens} **do**
    if $Y == coref$ **then**
        $\tilde{m}_1 = (S_{i-w}^{(1)}, ..., s_g, ..., S_{i+w}^{(1)})$
        $MP_{cf} = concat\{\tilde{m}_1,\ (S_{j-w}^{(2)}, ..., s_g, ..., S_{j+w}^{(2)})\}$
    else if $Y == not\ coref$ **then**
        $pre = LLM\left(PARA,\ (S_{j-w}^{(2)}, ..., S_{j-1}^{(2)})\right)$
        $suf = LLM\left(PARA,\ (S_{j+1}^{(2)}, ..., S_{j+w}^{(2)})\right)$
        $\tilde{m}_1 = concat\{pre, s_g, suf\}$
        $MP_{TIA} = concat\left\{\tilde{m}_1, \left(S_{j-w}^{(2)}, ..., S_{j+w}^{(2)}\right)\right\}$
    end if
end while
Add MP_{TIA} to the set D_{TIA}
return D_{TIA}

Algorithm 3 Generating Context Intervention Ablation (CIA) Data for the original mention pair

Input:
Original data MP=($S^{(1)}_{i-w}$...$S^{(1)}_{i}$...$S^{(1)}_{i+w}$, $S^{(2)}_{j-w}$...$S^{(2)}_{j}$...$S^{(2)}_{j+w}$) with label Y; Large language model LLM; Trigger terms of two mentions ($T^{(1)}$,$T^{(2)}$).
Prompt operators: Synonyms generator SYN; Coref events generator CE; Non-coref events generator NCE; Paraphraser $PARA$.
Output: Generated dataset: D_{CIA}

while sentence s in MP **do**
    if $Y == coref$ **then**
        if $s == S^{(1)}_{i}$ **then**
            $T^{(1)}_{syns} = LLM\left(SYN, T^{(1)}\right)$
            $S_{gens} = LLM\left(NCE, T^{(1)}_{syns}, S^{(1)}_{i}\right)$
        else
            continue
        end if
    else if $Y == not\ coref$ **then**
        if $s == S^{(2)}_{i}$ **then**
            $T^{(2)}_{syns} = LLM\left(SYN, T^{(2)}\right)$
            $S_{gens} = LLM\left(CE, T^{(2)}_{syns}, S^{(2)}_{j}\right)$
        else
            continue
        end if
    end if
end while
while sentence s_g in S_{gens} **do**
    if $Y == coref$ **then**
        $pre^{(1)} = LLM\left(PARA,\ (S^{(1)}_{i-w}, ..., S^{(1)}_{i-1})\right)$
        $suf^{(1)} = LLM\left(PARA,\ (S^{(1)}_{i+1}, ..., S^{(1)}_{i+w})\right)$
    else if $Y == not\ coref$ **then**
        $pre^{(1)} = LLM\left(PARA,\ (S^{(2)}_{j-w}, ..., S^{(2)}_{j-1})\right)$
        $suf^{(1)} = LLM\left(PARA,\ (S^{(2)}_{j+1}, ..., S^{(2)}_{j+w})\right)$
    end if
    $pre^{(2)} = LLM\left(PARA,\ (S^{(2)}_{j-w}, ..., S^{(2)}_{j-1})\right)$
    $suf^{(2)} = LLM\left(PARA,\ (S^{(2)}_{j+1}, ..., S^{(2)}_{j+w})\right)$
    $\tilde{m}_1 = concat\left\{pre^{(1)}, s_g, suf^{(1)}\right\}$
    $\tilde{m}_2 = concat\left\{pre^{(2)}, S^{(2)}_{j}, suf^{(2)}\right\}$
    $MP_{CIA} = concat\{\tilde{m}_1, \tilde{m}_2\}$
end while
Add MP_{CIA} to the set D_{CIA}
return D_{CIA}

Algorithm 4 Generating Temporal Commonsense Augmented Data (TAD) for the original mention pair

Input:
Original data MP=($S^{(1)}_{i-w}$...$S^{(1)}_{i}$...$S^{(1)}_{i+w}$, $S^{(2)}_{j-w}$...$S^{(2)}_{j}$...$S^{(2)}_{j+w}$) with label Y; Large language model LLM; Trigger terms of two mentions ($T^{(1)}$,$T^{(2)}$).
Prompt operators: Synonyms generator SYN; Coref events generator CE; Non-coref events generator NCE; Temporal commonsense generator TC.
Output: Generated dataset: D_{TAD}

while sentence s in MP **do**
  if $Y == coref$ **then**
    if $s == S^{(1)}_i$ **then**
      $T^{(1)}_{syns} = LLM\left(SYN, T^{(1)}\right)$
      $S_{gens} = LLM\left(NCE, T^{(1)}_{syns}, S^{(1)}_i\right)$
    else
      continue
    end if
  else if $Y == not\ coref$ **then**
    if $s == S^{(2)}_i$ **then**
      $T^{(2)}_{syns} = LLM\left(SYN, T^{(2)}\right)$
      $S_{gens} = LLM\left(CE, T^{(2)}_{syns}, S^{(2)}_j\right)$
    else
      continue
    end if
  end if
end while
while sentence s_g in S_{gens} **do**
  if $Y == coref$ **then**
    $pre^{(1)}, suf^{(1)} = LLM\left(TC,\ T^{(1)}_{syns}, s_g\right)$
  else if $Y == not\ coref$ **then**
    $pre^{(1)}, suf^{(1)} = LLM\left(TC,\ T^{(2)}_{syns}, s_g\right)$
  end if
  $pre^{(2)}, suf^{(2)} = LLM\left(TC,\ T^{(2)}, S^{(2)}_j\right))$
  $\tilde{m}_1 = concat\left\{pre^{(1)}, s_g, suf^{(1)}\right\}$
  $\tilde{m}_2 = concat\left\{pre^{(2)}, S^{(2)}_j, suf^{(2)}\right\}$
  $MP_{TAD} = concat\{\tilde{m}_1, \tilde{m}_2\}$
end while
Add MP_{TAD} to the set D_{TAD}
return D_{TAD}