Dr.ECI: Infusing Large Language Models with Causal Knowledge for Decomposed Reasoning in Event Causality Identification

Ruichu Cai[1], **Shengyin Yu**[1], **Jiahao Zhang**[1], **Wei Chen**[1], **Boyan Xu**[1*], **Keli Zhang**[2]
[1]School of Computer Science, Guangdong University of Technology
[2]Huawei Noah's Ark Lab
`{cairuichu, yushengyin2022, luckyanjooo, chenweidelight, hpakyim}@gmail.com`
`zhangkeli1@huawei.com`

Abstract

Despite the demonstrated potential of large language models (LLMs) in diverse NLP tasks, their causal reasoning capability appears inadequate when evaluated within the context of the Event Causality Identification (ECI) task. The ECI tasks pose significant complexity for LLMs and necessitate comprehensive causal priors for accurate identification. To improve the performance of LLMs for causal reasoning, we propose a multi-agent Decomposed reasoning framework for Event Causality Identification, designated as *Dr.ECI*. In the discovery stage, *Dr.ECI* incorporates specialized agents such as *Causal Explorer* and *Mediator Detector*, which capture implicit causality and indirect causality more effectively. In the reasoning stage, *Dr.ECI* introduces the agents *Direct Reasoner* and *Indirect Reasoner*, which leverage the knowledge of the generalized causal structure specific to the ECI. Extensive evaluations demonstrate the state-of-the-art performance of *Dr.ECI* compared with baselines based on LLMs and supervised training. Our implementation will be open-sourced at `https://github.com/DMIRLAB-Group/Dr.ECI`.

1 Introduction

Event Causality Identification (ECI) plays a crucial role in Natural Language Processing by identifying causal relationships among events within a context. Its importance extends to the improvement of machine reading capabilities (Berant et al., 2014), the prediction of future events (Hashimoto, 2019), and the improvement of question-answer systems (Oh et al., 2016). The availability of annotated data for ECI remains limited. Data augmentation techniques that use pre-trained language models have shown some progress in ECI (Zuo et al., 2020; Man et al., 2022; Shen et al., 2022a). However, existing solutions lack generalizability to unseen domains,

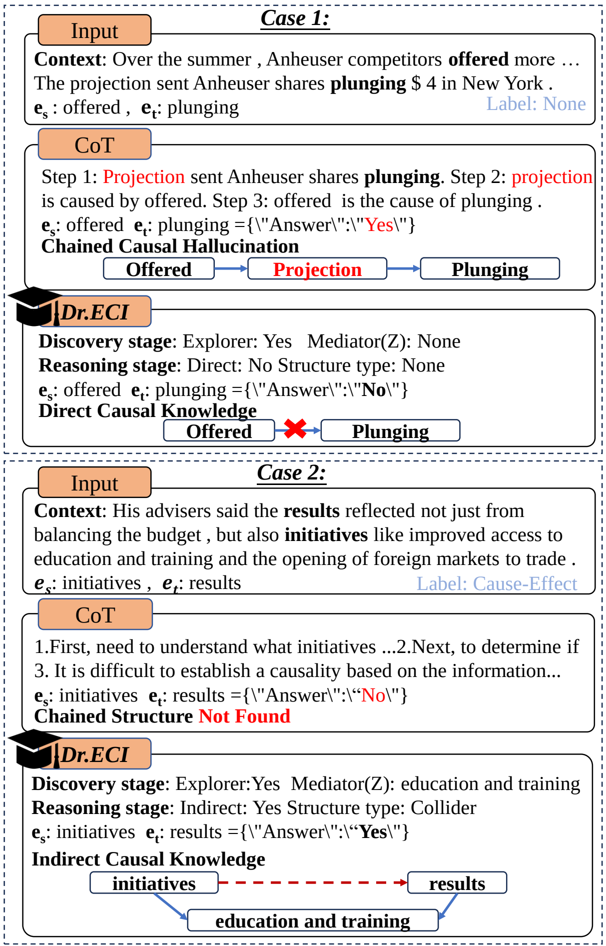


Figure 1: CoT vs *Dr.ECI*

underscoring the urgent need for generalization capabilities in the field of ECI.

The field of Natural Language Processing (NLP) has undergone substantial advancements due to the emergence of large language models (LLMs) such as ChatGPT (OpenAI, 2022) and PaLM (Chowdhery et al., 2023), renowned for their advanced language comprehension capabilities. The use of innovative prompting techniques, including chain-of-thought (CoT) (Wei et al., 2022) and in-context learning (ICL) (Brown et al., 2020), further en-

*Corresponding author, hpakyim@gmail.com

Proceedings of the 31st International Conference on Computational Linguistics, pages 9346–9375
January 19–24, 2025.

hances the reasoning abilities of LLMs. Recent studies have highlighted the notable deficiencies in ChatGPT's proficiency in ECI, as evidenced by its inaccuracies (Gao et al., 2023). These limitations can be attributed to the inherent reporting biases between causal and non-causal relationships in natural language, and they may be further influenced by CoT, resulting in what is referred to as "Causal Hallucinations." In this paper, we extensively investigate the causal reasoning limitations of LLMs in ECI, identifying a key factor: the insufficient causal knowledge possessed by LLMs. First, it is observed that LLMs excel in identifying explicit causality while tending to overlook implicit and indirect causality. Second, although CoT enhances the reasoning abilities of LLMs on chained structures, it falls short of supporting more complex causal structures. This limitation becomes evident in scenarios where non-chain structure causality between event "initiatives" and event "results" is present, as illustrated in Figure 1. Third, the utilization of CoT can lead to the emergence of fake conditions, exacerbating the occurrence of chained causal hallucinations. For instance, in the first case depicted in Figure 1, the chain causality between the entities event "offered" and event "plunging" is artificially constructed through the fake condition of event "projection."

To address the above issues, we propose a multi-agent decomposed reasoning framework for the ECI, designated as *Dr.ECI*, infusing the causal knowledge necessary to improve the causal reasoning ability of LLMs. *Dr.ECI* is composed of two stages: the discovery stage and the reasoning stage. During the discovery stage, *Causal Explorer* addresses the limited reasoning capabilities of LLMs by strategically using various natural language prompts, enabling effective exploration of both implicit and indirect causality within the given context. Subsequently, the *Mediator Detector* agent plays a critical role in detecting indirect causal relationships. It identifies potential mediators among relevant entities or events existing between pairs of events, facilitating the causal structure matching process for indirect causality reasoning. During the reasoning stage, we inject causal knowledge within the *Direct Reasoner* and *Indirect Reasoner* agents using Directed Acyclic Graphs (DAGs) to represent causal patterns within the ECI task. Leveraging the power of structured causal patterns, *Dr.ECI* offers precise identification of both direct and indirect causal relationships. Employing the aforementioned two stages, *Dr.ECI* is composed of four agents, collectively working to break down intricate causal reasoning problems into subproblems that can be effectively addressed utilizing causal structure knowledge. As shown in Figure 1, *Dr.ECI* accurately detects direct and indirect causality. Specifically, it identifies the failure to match direct causality in Case 1 and discerns the presence of indirect causality through the "Collider" structure in Case 2.

Our proposed *Dr.ECI* makes the following contributions:

- We summarize the necessary knowledge of causal structures in ECI through Directed Acyclic Graphs (DAGs).
- To enhance the causal reasoning capabilities of LLMs, we present the *Dr.ECI*, comprised of four agents. *Dr.ECI* effectively decomposes the intricate ECI task and utilizes the causal structure knowledge for addressing implicit causality and indirect causality challenges.
- We conduct extensive experiments to study the effectiveness of the proposed *Dr.ECI* framework. *Dr.ECI* achieves new state-of-the-art performance, as confirmed by quantitative and human evaluation.

2 Causal Priors for ECI

Event Causality Identification (ECI) aims to identify the causal relationships between a source event e_s and a target event e_t, within a defined context C. The input is denoted as $\mathcal{I} = (C, (e_s, e_t))$. And the output is a binary label $\mathcal{O}$, indicating whether the causal relationship exists between e_s and e_t. Using $\mathcal{I}$ as prompts for input to LLMs, we observe that LLMs typically employ either direct causal structures or chain causal structures (using CoT) in the reasoning process of $\mathcal{O}$. Chain causal structures are constructed by identifying additional events within the context C. However, the presence of other different causal structures, such as colliders, in ECI tasks poses challenges to LLMs, making it difficult for them to reason effectively.

In this paper, we employ causal graph theory to systematically summarize explicit and implicit causal structures under the ECI task via Directed Acyclic Graphs (DAGs). DAG is a graph $\mathcal{G}$ defined as $(\mathbf{V}, \mathbf{E})$ to represent causal relationships among variables. Each node $v_i \in \mathbf{V}$ corresponds to a Random Variable (RV) X_i that may be discrete or

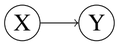
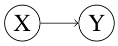
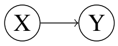
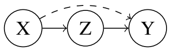
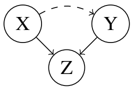
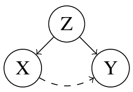
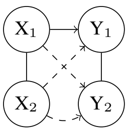
Causal pattern	Causal graph	Description	Natural language example
Explicit Words		There are obvious clue words indicating a causal relationship between X and Y: cause, because,...	The slowdown **raises[Y]** questions about the economy 's strength **because** spending **fueled[X]**much...
Implicit Words		There are clue words between X and Y that may indicate a causal relationship: spark, authorize,...	...the Security Council embargo **resolution[X]** did not specifically **authorize** such a naval **operation[Y]** .
Causal Order		Based on common sense knowledge, sequences of cause and effect can be identified.	Man **charged[Y]** with **murdering[X]** mother and sister in Cumbria.
Chain		In a causal chain, an intermediary event Z can link different events, where X causes Z in one pair, and Z subsequently causes Y in another.	Meanwhile, ISPs ... have been **scrambling[X]** to implement contingency **plans[Z]** to keep their customers **connected[Y]**.
Collider		There is an entity or event (Z) that is linked to both events X and Y, where X and Y each act upon Z, indicating a causal relationship between X and Y.	His advisers said the **results[Y]** reflected not just from balancing the budget, but also **initiatives[X]** like improved access to **education and training[Z]** ...
Fork		There is an entity or event Z that is linked to both events X and Y, where Z influences both X and Y, indicating a causal relationship between X and Y.	Vodafone Egypt said ... face slight **problems[X]** due to **damage[Z_1]** in. Vodafone, along with the country's three telecom providers has been **affected[Y]** by a **damage[Z_2]** in...
Coreference		Coreference events are defined here as events that have the same meaning but are expressed differently or have the same expression.	The violence... after **bombings[X_2]** were reported..., an **attack[X_1]** that provoked strong **condemnation[Y_1]** from ...

Table 1: An overview of the causal patterns in ECI. The causal graph distinguishes between direct and indirect causal relationships.

continuous. Each directed edge $e_{i,j} \in \mathbf{E}$ indicates a directional causality $X_i \rightarrow X_j$. In the context of ECI, such directed edges are essential for representing the direct causal relationships between e_s and e_t. Causal patterns within ECI are categorized into direct and indirect causal reasoning as detailed in Table 1:

- **Direct Causal Reasoning** includes patterns where causality is explicit or can be inferred, featuring types like Explicit words, Implicit words, and Causal Order.
- **Indirect Causal Reasoning** covers patterns with mediators or intermediate events, such as Chain, Collider, Fork, and Coreference patterns.

The summarized causal patterns aid in the understanding and identification of causal relationships in various contexts, enhancing LLMs' ability to process explicit and implicit causal information.

3 Multi-Agent Decomposed Reasoning

3.1 Overview

To integrate causal priors into the closed-source LLMs for the ECI task, our *Dr.ECI* is a multi-agent decomposed reasoning framework including the discovery stage and the reasoning stage, as shown in Figure 2. All agents operate independently based on LLMs. In the discovery stage, context C and the event pair (e_s, e_t) are provided as inputs. The *Causal Explorer* then responds based on the causal expressions $\mathcal{N}$. If a causal relationship is identified, the process proceeds to the *Mediator Detector*, which is designed to identify potential mediators

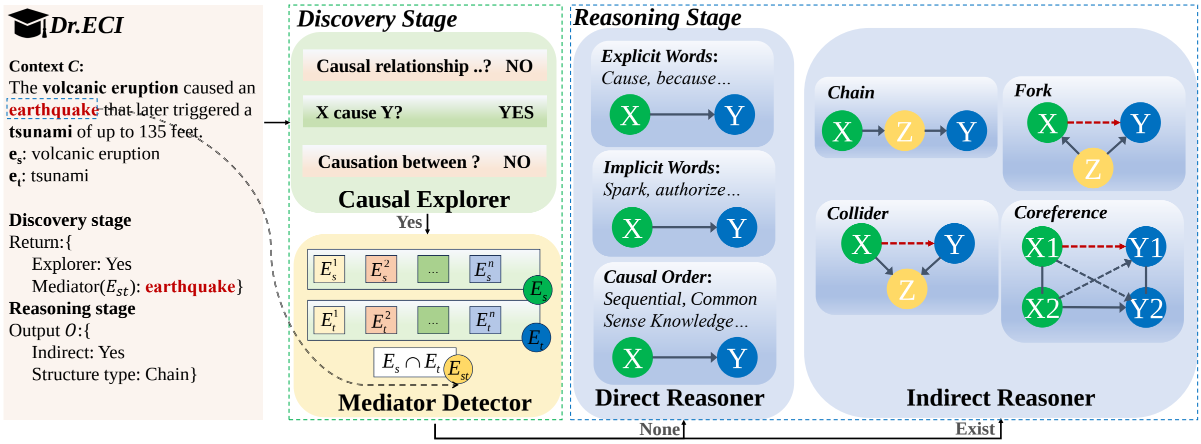


Figure 2: Dr.ECI comprises 4 agents: *Causal Explorer*, *Mediator Detector*, *Direct Reasoner* and *Indirect Reasoner*.

(E_{st}) that could act as causal intermediaries between the event pair (e_s, e_t). In the reasoning stage, if the mediator $E_{st} \neq \emptyset$, indicating the presence of common entities or events, the *Indirect Reasoner* utilizes indirect causal patterns from the causal priors, incorporating the potential mediator E_{st} for identification. Conversely, the *Direct Reasoner* employs direct causal patterns from the causal priors for identification. The framework's output, thus, reflects the intricate process of causal analysis, emphasizing the role of structured causal patterns in determining the nature of the causal relationship.

3.2 Causal Explorer

Causal Explorer relies on LLMs' understanding of the concept of "causality" to perform reasoning for the initial filtering of event pairs in ECI. The primary objective of the *Causal Explorer* is to filter out pairs that evidently lack a causal relationship while meticulously preserving samples that may embody implicit or indirect causality.

Acknowledging the varying sensitivity of LLMs to different natural language expressions of causality, we construct a varied set of causal expressions, $\mathcal{N} = \{n_1, n_2, \ldots, n_m\}$, where each n_i represents a distinct natural language formulation of causality. For each expression in $\mathcal{N}$, the *Causal Explorer* solicits a response, compiled into a set $\mathcal{R} = \{r_1, r_2, \ldots, r_n\}$, where each r_i indicates the presence ('Yes') or absence ('No') of a causal relationship between e_s and e_t. So the sets $\mathcal{N}$ and $\mathcal{R}$ have the same number of elements. The structure of the prompt is shown in Prompt 3.1, and full details can be found in Appendix Prompt 11. To synthesize these insights, a voting mechanism is employed:

$$q = \sum_{i=1}^{n} \begin{cases} 1, & \text{if } r_i \text{ indicates 'Yes'}, \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

Prompt 3.1: Causal Explorer

Instruction: Guide LLMs to identify whether a causal relationship exists based on the set of *Causal Expressions* $\mathcal{N} = \{n_1, n_2, \ldots, n_m\}$.

Causal Expression n_1: Is there a *causality* between *source event* e_s and *target event* e_t?...

Return: Response Set $\mathcal{R} = \{r_1, r_2, \ldots, r_n\}$, where each r_i represents a binary classification: Yes/No.

This aggregation reflects a collective judgment on the potential existence of causality. If q, the cumulative sum of affirmative responses, equals or exceeds a predetermined threshold—set to 1 based on our experimental findings—it suggests a probable causal link, warranting further analysis in subsequent reasoning stages. This threshold was chosen to ensure that even a single affirmative response (r_i = 'Yes') is sufficient to consider an event pair for deeper causal investigation.

This strategy is particularly effective in the context of our framework, where the introduction of causal priors in subsequent reasoning stages (e.g., by *Direct Reasoner* and *Indirect Reasoner*) further refines the assessment of causality.

3.3 Mediator Detector

The *Mediator Detector* is designed to unearth potential mediators E_{st} that may serve as links in the causal chain between event pairs (e_s, e_t). This detection process is crucial for understanding the indirect causality within a given context C, following these steps:

i **Extraction**: The prompt "Identify all entities or events (E_s for e_s and E_t for e_t)" is used as contextually associated with each event. This step involves parsing the textual content to gather entities or events that have a direct or indirect relation to e_s and e_t.

ii **Intersection**: To obtain potential mediators, use the prompt "Determine the common entities or events E_{st} between E_s and E_t." This process filters out the unique entities, focusing on those that share a relationship with both events.

The structure of the prompt is shown in Prompt 3.2.

> **Prompt 3.2: Mediator Detector**
>
> **Instruction**: Guide LLMs to detect *Mediators* E_{st} that may be links in the construction of causal patterns by *source event* e_s and *target event* e_t.
>
> **Extraction**: Identify all entities or events E_s for e_s and E_t for e_t. Answer me at word level.
>
> **Intersection**: Determine the common entities or events E_{st} between E_s and E_t. Answer me at word level.
>
> **Return**: *Mediators* E_{st}.

E_{st}, as the intermediary set between e_s and e_t, reveals the complex interactions influencing their causal relationship. Identifying E_{st} is pivotal for the subsequent analysis of indirect causality, as it highlights the existence of intermediary factors that could be influencing the observed causal dynamics.

3.4 Reasoners

Building upon the causal patterns outlined in the Causal Priors, the *Reasoners* are designed to discern causality between events e_s and e_t within a context C. These agents, the *Direct Reasoner* for direct causality and the *Indirect Reasoner* for scenarios involving mediators, leverage the structured causal patterns to systematically analyze and infer causal relationships.

Direct Reasoner: It employs direct causal patterns to establish causality between e_s and e_t. These patterns, including Explicit Words, Implicit Words, and Causal Order, enable *Direct Reasoner* to infer direct causality without E_{st}. The prompt is presented in Prompt 3.3.1, and additional details can be found in Appendix Prompt G.1.

Indirect Reasoner: Conversely, it is activated in the presence of potential mediators (E_{st}), applying indirect causal patterns identified from the causal graph. These patterns, such as Chain, Collider, Fork, and Coreference, guide *Indirect Reasoner* in uncovering complex causal structures. The prompt is presented in Prompt 3.3.2, and additional details can be found in Appendix Prompt G.1.

> **Prompt 3.3.1: Direct Reasoner**
>
> **Instruction**: Guide LLMs to choose a *Direct Causal Pattern* to identify whether *source event* e_s and *target event* e_t are causally related or not.
>
> **Pattern 1 (Explicit Words)**: There are verbs or verb phrases that clearly express causality. e.g., cause, lead,....
>
> **Pattern 2 (Implicit Words)**: There are implicit clue words in the text that imply a causal relationship. e.g., spark, arouse....
>
> **Pattern 3 (Causal Order)**: According to common sense knowledge, if *source event* e_s occurs before *target event* e_t, then there is a causal relationship between them.
>
> **Return**: Yes or No. If Yes, provide a list of *Direct Causal Pattern* candidates.

> **Prompt 3.3.2: Indirect Reasoner**
>
> **Instruction**: Guide LLMs to simultaneously choose an *Indirect Causal Pattern* and *Mediators* E_{st} to identify whether *source event* e_s and *target event* e_t are causally related or not.
>
> **Pattern 1 (Fork)**: Event e_{st} is related to *source event* e_s and *target event* e_t, and then it satisfies that e_{st} acts on *source event* e_s and e_{st} acts on *target event* e_t.
>
> **Pattern 2 (Collider)**: Event e_{st} is related to *source event* e_s and *target event* e_t, and then it satisfies that *source event* e_s acts on e_{st} and *target event* e_t acts on e_{st}.
>
> **Pattern 3 (Chain)**: Event e_{st} is related to *source event* e_s and *target event* e_t, then it satisfies that *source event* e_s acts on e_{st} and e_{st} acts on *target event* e_t.
>
> **Pattern 4 (Coreference of *source event* e_s)**: Event e_{st} is co-referred to the *source event* e_s, and e_{st} has a causal relationship with *target event* e_t.
>
> **Pattern 5 (Coreference of *target event* e_t)**: Event e_{st} is co-referred to the *target event* e_t, and e_{st} has a causal relationship with *source event* e_s.
>
> **Return**: Yes or No. If Yes, provide a list of *Indirect Causal Pattern* candidates and *Mediators* E_{st}.

Through the injection of causal patterns, the reasoner makes reasoning choices based on the patterns, which effectively improves the identification of complex causal relationships among events. Further, the selection of causal patterns by LLMs is analyzed experimentally to verify the effectiveness of Dr.ECI, see the Appendix Table 11 for details.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate the effectiveness of *Dr.ECI* on two widely used datasets in ECI. The Causal-TimeBank (Causal-TB) (Mirza, 2014) contains 184 documents and 6813 events, of which only 318 event pairs are causal. EventStoryLine (ESL) (Caselli and Vossen, 2017) consists of 258 documents covering 7275 annotated events and 5,519 plot links, representing generalized causal relationships. We follow the data disclosed in previous work (Zhang et al., 2023) for testing. In order to verify the effectiveness of our framework *Dr.ECI* in relationship extraction, we also conducted experiments on the classic dataset SemEval 2010 task 8 (Hendrickx et al., 2010), which contains 18 relationship types and 8k sentences for training and 2.72k sentences for testing. We evaluated all 2717 sentences in the test set, focusing specifically on the causal relationship types (cause-effect and effect-cause) among the 18 relationship types, which amounted to 328 sentences in total. MAVEN-ERE (Wang et al., 2022) contains 90 topics, 4,480 documents, 103,193 events, and 57,992 pairs of causal events. We follow previous work (Gao et al., 2023) to evaluate its performance on the development set.

Evaluation Metrics. To compare with previous works (Gao et al., 2023), we utilize Precision (P), Recall (R), and F1-score (F1) as the primary evaluation metrics.

Hyperparameters. We leveraged official APIs from OpenAI and Google to access the GPT-3.5-turbo[1] and PaLM2[2] models. Our primary focus was on PaLM2, with supplementary results from GPT-3.5-turbo detailed in Appendix D.1. We implemented a candidate-count parameter of 1 and a temperature setting of 0 for deterministic sampling. In our analysis, we used a pass threshold $q = 1$ for the causal explorer, aimed at minimizing errors in factual and inferential reasoning. For the few-shot setting, we created examples from ESL(Caselli and Vossen, 2017) and applied them across datasets.

[1]https://platform.openai.com/
[2]https://makersuite.google.com/

4.2 Baselines

PLM-based Baseline. We compare four PLM-based methods fine-tuned on the full training set: **GenECI** (Man et al., 2022) identifies event causality by jointly generating causal labels and dependency path words using the T5 model and REINFORCE algorithm. **DPJL** (Shen et al., 2022b) utilizes cue-based learning of pre-trained knowledge in language models to enhance event causality recognition.**KEPT** (Liu et al., 2023) uses knowledge-enhanced prompt tuning with external knowledge and attention mechanisms to improve Event Causality Identification. The latest SOTA method **CPATT** (Zhang et al., 2023) proposes a prompt-based approach with a Constrained Prefix ATTention mechanism for identifying causal and temporal relations between events.
LLM-based Baseline. We mainly compare LLM-based methods with different prompting: **Vanilla** and **CoT**. **Vanilla** aligns with the standard prompting method outlined in Wei et al. (2022) and Kojima et al. (2022), where no specific prompts are employed to elicit thoughts from the LLMs. Conversely, **CoT** guides the LLMs through a step-by-step thought process.We also compared the evaluation results of Gao et al. (2023) on ChatGPT. We added additional baselines, including **Self-ask** (Tang et al., 2023), **CaCo-CoT** (Press et al., 2022), and **L2M-CoT** (Zhou et al., 2022), as detailed in the Appendix D.3.

4.3 Main Results

Dr.ECI outperforms state-of-the-art PLM-based and LLM-based methods in the zero-shot setting, as Table 2 illustrates. Specifically, for the ESL's Intra-sentence dataset, it surpasses the top PLM baseline by 3.7% and the CoT baseline by 1.2% in F1-score. In the Causal-TB dataset, our method shows a 7% improvement over the PLM baseline and 5.5% over CoT in F1-score. The marginally lower performance of $Dr.ECI_{few}$ could be attributed to the predominance of ESL's examples in our selection. Under *Vanilla* and *CoT* baselines, we note low Precision but high Recall, indicating LLMs often misjudge non-causal events as causal, reflecting "Causal Hallucinations." Our *Dr.ECI*, through accurate causal structure matching, effectively reduces such misclassifications. Our method outperforms on most datasets in the few-shot setting, but CoT performs worse. Additional experiments are provided in the Appendix D.4.

As Table 3 illustrates, we rigorously evaluated the effectiveness of *Dr.ECI* in ESL's Inter-sentence and (Intra+Inter)-sentence tasks. Notably, *Dr.ECI*$_{few}$ achieved the highest F1-score in both PLM Baseline and LLM-based Baseline categories, surpassing the best PLM baseline by 9.7% in the Inter-sentence. This demonstrates the formidable reasoning capabilities of LLMs in identifying cross-sentence relationships, with our method further enhancing their causal reasoning prowess on LLMs platform.

Table 4 presents the results for the SemEval and MAVEN-ERE datasets. Our method outperforms the LLM-based baseline with the highest F1-score, underscoring *Dr.ECI*'s proficiency in detecting causal relationships in intricate datasets and its adeptness in detailed causal analysis.

Method	Intra			Causal-TB		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
GenECI†	59.5	57.1	58.8	60.1	53.3	56.5
DPJL†	65.3	70.8	67.9	63.6	66.7	64.6
KEPT†	50.0	68.8	57.9	48.2	60.0	53.5
CPATT†	**79.4**	81.3	80.4	<u>77.5</u>	73.2	75.2
Vanilla	71.2	**98.4**	82.6	63.6	**94.7**	76.1
CoT	71.7	<u>98.2</u>	<u>82.9</u>	64.7	<u>94.3</u>	76.7
Dr.ECI$_{zero}$	76.6	93.1	**84.1**	**80.2**	84.3	**82.2**
Dr.ECI$_{few}$	<u>77.6</u>	91.7	**84.1**	74.7	87.1	<u>80.4</u>

Table 2: Results on ESL's Intra-sentence and Causal-TB dataset. †: results from CPATT (Zhang et al., 2023) .The others come from our results in model PaLM2. The best result is shown in bold and the second best result is underlined.

Method	Inter			Intra+Inter		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
CPATT†	**74.9**	60.1	66.7	**76.5**	66.2	71.0
Vanilla	62.5	**94.2**	75.1	65.1	**95.5**	77.4
CoT	63.1	<u>93.4</u>	75.3	65.7	<u>94.9</u>	77.6
Dr.ECI$_{zero}$	<u>70.7</u>	81.3	<u>75.6</u>	<u>72.6</u>	85.0	<u>78.3</u>
Dr.ECI$_{few}$	68.4	86.6	**76.4**	70.1	88.2	**78.8**

Table 3: Results on ESL's Inter-sentence and (Intra+Inter)-sentence. †: results from CPATT (Zhang et al., 2023). The others come from our results in model PaLM2. The best result is shown in bold and the second best result is underlined.

4.4 Ablation Study

Our ablation study focuses on the *Causal Explorer*, the *Mediator*, and the *Reasoners* within our framework, as demonstrated in Tables 5 and 6. These components were rigorously tested across ESL and

Method	SemEval			MAVEN-ERE		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
ChatGPT†	-	-	-	19.9	85.8	32.3
Vanilla	15.8	**96.0**	27.2	21.3	<u>85.9</u>	34.1
CoT	16.8	<u>95.7</u>	28.6	20.4	**89.9**	33.3
Dr.ECI$_{zero}$	<u>16.9</u>	95.4	<u>28.8</u>	<u>22.0</u>	80.0	<u>34.5</u>
Dr.ECI$_{few}$	**20.3**	87.5	**32.9**	**22.9**	80.7	**35.6**

Table 4: Results on the SemEval and MAVEN-ERE. †: results from ChatGPT (Gao et al., 2023) . The others come from our results in model PaLM2. The best result is shown in bold and the second best result is underlined.

Causal-TB datasets. The *Explorer* represents the presence of a 'Yes.' *Reasoners* refers to the *Direct Reasoner* and *Indirect Reasoner* agents. The *Mediator Detector* only captures E_{st}, so it should be used with the *Reasoners* (as *Mediator*).

In the discovery stage, *Causal Explorer* can effectively explore implicit and indirect causal relationships in a given context through different natural language prompts. Table 5 shows that the *Causal Explorer* agent obtains the highest Recall on all datasets. This also illustrates the effectiveness of our *Causal Explorer* agent in the causal discovery stage. During the reasoning stage, our approach injects causal structure knowledge into the *Reasoners* agent to accurately identify direct and complex indirect causal relationships. Table 6 shows that the highest Precision was achieved on the ESL dataset.

In summary, through the collaboration of multiple agents, *Dr.ECI* successfully achieves the accurate identification of both direct and complex indirect causal relationships. This system delicately balances Precision and Recall, as reflected in its outstanding F1-score.

Method	Intra			Causal-TB		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
Vanilla	71.2	98.4	82.6	63.6	94.7	76.1
+ *Explorer*	70.7	**98.7**	82.4	63.2	**95.0**	75.9
+ *Reasoners*	**82.5**	84.1	83.3	**85.1**	75.2	79.8
+ *Mediator*	79.4	88.3	**83.7**	79.6	83.6	**81.6**

Table 5: The ablation study of *Dr.ECI* was conducted on the ESL's Intra-sentence and Causal-TB datasets.

4.5 Analysis

The results are shown in Table 7. The primary objective of this experiment was to evaluate the practicality of our framework, *Dr.ECI*+CoT, particularly in its applicability to reasoning methods akin

Method	Inter			Intra+Inter		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
Vanilla	62.5	94.2	75.1	65.1	95.5	77.4
+ *Explorer*	62.3	**94.8**	**75.2**	64.8	**96.0**	77.3
+ *Reasoners*	**75.7**	72.1	73.9	**78.0**	75.9	76.9
+ *Mediator*	74.1	75.7	74.9	75.8	79.7	**77.7**

Table 6: The ablation study of *Dr.ECI* was conducted on the ESL's Inter-sentence and (Intra+Inter)-sentence. The best result is shown in bold.

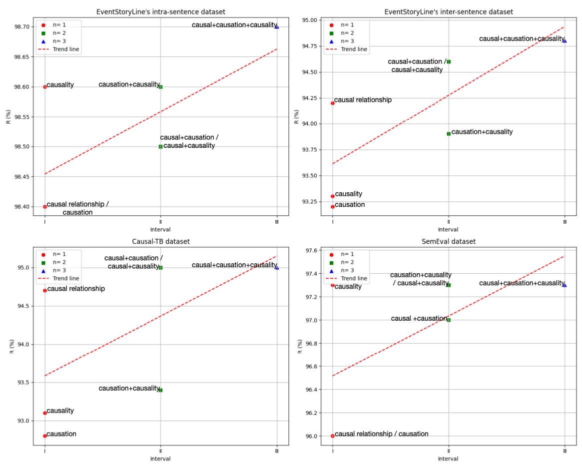


Figure 3: Recall results in the discovery stage where $\mathcal{N} = \{causal\ relationship, causation, causality\}$.

to CoT. The results clearly indicate that the integration of *Dr.ECI* with CoT significantly enhances the model's performance in identifying causal relationships within texts. Specifically, *Dr.ECI*+CoT achieves a state-of-the-art F1-score of 84.4% in ESL's Intra-sentence and 82.9% in the Causal-TimeBank dataset, surpassing the standalone implementations of both *Dr.ECI* and CoT. These results surpass those of standalone *Dr.ECI* and CoT implementations, underscoring the framework's adaptability and effectiveness in complex causality detection and reasoning tasks.

Method	Intra			Causal-TB		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
CoT	71.7	98.2	82.9	64.7	94.3	76.7
Dr.ECI	76.6	93.1	84.1	80.2	84.3	82.2
***Dr.ECI*+CoT**	76.6	94.0	**84.4**	79.4	86.8	**82.9**

Table 7: Adding CoT to our *Dr.ECI* achieves SOTA in F1-score. All results are zero-shot settings in model PaLM2. It shows that our framework is suitable for reasoning methods similar to CoT.

To assess the performance of the *Causal Explorer* agent, as illustrated in Figure 3, we analyzed the Recall across four datasets. Let's assume we have a set $\mathcal{N} = \{causal\ relationship, causation, causality\}$ that is divided into three distinct subsets, denoted as I, II, and III, corresponding to the number of causal expressions, which are 1, 2, and 3, respectively. Regardless of the subset, we set the threshold q to 1. For instance, in the EventStoryLine's Intra-sentence dataset, subset II, which corresponds to two causal expressions (n=2), is represented by the green points. The subsets $\{causal\ relationship, causation\}$ and $\{causal\ relationship, causality\}$ both have a Recall of 98.5%, while the subset $\{causation, causality\}$ has a slightly higher Recall of 98.6%. When n=3, represented by the blue points, the subset is $\{causal\ relationship, causation, causality\}$, and the Recall reaches its peak at 98.7%. The results show that the Recall increases with the number of terms used, suggesting a correlation between the LLMs' cognitive understanding of a concept and the number of its semantically similar expressions.

4.6 Case Study

In this section, we select two cases from the Causal-TB dataset to demonstrate the efficacy of our *Dr.ECI* compared to baseline CoT. in Table 8 is the first case. The first case involves an event pair (dispatching, stationed) that is not causally related, where both the *Causal Explorer* agent and CoT incorrectly predict a causal relationship as 'Yes.' However, *Dr.ECI*, through its causal patterns reasoning stage, accurately identifies the lack of causality, rectifying the initial misjudgments. In the second case, *Dr.ECI* correctly identifies the causal link by matching the 'Collider' structure; however, the result of the CoT is that there is no causal relationship. There are detailed cases in the Appendix Table 18.

5 Related Work

Feature-based approaches in Event Causality Identification have largely concentrated on lexical and syntactic patterns (Riaz and Girju, 2013, 2014a,b) along with statistical analysis (Hashimoto et al., 2014; Hu et al., 2017). However, recent advancements have shifted towards leveraging pre-trained language models. Innovations such as **DPJL** (Shen et al., 2022b), which enhances event causality recognition through cue-based learning from language models, **SemSIn** (Hu et al., 2023), improv-

Context C**:** In the first days after President Bush announced the dispatching of U.S. troops , they note , the Iraqi leader made several nationwide addresses indirectly – having them read by a television announcer . Three other carriers and their escort vessels already are stationed within striking distance of Iraq or are steaming toward the area. e_s:dispatching e_t:stationed
Label: NONE.

Causal Explorer:
1. Is there a causal relationship between dispatching and stationed ? *Answer:* Yes
2. Is there a causality between dispatching and stationed ? *Answer:* No
3. Is there a causation between dispatching and stationed ? *Answer:* Yes

Mediator Detector: **Answer:** Z: None.
1. Related entity or event sets E_s **and** E_t:
subquestion1: Entities or events related to "dispatching "? *Answer:* dispatching.
subquestion2: Entities or events related to "stationed "? *Answer:* stationed.
2. Commonly related entities or event setsE_{st}:
question: Entities or events associated with both events "dispatching" and "stationed"? *Answer:* None.

Direct Reasoner*:* **Answer:** No.
1. Input direct causal pattern:
Explicit Words: there are verbs or verb phrases that clearly express causality.eg:cause, lead,....
Implicit Words: There are implicit clue words in the text that imply a causal relationship.Eg:spark,arouse....
Causal Order: According to common sense knowledge, if event X occurs before event Y...
2. Reasoning Answer: There is no explicit causation words or implicit causation words between dispatching and stationed.There is no causal order between dispatching and stationed.So there is no causal relationship between dispatching and stationed.

Baseline:
1. Vanilla:
question: Is there a causal relationship between dispatching and stationed? *Answer:* Yes.
2. CoT:
question: Is there a causal relationship between dispatching and stationed, Let's think step by step?
Answer: Step 1: The first sentence says that the Iraqi leader made several nationwide addresses indirectly after President Bush announced the dispatching of U.S. troops.Step 2: The second sentence says that three other carriers and their escort vessels already are stationed within striking distance of Iraq or are steaming toward the area.Step 3: We can infer that the dispatching of U.S. troops is the cause of the Iraqi leader making several nationwide addresses indirectly.Step 4: We can also infer that the dispatching of U.S. troops is the cause of the three other carriers and their escort vessels being stationed within striking distance of Iraq or are steaming toward the area.Step 5: So there is a causal relationship between dispatching and stationed.

Table 8: A case from Causal-TimeBank where there is no causal relationship.

ing the detection of implicit causal relationships via explicit semantic structuring, and **CPATT** (Zhang et al., 2023), introducing a prompt-based technique with Constrained Prefix ATTention for identifying causal and temporal relations, signify this shift. Despite their effectiveness, these models are heavily reliant on high-quality, manually annotated datasets, limiting their adaptability to new scenarios.

Recent large language models such as PaLM (Chowdhery et al., 2023) and ChatGPT (OpenAI, 2022) have showcased substantial reasoning capabilities, particularly when guided by prompt learning. Studies (Qin et al., 2023; Chan et al., 2023) have demonstrated that ChatGPT exhibits powerful capabilities in processing multiple-choice causal datasets. Zhang et al. (2023) and Gao et al. (2023) both conducted evaluations on the ECI datasets, but the former significantly surpasses the latter. This variance in results can be attributed to the differences in their test set. However, these studies mainly focus on performance evaluation without delving deeply into enhancing the ICL capabilities for ECI tasks.

In our work, we reexamine the ECI task through a causal view, enhancing the causal reasoning abilities of LLMs by aligning them with causal graph structures. While some recent studies have explored counterfactual reasoning (Mu and Li, 2023) to reduce biases and improve LLMs' causal reasoning (Ban et al., 2023; Long et al., 2023; Kıcıman et al., 2023; Tang et al., 2023), our approach uses the Directed Acyclic Graphs (DAGs) structure in the causal graph to solve the ECI task.

6 Conclusion

In this paper, we introduce *Dr.ECI*, a novel framework designed to enhance the causal reasoning capabilities of large language models (LLMs) for Event Causality Identification (ECI). *Dr.ECI* incorporates multiple agents, namely *Causal Explorer*, *Mediator Detector*, *Direct Reasoner*, and *Indirect Reasoner*, strategically orchestrated within Directed Acyclic Graphs (DAGs) to refine the process of causality reasoning. Empirical evaluations conducted on four diverse datasets demonstrate that *Dr.ECI* outperforms existing methods. Additionally, when combined with the Chain of Thought (CoT) reasoning technique, *Dr.ECI* exhibits further improvements in performance. The compatibility of *Dr.ECI* with other augmented LLMs reasoning methods establishes its potential contribution to future research endeavors by effectively incorporating causal knowledge into LLMs.

Limitations

Our research introduces the *Dr.ECI* framework to enhance Event Causality Identification (ECI) in large language models (LLMs), yet it encounters specific limitations. Primarily, the framework's design is highly specialized for ECI tasks, potentially limiting its adaptability to broader Natural Language Processing challenges. Its effectiveness also closely aligns with the inherent capabilities and constraints of the underlying LLMs, such as gpt-3.5-turbo, suggesting a dependency that may affect its performance in diverse contexts. Moreover, the current evaluations, being dataset-specific, might not fully represent the framework's applicability across varied scenarios. While the integral components of *Dr.ECI*, the *Causal Explorer* and *Reasoner* agents are foundational to its operation, there exists a scope for further optimization to enhance their efficiency and context responsiveness. Additionally, the framework's reliance on LLMs' internal reasoning mechanisms, without extensive integration of external knowledge, might limit its comprehensive understanding in complex reasoning tasks.

Acknowledgments

This research was supported in part by National Science and Technology Major Project (2021ZD0111501), National Science Fund for Excellent Young Scholars (62122022), Natural Science Foundation of China (U24A20233, 62476163, 62206064, 62206061, 62406078), the major key project of PCL (PCL2021A12), Guangdong Basic and Applied Basic Research Foundation (2023B1515120020). This research was enabled by the computational resources and support of the High Performance Computing Platform at the School of Computer Science, Guangdong University of Technology.

References

Taiyu Ban, Lyvzhou Chen, Xiangyu Wang, and Huanhuan Chen. 2023. From query tools to causal architects: Harnessing large language models for advanced causal discovery from data. *arXiv preprint arXiv:2306.16902*.

Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Abby Vander Linden, Brittany Harding, Brad Huang, Peter Clark, and Christopher D Manning. 2014. Modeling biological processes for reading comprehension. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1499–1510.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Tommaso Caselli and Piek Vossen. 2017. The event storyline corpus: A new benchmark for causal and temporal relation extraction. In *Proceedings of the Events and Stories in the News Workshop*, pages 77–86.

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

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Jinglong Gao, Xiao Ding, Bing Qin, and Ting Liu. 2023. Is chatgpt a good causal reasoner? a comprehensive evaluation. *Preprint*, arXiv:2305.07375.

Chikara Hashimoto. 2019. Weakly supervised multilingual causality extraction from wikipedia. 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 2988–2999.

Chikara Hashimoto, Kentaro Torisawa, Julien Kloetzer, Motoki Sano, István Varga, Jong-Hoon Oh, and Yutaka Kidawara. 2014. Toward future scenario generation: Extracting event causality exploiting semantic relation, context, and association features. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 987–997, Baltimore, Maryland. Association for Computational Linguistics.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid O Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. *ACL 2010*, page 33.

Zhichao Hu, Elahe Rahimtoroghi, and Marilyn A Walker. 2017. Inference of fine-grained event causality from blogs and films. *arXiv preprint arXiv:1708.09453*.

Zhilei Hu, Zixuan Li, Xiaolong Jin, Long Bai, Saiping Guan, Jiafeng Guo, and Xueqi Cheng. 2023. Semantic structure enhanced event causality identification. *arXiv preprint arXiv:2305.12792*.

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. 2023. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Jintao Liu, Zequn Zhang, Zhi Guo, Li Jin, Xiaoyu Li, Kaiwen Wei, and Xian Sun. 2023. Kept: Knowledge enhanced prompt tuning for event causality identification. *Knowledge-Based Systems*, 259:110064.

Stephanie Long, Tibor Schuster, Alexandre Piché, ServiceNow Research, et al. 2023. Can large language models build causal graphs? *arXiv preprint arXiv:2303.05279*.

Hieu Man, Minh Van Nguyen, and Thien Huu Nguyen. 2022. Event causality identification via generation of important context words. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics (* SEM) at NAACL 2022*.

Paramita Mirza. 2014. Extracting temporal and causal relations between events. In *Proceedings of the ACL 2014 Student Research Workshop*, pages 10–17.

Feiteng Mu and Wenjie Li. 2023. Enhancing event causality identification with counterfactual reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 967–975.

Jong-Hoon Oh, Kentaro Torisawa, Chikara Hashimoto, Ryu Iida, Masahiro Tanaka, and Julien Kloetzer. 2016. A semi-supervised learning approach to why-question answering. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 3022–3029. AAAI Press.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. `https://openai.com/`. 2022.

Minh Tran Phu and Thien Huu Nguyen. 2021. Graph convolutional networks for event causality identification with rich document-level structures. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 3480–3490.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2022. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*.

Ruili Pu, Yang Li, Suge Wang, Deyu Li, Jianxing Zheng, and Jian Liao. 2023. Enhancing event causality identification with event causal label and event pair interaction graph. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10314–10322, Toronto, Canada. Association for Computational Linguistics.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

Mehwish Riaz and Roxana Girju. 2013. Toward a better understanding of causality between verbal events: Extraction and analysis of the causal power of verb-verb associations. In *Proceedings of the SIGDIAL 2013 Conference*, pages 21–30, Metz, France. Association for Computational Linguistics.

Mehwish Riaz and Roxana Girju. 2014a. In-depth exploitation of noun and verb semantics to identify causation in verb-noun pairs. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 161–170, Philadelphia, PA, U.S.A. Association for Computational Linguistics.

Mehwish Riaz and Roxana Girju. 2014b. Recognizing causality in verb-noun pairs via noun and verb semantics. In *Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL)*, pages 48–57, Gothenburg, Sweden. Association for Computational Linguistics.

Shirong Shen, Heng Zhou, Tongtong Wu, and Guilin Qi. 2022a. Event causality identification via derivative prompt joint learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2288–2299, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Shirong Shen, Heng Zhou, Tongtong Wu, and Guilin Qi. 2022b. Event causality identification via derivative prompt joint learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2288–2299.

Ziyi Tang, Ruilin Wang, Weixing Chen, Keze Wang, Yang Liu, Tianshui Chen, and Liang Lin. 2023. Towards causalgpt: A multi-agent approach for faithful knowledge reasoning via promoting causal consistency in llms. *arXiv preprint arXiv:2308.11914*.

Xiaozhi Wang, Yulin Chen, Ning Ding, Hao Peng, Zimu Wang, Yankai Lin, Xu Han, Lei Hou, Juanzi Li, Zhiyuan Liu, et al. 2022. Maven-ere: A unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction. *arXiv preprint arXiv:2211.07342*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Hang Zhang, Wenjun Ke, Jianwei Zhang, Zhizhao Luo, Hewen Ma, Zhen Luan, and Peng Wang. 2023. Prompt-based event relation identification with constrained prefix attention mechanism. *Knowledge-Based Systems*, 281:111072.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

Xinyu Zuo, Yubo Chen, Kang Liu, and Jun Zhao. 2020. KnowDis: Knowledge enhanced data augmentation for event causality detection via distant supervision. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1544–1550, Barcelona, Spain (Online). International Committee on Computational Linguistics.

A Contents

· B Details of Causal Priors

· C Multi-agent Cooperation for *Dr.ECI*

· D Experiments

· D.1 Experiment on gpt-3.5-turbo

· D.2 DAG structure prediction

· D.3 More Baselines

· D.4 Experiments in Few-shot Setting

· D.5 Experiments on different dataset settings

· E *Causal Explorer* and threshold q.

· E.1 Evaluating the Efficacy of *Causal Explorer*

· E.2 Impact of Threshold q on Performance Metrics

· F Different causal expressions

· G Prompts and some examples

B Details of Causal Priors

In our study, we define Event Causality Identification (ECI) as a set of causal diagram patterns, aligning ECI with structured causal diagrams. As illustrated in Table 1, each causal rule is represented by a distinct diagrammatic structure, accompanied by descriptions and natural language examples from various sources. The categorization of ECI into these specific diagram patterns facilitates a clearer understanding of the causal relationships within texts.

The coreference example in our analysis is adapted from Phu et al. (Phu and Nguyen, 2021), which demonstrates the identification of events with the same meaning but different expressions or identical expressions. The chain pattern, as highlighted in our study, is derived from Pu et al.(Pu et al., 2023), showcasing the sequential causality between events. The other examples used to illustrate Explicit Words, Implicit Words, Causal Order, Collider, and Fork are taken from two classic ECI datasets, Causal-TimeBank and EventStoryLine.

This structured classification and the integration of diverse examples from the literature and ECI datasets underscore our method's ability to align ECI tasks with various causal diagram structures, enhancing the understanding and interpretation of complex causal relationships in texts.

C Multi-agent Cooperation for *Dr.ECI*

To understand the agent workflow of our method, as shown in Alg. 1, we detail *Dr.ECI* using pseudocode. The algorithm takes several key inputs: All agents operate independently based on LLMs. Causal expression $\mathcal{N}$, combined with a causal expression n_i and the question Q for the input text $\mathcal{I}$, threshold q (which is the number of 'Yes' in the set $\mathcal{R}$ that has the same number of elements as the set $\mathcal{N}$), LLMs' output $\mathcal{O}$, *Causal Explorer* agent EXPLORE, *Mediator Detector* agent MD, *Indirect Reasoner* agent IR, *Direct Reasoner* agent DR.

Algorithm 1 *Dr.ECI*

```
Input: I = (C, (e_s, e_t))
Output: E_s, E_t, E_st, Yes, Causal Patterns/No
Define N = {n_1, n_2, ..., n_m}
Initialize q = 0                              ▷ Count of 'Yes'
for each n_i in N do
    Create question Q_i using n_i and (e_s, e_t)
    O_i = EXPLORE(I, Q_i)        ▷ O_i is 'Yes'/'No'
    if O_i = 'Yes' then
        q+ = 1
    end if
end for
if q ≥ 1 then
    Prompt1: Identify entities or events E_s / E_t.
    Prompt2: Determine the common parts E_st.
    O_Detector = MD(I, Prompt1, Prompt2)
    Return: Extract E_s, E_t, E_st from O_Detector
    if E_st ≠ ∅ then
        Prompt3: get Indirect Causal Patterns.
        O_IR = IR(I, Prompt3)
        Return:Yes, Indirect Causal Patterns/No.
    else
        Prompt4: get Direct Causal Patterns.
        O_DR = DR(I, Prompt4)
        Return: Yes, Direct Causal Patterns/No.
    end if
else
    Return: No
end if
```

D Experiments

D.1 Experiment on gpt-3.5-turbo

In the experiments shown in Tables 9 and 10 using the gpt-3.5-turbo model, we primarily focused on two classic datasets in the ECI task, namely the EventStoryLine and CausalTimeBank datasets.

Our research results indicate that compared to the LLM-based Baseline, our method achieved significant progress, outperforming both the Vanilla and CoT methods. Particularly noteworthy is the performance on the ESL dataset, where our method surpassed the current state-of-the-art CPATT technology in terms of the F1-score metric. This is especially evident in the ESL's Inter-sentence, where our method improved by 7.3% over CPATT. This significant enhancement highlights the effectiveness of our method in enhancing the model's cross-sentence understanding capabilities.

To optimize result acquisition, we utilized the following structured prompt format:

Prompt : Structured Answer Format

Instruction:...

Just answer me yes or no. Please return the Answer in a JSON format, such as {{"Answer":"Yes"}} or {{"Answer":"No"}}.

Method	Intra			Causal-TB		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
GenECI†	59.5	57.1	58.8	60.1	53.3	56.5
DPJL†	65.3	70.8	67.9	63.6	66.7	64.6
KEPT†	50.0	68.8	57.9	48.2	60.0	53.5
CPATT†	79.4	81.3	80.4	**77.5**	73.2	**75.2**
Vanilla	78.5	83.3	80.8	68.1	72.6	70.3
CoT	73.4	**84.1**	78.4	56.7	74.5	64.4
Dr.ECI$_{zero}$	**81.0**	82.4	**81.7**	68.0	**74.8**	71.3

Table 9: Results on EventStoryLine's Intra-sentence and Causal-TimeBank. †: results from CPATT (Zhang et al., 2023) .The others come from our results in model gpt-3.5-turbo.the best result is shown in bold and the second best result is underlined. *Dr.ECI* outperforms previous SOTA in F1-score.

Method	Inter			Intra+Inter		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
CPATT†	**74.9**	60.1	66.7	**76.5**	66.2	71.0
Vanilla	69.4	**78.6**	73.7	72.2	80.1	75.9
CoT	60.5	82.7	69.8	64.0	**83.1**	72.3
Dr.ECI$_{zero}$	69.6	79.1	**74.0**	72.9	80.1	**76.4**

Table 10: Results on EventStoryLine's Inter-sentence and (Intra+Inter)-sentence.†:results from CPATT (Zhang et al., 2023) .The others come from our results in model gpt-3.5-turbo.the best result is shown in bold and the second best result is underlined.

D.2 DAG structure prediction

In Table 1, we utilized Directed Acyclic Graphs (DAGs) structures—Direct, Chain, Collider, and Fork—from causal theory to construct natural language expressions suited for Event Causality Identification (ECI). We also incorporated Coreference relations from prior ECI research into our framework.

Experiments with these structured expressions demonstrate that our method, *Dr.ECI*, effectively predicts various DAG structures across multiple datasets, as shown in Table 11. These results highlight *Dr.ECI*'s ability to accurately predict complex causal patterns, reducing "Causal Hallucinations" and enhancing causal reasoning within large language models (LLMs).

Table 11 details the model's performance, emphasizing its capability in handling explicit and implicit causal terms, as well as complex patterns like Colliders, which are particularly notable in the ESL's Inter-sentence and MAVEN-ERE datasets. This efficiency suggests that *Dr.ECI* can significantly improve causal inference accuracy in Natural Language Processing. We found that the reason for the high number of Colliders is due to the order of the causal patterns in the prompt, with the Collider being placed first.

D.3 More Baselines

To further validate the effectiveness of our method, we compared *Dr.ECI* with other LLM-based methods, including those proposed by Tang et al. (2023), Press et al. (2022), and Zhou et al. (2022). These comparisons, particularly in zero-shot settings, underscore the superior performance of *Dr.ECI* in identifying causal relationships using prior knowledge.

The experimental results in Table 12 demonstrate that *Dr.ECI* achieves the highest F1-score across both the Causal-TB and MAVEN-ERE datasets. For the Causal-TB dataset, *Dr.ECI* outperforms other methods with an F1-score of 82.2%, significantly higher than the next best method, CoT, which scores 76.7%. On the MAVEN-ERE dataset, *Dr.ECI* achieves an F1-score of 34.5%, again surpassing CoT's 33.3%. These results highlight the robustness and accuracy of *Dr.ECI*, making it a highly effective approach for causal relationship identification in diverse contexts.

The experimental results displayed in Table 13 demonstrate the exceptional performance of

Causal pattern	SemEval	Causal-TB	Intra	Inter	MAVEN-ERE
Explicit Words	17	12	27	71	58
Implicit Words	3	2	13	26	50
Causal Order	75	96	593	934	278
Chain	10	24	192	328	131
Collider	172	122	613	1207	1976
Fork	3	11	47	65	51
Coreference X	3	1	12	21	36
Coreference Y	2	1	7	17	36

Table 11: Results indicate the number of DAG structures predicted by *Dr.ECI* across various datasets in the zero-setting on PaLM2.

Dr.ECI across ESL's Intra-sentence and Inter-sentence datasets, particularly with respect to the F1-score. Among all methods evaluated, only the Self-ask (Press et al., 2022) slightly outperformed our model, with a marginal difference of less than 0.01 in the F1-score. This minor discrepancy underscores the viability and competitive edge of our approach in handling this dataset.

D.4 Experiments in Few-shot Setting

Based on the performance of in-context learning on large language models, to demonstrate the efficacy of our method, we conducted experiments not only in a zero-shot setting but also in a few-shot setting. For each causal pattern we proposed, we provided one sample, leading us to construct eight samples manually based on these patterns. To align with this setting, we also configured eight samples for Chain of Thought (CoT), derived from results obtained in the zero-shot setting on the PaLM2 model. The experimental results, as shown in Table 14 and Table 15, reveal that our method significantly outperforms the LLM-based CoT approach in terms of F1-score.

However, contrary to our initial intuition, the few-shot experiment results showed a decrease in performance compared to the zero-shot setting, which suggests that LLMs may not always learn better reasoning paths with increased in-context learning as expected. This discrepancy highlights potential issues in the ECI task, such as the "Causal Hallucinations" observed with CoT, where providing samples actually exacerbated hallucinations. We hypothesize that one reason for this might be the model's preference for the distribution of input context samples, as all constructed samples involved event pairs (e_s, e_t) with causal relationships.

D.5 Experiments on different dataset settings

We noted that Gao et al. (2023) initially conducted a comprehensive evaluation of ChatGPT's causal reasoning capabilities. Here we conduct experiments on three datasets: ESL's Intra-sentence, Causal-TB, and MAVEN-ERE, following their publicly available data processing methods. It is important to highlight that there are differences in handling the ECI datasets between CPATT (Zhang et al., 2023) and Gao et al. (2023), especially with respect to datasets like ESL'Intra-sentence and Causal-TB. Our base model is PaLM2 and is compared with the ChatGPT performance metrics evaluated by Gao et al. (2023). The results, detailed in Table 16, demonstrate that our method surpasses all other baselines in F1-score, affirming its efficacy.

Significant variances in dataset sizes exist as outlined in the experimental settings by Gao et al. (2023), particularly observed across the ESL's Intra-sentence, Causal-TB, and MAVEN-ERE datasets. For instance, in experiments on the MAVEN-ERE dataset, our method achieved an F1-score 1.2% higher than CoT in zero-shot experiments, demonstrating the effectiveness of our approach. It is worth noting that this improvement may be even more significant than it appears at first glance. This is because, based on our statistics, the validation set of the MAVEN-ERE dataset contains 16,283 instances without causal relationships and 3,360 instances with causal relationships, with a ratio close to 5:1. It is very challenging to improve precision given such a disparity in data distribution, yet we outperformed CoT by 1.6%.

Method	Causal-TB			MAVEN-ERE		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
CoT(Wei et al., 2022)	64.7	94.3	76.7	20.4	**89.9**	33.3
Self-ask(Press et al., 2022)	55.0	**97.5**	70.3	20.1	88.7	32.8
CaCo-CoT(Tang et al., 2023)	63.0	81.8	71.1	20.5	76.7	32.3
L2M-CoT(Zhou et al., 2022)	75.6	57.5	65.4	**25.8**	44.1	32.6
Dr.ECI	**80.2**	84.3	**82.2**	22.0	80.0	**34.5**

Table 12: Results on the Causal-TB and MAVEN-ERE datasets.All results are in zero-shot setting on the PaLM2 model. The best result is shown in bold.

Method	Intra			Inter			Intra+Inter		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
GenECI†(Man et al., 2022)	59.5	57.1	58.8	-	-	-	-	-	-
DPJL†(Shen et al., 2022b)	65.3	70.8	67.9	-	-	-	-	-	-
KEPT†(Liu et al., 2023)	50.0	68.8	57.9	-	-	-	-	-	-
CPATT†(Zhang et al., 2023)	<u>79.4</u>	81.3	80.4	**74.9**	60.1	66.7	**76.5**	66.2	71.0
Vanilla(Kojima et al., 2022)	71.2	**98.4**	82.6	62.5	**94.2**	75.1	65.1	**95.5**	77.4
CoT (Wei et al., 2022)	71.7	<u>98.2</u>	82.9	63.1	<u>93.4</u>	75.3	65.7	<u>94.9</u>	77.6
Self-ask(Press et al., 2022)	76.4	95.4	**84.8**	65.4	92.0	**76.4**	68.5	93.1	**78.9**
CaCo-CoT(Tang et al., 2023)	75.3	85.2	80.0	64.9	84.3	73.4	67.9	84.6	75.3
L2M-CoT(Zhou et al., 2022)	**80.8**	72.3	76.5	<u>71.7</u>	71.2	71.5	<u>74.4</u>	71.7	73.0
Dr.ECI	76.6	93.1	<u>84.1</u>	70.7	81.3	<u>75.6</u>	72.6	85.0	<u>78.3</u>

Table 13: Results on ESL's Intra-sentence, Inter-sentence, and Intra+Inter-sentence datasets. †: Results based on Pre-trained models. The others come from our results in zero-shot setting on the PaLM2 model. The best result is shown in bold and the second best result is underlined.

E *Causal Explorer* and Threshold q

E.1 Evaluating the Efficacy of *Causal Explorer* in Exploring Implicit Causal Relationships

In the *Causal Explorer* part of our framework, we define a set of causal expressions:

$$\mathcal{N}_D = \{n_i, n_j, n_k\} \quad (2)$$

where n_i, n_j, and n_k represent '*causal relationship*', '*causation*', and '*causality*', respectively. These terms are ordered to reflect their respective relevance and contribution to the analysis of causal inference within text. This structured approach allows us to methodically assess the impact of each causal expression on the Recall, thereby providing a granular insight into the dynamics of causal reasoning.

The variable q, representing the threshold in the *Causal Explorer*, is crucial for this analysis. Specifically, when $q = 1$, the Recall pertains solely to the first element in the set $\mathcal{N}_D$, which is n_i. When q increases to 2, it is the comprehensive Recall that includes n_i and n_j. When q reaches 3, it reflects the Recall for all included elements—n_i, n_j, and n_k. In general, the setting of the threshold q depends on the number of elements in the set $\mathcal{N}_D$. And the *Causal Explorer* is explored by selecting the first element from the set $\mathcal{N}_D$ until the number of elements reaches the threshold q.This incremental inclusion of causal expressions is designed to test the robustness of our model in recognizing and reasoning over causal relationships embedded in varying linguistic constructs.

As demonstrated in Figures 4a, 4b, and 4c, across four different datasets, there is a consistent upward trend in Recall as the number of causal expressions, denoted by $\mathcal{N}_D$, included in the analysis increases. This indicates that the capacity of our proposed *Causal Explorer* to effectively explore both explicit and implicit causal relationships may enhance as the quantity of $\mathcal{N}_D$ increases. Such findings confirm the importance of the *Causal Explorer* in improving the causal reasoning capabilities of large language models (LLMs). The consistent conclusions across various datasets underscore the potential of our method to enhance the comprehensiveness and accuracy of causal relationship discovery across diverse datasets.

Furthermore, by comparing Figures 4a, 4b, and 4c, we observe that the order of elements within the set $\mathcal{N}_D$ also impacts the outcome Recall. For instance, on the Inter-sentence and Causal-TB datasets, at $q = 1$, the Recall in Figure 4a for $\mathcal{N}_{D1}$

Method	Intra			Inter			Intra+Inter		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
CoT$_{few-shot}$	70.5	99.2	82.4	57.5	97.6	72.3	61.1	98.1	75.3
Dr.ECI$_{few-shot}$	77.6	91.7	**84.1**	68.4	86.6	**76.4**	70.1	88.2	**78.8**

Table 14: Results on ESL's Intra-sentence, Inter-sentence, and Intra+Inter-sentence datasets. The results are from the few-shot setting of model PaLM2. The best result is shown in bold.

Method	Causal-TB			MAVEN-ERE		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
CoT$_{few-shot}$	51.2	98.7	67.5	18.7	96.3	31.3
Dr.ECI$_{few-shot}$	74.7	87.1	**80.4**	22.9	80.7	**35.6**

Table 15: Results on the MAVEN-ERE and Causal-TB datasets in the few-shot setting of the PaLM2 model. The best results are highlighted in bold.

exceeds 94%. However, it falls below 94% in Figures 4b and 4c. This disparity suggests that inappropriate selection of the first element in LLMs causal reasoning can diminish model performance. Yet, as previously analyzed, the Recall of our *Causal Explorer* increases with the addition of more $\mathcal{N}_D$ elements, thereby effectively mitigating the risk of selection errors.

E.2 Impact of Threshold q on Performance Metrics

When we observed that the threshold q impacts the experimental outcomes, we initiated experiments to investigate the properties of q. Based on the experimental results in Figures 5a and 5b, we analyzed the performance of the *Causal Explorer* on the ESL dataset in the discovery stage. As the threshold q increases, the Recall decreases significantly, while Precision increases, and the F1-score remains relatively stable. This indicates that higher thresholds lead to fewer detected causal relationships but with greater accuracy.

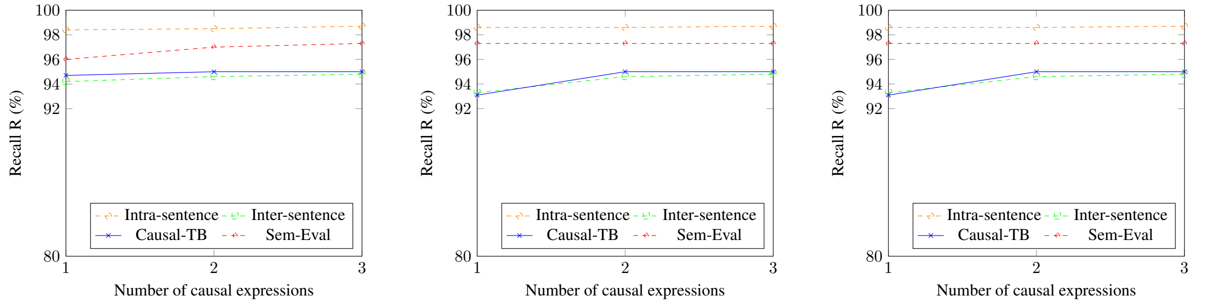
In the ESL's Inter-sentence dataset, Recall drops from 94.8% to 92.5% as q increases from 1 to 3, while Precision rises from 62.3% to 63.4%. The F1-score stays constant at around 75.2%. These results suggest that a higher threshold improves Precision by reducing false positives, though it also results in missing some true causal relationships, balancing the overall performance.

F Different causal expressions

In the discovery stage, the *Causal Explorer* we proposed has the ability to explore potential and implicit causal relationships. We know that for the same concept, there are different expressions in the language. Similarly, for causal relationships, compared with the previously mentioned $\mathcal{N}$ = {*causal relationship*, *causation*, *causality*}, we propose more causal expression types, as shown in the Table 17. The causal Explorer agent explores all these causal expressions separately to obtain more implicit causal event pairs. It is worth noting that the purpose of our stage is to explore more implicit relationships, so we restrict the model's answers here and add the instruction: Please you JUST answer me like {Answer:Yes} or {Answer:No}, in the format of JSON {Answer: }. Through this setting, the model only outputs the JSON format of {Answer:Yes} or {Answer:No}, which is convenient for direct judgment at this stage.

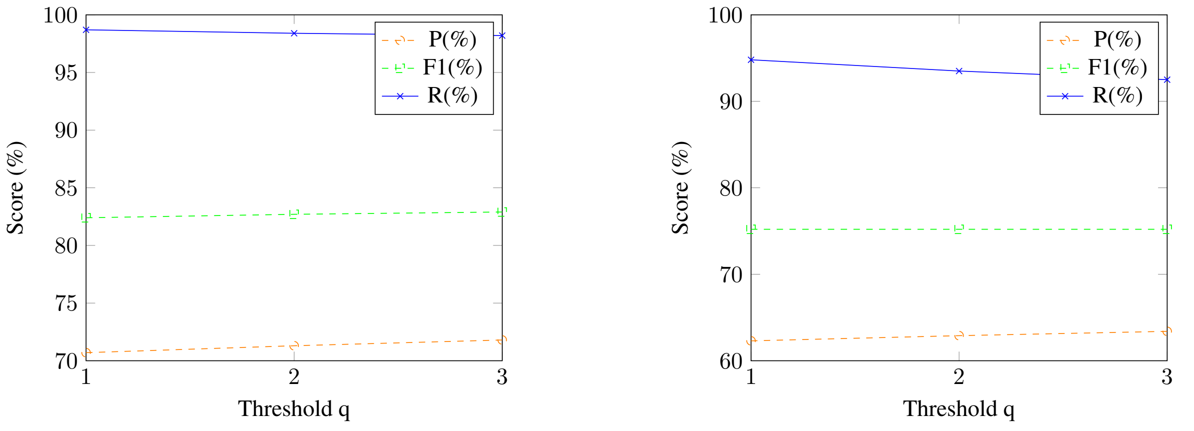
The experimental results presented in Figure 6 and Figure 7 illustrate the effectiveness of our Causal Explorer in identifying implicit causal relationships across various causal expressions. On Causal-TB dataset, our results show that the *Causal Explorer* accurately distinguishes between different types of causal relationships, with Recall ranging from 81.4% for "cause" to 95.9% for "change." This variance underscores the importance of using multiple templates to capture the nuances of different causal expressions.The high Recall across various expressions suggests that the *Causal Explorer* is robust and versatile, capable of adapting to different linguistic representations of causality. These findings support the utility of employing diverse causal expressions to uncover implicit relationships that may not be immediately apparent.

In summary, the experimental results confirm the efficacy of our approach in exploring and identifying implicit causal relationships using multiple



(a) $\mathcal{N}_{D1} = \{n_i, n_j, n_k\}$ (b) $\mathcal{N}_{D2} = \{n_j, n_k, n_i\}$ (c) $\mathcal{N}_{D3} = \{n_k, n_i, n_j\}$

Figure 4: Where, n_i = *causal relationship*, n_j = *causation*, and n_k = *causality* define different types of causal expressions. (c) Within the causal expressions set $\mathcal{N}_{D3} = \{n_k, n_i, n_j\}$:When the number of causal expressions is 1, it corresponds to the recall R for n_k.When the number of causal expressions is 2, it corresponds to the combined recall R for both n_k and n_i.When the number of causal expressions is 3, it corresponds to the recall R for all elements in the set, which are n_k, n_i, and n_j.



(a) ESL's Intra-sentence dataset (b) ESL's Inter-sentence dataset

Figure 5: Here $\mathcal{N} = \{$*causal relationship*, *causation*, *causality*$\}$, we examine the Precision (P), Recall (R), and F1-score metrics at different thresholds q. A threshold $q \geq 2$ indicates that within $\mathcal{N}$, there are at least two causal expressions assessed as positive.

Model	Method	Intra			Causal-TB			MAVEN-ERE		
		P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
T5	**FLAN-T5-Large**(Chung et al., 2024)	19.4	91.1	32.0	4.2	82.6	7.9	18.1	93.5	30.3
	- w/ RLHF	17.6	97.9	29.8	3.9	99.0	7.5	17.1	99.4	29.2
ChatGPT	**Vanilla**(Gao et al., 2023)	27.6	80.2	41.0	6.9	82.6	12.8	19.9	85.8	32.3
	open-ended A.3(Gao et al., 2023)	14.1	10.4	12.0	-	-	-	-	-	-
	open-ended B(Gao et al., 2023)	6.3	54.0	11.3	-	-	-	-	-	-
text-davinci-003	**CoT**(Gao et al., 2023)	16.6	32.8	33.1	-	-	-	-	-	-
PaLM2	**Vanilla**	27.2	82.2	40.8	6.5	74.5	11.9	21.3	85.9	34.1
	Dr.ECI	29.0	75.4	**41.9**	71.7	71.8	**13.0**	22.0	80.0	**34.5**

Table 16: Results on ESL's Intra-sentence, Causal-TB, and MAVEN-ERE datasets. Experiments on models ChatGPT, text-davinci-003, and PaLM2 are all zero-shot settings. The best result is shown in bold.

Table 17: In the discovery stage, templates for *Causal Explorer* in different causal expressions.

Causal Expression	**Causality Question Template**
causal relationship	Is there a causal relationship between {source} and {target} ?
causation	Is there causation between {source} and {target} ?
causality	Is there causality between {source} and {target} ?
lead	Can Does {source} lead to {target}?
contribute	Does {source} contribute to {target}?
change	Will changing {source} also change {target} ?
effect	Is {target} the effect of {source} ?
trigger	Does {source} trigger {target} ?
cause	Does {source} cause {target} ?

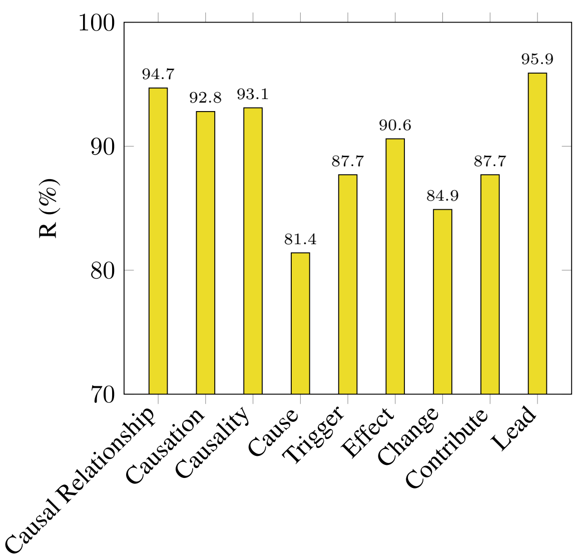


Figure 6: Results for R(%) across various causal expressions using the PaLM2 model on the Causal-TB dataset.

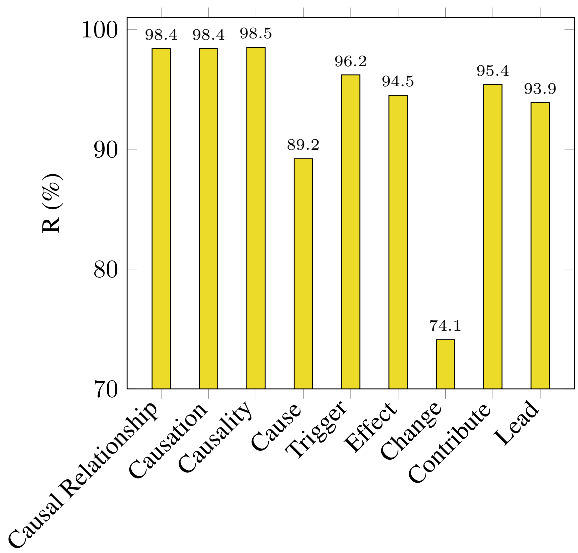


Figure 7: Results for R(%) across various causal expressions using the PaLM2 model on the ESL's Intra-sentence dataset.

expression templates. The use of a standardized output format further facilitated the analysis and validation of these relationships, demonstrating the model's potential for broader applications in causal relationship discovery.

G Prompts and some examples

The prompt in the Causal Priors section is in Prompt G.1. Prompts and exemplars of different datasets are shown in Prompt G.2, G.3, G.4, G.5, and G.6, respectively.

Task Input $\mathcal{I}$

context C: It wasn't until twenty years after the first astronauts were chosen that NASA finally included six women, and they were all scientists, not pilots. Finally today, we learned that the space agency has finally taken a giant leap forward.
e_s: chosen,
e_t: taken

User Query according to CoT

Question:Is there a causal relationship between chosen and taken,Let's think step by step?

Model Response

Step 1: The first astronauts were chosen twenty years before NASA finally included six women.Step 2: The space agency has finally taken a giant leap forward.Step 3: The space agency took a giant leap forward after including six women.Step 4: The space agency took a giant leap forward after the first astronauts were chosen.Step 5: The space agency took a giant leap forward because the first astronauts were chosen.Therefore, the answer is causal.

Labels.

None.

Figure 8: Examples of response from PaLM2 model on the Causal-TimeBank dataset. Task Input and User Query are inputs. This results from Chain-of-Thought(CoT) (Wei et al., 2022).

Task Input $\mathcal{I}$

context C: It wasn't until twenty years after the first astronauts were chosen that NASA finally included six women, and they were all scientists, not pilots. Finally today, we learned that the space agency has finally taken a giant leap forward.
e_s: chosen,
e_t: taken

User Query according to *Causal Explorer*

Question:Is there causation between chosen and taken?
Your final answer should be "Answer":"Yes" or "Answer":"No", in the format of JSON "Answer":" ", at the end of your reply.

Model Response

"Answer":"Yes"

Labels.

None.

Figure 9: Examples of response from PaLM2 model on the Causal-TimeBank dataset. Task Input and User Query are inputs. The causal expression is *causation*.

Task Input $\mathcal{I}$

context C: It wasn't until twenty years after the first astronauts were chosen that NASA finally included six women, and they were all scientists, not pilots. Finally today, we learned that the space agency has finally taken a giant leap forward.
e_s: chosen,
e_t: taken

User Query according to *Causal Explorer*

Question:Is there a causal relationship between chosen and taken?
Your final answer should be "Answer":"Yes" or "Answer":"No", in the format of JSON "Answer":" ", at the end of your reply.

Model Response

"Answer":"Yes"

Labels.

None.

Figure 10: Examples of response from PaLM2 model on the Causal-TimeBank dataset. Task Input and User Query are inputs. The causal expression is *Causal Relationship*.

Task Input $\mathcal{I}$

context C: It wasn't until twenty years after the first astronauts were chosen that NASA finally included six women, and they were all scientists, not pilots. Finally today, we learned that the space agency has finally taken a giant leap forward.
e_s: chosen,
e_t: taken

User Query according to *Causal Explorer*

Question:Is there causality between chosen and taken?
Your final answer should be "Answer":"Yes" or "Answer":"No", in the format of JSON "Answer":" ", at the end of your reply.

Model Response

"Answer":"Yes"

Labels.

None.

Figure 11: Examples of response from PaLM2 model on the Causal-TimeBank dataset. Task Input and User Query are inputs. The causal expression is *causality*.

Task Input $\mathcal{I}$

context C: It wasn't until twenty years after the first astronauts were chosen that NASA finally included six women, and they were all scientists, not pilots. Finally today, we learned that the space agency has finally taken a giant leap forward.
e_s: chosen,
e_t: taken

User Query according to *Mediator Detector*

When answering the question, first obtain the entity or event (E_{st}) that is associated with two events (e_s and e_t), and then use the provided rules wherever possible. Try not to invent rules by yourself unless necessary. **Find entities or events:** 1. Entities or events related to "chosen" in the text: answer me at word level, such as "stay", "begged", "locked up", "John", "person". 2. Entities or events related to "taken" in the text: answer me at word level, such as "stay", "begged", "locked up", "John", "person". 3. E_{st}(Entities or events associated with both events "chosen" and "taken"): answer me at word level, such as "stay", "begged", "locked up", "John", "person".

Model Response

Find entities or events: Entities or events related to "chosen" in the text: "twenty years", "first astronauts" Entities or events related to "taken" in the text: "today", "space agency" E_{st}(Entities or events associated with both "chosen" and "taken"): No.

Labels.

None.

Figure 12: Examples of response from PaLM2 model on the Causal-TimeBank dataset. Task Input and User Query are inputs. This results from *Mediator Detector*.

context C: Ralston Purina Co. reported a 47% decline in fourth-quarter earnings, reflecting restructuring costs as well as a more difficult pet food market.
e_s:decline
e_t:market
Label: Cause-Effect.

Causal Explorer:.
1. Q1: Is there a causal relationship between market and decline? *Answer:* *Yes*
2. Q2: Is there a causality between market and decline? *Answer:* *No*
3. Q3: Is there a causation between market and decline? *Answer:* *Yes*

Mediator Detector: **Answer:** E_{st}: pet food market.
1. Related entity or event sets E_s and E_t:
subquestion1: Entities or events related to "market"? *Answer:* pet food market, pet food market.
subquestion2: Entities or events related to "decline"? *Answer:* 47% decline in fourth-quarter earnings, pet food market.
2.Commonly related entities or event setsE_{st}:
question: Entities or events associated with both events "market" and "decline"? *Answer:*pet food market.

Indirect Reasoner: **Answer:** Yes.
1.Input indirect causal pattern:
Fork: There is an entity or event E_{st}. First, it is related to e_s and e_t, and then E_{st} acts on both e_s and e_t.
Collider: There is an entity or event E_{st}. First, it is related to e_s and e_t, and then e_s and e_t both act on E_{st}.
Chain: There is an entity or event E_{st}. First, it satisfies that it is related to e_s and e_t respectively, then it satisfies that e_s acts on E_{st} and E_{st} acts on e_t.
The Coreference event of event e_s: If an event with the similar meaning as the e_s can be found, and this similar event has a causal relationship with the e_t.
The Coreference event of event e_t: If an event with the similar meaning as the e_t can be found, and this similar event has a causal relationship with the e_s.
2.Reasoning Answer: 1. E_{st} is related to e_s (market) and e_t (decline). 2. e_s (market) acts on E_{st} (pet food market). 3. e_t (decline) acts on E_{st} (pet food market). 4. e_s (market), E_{st} (pet food market), and e_t (decline) form the Collider structure. Therefore, there is a causal relationship between market and decline. {"Answer": "Yes"}.

Baseline:
1. Vanilla:
question: Is there a causal relationship between market and decline? *Answer:* Yes.
2. CoT:
question: Is there a causal relationship between market and decline? Let's think step by step.
Answer:
1. Ralston Purina Co. reported a 47% decline in fourth-quarter earnings.
2. The decline in earnings could be... including ... pet food market.
3. The... pet food market could have contributed to the decline in earnings.
4. However, it is important to note that the statement does not explicitly state a causal relationship between the market and the decline in earnings.
5. Without further information, it is difficult to determine if the decline in earnings was solely caused by the more difficult pet food market or if other factors also played a role.
6. Therefore, it is not possible to definitively conclude a causal relationship between the market and the decline based on the given information.

Table 18: A case from Causal-TimeBank where there is an indirect causal relationship.

Task Input $\mathcal{I}$

context C: It wasn't until twenty years after the first astronauts were chosen that NASA finally included six women, and they were all scientists, not pilots. Finally today, we learned that the space agency has finally taken a giant leap forward.
e_s: chosen, e_t: taken
Find entities or events: Entities or events related to "chosen" in the text: "twenty years", "first astronauts" Entities or events related to "taken" in the text: "today", "space agency"E_{st}(Entities or events associated with both "chosen" and "taken"): No.

User Query according to *Reasoner*

If E_{st} exists, determine which of the following rules each E_{st} can satisfy with events e_s and e_t. If E_{st} does not exist, directly analyze what rules are satisfied between e_s and e_t.

Pattern 1 (Explicit Words)
In the text, there are verbs or verb phrases that clearly express causality, and they relate e_s and e_t. The relationship between them can be expressed as: e_s cause/lead/due to (Explicit causation words) e_t. So there is a causal relationship between e_s and e_t. Explicit causation words (In addition to e_s and e_t themselves, verbs or verb phrases that can clearly express causal relationships): "cause," "result," "lead," "lead to," "due to," "because of," "by," etc.

Pattern 2 (Implicit Words)
There are implicit clue words in the text that imply a causal relationship, and they associate e_s and e_t. These implicit clue words are expressed in natural language as being likely to cause event e_s and event e_t to have a causal relationship. The relationship between them can be expressed as: e_s aggravate/force/provoke (Implicit causation words) e_t. So there is a causal relationship between e_s and e_t. Implicit causation words (In addition to e_s and e_t themselves, verbs or verb phrases that implicitly express causal relationships can be used): "spark," "aggravate," "stir," "precipitate," "ignite," "pave," "exacerbate," "force," "arouse," "worsen," "result from," "reduce," "provoke," "inflict," "fuel," "stem," "prompt," "alleviate," "trigger," "increase with," "contribute to," etc.

Pattern 3 (Coreference of *source event* e_s)
In the text, if an event with the same or similar meaning as the e_s can be found, and this similar event has a causal relationship with the e_t event, then there is also a causal relationship between e_s and e_t.

Pattern 4 (Coreference of *target event* e_t)
In the text, if an event with the same or similar meaning as the e_t can be found, and this similar event has a causal relationship with the e_s, then there is also a causal relationship between e_s and e_t.

Pattern 5 (Causal Order)
According to common sense knowledge, if event e_s occurs before event e_t and can be expressed as: e_s causes e_t, then there is a causal relationship between e_s and e_t. If the expression "e_s causes e_t" does not follow common sense or is incorrectly expressed, then there is no causal relationship between them.

Pattern 6 (Chain)
In "text," there is an entity or event (E_{st}) that both events e_s and e_t are related to. First, it satisfies that it is related to e_s and e_t respectively, then it satisfies that e_s acts on E_{st} and E_{st} acts on e_t, and finally they form a causal chain: $e_s \to E_{st} \to e_t$. Therefore, it can be concluded that there is a causal relationship between e_s and e_t.

Pattern 7 (Collider)
In "text," there is an entity or event (E_{st}) that both events e_s and e_t are related to. First, it satisfies that it is related to e_s and e_t respectively, and then it satisfies that e_s acts on E_{st} and e_t acts on E_{st}. Finally, e_s, E_{st}, and e_t form the Collider structure. Therefore, it can be concluded that there is a causal relationship between e_s and e_t.

Pattern 8 (Fork)
In "text," there is an entity or event (E_{st}) that both events e_s and e_t are related to. First, it satisfies that it is related to e_s and e_t respectively, and then it satisfies that E_{st} acts on e_s and E_{st} acts on e_t. Finally, e_s, E_{st}, and e_t form the Fork structure. Therefore, it can be concluded that there is a causal relationship between e_s and e_t.

Model Response

Analyzing the rules: The text does not mention any causal relationship between "chosen" and "taken". {"Answer":"No"}

Figure 13: Examples of response from PaLM2 model on the Causal-TimeBank dataset. Task Input and User Query are inputs. This results from *Reasoner*. Label: None.

Prompt G.1: Causal Priors

Pattern 1 (Explicit Words)
In the text, there are verbs or verb phrases that clearly express causality, and they relate e_s and e_t. The relationship between them can be expressed as: e_s cause/lead/due to (Explicit causation words) e_t. So there is a causal relationship between e_s and e_t. Explicit causation words (In addition to e_s and e_t themselves, verbs or verb phrases that can clearly express causal relationships): "cause," "result," "lead," "create," "cause by," "lead to," "result in," "due to," "because of," "by," etc.

Pattern 2 (Implicit Words)
There are implicit clue words in the text that imply a causal relationship, and they associate e_s and e_t. These implicit clue words are expressed in natural language as being likely to cause event e_s and event e_t to have a causal relationship. The relationship between them can be expressed as: e_s aggravate/force/provoke (Implicit causation words) e_t. So there is a causal relationship between e_s and e_t. Implicit causation words (In addition to e_s and e_t themselves, verbs or verb phrases that implicitly express causal relationships can be used): "spark," "aggravate," "stir," "precipitate," "ignite," "pave," "exacerbate," "force," "arouse," "worsen," "result from," "reduce," "provoke," "inflict," "fuel," "stem," "prompt," "alleviate," "trigger," "increase with," "contribute to," etc.

Pattern 3 (Coreference of *source event* e_s)
In the text, if an event with the same or similar meaning as the e_s can be found, and this similar event has a causal relationship with the e_t event, then there is also a causal relationship between e_s and e_t.

Pattern 4 (Coreference of *target event* e_t)
In the text, if an event with the same or similar meaning as the e_t can be found, and this similar event has a causal relationship with the e_s, then there is also a causal relationship between e_s and e_t.

Pattern 5 (Causal Order)
According to common sense knowledge, if event e_s occurs before event e_t and can be expressed as: e_s causes e_t, then there is a causal relationship between e_s and e_t. If the expression "e_s causes e_t" does not follow common sense or is incorrectly expressed, then there is no causal relationship between them.

Pattern 6 (Chain)
In "text," there is an entity or event (E_{st}) that both events e_s and e_t are related to. First, it satisfies that it is related to e_s and e_t respectively, then it satisfies that e_s acts on E_{st} and E_{st} acts on e_t, and finally they form a causal chain: $e_s \rightarrow E_{st} \rightarrow e_t$. Therefore, it can be concluded that there is a causal relationship between e_s and e_t.

Pattern 7 (Collider)
In "text," there is an entity or event (E_{st}) that both events e_s and e_t are related to. First, it satisfies that it is related to e_s and e_t respectively, and then it satisfies that e_s acts on E_{st} and e_t acts on E_{st}. Finally, e_s, E_{st}, and e_t form the Collider structure. Therefore, it can be concluded that there is a causal relationship between e_s and e_t.

Pattern 8 (Fork)
In "text," there is an entity or event (E_{st}) that both events e_s and e_t are related to. First, it satisfies that it is related to e_s and e_t respectively, and then it satisfies that E_{st} acts on e_s and E_{st} acts on e_t. Finally, e_s, E_{st}, and e_t form the Fork structure. Therefore, it can be concluded that there is a causal relationship between e_s and e_t.

Prompt G.2: example of Causal-TimeBank. **Labels: NONE**

Input: context C: It wasn't until twenty years after the first astronauts were chosen that NASA finally included six women, and they were all scientists, not pilots. Finally today, we learned that the space agency has finally taken a giant leap forward. e_s: chosen, e_t: taken

Question:Is there a causal relationship between chosen and taken,Let's think step by step?
CoT: Step 1: The first astronauts were chosen twenty years before NASA finally included six women.Step 2: The space agency has finally taken a giant leap forward.Step 3: The space agency took a giant leap forward after including six women.Step 4: The space agency took a giant leap forward after the first astronauts were chosen.Step 5: The space agency took a giant leap forward because the first astronauts were chosen.Therefore, the answer is causal.

Question:Is there a causal relationship between chosen and taken?
Your final answer should be {"Answer":"Yes"} or {"Answer":"No"}, in the format of JSON {"Answer":" "}, at the end of your reply.
Causal Relationship:{"Answer":"Yes"}

Question:Is there causation between chosen and taken?
Your final answer should be {"Answer":"Yes"} or {"Answer":"No"}, in the format of JSON {"Answer":" "}, at the end of your reply.
Causation:{"Answer":"Yes"}

Question:Is there causality between chosen and taken?
Your final answer should be {"Answer":"Yes"} or {"Answer":"No"}, in the format of JSON {"Answer":" "}, at the end of your reply.
Causality:{"Answer":"Yes"}

Question: Instruction: When answering the question, first obtain the entity or event (E_{st}) that is associated with two events (e_s and e_t), and then use the provided rules wherever possible. Try not to invent rules by yourself unless necessary. If E_{st} exists, determine which of the following rules each E_{st} can satisfy with events e_s and e_t. If E_{st} does not exist, directly analyze what rules are satisfied between e_s and e_t. **Find entities or events:** 1. Entities or events related to "chosen" in the text: answer me at word level, such as "stay", "begged", "locked up", "John", "person". 2. Entities or events related to "taken" in the text: answer me at word level, such as "stay", "begged", "locked up", "John", "person". 3. E_{st}(Entities or events associated with both events "chosen" and "taken"): answer me at word level, such as "stay", "begged", "locked up", "John", "person". **Patterns:**... **Question:** Is there a causal relationship between "chosen" and "taken"? Your final answer should be {"Answer":"Yes"} or {"Answer":"No"}, in the format of JSON {"Answer":" "}, at the end of your reply.
***Dr.ECI*:** Find entities or events: Entities or events related to "chosen" in the text: "twenty years", "first astronauts". Entities or events related to "taken" in the text: "today", "space agency". E_{st}(Entities or events associated with both "chosen" and "taken"): No. Analyzing the rules: The text does not mention any causal relationship between "chosen" and "taken". {"Answer":"No"}

Prompt G.3: example of Causal-TimeBank. **Labels:Cause-Effect**

Input: context C: His advisers said the results reflected not just from balancing the budget, but also initiatives like improved access to education and training and the opening of foreign markets to trade.e_s: balancing, e_t: results

Question:Is there a causal relationship between balancing and results,Let's think step by step?
CoT: Step 1: The text states that the results reflected not just from balancing the budget.Step 2: So we can conclude that there is a causal relationship between balancing and results.

Question:Is there a causal relationship between balancing and results?
Your final answer should be {"Answer":"Yes"} or {"Answer":"No"}, in the format of JSON {"Answer":" "}, at the end of your reply.
Causal Relationship:{"Answer":"Yes"}

Question:Is there causation between balancing and results?
Your final answer should be {"Answer":"Yes"} or {"Answer":"No"}, in the format of JSON {"Answer":" "}, at the end of your reply.
Causation:{"Answer":"Yes"}

Question:Is there causality between balancing and results?
Your final answer should be {"Answer":"Yes"} or {"Answer":"No"}, in the format of JSON {"Answer":" "}, at the end of your reply.
Causality:{"Answer":"Yes"}

Question: Instruction: When answering the question, first obtain the entity or event (E_{st}) that is associated with two events (e_s and e_t), and then use the provided rules wherever possible. Try not to invent rules by yourself unless necessary. If E_{st} exists, determine which of the following rules each E_{st} can satisfy with events e_s and e_t. If E_{st} does not exist, directly analyze what rules are satisfied between e_s and e_t. **Find entities or events:** 1. Entities or events related to "balancing" in the text: answer me at word level, such as "stay", "begged", "locked up", "John", "person". 2. Entities or events related to "results" in the text: answer me at word level, such as "stay", "begged", "locked up", "John", "person". 3. E_{st}(Entities or events associated with both events "balancing" and "results"): answer me at word level, such as "stay", "begged", "locked up", "John", "person". **Patterns:**... **Question:** Is there a causal relationship between balancing and results? Your final answer should be {"Answer":"Yes"} or {"Answer":"No"}, in the format of JSON {"Answer":" "}, at the end of your reply.
Dr.ECI: {"Answer":"Yes"}

Prompt G.4: example of Causal-TimeBank. Labels:Cause-Effect

Input: context C: However, it is important to note that installing or implementing Internet technologies for technology's sake would not ensure instant success.e_s: implementing, e_t: success

Question:Is there a causal relationship between implementing and success,Let's think step by step?
CoT: Step 1: The text states that ïnstalling or implementing Internet technologies for technology 's sake would not ensure instant success.Step 2: This means that implementing Internet technologies does not guarantee success.Step 3: Therefore, there is no causal relationship between implementing and success.

Question:Is there a causal relationship between implementing and success?
Your final answer should be {"Answer":"Yes"} or {"Answer":"No"}, in the format of JSON {"Answer":" "}, at the end of your reply.
Causal Relationship:{"Answer":"Yes"}

Question:Is there causation between implementing and success?
Your final answer should be {"Answer":"Yes"} or {"Answer":"No"}, in the format of JSON {"Answer":" "}, at the end of your reply.
Causation:{"Answer":"No"}

Question:Is there causality between implementing and success?
Your final answer should be {"Answer":"Yes"} or {"Answer":"No"}, in the format of JSON {"Answer":" "}, at the end of your reply.
Causality:{"Answer":"Yes"}

Question: Instruction: When answering the question, first obtain the entity or event (E_{st}) that is associated with two events (e_s and e_t), and then use the provided rules wherever possible. Try not to invent rules by yourself unless necessary. If E_{st} exists, determine which of the following rules each E_{st} can satisfy with events e_s and e_t. If E_{st} does not exist, directly analyze what rules are satisfied between e_s and e_t. **Find entities or events:** 1. Entities or events related to "implementing" in the text: answer me at word level, such as "stay", "begged", "locked up", "John", "person". 2. Entities or events related to "success" in the text: answer me at word level, such as "stay", "begged", "locked up", "John", "person". 3. E_{st} (Entities or events associated with both events "implementing" and "success"): answer me at word level, such as "stay", "begged", "locked up", "John", "person". **Patterns:**... **Question:** Is there a causal relationship between implementing and success? Your final answer should be {"Answer":"Yes"} or {"Answer":"No"}, in the format of JSON {"Answer":" "}, at the end of your reply.
Dr.ECI: {"Answer":"Yes"}

Prompt G.5: example of EventStoryLine. **Labels:NONE**

Input: context C:Further restoration has continued throughout the weekend in cooperation with customers and partners and SEACOM will continue to actively work on securing additional capacity for all necessary services over the next day or two . Ä repair vessel has been notified of the call - out and mobilization will occur immediately once permits are received , S̈EACOM said in an official statement . e_s: work, e_t: received

Question:Is there a causal relationship between work and received,Let's think step by step?
CoT: Yes, there is a causal relationship between work and received.The text states that "SEACOM will continue to actively work on securing additional capacity for all necessary services over the next day or two". This means that SEACOM is working to get the necessary permits to repair the damage. Once the permits are received, the repair vessel will be mobilized to begin the repairs.

Question:Is there a causal relationship between between work and received?
Your final answer should be {"Answer":"Yes"} or {"Answer":"No"}, in the format of JSON {"Answer":" "}, at the end of your reply.
Causal Relationship:{"Answer":"Yes"}

Question:Is there causation between work and received?
Your final answer should be {"Answer":"Yes"} or {"Answer":"No"}, in the format of JSON {"Answer":" "}, at the end of your reply.
Causation:{"Answer":"Yes"}

Question:Is there causality between work and received?
Your final answer should be {"Answer":"Yes"} or {"Answer":"No"}, in the format of JSON {"Answer":" "}, at the end of your reply.
Causality:{"Answer":"Yes"}

Question: Instruction: When answering the question, first obtain the entity or event (E_{st}) that is associated with two events (e_s and e_t), and then use the provided rules wherever possible. Try not to invent rules by yourself unless necessary. If E_{st} exists, determine which of the following rules each E_{st} can satisfy with events e_s and e_t. If E_{st} does not exist, directly analyze what rules are satisfied between e_s and e_t. **Find entities or events:** 1. Entities or events related to "work" in the text: answer me at word level, such as "stay", "begged", "locked up", "John", "person". 2. Entities or events related to "received" in the text: answer me at word level, such as "stay", "begged", "locked up", "John", "person". 3. E_{st} (Entities or events associated with both events "work" and "received"): answer me at word level, such as "stay", "begged", "locked up", "John", "person". **Patterns:**... **Question:** Is there a causal relationship between work and received? Your final answer should be {"Answer":"Yes"} or {"Answer":"No"}, in the format of JSON {"Answer":" "}, at the end of your reply.
***Dr.ECI*:** {"Answer":"No"}. Entities or events related to "work" in the text: work. Entities or events related to "received" in the text: received. E_{st} (Entities or events associated with both events "work" and "received"): There is no E_{st} that satisfies the Chain, Collider, Fork rules. There is no explicit causation words or implicit causation words between work and received. There is no causal order between work and received. There is no coreference event of work or received. So the final answer is: No.

Prompt G.6: example of EventStoryLine. Labels:NONE

Input: context C:Police are continuing to investigate a fire at a Waitrose supermarket in Wellington . e_s: continuing, e_t: fire

Question:Is there a causal relationship between continuing and fire,Let's think step by step?
CoT: The answer is causal.The text states that police are continuing to investigate a fire. This means that the fire is still ongoing and that the police are still trying to figure out what happened. This is a causal relationship because the fire is causing the police to continue to investigate.

Question:Is there a causal relationship between continuing and fire?
Your final answer should be {"Answer":"Yes"} or {"Answer":"No"}, in the format of JSON {"Answer":" "}, at the end of your reply.
Causal Relationship:{"Answer":"Yes"}

Question:Is there causation between continuing and fire?
Your final answer should be {"Answer":"Yes"} or {"Answer":"No"}, in the format of JSON {"Answer":" "}, at the end of your reply.
Causation:{"Answer":"Yes"}

Question:Is there causality between continuing and fire?
Your final answer should be {"Answer":"Yes"} or {"Answer":"No"}, in the format of JSON {"Answer":" "}, at the end of your reply.
Causality:{"Answer":"Yes"}

Question: Instruction: When answering the question, first obtain the entity or event (E_{st}) that is associated with two events (e_s and e_t), and then use the provided rules wherever possible. Try not to invent rules by yourself unless necessary. If E_{st} exists, determine which of the following rules each E_{st} can satisfy with events e_s and e_t. If E_{st} does not exist, directly analyze what rules are satisfied between e_s and e_t. **Find entities or events:** 1. Entities or events related to "continuing" in the text: answer me at word level, such as "stay", "begged", "locked up", "John", "person". 2. Entities or events related to "fire" in the text: answer me at word level, such as "stay", "begged", "locked up", "John", "person". 3. E_{st} (Entities or events associated with both events "continuing" and "fire"): answer me at word level, such as "stay", "begged", "locked up", "John", "person". **Patterns:**... **Question:** Is there a causal relationship between continuing and fire? Your final answer should be {"Answer":"Yes"} or {"Answer":"No"}, in the format of JSON {"Answer":" "}, at the end of your reply.
***Dr.ECI*:** {"Answer":"No"}. 1. Entities or events related to "continuing" in the text: continue. 2. Entities or events related to "fire" in the text: fire. 3. E_{st} (Entities or events associated with both events "continuing" and "fire"): There is no E_{st} that satisfies the following Patterns: - Pattern 6: Chain - Pattern 7: Collider - Pattern 8: Fork. According to common sense knowledge, if event e_s occurs before event e_t and can be expressed as: e_s causes e_t, then there is a causal relationship between e_s and e_t. If the expression "e_s causes e_t" does not follow common sense or is incorrectly expressed, then there is no causal relationship between them. In the text, "continuing" and "fire" are not related to each other. So there is no causal relationship between "continuing" and "fire".