Debate with Myself: Zero-Shot Event Causality Identification with Adversarial Evidence Integration via Large Language Models

Zefan Zeng[1], **Yuehang Si**[1], **Xingchen Hu**[2], **Qing Cheng**[1], **Jiakun Liu**[3*] and **Zhong Liu**[1]

[1]Laboratory for Big Data and Decision, National University of Defense Technology
[2]College of Computer Science and Technology, National University of Defense Technology
[3]Faculty of Computing, Harbin Institute of Technology
{zengzefan, siyuehang}@nudt.edu.cn, xhu4@ualberta.ca, chengqing@nudt.edu.cn,
jiakunliu@hit.edu.cn, liuzhong@nudt.edu.cn

Abstract

Event Causality Identification (ECI) is a crucial task in knowledge discovery that extracts structured causal relationships between annotated event mentions from unstructured text. However, existing approaches typically rely on extensive labeled data, which is scarce for specialized domains and topics. Although Large Language Models (LLMs) show strong promise for few-shot and zero-shot information extraction, they are prone to "causal hallucination," generating unreliable and spurious causal links. To address these limitations, we propose LLM-SD (Large Language Model Self-Debate), a novel framework that formulates ECI as a structured debate among multiple identical instances of a single LLM. Causality is determined through the integration of adversarial evidence. The framework employs LLMs in distinct roles: an affirmative team argues for the existence of causality, a negative team argues against it, and an adjudication committee evaluates the evidence for determination. An evidence strength grading rule guides the quantification and integration of adversarial evidence. The automatic and LLM-driven verification finally produce a reasoned verdict for event causality. LLM-SD reduces spurious causal links resulting from causal hallucination in LLMs and identifies more long-distance causalities by promoting a balanced evaluation of arguments. Extensive experiments on three benchmark datasets demonstrate that LLM-SD achieves state-of-the-art performance in a zero-shot setting.

1 Introduction

Event Causality Identification (ECI) is a critical task in Natural Language Processing (NLP) that systematically identifies and validates causal relationships among events in unstructured text, facilitating automated extraction of structured knowledge [Cheng *et al.*, 2025]. Revealing causalities between events in text is of significant importance for advancing automated knowledge extraction and decision-making processes in complex domains, such as military intelligence [Cockburn, 2010], scientific research [Gopalakrishnan *et al.*, 2025], and social development [Radinsky *et al.*, 2012]. In the ECI task, events are represented by their triggers, referred to as "event mentions." Event causality in text is classified into two types according to the distribution of event mentions: intra-sentence causality (or sentence-level causality), in which all related event mentions appear within a single sentence, and inter-sentence causality (or document-level causality), in which event mentions may span multiple sentences.

Early ECI relied on lexical, temporal, and co-occurrence features [Chan *et al.*, 1999; Mirza, 2014; Do *et al.*, 2011]. With deep learning and Pre-trained Language Models (PLMs) [Devlin *et al.*, 2019], performance improved through richer semantic representations [Zhao *et al.*, 2024; Li *et al.*, 2024; Hu *et al.*, 2023; Kadowaki *et al.*, 2019; Liu *et al.*, 2021]. However, these supervised methods require large-scale annotated datasets, which are often inaccessible for domain-specific texts. Large Language Models (LLMs) exhibit strong few-shot and zero-shot capabilities [Kiciman *et al.*, 2024], yet they face significant challenges in ECI tasks, primarily: (1) **Causal hallucination**, where LLMs systematically overestimate causal links due to architectural and pre-training biases [Gao *et al.*, 2023]; and (2) **Underestimation of long-distance causalities**, as cross-sentence event pairs are often overlooked due to lexical distance and sequence modeling constraints. Figure 1 intuitively illustrates the two challenges mentioned above.

We argue that zero-shot ECI should **return to the essence of causality, making judgments based on definitions**. Although predictions from an individual LLM may be unreliable, their reasoning exhibits consistent internal patterns [Wang *et al.*, 2023], suggesting that systematic consensus from multiple perspectives can lead to more reliable causal inference [Cai *et al.*, 2025]. Inspired by this, we propose **LLM Self-Debate (LLM-SD)**. Unlike recent multi-agent methods that deploy different LLMs for the same task, our approach employs **multiple identical LLM agents**. Each agent is assigned a unique, non-overlapping role to evaluate a specific dimension of causality, with the final decision emerging from the adversarial integration of their heterogeneous evidence, not from model diversity. Specifically, the framework formalizes ECI as a structured debate process, consisting of af-

*Corresponding author

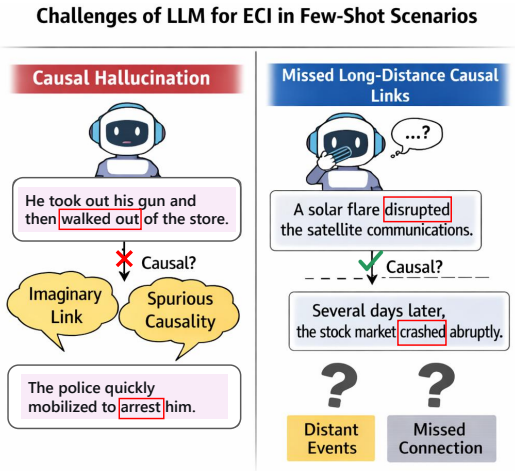

Figure 1: The primary challenges of directly applying LLMs to the ECI task.

firmative and negative teams that collect multi-dimensional evidence, and an adjudication committee that evaluates the evidence to deliver a final verdict. This **role-decomposed design** is key to reducing causal hallucination while improving the recall of long-distance causal relations. It effectively mitigates spurious causal links and false negatives caused by single-evidence errors, as illustrated in Figure 2.

Our contributions are threefold:

- We systematically summarize prerequisite conditions for causal relations and their supporting/rebuttal evidence based on causality definitions, providing a comprehensive argumentative foundation for multi-evidence causal reasoning.
- We innovatively reformulate ECI as an LLM debate, where identical LLMs act as debaters and adjudicators to present, evaluate, and adjudicate causal evidence.
- Extensive experiments on three benchmarks show that LLM-SD achieves state-of-the-art zero-shot performance, with notable robustness in long-text and long-distance ECI tasks.

2 Related Works

2.1 Event Causality Identification

Early research on ECI primarily focused on sentence-level ECI (SECI) [Liu *et al.*, 2023; Liu *et al.*, 2021; Zuo *et al.*, 2020]. However, document-level ECI (DECI) has recently gained significant attention and is an emerging research area. Current DECI strategies include: (1) **Prompt-based fine-tuning** models [Xiang *et al.*, 2023; Yuan *et al.*, 2023], which adapt pre-trained models through innovative prompt engineering. (2) **Event graph reasoning** methods [Liu *et al.*, 2024; Chen *et al.*, 2022], which model causality as graph inference problems using structured event representations. (3) **Contextual encoding** techniques [Man *et al.*, 2024; Huang *et al.*, 2024], which leverage deep learning architectures to capture semantic and contextual features. In the past two years, **LLM-based** zero-shot ECI models have been explored but are limited by causal hallucination. Models such as KnowQA [Wang *et al.*, 2024] and Dr.ECI [Cai *et al.*, 2025] proposed targeted solutions: the former utilizes event argument structures to enhance reasoning, while the latter addresses diverse causal structures by incorporating prior causal knowledge.

However, existing LLM-based ECI models oversimplify causality analysis by neglecting essential causal definitions and supporting/contradictory evidence, leading to persistent hallucination. Our **LLM-SD** mitigates this by simulating evidence-based human reasoning through LLMs' self-consistent causal understanding.

2.2 Multiple LLM Agents Debate

Multiple LLM agents debating has emerged as a promising approach to enhance reasoning and factual accuracy by letting models challenge and refine each other's outputs. [Estornell and Liu, 2025] provide a formal framework for debate dynamics and propose interventions to overcome majority convergence. [Ku *et al.*, 2025] demonstrate that structured two-agent debates can effectively recover implicit premises, outperforming single-agent models. MRBalance[Zou *et al.*, 2025] employs multiple LLM agents in a joint question-answering framework for ECI. However, it mainly aggregates agents that repeatedly reason about causality at a coarse granularity. More broadly, most of current multi-agent LLM paradigms rely on *heterogeneous* agents (different models or differently tuned agents) to solve the *same* task instance, expecting diversity to improve accuracy.

By contrast, **LLM-SD** uses *identical* LLM instances and achieves diversity through *role decomposition*: each agent is constrained to produce a specific type of evidence. Therefore, LLM-SD is not "multiple models doing the same reasoning," but rather "one model performing multiple complementary causal tests," enabling adversarial evidence integration that is directly grounded in the causality definition.

3 Preliminaries

3.1 Problem Formulation

Given an input document $\mathcal{D}$ comprising multiple sentences $\{S_1, S_2, \ldots, S_N\}$, containing a set of annotated event mentions $\{e_i, e_j, \ldots, e_M\}$, ECI seeks to identify all causal event pairs $\{p_1, p_2, \ldots\}$, where each pair $p_l = (e_i, e_j)$ denotes a causal relationship $\mathcal{C}(p_l) \in \{0, 1\}$, where $i, j \in 1, 2, \ldots, M$.

3.2 Event Causality Definition

Grounding in both logical/commonsense causality definitions [Cheng *et al.*, 2025] and established annotation protocols from prior work [Caselli and Vossen, 2017; Wang *et al.*, 2022; Lai *et al.*, 2022], we define event causality as follows:

Definition 1. *For a given context $\mathcal{D}$, causality $e_i \rightarrow e_j$ between two events $e_i, e_j \in \mathcal{D}$ exists if and only if they satisfy **at least one** of the following three conditions:*

- ***Enablement/Prevention**: e_i occurs before e_j and has effect on (enables or prevents) the occurrence of e_j;*
- ***Counterfactual Dependence**: e_j would not occur if e_i did not occur (counterfactual necessity);*
- ***Deterministic Implication**: The occurrence of e_i necessarily entails the occurrence of e_j (logical sufficiency).*

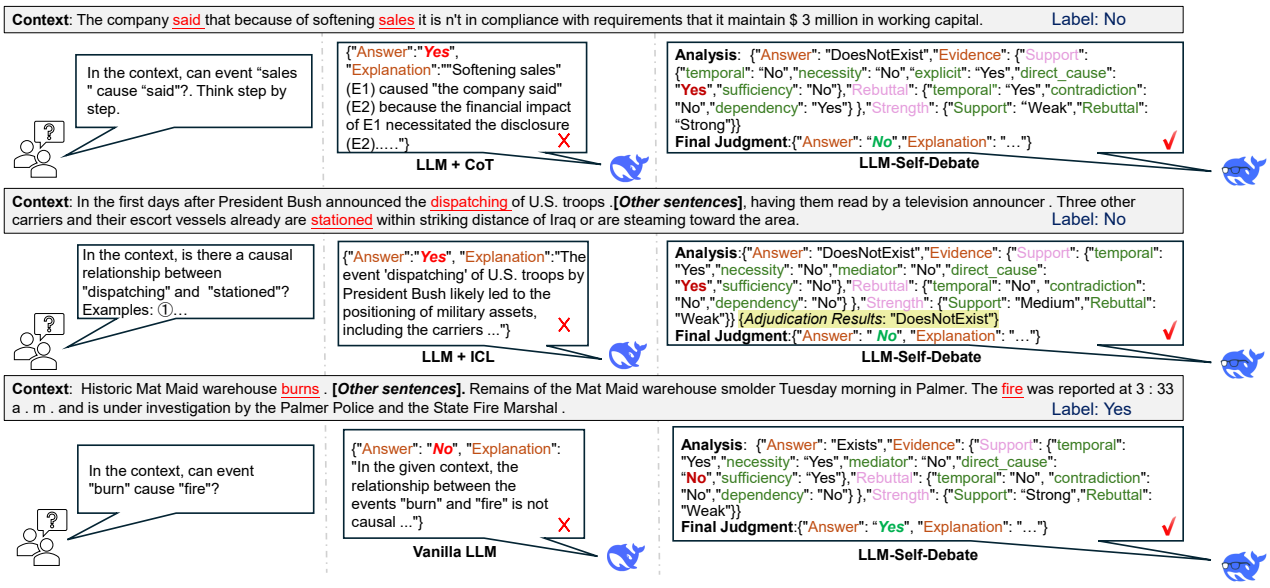

Figure 2: Case studies comparing ECI approaches: (a) simple LLM-based methods (vanilla LLM, LLM + CoT, LLM + ICL) versus (b) LLM-Self-Debate framework, implemented using DeepSeek-Chat as backbone model. Our LLM-SD reduces the spurious causalities led by causal hallucination and mitigates false negatives (with long lexical distance) caused by errors in single-evidence analysis via adversarial evidence integration.

4 Methodology

LLM-SD consists of three modules: Affirmative Team (AT), Negative Team (NT), and Adjudication Committee (AC), each comprising identical LLM agents to collect different evidence or issue a verdict across two distinct reasoning phases (see Figure 3).

4.1 Reasoning Phrases

The first phase is the Evidence Presentation Phase, AT and NT deploy LLM agents to collect supporting or rebuttal evidence for the existence of causality. In this phase, the LLM agents function as "debaters," articulating their positions by leveraging allocated reasoning evidence to support their claims. This step parallels the *Constructive Speech* stage in formal debates, where teams systematically present their core arguments.

The second phase is the Adjudication Phase, which involves evaluating the strength (reliability) of both supporting and rebuttal evidence in the first phase through two adjudication mechanisms: 1) An automatic adjudication method that applies pre-determined rules to cases where the evidence strength demonstrates a clear disparity; and 2) A panel adjudication method where the collected evidence is submitted to AC for contradiction analysis and final verdict. This phase resembles the *Open Argumentation* stage in formal debates, where the debaters present and rebut claims, the adjudication mechanisms systematically weigh opposing evidence. The automatic adjudication handles clear-cut cases efficiently, while the panel adjudication resembles moderated deliberation.

4.2 Debate Teams

Aligned with standard debate formats, LLM-SD comprises two opposing teams: AT (arguing for causality) and the NT (rejecting it). For an ordered query event pair (e_i, e_j), these two teams target for collecting and presenting evidence that supports/rebuts the causality $e_i \rightarrow e_j$ (see Figure 4).

Supporting Evidence of Affirmative Team

Building upon formal definitions of causality, our analysis employs six categories of **supporting** evidence collected through LLM agents: (1) direct causal question answering result E^s_{dir}, (2) temporal ordering validation result E^s_{tem}, (3) necessity testing result via counterfactuals E^s_{nec}, (4) sufficiency analysis result E^s_{suf} (if e_i occurs, e_j must occur), and (5) explicit lexical evidence E^s_{exp} identifying causal markers that linguistically connect the events in context. (6) mediating factor evidence result E^s_{med}: whether there exists a mediator e_m that satisfies $e_i \rightarrow e_m$ and $e_m \rightarrow e_j$. The supporting evidence can be formalized as:

$$\mathcal{E}^s_{intra} = \{E^s_{dir}, E^s_{tem}, E^s_{nec}, E^s_{suf}, E^s_{exp}\}, \quad (1)$$

$$\mathcal{E}^s_{inter} = \{E^s_{dir}, E^s_{tem}, E^s_{nec}, E^s_{suf}, E^s_{med}\}. \quad (2)$$

Explicit lexical evidence E^s_{exp} is used only for intra-sentence event pairs due to the difficulty of accurately identifying causal markers across sentences. Mediating factor evidence E^s_{med} is applied only to inter-sentence event pairs, as mediators are less common within single sentences and prone to confusion in intra-sentence contexts.

Rebuttal Evidence of Negative Team

The opposing debater agents gather the following three categories of **rebuttal** evidence: (1) reverse temporal evidence

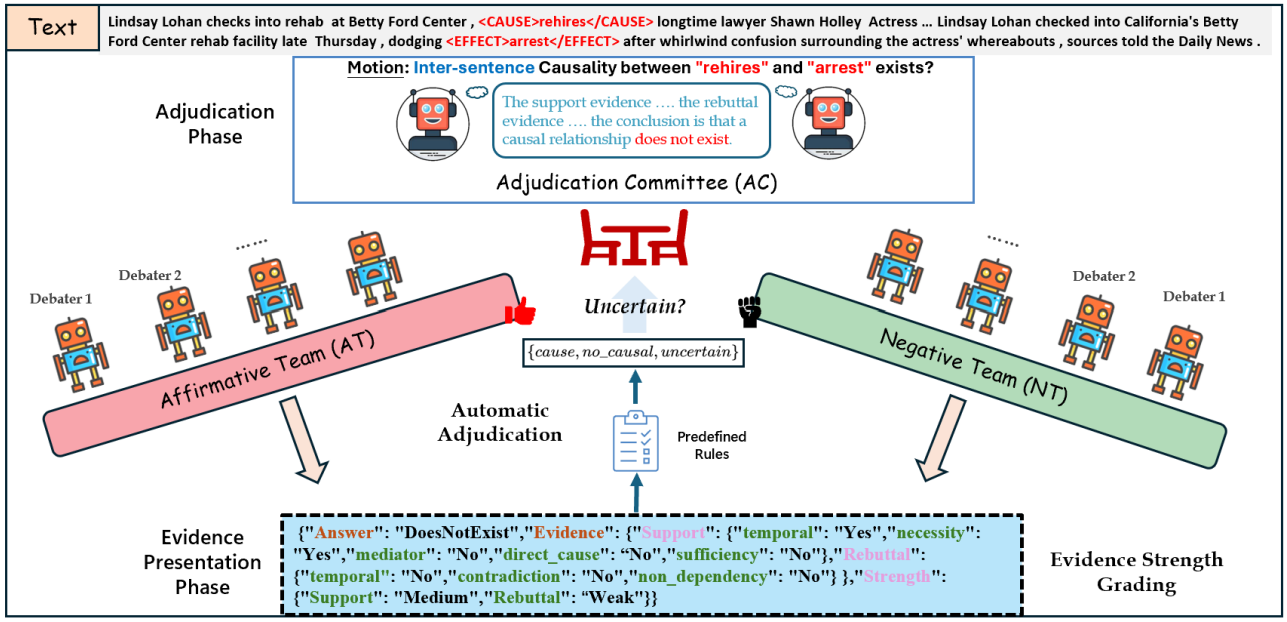

Figure 3: Framework of LLM-Self-Debate. All debate and adjudication agents in the framework are **identical** LLMs.

E^r_{tem}, which demonstrates that the cause event does not precede the effect event; (2) non-dependency evidence E^r_{dep}, indicating that the change of e_i has no effect on e_j; (3) contradictory relation evidence E^r_{con}, where the relationship r_{ab} between the two events contradicts causality, such as coreferential events or inclusive (sub-event) relations. The rebuttal evidence can be formalized as:

$$\mathcal{E}^r = \{E^r_{tem}, E^r_{dep}, E^r_{con}\}. \tag{3}$$

We only selects reverse temporal evidence for rebuttal and does not select the evidence that contradicts necessity and sufficiency, because according to Definition 1, sufficiency and necessity are not prerequisites for the existence of causality. The purpose of non-dependency evidence is to prevent LLMs from engaging in excessive reasoning.

Evidence Strength Grading

As the affirmative and negative teams may present perspectively divergent evidence, we evaluate the strength of both supporting and rebuttal evidence via a predefined rule-based approach. Strength grading rules are designed according to the importance and decisiveness of different evidence types (denoted f_s for support and f_r for rebuttal). Temporal precedence (E^s_{tem}) is treated as a near-necessary constraint and combined with complementary tests—direct causal judgment (E^s_{dir}) and counterfactual necessity (E^s_{nec})—to avoid over-reliance on any single cue. Strong support is assigned only under multi-test agreement accompanied by a decisive connector. On the rebuttal side, higher strength grades require multiple independent refutation cues, enhancing robustness against noisy single-evidence rebuttals.

Table 1 presents the grading rules[1]. These rules are consistent with the fundamental causality logic in Definition 1, and the multi-evidence grading reduces misidentification risks caused by individual misinterpretations and overestimation biases.

[1]Where 1 denotes 'TRUE' and 0 denotes 'FALSE' for the evidence.

Evidence Adjudication

Based on the supporting and rebuttal evidence strengths, $G_s = f_s(\mathcal{E}^s)$ and $G_r = f_r(\mathcal{E}^r)$, we employ two adjudication methods: automatic adjudication and LLM-based (AC) adjudication. The automatic method follows the rules in Table 2. If it cannot reach a conclusion, the evidence is submitted to AC, which then reasons comprehensively to deliver a final verdict. The causality verdict $\mathcal{C}(e_i, e_j)$ thus integrates rule-based (Φ) and LLM-based (Ψ) adjudication via Equation 4.

$$\mathcal{C}(e_i, e_j) = \begin{cases} 1, & \text{if } \Phi(G_s, G_r) = \text{“Exists”}, \\ 0, & \text{if } \Phi(G_s, G_r) = \text{“Not Exist”}, \\ \Psi(\mathcal{E}^s, \mathcal{E}^r), & \text{if } \Phi(G_s, G_r) = \text{“Uncertain”}. \end{cases} \tag{4}$$

Here $\Phi(\cdot)$ denotes leveraging prompt template to guider LLM for causal determination according to support evidence and rebuttal evidence.

5 Experiments

To evaluate the effectiveness and advancement of LLM-SD and verify the rationality of the framework design, we conducted extensive comparative experiments on three benchmark datasets. In our experiments, we primarily investigate the following five research questions (RQs):

- **RQ1**: Can LLM-SD effectively mitigate the excessive false positives caused by causal hallucination?
- **RQ2**: Are all "debaters" (i.e., LLM agents for evidence collection) in the affirmative and opposing teams necessary?

Strength Level	Supporting Evidence Satisfied	Rebuttal Evidence Satisfied
Strong	$E^s_{tem} + E^s_{dir} + E^s_{nec} \geq 2$ AND $(E^s_{exp}/E^s_{med} = 1$ OR $E^s_{suf} = 1)$	$E^r_{tem} + E^r_{con} + E^r_{dep} \geq 2$
Medium	$E^s_{tem} + E^s_{dir} + E^s_{nec} \geq 2$	$E^r_{tem} + E^r_{con} + E^r_{dep} = 1$
Weak	$E^s_{tem} + E^s_{dir} + E^s_{nec} \leq 1$	$E^r_{tem} + E^r_{con} + E^r_{dep} = 0$

Table 1: Supporting and rebuttal evidence strength grading rules.

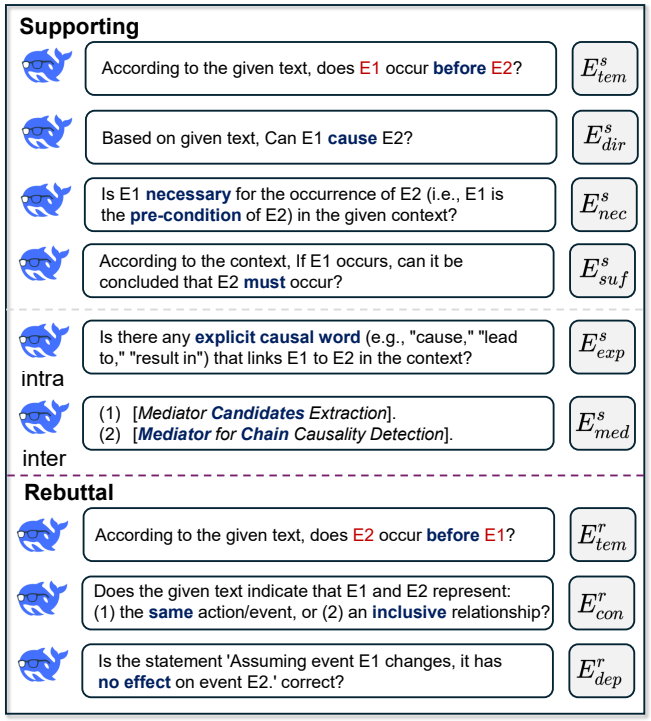

Figure 4: Simplified prompt templates for LLM agents to query supporting/rebuttal evidence. The mediator evidence queries employ a two-stage prompt design.

Supporting	Rebuttal	Conclusion
Strong	Weak	Exists
Strong	Medium	Uncertain
Strong	Strong	Not Exist
Medium	Strong/Medium	Not Exist
Medium	Weak	Uncertain

Table 2: Automatic adjudication rules.

• **RQ3**: Are the two adjudication mechanisms necessary for the model?

• **RQ4**: Are the evidence strength grading rules effective and necessary for the model?

• **RQ5**: How robust is LLM-SD with respect to event distance and text length?

5.1 Experimental Setup

Datasets

We selected three datasets for comparative validation.

• **Causal-TimeBank (CTB)** [Mirza *et al.*, 2014] focuses on explicit causality. It is commonly used for evaluating intra-sentence ECI performance. We conduct 10 independent trials on CTB, each time randomly selecting 10% of the dataset as the test set, and report the average results.

• **EventStoryLine Corpus (ESL)** [Caselli and Vossen, 2017] is designed for cross-document reasoning and narrative generation. We perform 5 independent trials, each excluding 2 topics and taking 20% of the remaining data as the test set, then report the average results.

• **MAVEN-ERE** [Wang *et al.*, 2022] is a large-scale event relation extraction dataset covers coreference, temporal, causal, and subevent relations. Following the original split, we directly use its development set as the test set for evaluation.

Baselines

Given that we focus on model performance in zero-shot scenarios, we employed different reasoning methods based on LLMs as unsupervised baselines. Specifically, we selected five prompt-based reasoning paradigms for comparison:

• **Simple Prompt (SP)** [Brown *et al.*, 2020]: We use a simple prompt to guide the LLM to directly judge and answer whether a causal relationship exists based on the context and event pair.

• **In-Context Learning (ICL)** [Min *et al.*, 2022]: In addition to the simple prompt, we provide positive and negative examples for the LLM to perform in-context learning. Following the findings in [Gao *et al.*, 2023], we selected 2 positive and 4 negative examples as demonstrations.

• **Zero-Shot CoT (ZSCoT)** [Wei *et al.*, 2022]: Based on the simple prompt, we add the instruction "*think step-by-step*" to require the LLM to reason step-by-step.

• **Self-Ask** [Press *et al.*, 2023]: Following the causal judgment logic, we crafted a prompt template to guide the LLM to reason sequentially from perspectives of temporality, sufficiency, and necessity.

• **Definition Prompt**: We formulated the definition of event causality proposed in this paper into a prompt, requiring the LLM to judge based on this definition and select the corresponding causal relation type.

In addition, we compared LLM-SD with two advanced unsupervised LLM-based models, KnowQA [Wang *et al.*, 2024] and Dr.ECI [Cai *et al.*, 2025]. To quantify the gap between unsupervised and supervised models, we also selected multiple supervised models for comparison, including BERT [Devlin *et al.*, 2019], RoBERTa [Zhuang *et al.*, 2021],

Model	CTB Intra			ESL Intra			ESL Inter			ESL Overall		
	P/%	R/%	F1/%	P/%	R/%	F1/%	P/%	R/%	F1/%	P/%	R/%	F1/%
BERT (2019)	47.6	55.1	51.1	47.8	57.2	52.1	36.8	29.2	32.6	41.3	38.3	39.7
RoBERTa (2021)	39.9	60.9	48.2	59.7	38.0	46.4	31.3	34.2	32.7	37.3	35.5	36.4
LongFormer (2020)	29.6	68.6	41.4	59.0	40.5	48.0	35.2	30.5	32.7	41.6	33.8	37.3
ERGO (2022)	62.1	61.3	61.7	57.5	72.0	63.9	51.6	43.3	47.1	48.6	53.4	50.9
KADE (2023)	56.8	70.6	66.7	61.5	73.2	66.8	51.2	74.2	60.5	51.9	70.6	59.8
Unsupervised												
davinci-002 (2023)	5.0	75.2	9.3	23.2	80.0	36.0	-	-	-	-	-	-
davinci-003 (2023)	8.5	64.4	15.0	33.2	74.4	45.9	-	-	-	-	-	-
GPT-3.5 (2023)	7.0	82.6	12.8	27.6	80.2	41.0	-	-	-	-	-	-
GPT-4 (2023)	6.1	**97.4**	11.5	27.2	94.7	42.2	-	-	-	-	-	-
Dr. ECI (2025)	11.7	71.8	13.0	29.0	75.4	41.9	-	-	-	-	-	-
gpt-4o-mini												
SP (2020)	11.4	65.8	13.5	29.1	62.7	31.8	2.9	34.7	6.2	10.6	44.3	18.5
ICL (2022)	10.7	67.7	14.4	34.6	60.4	35.3	3.6	30.6	6.9	12.5	46.9	20.7
ZSCoT (2022)	9.2	74.9	13.1	26.0	80.7	37.0	4.2	48.0	8.6	8.1	62.8	18.3
Self-Ask (2023)	11.3	59.8	16.4	21.2	74.4	38.3	4.9	45.1	9.5	8.7	58.5	18.2
Definition (ours)	8.8	79.5	12.7	22.5	83.1	40.2	4.5	39.3	8.6	9.2	60.5	17.3
LLM-SD	16.3	81.5	23.7	39.6	75.5	41.5	11.1	74.2	14.4	21.2	74.7	28.5
improved	*4.9*	*6.6*	*7.3*	*5.0*	*2.2*	*6.3*	*6.2*	*26.2*	*4.9*	*8.7*	*11.9*	*7.8*
deepseek-chat												
SP (2020)	11.3	90.6	16.0	28.4	91.6	40.2	9.7	37.7	13.7	21.9	66.6	30.7
ICL (2022)	12.2	75.0	16.4	29.3	77.1	42.6	8.8	39.4	12.8	22.7	73.9	30.6
ZSCoT (2022)	9.6	91.7	13.8	27.0	90.0	37.2	8.4	53.1	12.6	16.2	78.9	25.3
Self-Ask (2023)	12.9	89.2	17.3	34.3	87.6	42.2	7.1	56.1	11.8	16.0	77.6	27.2
Definition (ours)	9.7	96.9	16.3	26.0	**94.0**	40.5	8.0	41.5	11.7	19.1	66.3	27.6
LLM-SD	**15.5**	85.0	**26.2**	**42.4**	87.3	**49.1**	**18.3**	**86.4**	**28.0**	**24.9**	**87.1**	**35.7**
improved	*2.6*	-	*8.9*	*8.1*	-	*6.5*	*8.6*	*30.3*	*14.3*	*2.2*	*8.2*	*5.0*

Table 3: Experimental results of LLM-SD and baselines on CTB and ESL. In the table, **bold** values indicate the best performance among unsupervised models. Underlined values indicate the best overall performance among all models.

Model	Intra			Inter		
	P/%	R/%	F1/%	P/%	R/%	F1/%
Supervised						
BERT (2019)	46.8	50.3	48.5	43.0	46.8	44.8
RoBERTa (2021)	45.9	48.8	47.2	42.3	45.1	43.6
LongFormer (2020)	39.3	55.9	44.6	35.8	46.0	39.2
ERGO (2022)	63.1	65.3	64.2	48.7	62.0	54.6
KADE (2023)	58.7	64.4	60.0	43.7	59.3	48.0
iLIF (2024)	74.4	51.5	60.9	67.1	49.2	56.8
Unsupervised						
davinci-002 (2023)	19.6	92.9	32.4	-	-	-
davinci-003 (2023)	25.0	75.1	37.5	-	-	-
GPT-3.5 (2023)	19.9	85.8	32.3	-	-	-
GPT-4 (2023)	22.5	92.4	36.2	-	-	-
KnowQA (2024)	29.2	62.1	39.7	-	-	-
Dr. ECI (2025)	22.0	80.0	34.5	-	-	-
deepseek-chat						
Self-Ask (2023)	22.5	**87.7**	35.3	22.6	53.4	32.6
Definition (ours)	23.0	87.5	31.0	24.5	65.8	34.4
LLM-SD	**34.2**	61.2	**43.9**	**27.7**	**69.2**	**38.2**
improved	*11.7*	-	*4.2*	*3.2*	*3.4*	*3.8*

Table 4: Experimental results of LLM-SD and baselines on MAVEN-ERE.

Longformer [Beltagy *et al.*, 2020], KADE [Wu *et al.*, 2023], ERGO [Chen *et al.*, 2022] and iLIF [Liu *et al.*, 2024].

Evaluation Metrics

Consistent with the previous works, we use Precision (P), Recall (R), and F1-score (F1) to evaluate model performance.

Implementation Details

We selected two advanced LLMs as the backbone models: GPT-4o-mini[2] and Deepseek-V3 (DeepSeek-Chat)[3] for main experiment. In Section 5.6, we compared the performance of LLM-SD on different LLMs.

In experiments, we utilized Python's concurrent. futures.ThreadPoolExecutor[4] to implement multi-threaded parallel calls to all LLM agents, thereby saving inference time. We set $temperature = 0$ to reduce the randomness of generated outputs while maintaining logical consistency.

5.2 Main Results (RQ1)

Tables 3–4 show that LLM-SD achieves state-of-the-art performance among unsupervised models across all benchmarks. It significantly improves precision by reducing false positives from causal hallucination and enhances recall on inter-sentence ECI through multi-evidence reasoning. Compared to supervised baselines, LLM-SD remains competi-

[2]https://openai.com/api/

[3]https://api.deepseek.com/v1

[4]https://docs.python.org/3/library/concurrent.futures.html

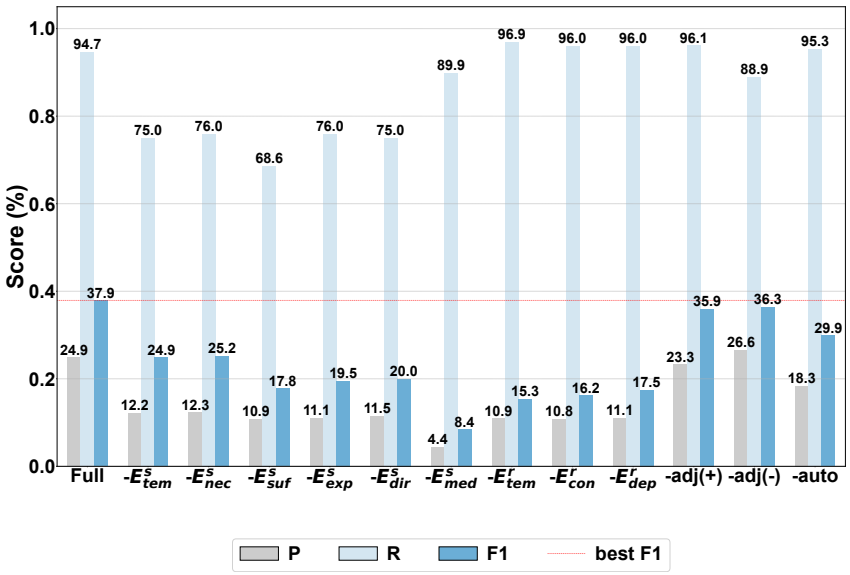

Figure 5: Results of ablation study on ESL.

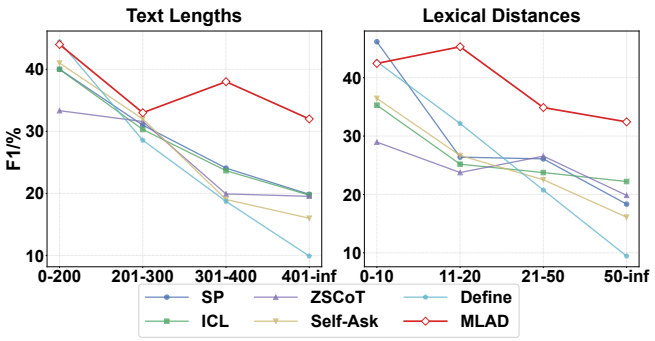

Figure 6: Performance of LLM-SD and baseline models as a function of event distance and document length.

Grading Strategy	Intra-sentence			Inter-sentence		
	P/%	R/%	F1/%	P/%	R/%	F1/%
Count-based	18.0	70.0	28.6	11.8	20.3	15.0
Simple-voting ($k = 1$)	22.3	56.4	36.4	12.8	28.2	17.6
Simple-voting ($k = 2$)	30.5	43.6	37.0	15.0	20.5	18.5
Simple-voting ($k = 3$)	41.0	44.4	42.7	17.8	18.0	17.9
Simple-voting ($k = 4$)	42.7	20.5	32.0	18.3	15.4	16.0
Ours	**42.4**	87.3	**49.1**	**18.3**	**86.4**	**28.0**

Table 5: Validation of evidence strength grading rules on ESL.

tive and even outperforms several supervised models on ESL, highlighting its strong potential. Simpler strategies like Self-Ask or providing the Definition yield only marginal gains, confirming that explicit instructions alone cannot fix LLMs' consistency errors in causal reasoning. The consistent results across different LLMs (GPT-4o-mini and DeepSeek-Chat) further validate the robustness of our approach.

5.3 Ablation Studies (RQ2-RQ3)

This section presents ablation studies from three perspectives to verify the effectiveness of each evidence type and module. All experiments were conducted on a test subset of the ESL dataset, with LLM parameters kept identical to the main experiments. The ablation experiments results are shown in Figure 5.

Evidence Ablation To systematically evaluate the independent contribution and indispensability of each type of evidence, an ablation study on evidence was conducted. Specifically, according to the rules in Table 1, when a specific evidence-collection agent is ablated, the corresponding term is excluded from the evidence aggregation formula (with the threshold kept unchanged), and the assessment outcome of causal relations is re-determined. The experimental results in Figure 5 demonstrate that removing any supportive or rebuttal evidence leads to a decline in overall model performance, indicating that all the aforementioned evidence is essential for the framework.

LLM Adjudication Mechanism Ablation We further ablated the LLM-based adjudication mechanism (`AC`) by forcing uncertain cases to be judged as either always non-existent (-adj(-)) or always existent (-adj(+)). Both ablation settings lead to performance degradation, confirming that the LLM adjudicator is an indispensable component of the framework. This indicates that while our rule-based automatic adjudication is logically grounded, it benefits from the flexible judgment of an LLM to handle nuanced and complex scenarios that pre-defined rules may not fully cover. It also compensates for potential errors arise from the subjectivity of rules.

Automatic Adjudication Mechanism Ablation We removed the predefined adjudication rules and relied solely on the LLM adjudicator (`AC`) to judge all cases. As shown in Figure 5, this leads to a significant performance drop, especially in precision. This result demonstrates that relying exclusively on LLM judgment is unreliable; without structured

rules to guide the synthesis of multiple evidences, the LLM tends to make simplistic decisions based on evidence counts, which increases false positives and confirms the necessity of our hybrid adjudication design.

5.4 Validating the Evidence Grading Rules (RQ4)

To further examine whether the evidence strength grading rules in Table 1 are reasonable and not overly tailored, we keep the debater agents and evidence extraction prompts unchanged, and only replace the mapping from evidences to (G_s, G_r) and the subsequent adjudication. We consider the following alternatives:

(1) **Count-based grading**, which assigns strength levels based solely on the count of satisfied evidences:

$$G_s = \begin{cases} \text{“Strong”}, & \text{if } \sum \mathcal{E}^s \geq 4, \\ \text{“Medium”}, & \text{if } 2 \leq \sum \mathcal{E}^s \leq 3, \\ \text{“Weak”}, & \text{otherwise,} \end{cases} \quad G_r = \begin{cases} \text{“Strong”}, & \text{if } \sum \mathcal{E}^r = 3, \\ \text{“Medium”}, & \text{if } 1 \leq \sum \mathcal{E}^r \leq 2, \\ \text{“Weak”}, & \text{otherwise.} \end{cases} \tag{5}$$

(2) **Simple-voting grading**, which determines causality directly by comparing the counts of supporting and rebuttal evidences with a threshold k:

$$\mathcal{C}(e_i, e_j) = \begin{cases} 1, & \text{if } \sum \mathcal{E}^r = 0 \text{ and } \sum \mathcal{E}^s \geq k \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

In this method, k serves as a hyperparameter representing the minimum number of supporting evidences required when no rebuttal evidence is present.

Table 5 shows that our evidence grading rules significantly outperform the other two alternatives, confirming their effectiveness. The count-based method performs worst, as merely tallying evidence amplifies the LLM's systematic errors and yields low precision. Simple-voting grading strongly depends on threshold k, with precision rising and recall falling as k increases, making it difficult to achieve an effective balance.

5.5 Impact of Lexical Distance and Text Length (RQ5)

Prior research suggests that LLM-based ECI performance degrades with increasing document length and event distance. To assess LLM-SD's robustness, we analyzed its performance across different lexical distance and document length intervals on CTB and ESL datasets (Figure 6).

The results confirm that all unsupervised baselines suffer significant performance drops as distance and length increase, with the Definition baseline showing a near-linear decline. In contrast, LLM-SD exhibits strong resilience, maintaining stable performance (F1 > 0.3) with only minor degradation. This demonstrates LLM-SD's superior robustness in handling long-range dependencies and confirms the effectiveness of its multi-evidence reasoning approach for complex ECI tasks.

5.6 Evaluating LLM-SD with Different LLM Architectures

To assess the sensitivity and practical deployability of the LLM-SD framework across different LLM architectures, we conduct additional experiments on the ESL dataset using the same zero-shot prompts, identical evidence role assignments, and consistent decoding configurations (temperature $= 0$). We evaluate several widely adopted LLMs with distinct architectures, including **LLaMA3-8B**, **GPT-3.5**, **Qwen3-max**, and **GLM4-32B**, and compare their performance with **DeepSeek-Chat** (our backbone).

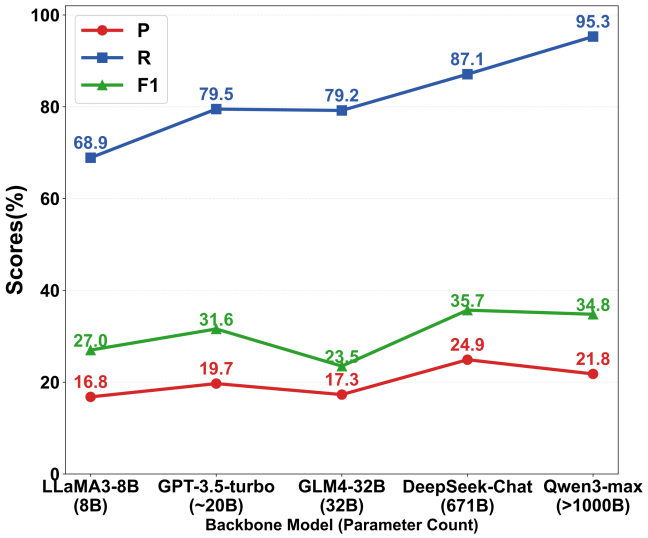

Figure 7: Performance comparison of different backbone models on the ESL dataset (overall metrics).

Figure 7 reveals two key findings. First, all LLM-SD variants outperform unsupervised baselines using the same corresponding LLM (overall F1 < 0.2), confirming that its role-decomposition framework structurally reduces causal hallucination and enhances performance independently of the backbone model. Second, performance does not scale monotonically with parameter count: DeepSeek-Chat performs best, while the larger Qwen3-max scores lower; the smaller GPT-3.5-turbo outperforms the larger GLM4-32B. This indicates that ECI capability depends on factors beyond scale, such as pre-training data, architecture, and specialized tuning. The backbone model's inductive biases and training alignment are also crucial for effective deployment within LLM-SD.

6 Conclusion

We propose LLM-SD, a framework that casts ECI as a structured debate among specialized LLM agents. LLM-SD employs Affirmative and Negative Teams to collect supporting and rebuttal evidence, evaluated by an Adjudication Committee using rule-based and LLM-driven methods. Experiments on three benchmarks show state-of-the-art zero-shot performance, with improved precision and robust recall. However, the reliance on multiple LLM invocations raises cost, and performance on inter-sentence causality remains below that of supervised models. Future research could explore fine-tuning strategies or alignment mechanisms that constrain LLM outputs to further mitigate hallucinations, enhance evidence consistency, and improve multi-hop reasoning capabilities.

Acknowledgments

Zefan Zeng and Yuehang Si contributed equally to this work.

References

[Beltagy *et al.*, 2020] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

[Brown *et al.*, 2020] Tom B. Brown, Benjamin Mann, Nick Ryder, et al. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2020.

[Cai *et al.*, 2025] Ruichu Cai, Shengyin Yu, Jiahao Zhang, Wei Chen, Boyan Xu, and Keli Zhang. Dr.ECI: Infusing large language models with causal knowledge for decomposed reasoning in event causality identification. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9346–9375, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics.

[Caselli and Vossen, 2017] Tommaso Caselli and Piek Vossen. 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, August 2017.

[Chan *et al.*, 1999] Syin Chan, Yun Niu, Alyssa Ang, and Christopher S. G. Khoo. A method for extracting causal knowledge from textual databases. *Singapore Journal of Library & Information Management*, pages 48–63, 1999.

[Chen *et al.*, 2022] Meiqi Chen, Yixin Cao, Kunquan Deng, Mukai Li, Kun Wang, Jing Shao, and Yan Zhang. ERGO: Event relational graph transformer for document-level event causality identification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2118–2128, October 2022.

[Cheng *et al.*, 2025] Qing Cheng, Zefan Zeng, Xingchen Hu, Yuehang Si, and Zhong Liu. A survey of event causality identification: Taxonomy, challenges, assessment, and prospects. *ACM Computing Surveys*, 58(3), 2025.

[Cockburn, 2010] Cynthia Cockburn. Gender relations as causal in militarization and war. *International Feminist Journal of Politics*, 12(2):139–157, 2010.

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

[Do *et al.*, 2011] Quang Do, Yee Seng Chan, and Dan Roth. Minimally supervised event causality identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 294–303, July 2011.

[Estornell and Liu, 2025] Andrew Estornell and Yang Liu. Multi-llm debate: framework, principals, and interventions. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, 2025.

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

[Gopalakrishnan *et al.*, 2025] Seethalakshmi Gopalakrishnan, Luciana Garbayo, and Wlodek Zadrozny. Causality extraction from medical text using large language models (llms). *Information*, 16(1), 2025.

[Hu *et al.*, 2023] Zhilei Hu, Zixuan Li, Xiaolong Jin, Long Bai, Saiping Guan, Jiafeng Guo, and Xueqi Cheng. Semantic structure enhanced event causality identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10901–10913, July 2023.

[Huang *et al.*, 2024] Peixin Huang, Xiang Zhao, Minghao Hu, Zhen Tan, and Weidong Xiao. Distill, fuse, pretrain: Towards effective event causality identification with commonsense-aware pre-trained model. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5029–5040, May 2024.

[Kadowaki *et al.*, 2019] Kazuma Kadowaki, Ryu Iida, Kentaro Torisawa, Jong-Hoon Oh, and Julien Kloetzer. Event causality recognition exploiting multiple annotators' judgments and background knowledge. 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 5816–5822, November 2019.

[Kiciman *et al.*, 2024] Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research*, 2024.

[Ku *et al.*, 2025] Harvey Bonmu Ku, Jeongyeol Shin, Hyoun Jun Lee, Seonok Na, and Insu Jeon. Multi-agent LLM debate unveils the premise left unsaid. In *Proceedings of the 12th Argument mining Workshop*, pages 58–73, July 2025.

[Lai *et al.*, 2022] Viet Dac Lai, Amir Pouran Ben Veyseh, Minh Van Nguyen, Franck Dernoncourt, and Thien Huu Nguyen. MECI: A multilingual dataset for event causality identification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2346–2356, October 2022.

[Li *et al.*, 2024] Haoran Li, Qiang Gao, Hongmei Wu, and Li Huang. Advancing event causality identification via heuristic semantic dependency inquiry network. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1467–1478, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[Liu *et al.*, 2021] Jian Liu, Yubo Chen, and Jun Zhao. Knowledge enhanced event causality identification with mention masking generalizations. In *Proceedings of the*

Twenty-Ninth International Joint Conference on Artificial Intelligence, 2021.

[Liu *et al.*, 2023] Jintao Liu, Zequn Zhang, Zhi Guo, Li Jin, Xiaoyu Li, Kaiwen Wei, and Xian Sun. KEPT: Knowledge Enhanced Prompt Tuning for event causality identification. *Knowledge-Based Systems*, 259:110064, January 2023.

[Liu *et al.*, 2024] Cheng Liu, Wei Xiang, and Bang Wang. Identifying while learning for document event causality identification. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3815–3827, August 2024.

[Man *et al.*, 2024] Hieu Man, Franck Dernoncourt, and Thien Huu Nguyen. Mastering Context-to-Label Representation Transformation for Event Causality Identification with Diffusion Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):18760–18768, March 2024. Number: 17.

[Min *et al.*, 2022] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, 2022.

[Mirza *et al.*, 2014] Paramita Mirza, Rachele Sprugnoli, Sara Tonelli, and Manuela Speranza. Annotating causality in the TempEval-3 corpus. In *Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL)*, pages 10–19, April 2014.

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

[Press *et al.*, 2023] Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711, 2023.

[Radinsky *et al.*, 2012] Kira Radinsky, Sagie Davidovich, and Shaul Markovitch. Learning causality for news events prediction. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, page 909–918, New York, NY, USA, 2012. Association for Computing Machinery.

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

[Wang *et al.*, 2023] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

[Wang *et al.*, 2024] Zimu Wang, Lei Xia, Wei Wang, and Xinya Du. Document-level causal relation extraction with knowledge-guided binary question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16944–16955, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[Wei *et al.*, 2022] Jason Wei, Xuezhi Wang, Dale Schuurmans, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837, 2022.

[Wu *et al.*, 2023] Sifan Wu, Ruihui Zhao, Yefeng Zheng, Jian Pei, and Bang Liu. Identify Event Causality with Knowledge and Analogy. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):13745–13753, June 2023. Number: 11.

[Xiang *et al.*, 2023] Wei Xiang, Chuanhong Zhan, and Bang Wang. Daprompt: Deterministic assumption prompt learning for event causality identification. *CoRR*, abs/2307.09813, 2023.

[Yuan *et al.*, 2023] Changsen Yuan, Heyan Huang, Yixin Cao, and Yonggang Wen. Discriminative reasoning with sparse event representation for document-level event-event relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16222–16234, July 2023.

[Zhao *et al.*, 2024] Xiaoyan Zhao, Yang Deng, Min Yang, Lingzhi Wang, Rui Zhang, Hong Cheng, Wai Lam, Ying Shen, and Ruifeng Xu. A comprehensive survey on relation extraction: Recent advances and new frontiers. *ACM Computing Surveys*, 56(11), July 2024.

[Zhuang *et al.*, 2021] Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, August 2021.

[Zou *et al.*, 2025] Xiang Zou, Xuanhong Li, Po Hu, and Ming Dong. Mrbalance: A framework for enhancing event causality identification in multi-agent debates via role assignment. *Knowledge-Based Systems*, 330:114470, 2025.

[Zuo *et al.*, 2020] Xinyu Zuo, Yubo Chen, Kang Liu, and Jun Zhao. 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, December 2020.