Hypothetical-Deductive Reasoning for Event Causality Identification

Anonymous ACL submission

Abstract

Event Causality Identification (ECI) is the task of identifying causal relations between two events. Most works mainly enhance event encoding with pre-trained language models (PLMs), often neglecting the implicit and long-text reasoning capabilities needed for ECI tasks. Large language models (LLMs) have recently revealed substantial reasoning potential through chain-of-thought (CoT). Inspired by Pearl's Causal Hierarchy, we first introduce CoT into the ECI task and propose Causal Progressive Reasoning CoT (CPR). CPR uses a progressive reasoning approach, guiding the model step by step to explore the causal relation between two events. More importantly, we find that CoT may generate incorrect intermediate steps that propagate to the next ones, leading to error results. To deal with this problem, we propose a Hypothetical-Deductive Reasoning framework (HYDRO). HYDRO is based on hypothetical-deductive reasoning, where each step is independently reasoned. Extensive experiments have demonstrated that our methods achieve state-of-the-art performance (17.8% and 6.8% F1 score gains on EventStoryLine and Causal-TimeBank) on two benchmark datasets. Additionally, it exhibits significant advantages only using Flan-T5-Base (250M) in zero-shot settings.

1 Introduction

Event Causality Identification (ECI) aims to determine whether a causal relation exists between two events. For example, in Figure 1, event *tornadoes* cause event *declaration*. The ECI model needs to identify such causal relations, which is beneficial for various NLP applications such as question answering (Sui et al., 2022; Shi et al., 2021) and future event prediction (Mathur et al., 2024).

Existing ECI research can be categorized into two types: sentence-level ECI (**SECI**) (Liu et al., 2021) aims to identify causal relations between two events within a single sentence, and Document-level ECI (**DECI**) (Phu and Nguyen, 2021) aims to identify causal relations between events across the entire document. Existing works have focused on enhancing encoder performance. For example, (Shen et al., 2022) employed joint learning to supervise the quality of event representations, but this requires additional annotated information. (Wu et al., 2023) utilized ConceptNet for event analogy to enhance ECI performance, achieving the current state-of-the-art (SOTA) results.

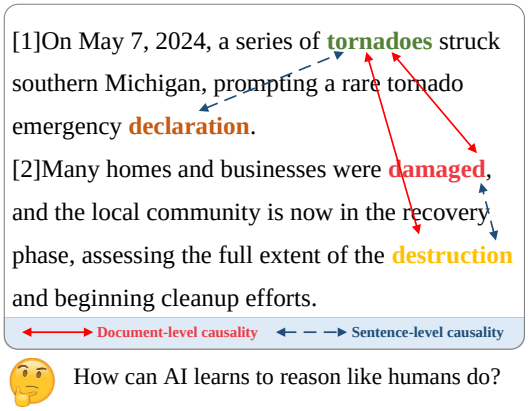

Figure 1: An example of ECI. Each double arrow indicates a causal relation between two events.

However, these works are based on encode-style models, which primarily rely on external knowledge to enhance event encoding quality, and their performance remains at a similar level. This indicates that encode-style models have reached a performance bottleneck in ECI tasks. The main reasons limiting the improvement of ECI performance are as follows: Firstly, in the ECI task, events often lack explicit causal clue words (e.g., "causes", "results in"). For example, in Figure 1, event *tornadoes* cause event *damaged*, but there are no direct clue words between these two events. This challenges the model's implicit reasoning ability. Secondly, in DECI, documents are much longer,

requiring the model to have robust long-text reasoning capabilities. While improving the quality of event representations through prompts can benefit ECI performance, ECI requires more implicit and long-text reasoning capabilities. Fortunately, large language models (LLMs) with rich embedded knowledge for chain reasoning have revealed the significant reasoning potential (Wang et al., 2022; Zhang et al., 2023), providing a new paradigm for solving reasoning problems.

In this paper, we innovatively introduce the CoT into the ECI task. Based on the first level of Pearl Causal Hierarchy (PCH) theory (Pearl, 2001), we propose the Causal Progressive Reasoning CoT framework (**CPR**). Technically, we design a four-hop reasoning framework where each subsequent step of reasoning is based on the answer from the previous step. This progressive reasoning breaks down complex causal inference into multiple smaller questions, guiding the model to explore implicit causality. Meanwhile, it can identify the causality only relying on the answer of reasoning, which can effectively reduce the length of the text.

More importantly, we find incorrect intermediate reasoning steps may occur during the CoT's reasoning process and propagate to the next step, leading to erroneous results. To address this issue, we propose a Hypothetical-Deductive Reasoning (**HYDRO**) framework for ECI. Unlike CoT, HYDRO performs each step independently without relying on the previous step's answer. Technically, HYDRO employs a two-stage reasoning framework and incorporates hypothetical-deductive reasoning. In the first stage, we propose three hypotheses about causal relations between events, and HYDRO judges whether these causal hypotheses hold. The second stage considers these answers in the first stage to make a final judgment based on the principles of hypothetical-deductive reasoning: if any hypothesis fails, there is no causal relation between events A and B. This approach significantly reduces dependency on the previous step's answers.

To supervise the correctness of the model's hypothetical reasoning, we introduce supervised reasoning correction to more rigorously supervise the answers of the first stage in hypothetical-deductive reasoning. During training, these training gold labels continuously adjust the model, correcting each hop's hypothesis to produce more accurate reasoning.

In summary, our contributions can be summarized as follows:

- We introduce the Chain-of-Thought(CoT) into the ECI task and propose a Causal Progressive Reasoning CoT (**CPR**) for the ECI task, enabling progressive reasoning to uncover causal relations between events. To the best of our knowledge, we are the first to introduce the causal Chain-of-Thought into the ECI task.

- More importantly, to address error propagation in CoT, we further propose a new Hypothetical Deductive Reasoning (**HYDRO**) framework, which is different from CoT and prevents error propagation. The HYDRO is a completely new reasoning framework different from CoT in that each step of reasoning is independent.

- Experimental results show that the HYDRO achieves an F1 score improvement of 20.5% on EventStoryLine and 6.8% on Causal-TimeBank. In zero-shot settings, compared to ChatGPT, our approach outperforms by 16.1% on EventStoryLine and 4.6% on Causal-TimeBank in F1 scores only using the Flan-T5-Base with 250 million parameters.

2 Methodology

Given a document D and a set of events E, SECI aims to identify causal relations between two events within a single sentence. DECI aims to predict whether there is a causal relation between events e_i and e_j mentioned in different sentences within the document. As shown in Figure 2 and Figure 3, we respectively illustrate the Causal Progressive Reasoning CoT (CPR) and the Hypothetical-Deductive Reasoning CoT (HYDRO). CPR adopts a four-step reasoning framework, using a progressive thinking approach where each reasoning step builds upon the previous step. In contrast, HYDRO employs a two-stage reasoning framework, incorporating hypothetical-deductive reasoning to derive the final answer.

2.1 Causal Progressive Reasoning CoT (CPR)

We construct CPR based on the first level of Pearl Causal Hierarchy (PCH) (Pearl, 2001), focusing on event correlations. At this level, the subject passively observes the world to identify patterns without intervening. A typical question at this level is: "With many dark clouds in the sky, what is the probability of rain?" CPR requires the model to

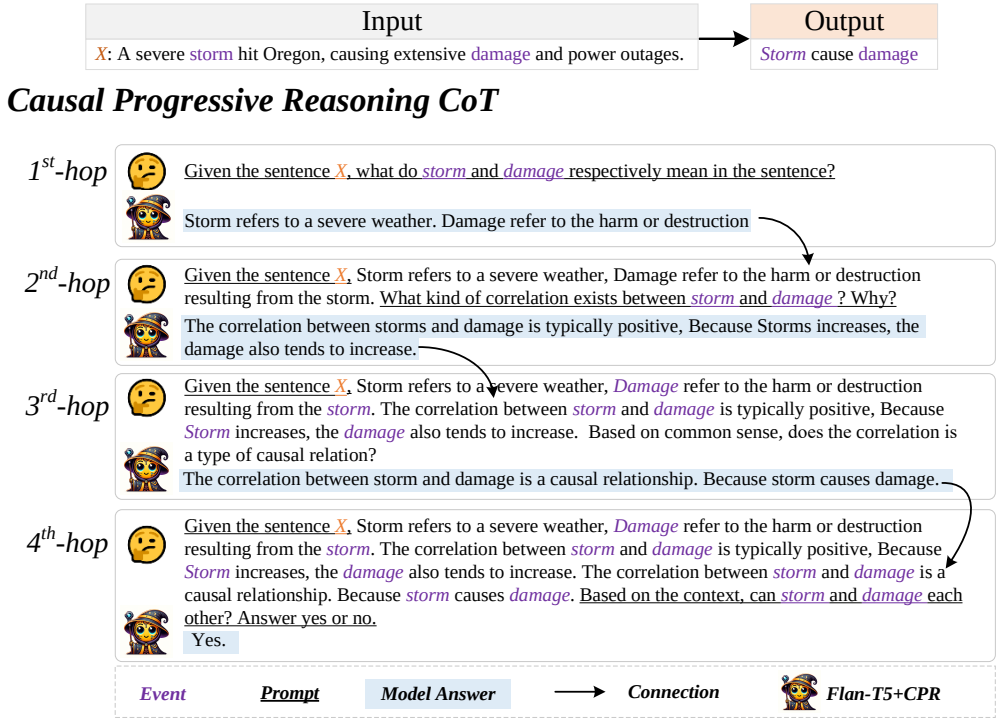

Figure 2: An illustration of Causal Progressive Reasoning CoT for ECI task. Dashed rectangular boxes represent the graphic symbols.

reason about the correlation between events, making judgments about the causal relationships of events based on the model's understanding of the events themselves. Based on the first level of event correlations, we propose the Causal Progressive Reasoning CoT (CPR) for ECI (Figure 2). Given the context prompt T**: Given the sentence.**

1st**-Hop:** our first-hop is to prompt the LLM M to consider what the two events are, we ask the M using the Prompt T_1:

T_1**:** T**. What do** e_i **and** e_j **respectively mean in the sentence?**

This step can be formally expressed as: $A = M(T_1)$, where A denotes the model's explanations of e_i and e_j.

2nd**-Hop:** The second-hop, based on the answer A generated by the LLM's understanding of e_i and e_j, we ask the LLM to continue answering the correlation between e_i and e_j and provide the corresponding explanation. The Prompt T_2 is as follows:

T_2**: [**T, A**]. What kind of correlation exists between** e_i **and** e_j**? Why?**

Here, [ ] represents the concatenation of context. Similarly, after feeding T_2 to the LLM, we obtain response B which represents the LLM's answer and explanation regarding the correlation between e_i and e_j.

3rd**-Hop:** The third-hop involves asking the LLM to determine whether the correlation between e_i and e_j constitutes a causal relation. The Prompt T_3 is as follows:

T_3**: [**T, A, B**]. Based on common sense, does the correlation is a type of causal relation?**

We feed T_3 to the LLM, obtaining answer C. C represents the LLM's answer regarding whether the correlation is a causal relation.

4th**-Hop:** Based on the previous three steps of progressive reasoning, the fourth-hop is about asking the LLM to make a final judgment on whether e_i and e_j have a causal relation, based on the previous three steps of progressive reasoning. The Prompt T_4 is as follows:

T_4**: [**T, A, B, C**]. Based on the context, can** e_i **and** e_j **cause each other? Answer yes or no.**

This can be expressed with the formula:

$$\hat{y} = M(y|T, A, B, C), \quad (1)$$

where y represents golden labels, $\hat{y}$ represents the model's predicted answer.

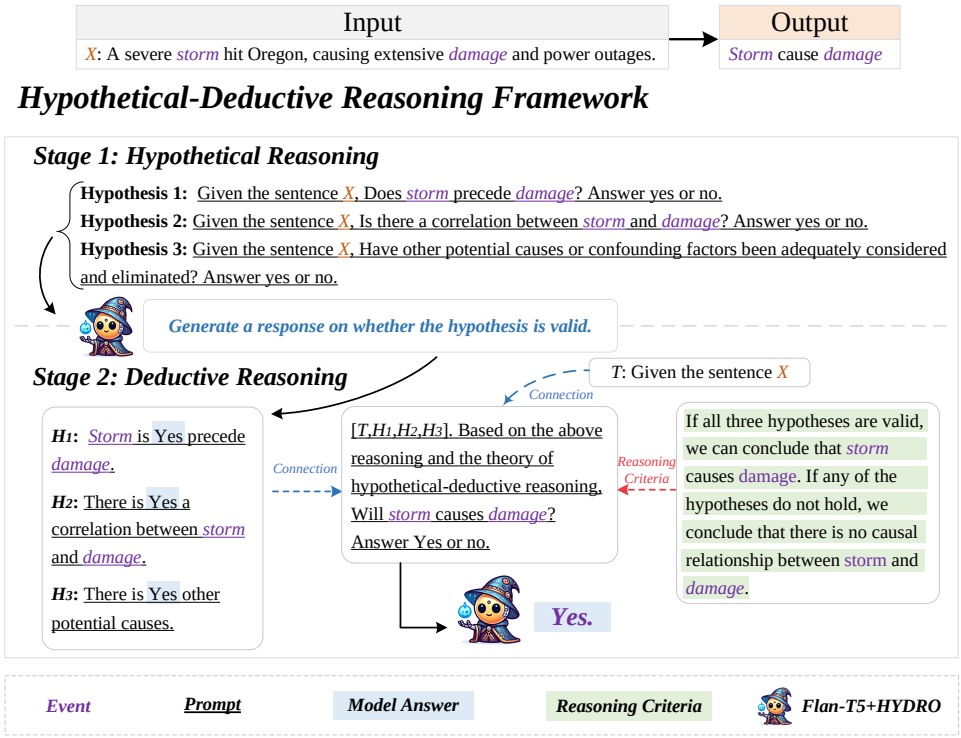

Figure 3: An illustration of Hypothetical-Deductive Reasoning framework for ECI task. Hydro means "water" in the Greek language. Dashed rectangular boxes represent the graphic symbols.

2.2 Hypothetical-Deductive Reasoning framework

While CPR can enhance performance, the errors in intermediate reasoning steps can propagate to subsequent steps, leading to cumulative mistakes and ultimately incorrect predictions. To address this issue, we propose HYDRO. The hypothesis construction for HYDRO is based on the third level of Judea Pearl's Causal Hierarchy(Pearl, 2001), which deals with counterfactuals. A typical question at this level is: "What would have happened if I had...?" This involves comparing the observed world with a counterfactual world and assessing the feasibility of proposed hypotheses. Based on this, we propose three hypotheses to verify causal relationships between events. These hypotheses may be further refined, but adding more would require more computational resources and time. As illustrated in Figure 3, HYDRO combines hypothetical-deductive reasoning with a two-stage reasoning framework. Specifically, the two-stage reasoning framework operates as follows:

2.2.1 Stage 1: Hypothetical Reasoning

We establish three causal hypotheses to determine whether there is a causal relation between events. First, we provide the reasoning context prompt T: **Given the sentence X.**

Hypothesis 1: We hypothesize that e_i occurs before e_j. If there is a causal relation between the two events, there will necessarily be a temporal order.

Hypothesis 2: We hypothesize that there is a correlation between e_i and e_j. If the events are causally related, they must be correlated.

Hypothesis 3: We hypothesize that when determining the causal relation between e_i and e_j, the model has eliminated other potential factors.

The prompts $P = \{P_1, P_2, P_3\}$ for the three hypotheses are as follows:

P_1: Given the sentence X. Does e_i precede e_j? Answer yes or no.

P_2: Given the sentence X. Is there a correlation between e_i and e_j? Answer yes or no.

P_3: Given the sentence X. Have other potential causes or confounding factors been adequately considered and eliminated? Answer yes or no.

We input these three hypothesis reasoning prompts into the LLM to obtain the answers for each hypothesis. This process can be described using the following formula: $U = M(y|P)$, where

U represents the LLM's answer. Integrate U into our designed Prompt H. $H = \{H_1, H_2, H_3\}$, and H_1, H_2, H_3 represent the model's reasoning answers regarding temporal order, correlation, and consideration of other factors, respectively. Refer to Figure 3 for the H Prompt.

Stage 2: Deductive Reasoning

According to the theory of hypothetical-deductive reasoning: if all three hypotheses hold, we can conclude that e_i causes e_j; if any of the hypotheses do not hold, we conclude that there is no causal relation between e_i and e_j. Based on this criterion, we then ask the LLM whether there is a causal relation between e_i and e_j. Our prompt F is as follows:

$[T, H_1, H_2, H_3]$**. Based on the above reasoning and the theory of hypothetical-deductive reasoning, does** e_i **cause** e_j**? Answer yes or no.**

This prompt guides the LLM to evaluate the causal relation based on the outcomes of the three hypotheses.

2.3 Hypothetical Reasoning Supervision

To enhance the correctness of the model's reasoning for the three hypotheses in **Stage 1**, we use the ground truth causal relation labels from the original dataset. Each hypothesis answer is input into the LLM, prompting it to predict the final relation label. Since the reasoning criteria is that if there is a causal relationship between two events, all three hypotheses are valid. Therefore, they are consistent with the ground truth causal relation labels. This reasoning structure allows us to supervise the model's hypothesis answers without additional annotation. By continuously refining its hypothetical reasoning content, the model aligns its deductive reasoning results with the ground truth labels in **Stage 2**. This process improves the model's hypothetical reasoning ability. It is a straightforward and efficient method that does not require additional annotations.

3 Experiments

3.1 Datasets and Evaluation Metrics

We evaluate CPR and HYDRO on two widely used datasets.

EventStoryLine (Caselli and Vossen, 2017) contains 22 topics, 258 documents, and 5,334 events. Among these, 1,770 pairs of intra-sentence event pairs and 3,885 pairs of inter-sentence event pairs are annotated with causal relations. Following the previous work (Gao et al., 2019), we use the documents from the last two topics as development data, while the documents from the remaining 20 topics are used for 5-fold cross-validation.

Causal-TimeBank (Mirza, 2014) contains 183 documents, 6,811 event mentions, and 7,608 intra-sentence event pairs (308 of which have causal relations). Following the previous works (Chen et al., 2022; Liu et al., 2023), we evaluate intra-sentence event pairs using 10-fold cross-validation.

Evaluation Metrics We use precision (P), recall (R), and F1 score (F1) as evaluation metrics to ensure comparability with previous works (Chen et al., 2022; Phu and Nguyen, 2021).

3.2 Implementation Details

Because encoder-style models cannot generate text that supports chains of thought, we use the encoder-decoder architecture of Flan-T5 as our main LLM. The model is optimized using AdamW (Loshchilov and Hutter, 2017) with a learning rate of 1e-4 and a weight decay of 0.01. We clip the gradients of model parameters to a max norm of 1.0. We adopt a negative sampling rate of 0.6 for training our model. The model is trained for 10 epochs, and we select the best checkpoint on the development set for testing. Our experiments are conducted with 4 NVIDIA RTX A100 GPUs.

3.3 Compared Baselines

SECI: We compare the following methods with HYDRO and CPR on SECI: 1) **KMMG** (Liu et al., 2021), which utilized external knowledge and proposes a mention masking generalization method for accurate inference. 2) **LSIN** (Cao et al., 2021), which used a descriptive graph induction module to leverage external structural knowledge. 3) **DPJL** (Shen et al., 2022), which utilized joint prompt learning and incorporates two derivative recognition tasks.

ECI (Includes SECI and DECI): We compare the following methods with HYDRO and CPR on ECI: 1) **ERGO** (Chen et al., 2022) designed an event relation graph and transformed event causality identification into a node classification framework. 2) **CHEER** (Chen et al., 2023), which proposed a reasoning network centered around perceiving key events for global reasoning. 3) **PPAT** (Liu et al., 2023) utilized pairwise attention to capture inference chains on the event relation graph at sentence boundaries. 4) **SENDIR** (Yuan et al., 2023) em-

Model	EventStoryLine (SECI)			EventStoryLine (DECI)			EventStoryLine (Overall)		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
KMMG (Liu et al., 2021)	41.9	62.5	50.1	-	-	-	-	-	-
LearnDA (Zuo et al., 2021b)	42.2	69.8	52.6	-	-	-	-	-	-
LSIN (Cao et al., 2021)	47.9	58.1	52.5	-	-	-	-	-	-
DPJL (Shen et al., 2022)	65.3	70.8	67.9	-	-	-	-	-	-
SemSIm (Hu et al., 2023)	64.2	65.7	64.9	-	-	-	-	-	-
RichGCN (Phu and Nguyen, 2021)	49.7	63.0	55.2	39.2	45.7	42.2	42.6	51.3	46.6
ERGO (Chen et al., 2022)	57.5	72.0	63.9	51.6	43.3	47.1	48.6	53.4	50.9
CHEER (Chen et al., 2023)	56.9	69.6	62.6	45.2	52.1	48.4	49.7	53.3	51.4
PPAT (Liu et al., 2023)	60.7	70.5	65.2	48.9	49.8	49.3	52.9	56.3	54.5
SENDIR (Yuan et al., 2023)	65.8	66.7	66.2	33.0	**90.0**	48.3	37.8	82.8	51.9
KADE (Wu et al., 2023)	61.5	73.2	66.8	51.2	74.2	60.5	51.9	70.6	59.8
iLIF (Liu et al., 2024)	76.8	66.3	71.2	53.5	65.9	59.1	59.2	66.1	62.5
DiffusECI (Man et al., 2024)	65.8	78.3	71.4	61.9	59.9	60.9	63.0	64.1	63.5
BART-Large (400M)	45.3	76.5	56.9	51.2	80.3	62.5	49.1	79.3	60.6
BART-Large+CPR (400M) (ours)	58.3	86.0	69.5	67.0	86.3	75.4	63.9	86.2	73.4
BART-Large+HYDRO (400M) (ours)	64.8	**86.7**	74.1	69.8	84.6	76.5	68.0	85.3	75.7
Flan-T5-Base (250M)	55.0	78.8	64.8	67.9	77.2	72.2	63.6	77.7	69.9
Flan-T5-Base (3B)	61.0	81.9	69.3	64.3	84.7	73.1	67.8	75.5	71.4
Flan-T5-Base+CPR (250M) (ours)	62.6	85.5	72.2	69.2	84.7	76.2	67.0	85.0	74.9
Flan-T5-Base+CPR (3B) (ours)	68.1	84.7	75.5	69.4	88.0	77.7	69.0	87.0	77.0
Flan-T5-Base+HYDRO (250M) (ours)	66.1	85.3	73.9	72.8	84.8	77.5	69.7	84.9	76.5
Flan-T5-XL+HYDRO (3B) (ours)	**72.1**	85.2	**78.0**	**77.8**	85.6	**81.5**	**75.9**	85.4	**80.3**
w/o Hypothetical Reasoning Supervision	66.4	85.3	74.7	75.1	88.3	81.2	72.2	**87.4**	79.1

Table 1: Compare different methods on EventStoryLine. The best results are in **bold** and the second-best results are in underlined.

ployed a novel discriminative reasoning method with sparse event representations. 5) **KADE** (Wu et al., 2023) used external knowledge and event analogy. 6) **iLIF** (Liu et al., 2024) used an iterative learning and identifying framework. 7)**DiffusECI** (Man et al., 2024)refined event context representations into causal label representations. 8) In our zero-shot settings, we compare our method with four progressive SOTA versions of ChatGPT (**GPT-3.5-turbo, GPT-4, text-DaVinci-002, text-DaVinci-003**) and another popular large model: **LLaMA-2**.

3.4 Overall Performance

Due to the limited number of inter-sentence causal event pairs in Causal-TimeBank (only 20 of 252,084 inter-sentence event pairs), we only evaluate SECI performance on Causal-TimeBank (Wu et al., 2023). Table 1 and Table 2 present the experimental results for EventStorline and Causal-TimeBank, respectively. From these results, we have the following observations:

(1) Both proposed reasoning chains significantly outperform all baselines on both benchmarks, achieving SOTA in SECI and DECI. Compared to DiffusECI (previous work's SOTA), our CPR CoT improves the F1 score on EventStoryLine's SECI and DECI by 4.1% and 16.8%, respectively. On Causal-TimeBank's SECI improves by 6.8%. HYDRO improves the F1 score on EventStoryLine's SECI and DECI by 6.6% and 17.0%, respectively, and on Causal-TimeBank's SECI by 6.8%. After applying CPR and HYDRO, the causal reasoning performance of both BART and Flan-T5 has significantly improved. This demonstrates that when prompted with our CPR and HYDRO reasoning frameworks, LLM exhibits strong causal reasoning abilities. Although SENDIR has a high Recall, its

Model	Causal-TimeBank (SECI)		
	P(%)	R(%)	F1(%)
KMMG (Liu et al., 2021)	36.6	55.6	44.1
LSIN (Cao et al., 2021)	51.5	56.2	52.9
DPJL (Shen et al., 2022)	63.6	66.7	64.6
ERGO (Chen et al., 2022)	62.1	61.3	61.7
CHEER (Chen et al., 2023)	56.4	69.5	62.3
PPAT (Liu et al., 2023)	7.9	64.6	66.2
SENDIR (Yuan et al., 2023)	65.2	57.7	61.2
KADE (Wu et al., 2023)	56.8	70.6	66.7
GenSORL (Chen et al., 2024)	66.2	57.0	60.9
KIGP (Hu et al., 2025)	61.3	63.4	62.3
BART-Large (400M)	62.5	45.5	52.6
BART-Large+CPR (400M) (ours)	56.5	61.9	59.9
BART-Large-HYDRO (400M) (ours)	57.1	66.7	61.5
Flan-T5-Base (250M)	73.3	47.8	57.9
Flan-T5-Base (3B)	56.5	61.9	59.9
Flan-T5-Base+CPR (250M) (ours)	71.8	64.9	67.5
Flan-T5-Base+CPR (3B) (ours)	69.6	66.7	68.0
Flan-T5-Base+HYDRO (250M) (ours)	69.6	**74.2**	71.2
Flan-T5-XL+HYDRO (3B) (ours)	**78.2**	70.3	**73.5**
w/o Hypothetical Reasoning Supervision	78.0	68.0	71.5

Table 2: Compare different methods on Causal-TimeBank. The best results are in **bold** and the previous work's best results are in underlined.

precision is lower. This indicates that SENDIR tends to predict all answers as positive samples, which does not demonstrate good model performance.

(2) At the same parameter level, HYDRO outperforms CPR on both EventStoryLine and Causal-TimeBank. This demonstrates the effectiveness of hypothetical-deductive reasoning. The advantage of HYDRO lies in its independent evaluation of the three hypotheses, each unaffected by the others. We also observe that both CPR and HYDRO perform better in DECI than in SECI. We attribute this to the multi-step reasoning process in the chains, which guides the model in inferring implicit relations between events and reasoning information from long texts.

(3) Despite achieving excellent results with the Flan-T5-Base+HYDRO (250M) model, we observe even more significant performance gains when using a larger model. Compared to Flan-T5-Base+HYDRO, Flan-T5-XL+HYDRO improved by 4.1% and 4.0% in SECI and DECI tasks on EventStoryLine respectively on F1. In Causal-TimeBank's SECI task, there is a 2.3% improvement. This indicates that the model's reasoning ability strengthens with an increase in parameters. Furthermore, removing hypothesis reasoning supervision leads to a decline in performance, demonstrating its effectiveness.

3.5 Influence of Different Model Sizes of LLMs

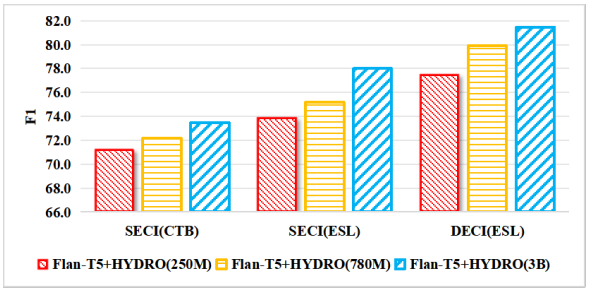

Figure 4: Performance of different parameter sizes of FLAN-T5+HYDRO on two benchmark datasets. ESL and CTB represent EventStoryLine and Causal-TimeBank respectively.

To investigate the impact of different LLM scales. In Figure 4, It can be seen that as the model scale increases, the effectiveness of our hypothesis deduction method is gradually amplified. This aligns well with existing research on CoT, indicating that as the number of parameters increases, LLMs' multi-hop reasoning abilities experience significant improvements.

3.6 Error Analysis

To understand the performance of LLMs in causal multi-hop reasoning, we randomly select 100 incorrect reasoning samples and analyze the reasons behind these errors. We use the Flan-T5+CPR and Flan-T5+HYDRO models, which have been trained under supervision. Based on where the errors occur in the different reasoning stages (or hops), we categorize the errors into three types: logic (logic inconsistency), commonsense (incorrect responses when it needs to combine commonsense), and summarization (summary errors based on the context). Figure 5 shows the distribution of these error types. Additionally, in the appendix A, we provide examples corresponding to each of the three categories.

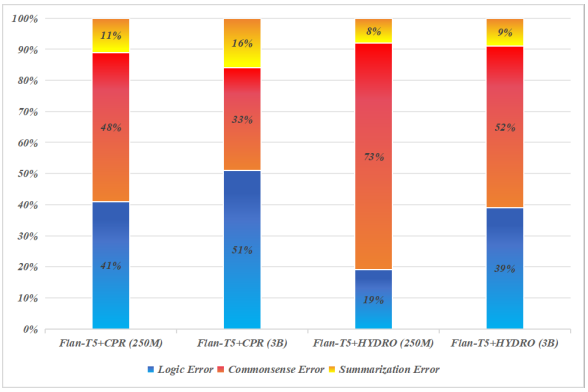

Figure 5: The percentage of three reasoning error types in the two reasoning frameworks.

We find that under the CPR, logic errors and commonsense errors are the most common; under the HYDRO, commonsense errors are the most prevalent. Summarization errors are relatively rare in both models. We can observe that as the number of model parameters increases, the commonsense errors in both methods significantly decrease. logic error occurs more frequently in forward reasoning like CPR. This indicates that increasing the number of model parameters can improve the model's commonsense understanding, thereby enhancing its performance.

3.7 Zero-shot Setting

To further validate the effectiveness of our two reasoning frameworks in a zero-shot setting, we conduct experiments on the EventStoryline and Causal-TimeBank benchmark datasets. Table 3 and Table 4 show that our method significantly out-

Model	EventStoryLine (SECI)			EventStoryLine (DECI)			EventStoryLine (Overall)		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
LLaMa-2 (7B) (Liu et al., 2024)	26.9	29.3	28.0	10.8	31.9	16.1	13.2	31.1	18.5
text-DaVinci-002 (Liu et al., 2024)	23.2	80.0	36.0	-	-	-	-	-	-
text-DaVinci-003 (Liu et al., 2024)	33.2	74.4	**45.9**	-	-	-	-	-	-
GPT-3.5-Turbo (Liu et al., 2024)	27.6	80.2	41.0	-	-	-	-	-	-
GPT-4 (Liu et al., 2024)	27.2	**94.7**	42.2	-	-	-	-	-	-
Flan-T5-Base (250M)	19.1	85.0	31.2	7.4	83.7	13.6	9.2	84.1	16.6
Flan-T5-Base+CPR (250M) (ours)	30.0	87.3	44.7	13.1	**85.6**	22.7	16.0	**86.1**	26.9
Flan-T5-Base+HYDRO (250M) (ours)	**44.3**	46.2	45.2	**22.7**	52.5	**31.6**	**26.4**	50.2	**34.6**

Table 3: In the zero-shot setting, Compare different methods on EventStoryLine. The best results are in **bold** and the second-best results are in underlined.

Model	Causal-TimeBank (SECI)		
	P(%)	R(%)	F1(%)
text-DaVinci-002 (Liu et al., 2024)	5.0	75.2	9.3
text-DaVinci-003 (Liu et al., 2024)	8.5	64.4	15.0
GPT-3.5-Turbo (Liu et al., 2024)	6.9	82.6	12.8
GPT-4 (Liu et al., 2024)	6.1	**97.4**	11.5
Flan-T5-Base (250M)	6.5	57.8	11.7
Flan-T5-Base+CPR (250M) (ours)	5.5	84.0	9.5
Flan-T5-Base+HYDRO (250M) (ours)	**12.8**	42.2	**19.6**

Table 4: In the zero-shot setting, Compare different methods on Causal-TimeBank. The best results are in **bold** and the second-best results are in underlined.

performs multiple versions of ChatGPT in ECI. Despite text-DaVinci-003's (175B) parameter volume being **700 times** that of Flan-T5+HYDRO (250M), we still demonstrate superior performance in zero-shot scenarios, achieving nearly comparable levels on EventStoryline and surpassing it by 4.6% in F1 score on CTB. This demonstrates that the HYDRO two-stage reasoning framework can effectively enhance the model's reasoning capabilities even with low-resource and low-size models. CPR's performance in zero-shot settings is not as impressive because the low-size LLM produces more errors in the initial hops of reasoning, which propagates to subsequent hops.

4 Related work

Early ECI mainly focused on the SECI task, leveraging sentence features to enhance performance, such as lexical patterns (Hidey and McKeown, 2016), and causal patterns (Riaz and Girju, 2014; Hu et al., 2017), syntactic structures (Mirza, 2014). Later, due to the success of deep learning, some work shifted towards using pre-trained language models (PLMs) to obtain high-quality event contexts, achieving good performance (Kadowaki et al., 2019; Liu et al., 2021; Zuo et al., 2021a). For instance, Shen et al. (2022) used prompt-based joint learning, incorporating causal keyword information and event information, demonstrating excellent performance on the SECI task.

As SECI performance has improved, DECI has posed new challenges for the model's reasoning capabilities. Gao et al. (2019) used integer linear programming to model global causal relations. Graph neural networks have also played a positive role in DECI. ERGO (Chen et al., 2022) achieved performance improvement through graph transformers on event relation graphs. Liu et al. (2023) proposed PPAT for incremental reasoning on event relation graphs at the sentence boundary. Recent Work integrating external knowledge has also shown excellent performance in causality reasoning. Chen et al. (2023) manually annotated the central events of documents, considering the centrality of events. Wu et al. (2023) introduced ConceptNet to retrieve relevant knowledge and then compared the given events with other events in memory.

5 Conclusion

In this paper, we emphasize the importance of multi-hop reasoning in ECI tasks. We first introduce the Causal Progressive Reasoning (CPR) chain, which guides LLMs through a step-by-step reasoning process to derive predictions. The key to CPR is breaking down complex causal reasoning into manageable steps. However, considering the error propagation in CoT, we propose HYDRO, which is based on hypothetical-deductive reasoning. The HYDRO is a completely new reasoning framework different from CoT that each step of reasoning is independent. Our extensive experiments demonstrate that both reasoning chains achieve SOTA performance on two ECI benchmark datasets. Additionally, in zero-shot settings, Flan-T5-Base (250M) with HYDRO surpasses ChatGPT's performance.

Limitations

Due to limited computational resources, HYDRO could only be fine-tuned on Flan-T5-xl (3B) and not on larger LLMs, which somewhat restricts its performance. This also indicates that HYDRO's effectiveness is also constrained by the scale of the LLMs. Additionally, in Zero-shot settings, HYDRO is applied only to Flan-T5-Base (250M) and not to the ChatGPT series of models.

References

Pengfei Cao, Xinyu Zuo, Yubo Chen, Kang Liu, Jun Zhao, Yuguang Chen, and Weihua Peng. 2021. Knowledge-enriched event causality identification via latent structure induction networks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4862–4872.

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.

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

Meiqi Chen, Yixin Cao, Yan Zhang, and Zhiwei Liu. 2023. Cheer: Centrality-aware high-order event reasoning network for document-level event causality identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10804–10816.

Mingliang Chen, Wenzhong Yang, Fuyuan Wei, Qicai Dai, Mingjie Qiu, Chenghao Fu, and Mo Sha. 2024. Event causality identification via structure optimization and reinforcement learning. *Knowledge-Based Systems*, 284:111256.

Lei Gao, Prafulla Kumar Choubey, and Ruihong Huang. 2019. Modeling document-level causal structures for event causal relation identification. 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 1808–1817.

Christopher Hidey and Kathleen McKeown. 2016. Identifying causal relations using parallel wikipedia articles. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1424–1433.

Ruijuan Hu, Jian Li, Haiyan Liu, Guilin Qi, and Yuxin Zhang. 2025. Knowledge interaction graph guided prompting for event causality identification. *Applied Intelligence*, 55(2):1–14.

Zhichao Hu, Elahe Rahimtoroghi, and Marilyn Walker. 2017. Inference of fine-grained event causality from blogs and films. In *Proceedings of the Events and Stories in the News Workshop*, pages 52–58.

Zhilei Hu, Zixuan Li, Xiaolong Jin, Long Bai, Saiping Guan, Jiafeng Guo, and Xueqi Cheng. 2023. 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.

Kazuma Kadowaki, Ryu Iida, Kentaro Torisawa, Jong-Hoon Oh, and Julien Kloetzer. 2019. 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.

Cheng Liu, Wei Xiang, and Bang Wang. 2024. 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, Bangkok, Thailand. Association for Computational Linguistics.

Jian Liu, Yubo Chen, and Jun Zhao. 2021. Knowledge enhanced event causality identification with mention masking generalizations. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence*, pages 3608–3614.

Zhenyu Liu, Baotian Hu, Zhenran Xu, and Min Zhang. 2023. Ppat: progressive graph pairwise attention network for event causality identification. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 5150–5158.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Hieu Man, Franck Dernoncourt, and Thien Huu Nguyen. 2024. Mastering context-to-label representation transformation for event causality identification with diffusion models. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 18760–18768. AAAI Press.

Puneet Mathur, Vlad I Morariu, Aparna Garimella, Franck Dernoncourt, Jiuxiang Gu, Ramit Sawhney, Preslav Nakov, Dinesh Manocha, and Rajiv Jain.

2024. Docscript: Document-level script event prediction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5140–5155.

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

Judea Pearl. 2001. Causality: Models, reasoning, and inference. *Neural computing*, 41(1-4):189–190.

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.

Mehwish Riaz and Roxana Girju. 2014. 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.

Shirong Shen, Heng Zhou, Tongtong Wu, and Guilin Qi. 2022. Event causality identification via derivative prompt joint learning. In *Proceedings of the 29th international conference on computational linguistics*, pages 2288–2299.

Jiaxin Shi, Shulin Cao, Lei Hou, Juanzi Li, and Hanwang Zhang. 2021. Transfernet: An effective and transparent framework for multi-hop question answering over relation graph. *arXiv preprint arXiv:2104.07302*.

Yuan Sui, Shanshan Feng, Huaxiang Zhang, Jian Cao, Liang Hu, and Nengjun Zhu. 2022. Causality-aware enhanced model for multi-hop question answering over knowledge graphs. *Knowledge-Based Systems*, 250:108943.

Boshi Wang, Xiang Deng, and Huan Sun. 2022. Iteratively prompt pre-trained language models for chain of thought. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2714–2730.

Sifan Wu, Ruihui Zhao, Yefeng Zheng, Jian Pei, and Bang Liu. 2023. Identify event causality with knowledge and analogy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13745–13753.

Changsen Yuan, He-Yan Huang, Yixin Cao, and Yonggang Wen. 2023. 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.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Xinyu Zuo, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Weihua Peng, and Yuguang Chen. 2021a. Improving event causality identification via self-supervised representation learning on external causal statement. *arXiv preprint arXiv:2106.01654*.

Xinyu Zuo, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Weihua Peng, and Yuguang Chen. 2021b. Learnda: Learnable knowledge-guided data augmentation for event causality identification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3558–3571.

A Error Type

In CPR, the errors occurring in the 1^{st} and 3^{rd} hops are categorized as commonsense errors, as they require commonsense knowledge to make inferences. The Errors in the 2^{nd} hop are categorized as logical errors because this step involves reasoning about the relation between two entities, which necessitates a certain level of logical capability. The Errors in the 4^{th} hop are categorized as summarization errors, as they involve summarizing the context to provide an answer.

In HYDRO, errors in Hypothesis 1 and Hypothesis 2 during the hypothesis reasoning stage are categorized as logical errors, while errors in Hypothesis 3 are categorized as commonsense errors. Errors occurring in the deductive phase are categorized as summarization errors.

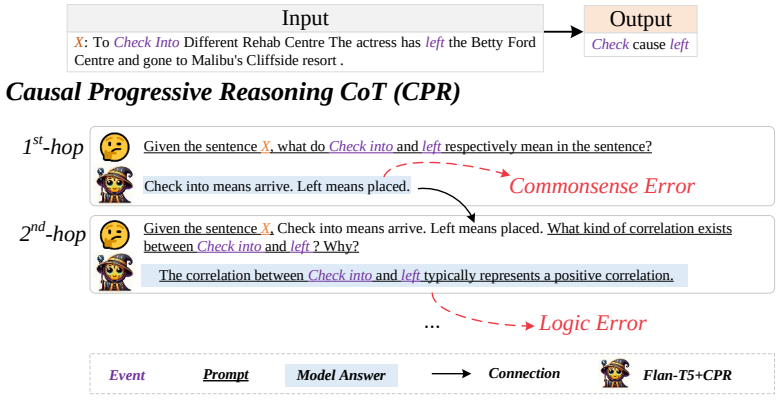

Figure 6: The case of Commonsense Error and Logic Error

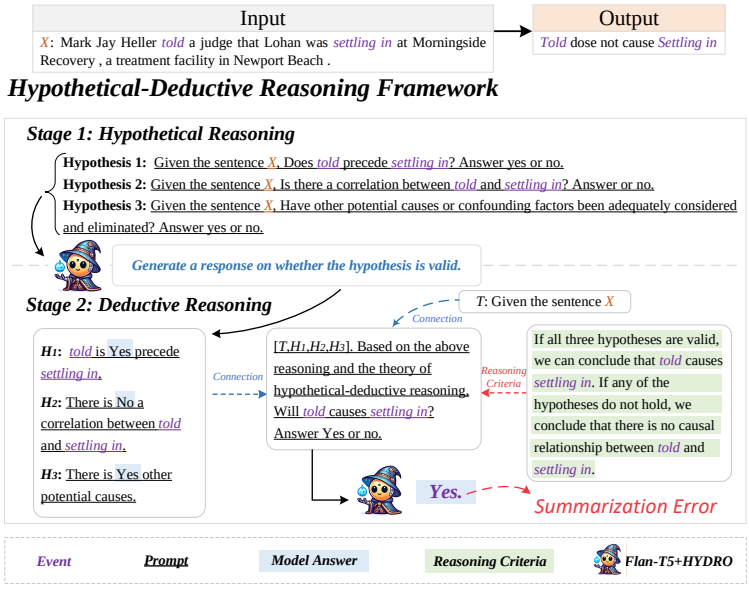

Figure 7: The case of Summarization Error