Mitigating Causal Bias in LLMs via Potential Outcomes Framework and Actual Causality Theory

Yiheng Zhao[1], Yuanliang Li[1], Shreya Prithviraj Savant[1], Jun Yan[1, *†],
[1]Concordia University, Montreal, Canada

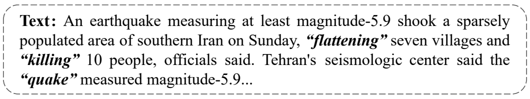
†**Correspondence:** jun.yan@concordia.ca

Abstract

Event Causality Identification (ECI) aims to identify causal relationships between events, which is essential for root cause analysis. While recent studies reveal that Large Language Models (LLMs) exhibit significant causal hallucination, a systematic evaluation of their document-level ECI performance across varied structural characteristics and a corresponding dataset is currently lacking. To fill this gap, we first construct a structure-controlled dataset to comprehensively assess their document-level ECI performance across texts with various structural characteristics that influence the causal behaviors in ECI. We find that different LLMs exhibit divergent causal bias across texts with varied structures, ranging from consistent hallucination or neglect to structure-dependent shifts between the two. To mitigate the bias, furthermore, we formulate ECI as a causal inference problem and propose a causality identification framework grounded in the potential outcomes and the Halpern–Pearl (HP) definition of actual causality theory. Experimental results demonstrate that our framework significantly reduces the causal bias associated with directly using LLMs on ECI, while also achieving superior performance.

1 Introduction

Event causality identification (ECI) is a crucial task in natural language processing (NLP) that aims to determine whether a causal relationship exists between two identified events in the text. With a better understanding of event causality, ECI can facilitate root cause analysis, such as delays, cost overruns, and potential safety risks, in the engineering projects (Assaf and Al-Hejji, 2006; Cooper et al., 2008; Shim et al., 2016). The ECI can be divided into sentence-level, where two events appear within the same sentence, and document-level, where the two events occur within the same sentence or different sentences, as shown in Figure 1.



Text: An earthquake measuring at least magnitude-5.9 shook a sparsely populated area of southern Iran on Sunday, ***"flattening"*** seven villages and ***"killing"*** 10 people, officials said. Tehran's seismologic center said the ***"quake"*** measured magnitude-5.9...

Figure 1: An example of ECI, where ("flattening", "killing") is an intra-sentence event pair, and ("flattening", "quake") and ("killing", "quake") are inter-sentence event pairs.

Recently, as large language models (LLMs) have exhibited remarkable reasoning capabilities, several studies (Liu et al., 2024; Tao et al., 2024; Gao et al., 2023; Su et al., 2025) have explored their ECI performance. Some results reveal that LLMs exhibit causal hallucination, tending to assume causal relationships between event pairs regardless of whether such relationships exist, which makes their output less trustworthy. However, a systematic evaluation of LLM performance on document-level ECI across texts with various structural characteristics that impact task difficulty, and a corresponding dataset, is currently lacking.

To fill this gap, we design a structure-controlled dataset and use it to evaluate LLMs' document-level ECI performance comprehensively. Unlike existing datasets (Caselli and Vossen, 2017; Mirza and Tonelli, 2014; Wang et al., 2022a; Lai et al., 2022), our dataset enables the evaluation of LLMs' ECI performance across texts with varied structural characteristics such as text length, event count, and event distance, while ensuring that each factor is assessed independently without confounding influence from the others. Based on our evaluation, we find that different LLMs exhibit divergent causal bias across texts with various structural characteristics, ranging from consistent hallucination or neglect to structure-dependent shifts between the two. Figure 2 presents some of our evaluation results on the text length part of our dataset. To mitigate bias, we further formulate ECI as a causal inference problem and propose a causality

Findings of the Association for Computational Linguistics: EACL 2026, pages 4212–4222
March 24-29, 2026 ©2026 Association for Computational Linguistics

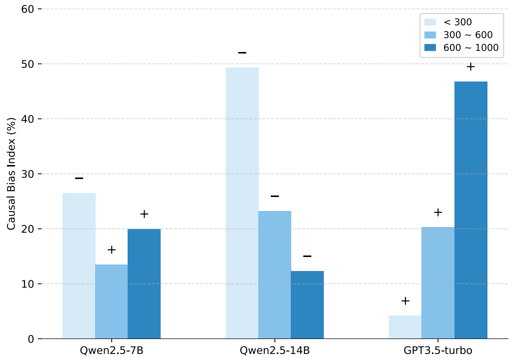


Figure 2: Results on the text length part of our dataset. Causal Bias Index (CBI) measures the absolute performance gap of models between causal and non-causal event pairs. "+" indicates better performance on causal pairs (causal hallucination), while "–" indicates the opposite (causal neglect). "$<$ 300", "300 $\sim$ 600", and "600 $\sim$ 1000" denote the text length ranges in which the event pairs appear. These results demonstrate the severity of causal biases in LLMs.

identification framework grounded in the potential outcomes framework and the Halpern–Pearl (HP) definition from actual causality theory. Instead of directly allowing LLMs to decide whether a causal relationship exists between two events, our framework requires the models to separately estimate the likelihood of the subsequent event under both factual and counterfactual scenarios. If the subsequent event is more likely to occur when the preceding event happens, our method infers a causal relationship; otherwise, it does not. The experiments demonstrate that our method not only exhibits lower causal bias but also achieves higher accuracy.

In summary, our contributions can be summarized as follows: (1) We construct a structure-controlled dataset to comprehensively assess LLMs' document-level ECI performance across texts with various structural characteristics. (2) We propose a causality identification framework grounded in the potential outcomes framework and the HP definition of actual causality theory to produce more reliable causal judgments. (3) Experimental results demonstrate that our framework significantly reduces the causal bias associated with directly using LLMs on ECI, as quantified by a substantially lower Causal Bias Index (CBI), while also achieving superior Overall Accuracy (OA). For GPT3.5-turbo (Ouyang et al., 2022), the CBI was reduced by 12.2% (OA up 4.2%), 15.6% (OA up 5.8%), and a maximum of 19.4% (OA up 5.9%) across the text length, event count, and event distance parts, respectively. Similarly, for Qwen2.5-14B (Yang et al., 2024), the CBI was lowered by 16.7% (OA up 3.2%), 13.5% (OA up 6.5%), and 5.4% (OA up 4.2%) across the same respective parts.

2 Preliminaries

In this section, we present the two foundations of our study: the potential outcomes framework and the Halpern-Pearl (HP) definition from actual causality theory.

2.1 Potential outcomes framework

The potential outcomes framework (Rubin, 1974) is a foundational approach for identifying and quantifying causal effects in observational data. Here, we present its fundamental concepts.

Unit The basic analysis entity (e.g., a person or object).

Treatment An intervention applied to a unit. Given treatments $T = \{1, 2, \ldots, N\}$, the potential outcome under treatment T_i is denoted as $Y(T_i)$.

Treatment effect Defined as the difference between potential outcomes under different treatments. It can be measured at multiple levels, including the population (Average Treatment Effect, ATE), subgroup (Conditional Average Treatment Effect, CATE), and individual (Individual Treatment Effect, ITE).

Suppose we aim to measure the treatment effect of a treatment $T = 1$. The ATE is defined as:

$$\mathbb{E}\left[Y(T=1) - Y(T=0)\right] \tag{1}$$

where $Y_{T=1}$ and $Y_{T=0}$ denote the potential outcomes under treatment and control, respectively. The CATE is defined as:

$$\mathbb{E}[Y(T=1) \mid x] - \mathbb{E}[Y(T=0) \mid x] \tag{2}$$

where $\mathbb{E}[Y(T=1) \mid x]$ and $\mathbb{E}[Y(T=0) \mid x]$ are the potential treated and control outcomes of the subgroup with x. When x corresponds to an individual unit; ITE is treated the same as CATE.

2.2 The Halpern-Pearl Definition

The HP definition from actual causality theory is used to determine the actual causes of events (Halpern, 2016). Following, we present its modified version:

$\vec{X} = \vec{x}$ is an actual cause of event φ, in the causal setting $(M, \vec{u})$, the specific scenario, if the following three conditions hold:

AC1. (M, $\vec{u}$) $\models$ ($\vec{X} = \vec{x}$) and (M, $\vec{u}$) $\models \varphi$.

AC2. There exist two disjoint subsets $\vec{Z}$ and $\vec{W}$ in $\vec{V}$ ($\vec{Z} \cap \vec{W} = \emptyset$) with $\vec{X} \wedge \vec{C} \subseteq \vec{Z}$ ($\vec{X} \cap \vec{C} = \emptyset$) and a setting $\vec{x}'$, $\vec{w}$ and $\vec{c}$ of the variables in $\vec{X}$, $\vec{W}$ and $\vec{C}$, respectively, such that

$$(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}', \vec{W} \leftarrow \vec{w}, \vec{C} \leftarrow \vec{c}]\neg\varphi.$$

AC3. $\vec{X}$ is minimal.

AC1 requires that $\vec{X} = \vec{x}$ and φ both actually occur in the causal setting $(M, \vec{u})$, the specific scenario. AC2 establishes a permissive counterfactual dependence of φ on $\vec{X}$ by fixing the variables in $\vec{W}$, which are unrelated to the occurrence of φ, and the variables in $\vec{C}$, which could otherwise cause φ if $\vec{X} = \vec{x}'$, at their actual values. $\vec{X} = \vec{x}'$ means $\vec{X}$ didn't happen. AC3 ensures that only essential components are included. Overall, the HP definition means: $\vec{X} = \vec{x}$ is a but-for cause of φ if $\vec{X} = \vec{x}'$ and when assuming $\vec{X} = \vec{x}$ didn't happen, only events influenced by $\vec{X}$ can change.

3 Data

In this section, we first introduce our data construction method and then provide statistical descriptions of the constructed dataset.

3.1 Data construction

Our data construction method is implemented through a controlled partitioning of the EventStoryLine (Caselli and Vossen, 2017), which is built from news documents and annotated with specific events as well as the causal relations between them, supporting research on event causality identification. The process consists of three main steps: (1) Structural characteristics identification, (2) Data partitioning, and (3) Data filtering.

Structural characteristics identification The purpose of this step is to identify the textual structural Characteristics that influence the difficulty of ECI. Based on our analysis, we identify three such characteristics: text length, the number of events, and the lexical distance between events, because longer texts and more events introduce more information, while a greater lexical distance between events requires a stronger ability to capture long-range dependencies.

Data partitioning The goal of this step is to partition event pairs into subsets of varying difficulty levels based on each textual structural characteristic. To ensure that each subset contains sufficient data, we divide the data into three subsets for each characteristic. For text length, event pairs are grouped into three subsets depending on the length of the document where they occur: 0 to 300, 300 to 600, and 600 to 1,000 words. This range is chosen because most documents in the EventStoryLine corpus are under 1,000 words. Therefore, we filter out samples where the document length exceeds 1,000 words. For the number of events, we assign them to three different subsets: those occurring in texts containing 0 to 20, 20 to 40, and 40 to 60 events. The upper limit of 60 is chosen because our analysis of the filtered EventStoryLine shows that the maximum number of events in a single document is 58. For the lexical distance between events, we classify event pairs into three subsets: those with a lexical distance of 0 to 50, 50 to 200, and 200 to 1,000. This classification follows the standard introduced by MAVEN-ERE. (Wang et al., 2022a).

Data filtering This step aims to control for confounding variables and ensure that, when evaluating the ECI performance of LLMs across different subsets of a textual structural characteristic, the results are not influenced by variations in other factors. For text length, we first examine the distributions of event count and event distance within each text-length subset. Based on this analysis, we determine the minimal overlapping range across subsets and filter the data accordingly, ensuring that comparisons across text length subsets are not driven by variations in other structural characteristics. We apply the same procedure for the number of events and the lexical distance between event subsets. And each of our subsets contains an equal number of causal and non-causal event pairs to ensure a fair evaluation.

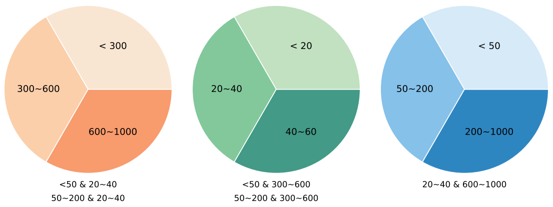
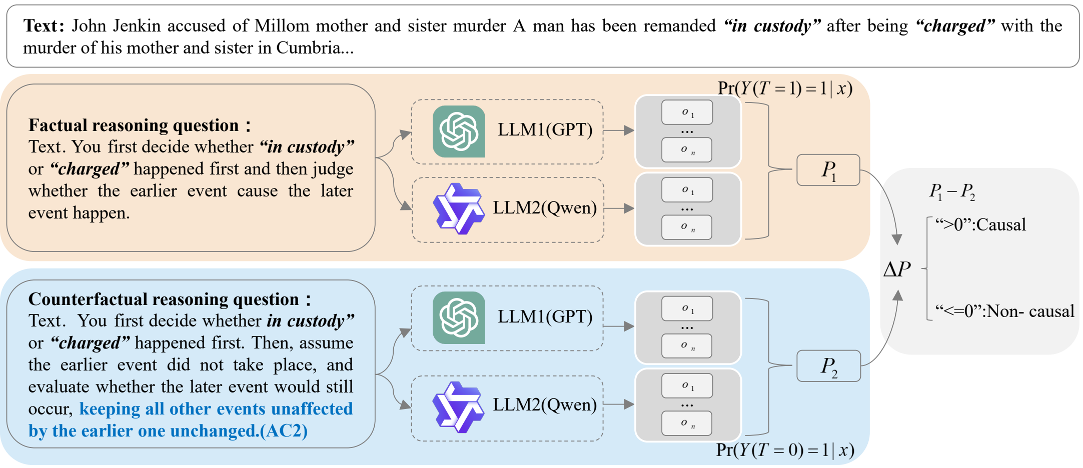


Figure 3: Illustration of our dataset distribution. The distributions of text length, event count, and event distance in our dataset are shown from left to right. Each part is divided into three groups, with consistent data distributions and equal sample sizes.



Text: John Jenkin accused of Millom mother and sister murder A man has been remanded ***"in custody"*** after being ***"charged"*** with the murder of his mother and sister in Cumbria...

Factual reasoning question :
Text. You first decide whether ***"in custody"*** or ***"charged"*** happened first and then judge whether the earlier event cause the later event happen.

Counterfactual reasoning question :
Text. You first decide whether ***in custody"*** or ***"charged"*** happened first. Then, assume the earlier event did not take place, and evaluate whether the later event would still occur, **keeping all other events unaffected by the earlier one unchanged.(AC2)**

Figure 4: Illustration of our ECI framework. In this framework, P_1 represents the probability that the later event ("in custody") occurs given the earlier event ("charged") happens; P_2 represents the probability that the later event ("in custody") occurs assuming the earlier event ("charged") did not happen. $\Delta P = P_1 - P_2$. LLM1 and LLM2 denote GPT3.5-turbo and Qwen2.5-14B, respectively.

3.2 Data statistics

Following the setup above, our constructed dataset consists of three parts. The three parts are used to evaluate LLMs' ECI performance under different controlled conditions, including different text lengths, the number of events, and event distances. Each part consists of three subsets, respectively. The overall data distribution is illustrated in Figure 3. In each group of our text length part, half of the event pairs have a lexical distance of fewer than 50 words and occur in texts containing 20 to 40 events, while the other half have a distance ranging from 50 to 200 words within the same event count range. In each group of our event count part, half of the event pairs have a lexical distance of fewer than 50 words and are located in texts of 600 to 1,000 words in length; the other half fall in the 50 to 200 word range, also within the same text length range. In each group of our event distance part, all event pairs are drawn from texts containing 20 to 40 events and ranging in length from 600 to 1,000 words. Additionally, each subset, comprising text length, event count, and event distance parts, contains 800, 560, and 240 samples, respectively.

4 Methodology

In this section, we introduce our proposed causality identification framework, as illustrated in Figure 4. Considering the risk of causal hallucination and neglect when prompting LLMs to identify causal relations directly, our framework avoids relying on binary causal judgments from LLMs. Instead, we formulate ECI as a causal inference problem. Let the unit be a document x with two marked events: an earlier event T (treatment; $T \in \{0,1\}$) and a later event Y (outcome; $Y \in \{0,1\}$). We estimate P_1, $Pr(Y(T=1)=1 \mid x)$, and P_2, $Pr(Y(T=0)=1 \mid x)$, and compute a potential-outcomes contrast ΔP, $P_1 - P_2$, to infer causality. HP guides how we instantiate the counterfactual $Y(T=0)=1$: we intervene $T \leftarrow 0$ while holding variables unaffected by T at their factual values (AC2 in HP). Overall, the framework has three components: (1) factual reasoning to estimate P_1, (2) counterfactual reasoning to estimate P_2, and (3) causal decision to compute ΔP and predict "causal" if $\Delta P > 0$ (otherwise "non-causal").

Factual reasoning In the factual reasoning stage, we prompt the LLM to perform factual reasoning. Given an input passage with two target events explicitly marked, the model is asked to judge whether the earlier event causes the later one to happen. To obtain a more accurate estimate of the model's belief, we adopt parallel sampling (Wang et al., 2022b)—generating multiple outputs for the same prompt and computing the empirical probability. Compared to relying on raw logits or verbalized output probability, parallel sampling provides a more calibrated and robust estimate of the model's actual confidence (Lyu et al., 2025). In our framework, the number of samples N is set to 5. In addition, to improve the estimate of the later event occurrence probability, we ensemble two LLMs with complementary causal bias profiles and aggregate their samples. Accordingly, P_1 is defined as the proportion of the 10 combined factual responses that assert the later event occurs given that the earlier event has happened.

Counterfactual reasoning In the counterfactual reasoning stage, we prompt the LLM to perform counterfactual reasoning. Given an input passage with two target events explicitly marked, the model is asked to judge whether the later event would still occur if the earlier event had not happened. The counterfactual prompt is carefully designed based on the HP definition of causality. The core of the counterfactual prompt is to apply a critical constraint: that upon assuming the earlier event did not happen, we must simultaneously hold all other events unaffected by T. Similarly, here we adopt the same parallel sampling and two-LLM ensemble strategy as in the factual reasoning. And P_2 is defined as the proportion of the 10 combined factual responses that assert the later event occurs given the earlier event has not happened.

Causal decision In this final stage, we use the probabilities obtained from the previous two stages to make a causal judgment. Specifically, we compute the difference $\Delta P = P_1 - P_2$; if $\Delta P > 0$, we conclude that the earlier event has a causal effect on the later one; otherwise, we determine that no causal relation exists.

5 Experiments

5.1 Settings

Implementation details We evaluate models from several LLM families, including Qwen2.5 (Yang et al., 2024), LLaMA3 (Grattafiori et al., 2024), a distilled variant of DeepSeek-R1 (Guo et al., 2025) and GPT (Ouyang et al., 2022; Achiam et al., 2023). We set the temperature to 0 when evaluating their ECI performance, but used 0.7 to enable diverse sampling in our framework. At a temperature setting of 0.7, the results are averaged over three independent trials. All API utilize official provider APIs; additionally, DeepInfra is employed for LLaMA3.1-8B.

Evaluation metrics To evaluate the performance differences on causal versus non-causal event pairs for these LLMs and our method, we propose an evaluation metric called the Causal Bias Index (CBI). The CBI is formulated as follows:

$$\mathrm{CBI} = \left|\mathrm{Acc}_{\text{causal}} - \mathrm{Acc}_{\text{non-causal}}\right|, \tag{3}$$

where $\mathrm{Acc}_{\text{causal}}$ and $\mathrm{Acc}_{\text{non-causal}}$ denote the model's accuracy on causal and non-causal event pairs, respectively. A higher CBI indicates a stronger tendency of the model to favor one type over the other, suggesting a causal bias in its reasoning behavior on ECI. Therefore, a lower CBI is preferred. We also adopt overall accuracy (OA) as our evaluation metric for two main reasons. First, we aim not only to minimize the performance gap between causal and non-causal event pairs when LLMs exhibit severe causal bias, but also expect LLMs to perform well on both types of data. Second, OA serves as an appropriate metric since our dataset is balanced.

5.2 Overall results

Tables 1, 2, and 3 present the evaluation results of LLMs and our proposed method on our dataset. In addition, since advanced non-LLM methods including GenECI (Man et al., 2022), DPJL (Shen et al., 2022), KEPT (Liu et al., 2023), and CPATT (Zhang et al., 2023) are not fully open-sourced and thus cannot be trained on our dataset, we exclude them from our comparisons. Furthermore, Cai et al. (Cai et al., 2025) have shown that GPT3.5-turbo already outperforms these approaches, which justifies using GPT3.5-turbo as a baseline. The results and more details are shown in Appendix A.

As shown in Tables 1 and 2, our framework achieves the lowest average CBI, indicating the least causal bias among all evaluated models on text length and event count subsets. Meanwhile, it also attains the highest overall accuracy (OA), demonstrating that reducing causal bias does not come at the cost of performance; instead, it leads to more reliable causal reasoning. Additionally, as shown in Table 3, although our method achieves the best OA on the event distance subset, its CBI is also competitive, being slightly higher than that of GPT-4.1-nano, which comes at the cost of significantly lower OA.

From Tables 1, 2, and 3, we can further observe that GPT3.5-turbo and Dsk-R1-Qwen-14B consistently exhibit causal hallucination across all data subsets, while Qwen2.5-14B consistently shows causal neglect. Moreover, even when these models consistently exhibit either hallucination or neglect, the severity of their biases still varies across different subsets. This indicates that the causal behaviors of these LLMs are not static, but are influenced by the textual characteristics. Notably, compared to the consistent neglect exhibited by the original Qwen2.5-14B, this shift may be attributed to the fact that Dsk-R1-Qwen-14B is a distilled version of Qwen2.5-14B using DeepSeek-R1 (Guo et al.,

Methods	**< 300**		**300 ∼ 600**		**600 ∼ 1000**		**Average**	
	CBI	**OA**	**CBI**	**OA**	**CBI**	**OA**	**CBI**	**OA**
GPT3.5-turbo	4.2^{+}	70.6	20.3^{+}	70.3	46.8^{+}	64.1	23.8	68.3
GPT3.5-turbo (CoT)	46.5^{-}	67.2	31.7^{-}	67.8	5.7^{+}	66.8	28.0	67.3
GPT4.1-nano	41.5^{-}	63.8	5.7^{-}	68.3	8.7^{+}	67.3	18.6	66.5
LLaMA3.1-8B	8.0^{-}	64.5	8.0^{+}	63.7	29.3^{+}	65.3	15.1	64.5
Qwen2.5-7B	26.5^{-}	67.0	13.5^{+}	70.8	20.0^{+}	69.7	20.0	69.1
Qwen2.5-14B	49.3^{-}	64.3	23.2^{-}	72.1	12.3^{-}	71.3	28.3	69.3
Qwen2.5-14B (CoT)	54.5^{-}	61.8	29.2^{-}	72.1	17.8^{-}	70.6	33.8	68.2
Dsk-R1-Qwen-14B	35.6^{+}	58.2	41.8^{+}	56.3	44.7^{+}	53.4	40.7	55.9
Ours	7.2^{-}	73.9	8.2^{+}	72.4	19.5^{+}	71.2	11.6	72.5

Table 1: Comparison experiment results on text length subsets. "+" denote causal hallucination, "–" denote causal neglect.

Methods	**< 20**		**20 ∼ 40**		**40 ∼ 60**		**Average**	
	CBI	**OA**	**CBI**	**OA**	**CBI**	**OA**	**CBI**	**OA**
GPT3.5-turbo	40.6^{+}	66.0	25.7^{+}	68.8	17.8^{+}	77.8	28.0	70.9
GPT3.5-turbo (CoT)	15.2^{-}	68.3	32.0^{-}	65.0	37.5^{-}	68.8	28.2	67.4
GPT4.1-nano	3.3^{+}	66.4	12.3^{+}	66.7	30.5^{-}	68.2	15.4	67.1
LLaMA3.1-8B	19.3^{+}	63.0	15.1^{+}	64.5	8.8^{-}	67.7	14.4	65.1
Qwen2.5-7B	24.2^{+}	70.3	19.3^{+}	69.9	14.3^{-}	78.2	19.3	72.8
Qwen2.5-14B	1.0^{-}	74.0	16.0^{-}	71.4	60.7^{-}	65.0	25.9	70.1
Qwen2.5-14B (CoT)	10.8^{-}	73.5	20.8^{-}	72.1	65.7^{-}	62.9	32.4	69.5
Dsk-R1-Qwen-14B	41.4^{+}	57.4	34.0^{+}	58.7	27.9^{+}	66.8	34.4	60.9
Ours	23.9^{-}	77.2	8.5^{-}	72.8	4.8^{-}	80.1	12.4	76.7

Table 2: Comparison experiment results on event count subsets. "+": causal hallucination; "–": causal neglect.

2025), which we find to exhibit severe causal hallucination. Table 4 presents a detailed evaluation of DeepSeek-R1 on the $<$ 300 text length subset of our dataset. The remaining models, including our method, show structure-dependent causal behaviors, with hallucination or neglect varying across different textual subsets. This also indicates that textual characteristics influence the causal behaviors of these LLMs and our method.

Additionally, we observe that Chain-of-Thought (CoT) prompting does not effectively reduce causal bias; instead, it tends to amplify the model's tendency toward causal neglect. For example, as shown in Table 1, when applying CoT prompting to Qwen2.5-14B and GPT3.5-turbo, we observe that the causal neglect of Qwen2.5-14B becomes more pronounced. Meanwhile, GPT3.5-turbo begins to shift from causal hallucination toward neglect, especially in the $<$ 300 and 300∼600 text length segments, where the shift appears to be complete.

5.3 Ablation Study

We also conducted an ablation study to demonstrate the contribution of our framework design, and the results are presented in Table 5. All results in Table 5 are averaged over our text length, event count, and event distance subsets. To further validate the effectiveness of our framework design, we compare its performance with a simplified baseline that directly applies majority voting over the outputs of GPT3.5-turbo and Qwen2.5-14B. This comparison again yields higher CBI and lower OA scores, confirming that our framework plays a crucial role in ensuring reliable causal identification.

6 Analysis

In this section, we investigate (1) why LLMs exhibit causal hallucination or neglect, and (2) whether scaling the number of LLMs enhances causal identification of our method.

6.1 Why do LLMs exhibit causal hallucination or neglect?

Causal hallucination For models exhibiting causal hallucination, we primarily analyze GPT3.5-turbo. By analyzing reasoning processes as reflected in its complete responses, we find that GPT3.5-turbo exhibits causal hallucination by mis-

Methods	< 50		50 ∼ 200		200 ∼ 1000		Average	
	CBI	**OA**	**CBI**	**OA**	**CBI**	**OA**	**CBI**	**OA**
GPT3.5-turbo	46.6^{+}	63.3	45.8^{+}	66.2	12.5^{+}	69.5	35.0	66.3
GPT3.5-turbo (CoT)	29.5^{-}	65.2	16.9^{-}	69.0	46.9^{-}	64.0	31.1	66.1
GPT4.1-nano	11.3^{+}	65.9	11.4^{+}	69.3	13.1^{-}	60.0	11.9	65.1
LLaMA3.1-8B	26.6^{+}	63.3	30.0^{+}	67.5	15.0^{+}	63.3	23.9	64.7
Qwen2.5-7B	22.5^{+}	72.0	17.8^{+}	71.4	25.5^{-}	66.3	21.9	69.9
Qwen2.5-14B	15.0^{-}	72.5	8.1^{-}	65.0	40.0^{-}	66.6	21.0	68.0
Qwen2.5-14B (CoT)	23.0^{-}	71.8	13.7^{-}	65.7	48.2^{-}	63.4	28.3	67.0
Dsk-R1-Qwen-14B	45.9^{+}	54.5	50.0^{+}	53.3	43.3^{+}	57.4	46.4	55.1
Ours	17.5^{+}	72.0	21.7^{+}	71.7	7.5^{-}	72.8	15.6	72.2

Table 3: Comparison experiment results on event distance subsets. "+" denote causal hallucination, "–" denote causal neglect.

Methods	CBI	OA
< 300 subset of Text Length		
Deepseek-R1	29.5^{+}	68.2
Qwen2.5-14B	49.3^{-}	64.3
Dsk-R1-Qwen-14B	35.6^{+}	58.2

Table 4: Experiment results of DeepSeek-R1. Superscripts: "+": causal hallucination; "–": causal neglect.

Methods	CBI	OA
Text Length		
Voting	28.7	69.1
Ours	11.6	72.5
Event Count		
Voting	25.7	71.1
Ours	12.4	76.7
Event Distance		
Voting	20.6	68.2
Ours	15.6	72.2

Table 5: Ablation experiment results on our dataset.

interpreting linguistic co-occurrence as causal linkage. This suggests that the causal hallucination behavior exhibited by GPT3.5-turbo is primarily driven by flawed internal causal knowledge, which often conflates correlation with causation.

Causal neglect For models exhibiting causal neglect, we primarily analyze Qwen2.5-14B. We find that Qwen2.5-14B often refrains from making causal judgments, frequently generating responses such as "the contextual information is limited and requires further confirmation," and thus defaults to a "non-causal" conclusion. However, our manual evaluation reveals that humans, leveraging their own causal knowledge, can make accurate judgments based on the same context. This suggests that the causal neglect behavior exhibited by Qwen2.5-14B is primarily due to insufficient internal causal knowledge.

To support this conclusion, we evaluate Qwen2.5-14B and GPT3.5-turbo via causal intervention: an adversarial prompt is injected with an incorrect causal relation, after which each model is asked to identify the causal relationships. The degradation results of OA are shown in Figure 5. Compared to the original results without intervention, we observe a sharp drop in OA after introducing incorrect causal relationship interventions. Moreover, we find that the performance drop is more pronounced for both models on causal event pairs than on non-causal ones. The effect is particularly pronounced in Qwen2.5-14B, where the OA on causal event pairs drops to as low as 1%. Details of the prompts are presented in Appendix B.

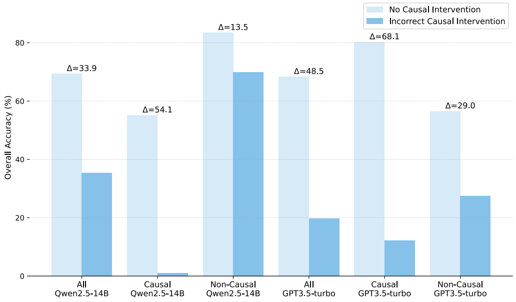


Figure 5: OA measured under incorrect causal intervention on the text length part. "all" denotes all event pairs, while "Causal" and "Non-Causal" refer to causal and non-causal event pairs, respectively. Δ represents the accuracy difference between no intervention and incorrect intervention.

Methods	CBI	OA
< 300		
w/ LLaMA3.1-8B	0.8^{-}	74.1
Original	7.2^{-}	73.9
300 ∼ 600		
w/ LLaMA3.1-8B	9.7^{+}	73.7
Original	8.2^{+}	72.4
600 ∼ 1000		
w/ LLaMA3.1-8B	24.7^{+}	73.2
Original	19.5^{+}	71.2
Average		
w/ LLaMA3.1-8B	11.7	73.6
Original	11.6	72.5

Table 6: Comparison between our original framework and the version incorporating LLaMA3.1-8B on text length subsets. Superscripts "+" denote causal hallucination, "–" denote causal neglect.

6.2 Can scaling up LLMs enhance causal identification?

To investigate whether scaling up LLMs can enhance the performance of our framework, we integrate LLaMA-3.1-8B into our framework and compare it with the original version without LLaMA3.1-8B on subsets of text lengths. The results are presented in Table 6. Incorporating LLaMA3.1-8B into our framework yields improved overall performance, with a higher average OA and comparable CBI. This suggests that scaling up LLMs is beneficial to our framework. The slight increase in average CBI is primarily attributed to longer texts, where scaling up LLMs improves OA but may also amplify causal bias. In contrast, for the 0-300 words subset, both CBI and OA show significant improvement. Specifically, CBI drops from 7.2% to 0.8%, and OA increases, indicating that scaling up LLMs is particularly beneficial to our framework FOR short texts.

6.3 How do sample sizes influence causal identification?

Additionally, we investigate how varying sample sizes affect the causal identification of our method. We observe that OA increases monotonically with the sample size, indicating that more samples contribute to improved overall performance. Conversely, although CBI increases with sample size initially, indicating a rising causal bias, it later declines, implying that further sampling helps reduce bias. Notably, even when sampling only twice each model, our framework consistently outperforms all individual LLMs in terms of average CBI and OA across the text-length subsets. The changes in CBI with respect to sample size are shown in Figure 6, while the trend of overall accuracy is presented in Figure 7.

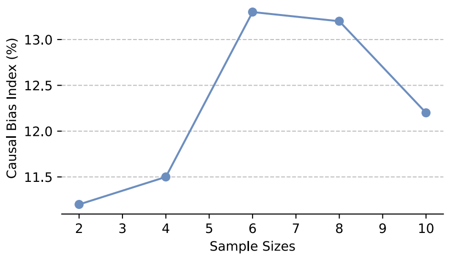


Figure 6: CBI measured across varying sample sizes on our text length subsets.

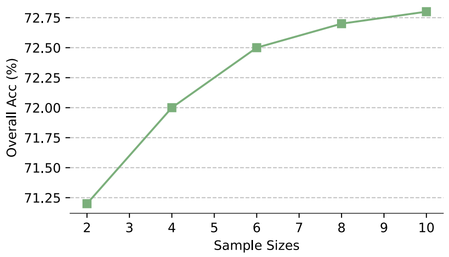


Figure 7: OA measured across varying sample sizes on our text length subsets.

7 Related work

Prior to the rise of LLM-based approaches, the study of ECI had already been evolving toward the use of pre-trained language models. For instance, DPJL (Shen et al., 2022), which enhances event causality recognition via cue-based learning from language models. CPATT (Zhang et al., 2023), which introduces a prompt-based technique with constrained prefix attention. Despite their effectiveness, these models heavily rely on high-quality and manually annotated datasets, which limits their adaptability to new scenarios.

Recently, with the remarkable reasoning capabilities demonstrated by LLMs, several studies (Liu et al., 2024; Tao et al., 2024; Gao et al., 2023) have investigated their ECI performance. These works investigate models such as LLaMA2 (Roumeliotis et al., 2023), GPT3.5, and GPT4, consistently revealing a tendency toward causal hallucination: LLMs often predict causal relationships even when none exist, raising concerns about their reliability in ECI. However, these studies focus exclusively on sentence-level ECI, leaving document-level ECI underexplored, and they do not propose methods to mitigate the observed causal bias.

8 Conclusion

In this work, we first design a novel structure-controlled dataset to comprehensively evaluate the ECI performance of LLMs across texts with varied structural characteristics. Our analysis reveals that different LLMs exhibit divergent causal behaviors across structurally varied texts, ranging from consistent hallucination or neglect to structure-dependent shifts between the two. We then propose a causal identification framework grounded in the potential outcomes framework and the HP definition, aiming to generate more reliable causal judgments. Experiments demonstrate that our method not only exhibits lower causal bias but also achieves higher accuracy. Furthermore, we investigate how sample sizes and model scaling affect causal identification, analyzing why LLMs exhibit causal hallucination or neglect and whether scaling the number of LLMs enhances our method's ECI performance.

Limitations

Our framework relies on a two-LLM ensemble with parallel sampling to determine whether one event causes another. While this stabilizes estimates, it increases inference cost and latency. In future, we aim to improve the model's intrinsic reasoning to deliver reliable causal judgments.

Acknowledgments

We gratefully acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) under Grant ALLRP 585937-23 and Mitacs under Grant IT35587.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Sadi A Assaf and Sadiq Al-Hejji. 2006. Causes of delay in large construction projects. *International journal of project management*, 24(4):349–357.

Ruichu Cai, Shengyin Yu, Jiahao Zhang, Wei Chen, Boyan Xu, and Keli Zhang. 2025. 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.

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.

Rachel Cooper, Ghassan Aouad, Angela Lee, Song Wu, Andrew Fleming, and Michail Kagioglou. 2008. *Process management in design and construction*. John Wiley & Sons.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Joseph Y Halpern. 2016. *Actual causality*. MiT Press.

Viet Dac Lai, Amir Pouran Ben Veyseh, Minh Van Nguyen, Franck Dernoncourt, and Thien Huu Nguyen. 2022. Meci: A multilingual dataset for event causality identification. In *Proceedings of the 29th international conference on computational linguistics*, pages 2346–2356.

Cheng Liu, Wei Xiang, and Bang Wang. 2024. Identifying while learning for document event causality identification. *arXiv preprint arXiv:2405.20608*.

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.

Qing Lyu, Kumar Shridhar, Chaitanya Malaviya, Li Zhang, Yanai Elazar, Niket Tandon, Marianna Apidianaki, Mrinmaya Sachan, and Chris Callison-Burch. 2025. Calibrating large language models with sample consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 19260–19268.

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 and Sara Tonelli. 2014. An analysis of causality between events and its relation to temporal information. In *Proceedings of COLING 2014,* the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2097–2106.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Konstantinos I Roumeliotis, Nikolaos D Tselikas, and Dimitrios K Nasiopoulos. 2023. Llama 2: Early adopters' utilization of meta's new open-source pretrained model.

Donald B Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

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.

Euysup Shim, Brad Carter, and Seongchan Kim. 2016. Request for information (rfi) management: a case study. In *Proceedings of the 52nd ASC Annual International Conference Proceedings, Provo, UT, USA*, pages 13–16.

Ya Su, Hu Zhang, Guangjun Zhang, Yujie Wang, Yue Fan, Ru Li, and Yuanlong Wang. 2025. Enhancing event causality identification with llm knowledge and concept-level event relations. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7403–7414.

Zhengwei Tao, Zhi Jin, Yifan Zhang, Xiancai Chen, Xiaoying Bai, Yue Fang, Haiyan Zhao, Jia Li, and Chongyang Tao. 2024. A comprehensive evaluation on event reasoning of large language models. *arXiv preprint arXiv:2404.17513*.

Xiaozhi Wang, Yulin Chen, Ning Ding, Hao Peng, Zimu Wang, Yankai Lin, Xu Han, Lei Hou, Juanzi Li, Zhiyuan Liu, and 1 others. 2022a. Maven-ere: A unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction. *arXiv preprint arXiv:2211.07342*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022b. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

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.

A GPT3.5-turbo vs non-LLMs methods

As shown in Table 7 and Table 8, GPT3.5-turbo achieves the optimal F1 score compared to non-LLM methods, regardless of whether the event pairs are Intra-Sentence or Inter-Sentence.

Method	**Intra-Sentence**		
	P (%)	R (%)	F1 (%)
DPJL	65.3	70.8	67.9
KEPT	50.0	68.8	57.9
CPATT	79.4	81.3	80.4
GPT3.5-turbo	78.5	83.3	**80.8**

Table 7: Comparison experiment results between GPT3.5-turbo and non-LLMs methods on intra-sentence event pairs of EventStoryLine

Method	**Inter-Sentence**		
	P (%)	R (%)	F1 (%)
CPATT	74.9	60.1	66.7
GPT3.5-turbo	69.4	78.6	**73.7**

Table 8: Comparison experiment results between GPT3.5-turbo and non-LLMs methods on inter-sentence event pairs of EventStoryLine

B Non-adversarial and adversarial prompts

Figures 8 and 9 illustrate the non-adversarial and adversarial prompts used for both causal and non-causal event pairs, respectively. The non-adversarial prompts are identical for both pair types. For the adversarial prompts, a false causal relationship is introduced to serve as a misleading context for the LLMs.

Passage: Lawyer: Lindsay Lohan checks into California rehab May 3, 2013| 8: 20 am Lindsay Lohan’ s attorney said Thursday the actress checked into a California rehab facility, but a state official said it is unlicensed to perform the type of restrictive in- patient treatment a judge required the actress to receive in a misdemeanor driving case. Mark Jay Heller told a judge that Lohan was settling in at Morningside Recovery and argued that the actress should be allowed to stay until a judge approves her <placement>.“ My client is ensconced in the bosom of that facility right now,” Heller argued after a prosecutor objected to Lohan’ s choice of facilities.“ She’ s in <rehab> right now. Nothing bad is going to happen.” Celebrity website TMZ reported Lohan was shopping at an electronics store while her attorney was in court and that she never entered Morningside. Superior Court Judge James R. Dabney agreed in the hearing that Lohan should remain at Morningside, although the actress’ whereabouts were unknown.
Question(Non-adversarial): Is there a causal relationship between <placement> and <rehab>?
Question(Adversarial): Is there a causal relationship between <placement> and <rehab>? **You may refer to the provided information that there is not a causal relationship between <problems> and <internet and mobile services>**

Figure 8: Non-adversarial and adversarial prompts on causal event pair.

Passage: An undersea telecommunications cable cut on Tuesday partially blocked internet and mobile services in Alexandria and some other governorates. The cable cut happened around the same time that a different cable, owned by Telecom Egypt, was cut that also affected services, Egyptian authorities said. Mahmoud al- Goweiny, a member of the National Telecommunications RegulatoryAuthority's board of directors, told Al- Masry Al- Youm that this particular undersea cable has frequently malfunctioned since late 2007. Ahmed Osama, deputy executive director of Telecom Egypt, said the cable passes through Egypt from Europe and reaches Asia. Osama said thecut's effect in Egypt was limited, and therehaven't been clear reports about a cause."Some internet users suffered <problems> with <internet and mobile services>, Osama noted. Osama attributed these problems to the Telecom Egypt cable cut, which affected Etisalat Egypt and LINKdotNET, two of Egypt's largest internet services providers. No further details were reported about the effects of the cuts.
Question(Non-adversarial): Is there a causal relationship between <problems> and <internet and mobile services>?
Question(Adversarial): Is there a causal relationship between <problems> and <internet and mobile services>? **You may refer to the provided information that there is a causal relationship between <problems> and <internet and mobile services>**

Figure 9: Non-adversarial and adversarial prompts on a non-causal event pair.