Generating Effective CoT Traces for Mitigating Causal Hallucination

Yiheng Zhao[1,†], **Jun Yan**[1]
[1]Concordia University, Montreal, Canada
† Corresponding author: yiheng.zhao@mail.concordia.ca

Abstract

Although large language models (LLMs) excel in complex reasoning tasks, they suffer from severe causal hallucination in event causality identification (ECI), particularly in smaller models ($\leq$1.5B parameters). A promising approach to address this issue is to fine-tune them with Chain-of-Thought (CoT) traces. However, there is currently a lack of CoT trace dataset available for ECI. In this paper, we first investigate the essential criteria that effective CoT traces should possess to mitigate causal hallucination in smaller models. We then design a pipeline to generate CoT traces that meet these criteria. Moreover, since there is currently no metric for quantifying causal hallucination, we also introduce a new metric, the Causal Hallucination Rate (CHR), to quantify causal hallucination, guide the formulation of effective CoT trace criteria, and validate the effectiveness of our pipeline. Our experiments show that fine-tuning with the CoT traces generated by our pipeline not only substantially reduces causal hallucination in smaller LLMs but also improves mean accuracy. Moreover, the fine-tuned models exhibit strong cross-dataset and cross-difficulty generalization, as well as robustness under misleading intervention prompts.

1 Introduction

Large language models (LLMs) have demonstrated impressive performance across a wide range of complex reasoning tasks, including mathematics, coding and deep research (Achiam et al., 2023; Yang et al., 2025; Guo et al., 2025). Despite these successes, recent studies reveal that they still suffer from severe causal hallucination (Gao et al., 2023; Cheng et al., 2025) in event causality identification (ECI), which aims to identify whether there is a causal relationship between two events in a text. Causal hallucination refers to a model's tendency to assume causal relationships between event pairs regardless of whether such relationships actually exist. We find this issue to be particularly pronounced in smaller models (Figure 1); as smaller models are essential for real-world deployment, where efficiency matter, mitigating causal hallucination in these smaller models is critical.

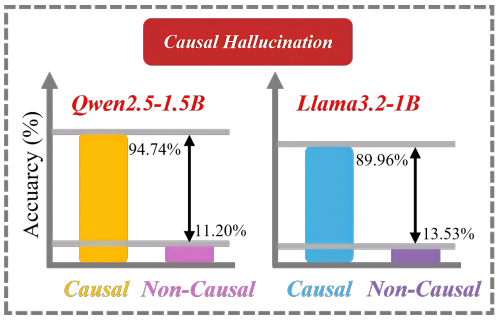

Figure 1: Illustration of causal hallucination in smaller models ($\leq$1.5B parameters) on EventStoryLine (Caselli and Vossen, 2017), where a large accuracy gap exists between causal and non-causal event pairs. Causal event pairs denote pairs with a causal relationship, while non-causal pairs do not. Accuracy is computed separately for each category as the proportion of correctly predicted instances.

A promising direction for addressing this issue is to fine-tune them using Chain-of-Thought (CoT) traces (Li et al., 2026a,b). However, current ECI datasets (Mirza and Tonelli, 2016; Caselli and Vossen, 2017; Wang et al., 2022; Lai et al., 2022) contain only binary labels, which lack intermediate reasoning steps and are therefore inadequate for reducing causal hallucination. Moreover, existing LLM-based ECI studies primarily focus on prompt design and inference-time strategies, and there is currently no CoT trace dataset available for ECI.

In this paper, we first investigate the essential criteria that effective CoT traces should satisfy to mitigate causal hallucination in smaller models, focusing on three factors: perplexity, CoT trace length, and the distribution gap between CoT traces and the target models being fine-tuned. Prior works (Zhang

Findings of the Association for Computational Linguistics: ACL 2026, pages 5352–5362

et al., 2025; Li et al., 2025; Yang et al., 2024) have proposed several criteria: (1) the one with the lowest perplexity should be preferred when multiple traces are available for a sample; (2) smaller models struggle to learn from long CoT traces; and (3) rewriting CoT traces with the target model to reduce the distribution gap helps to learn more effectively. However, we find that these criteria do not hold for ECI. Our analysis reveals three new findings: (1) perplexity is not a reliable selection criterion, as trace length has a stronger impact; (2) smaller models actually benefit more from long CoT traces; and (3) rewriting CoT traces is beneficial only when it does not increase perplexity.

Based on these findings, we define that effective CoT traces for mitigating causal hallucination in smaller models should satisfy two criteria: (1) providing sufficiently long reasoning traces enriched with semantic explanations and intermediate steps; and (2) maintaining a small distribution gap and alignment with the target model. We then design the first data generation pipeline for ECI that generates fine-tuning data that meet these criteria, enabling smaller models to identify causal relationships more reliably. In addition, to properly quantify causal hallucination, we introduce the Causal Hallucination Rate (CHR), a new metric that quantifies causal hallucination, guides the formulation of effective CoT trace criteria, and validates the effectiveness of our pipeline.

Experiments on EventStoryLine reveal that fine-tuning with the CoT traces generated by our pipeline yields significant reductions in causal hallucination while improving mean accuracy. Ablation analyses highlight the importance of rewriting CoT traces to match the target model's distribution. The fine-tuned models further exhibit strong cross-dataset generalization on Causal-TimeBank and MAVEN-ERE, as well as cross-difficulty generalization from sentence-level to document-level settings. Finally, robustness tests show that the models resist misleading intervention prompts, indicating that they have acquired more stable and reliable causal reasoning behavior. Together, these results demonstrate the effectiveness and broad applicability of our proposed pipeline for mitigating causal hallucination in smaller language models.

2 Related Work

ECI with LLMs Existing LLM-based approaches primarily focus on inference-time strategies to enhance ECI performance of LLMs. Dr.ECI (Cai et al., 2025) constructs causal prompts grounded in causal-inference principles; MRBalance (Zou et al., 2025) employs multi-agent debates with role-specific prompts to elicit complementary evidence; and MEFA (Zeng et al., 2025) decomposes ECI into six sub-tasks and designs dedicated prompts for each. In addition, Zhao et al. (Zhao and Yan, 2025) investigate whether debate can improve the factuality and reasoning of LLMs on ECI.

Building Effective CoT Traces Recent studies have explored how to construct effective fine-tuning data for improving LLM reasoning. However, these methods have been developed primarily in mathematical reasoning rather than for ECI. Zhang et al. (Zhang et al., 2025) propose a perplexity-based selection strategy that chooses instruction–response pairs best aligned with the target model's own distribution. Li et al. (Li et al., 2025) show that small models cannot effectively learn from large teachers or long, complex reasoning traces, and thus introduce a mixed-distillation strategy for constructing more suitable fine-tuning data. Yang et al. (Yang et al., 2024) further propose a rewriting-based self-distillation approach that generates model-aligned training data to narrow the distribution gap between downstream fine-tuning data and the target model.

3 Methodology

In this section, we first introduce the proposed causal hallucination metric. We then present the criteria for constructing effective CoT traces and describe our data generation pipeline used to produce CoT traces that meet the criteria.

3.1 Causal Hallucination Rate (CHR)

To mitigate causal hallucination, we must first quantify it. However, there is currently no metric for measuring causal hallucination in ECI. To address this gap, we propose CHR. This metric not only quantifies causal hallucination but also guides the formulation of effective CoT trace criteria and evaluates the effectiveness of our pipeline. The CHR is formulated as follows:

$$\text{CHR} = \text{Acc}_{\text{causal}} - \text{Acc}_{\text{non-causal}}, \quad (1)$$

where $\text{Acc}_{\text{causal}}$ and $\text{Acc}_{\text{non-causal}}$ denote the model's accuracy on causal and non-causal event pairs, respectively. A CHR value greater than 0 indicates causal hallucination, and smaller values

reflect weaker hallucination. If CHR becomes negative, it means the model tends to overpredict non-causal relations, with larger deviations indicating stronger bias in that direction.

3.2 Effective CoT Trace Criteria

To investigate the criteria that effective CoT traces should satisfy to mitigate causal hallucination in smaller models, we first conduct a comprehensive analysis across three key factors: perplexity, CoT trace length, and the distribution gap between CoT traces and target models.

Perplexity Prior work has shown that when multiple CoT traces exist for a sample, the one with the lowest perplexity should be selected. To verify whether this criterion holds in ECI, we conduct the following experiments. We first use two models (Qwen2.5-7B (Hui et al., 2024), and Llama3.1-8B (Dubey et al., 2024)) to generate CoT traces and retain only those that produce correct answers. We choose models at the 7B–8B scale to avoid an excessively large distribution gap from the target models. We then fine-tune Qwen2.5-1.5B using (1) the traces from each model and (2) the lowest-perplexity traces selected across models. As shown in Table 1, although the perplexity-based selection achieves the lowest perplexity, it does not yield the lowest CHR. Instead, longer traces (e.g., those generated by Llama) lead to substantially lower causal hallucination, indicating that length has a stronger influence than perplexity when multiple traces are available. To understand why longer traces are more effective, we examine their content and find that they contain richer semantic explanations, suggesting that semantic elaboration in CoT traces is the primary factor mitigating causal hallucination. Here, semantic explanations refer to additional explanations about the passage. To further validate this, we remove Qwen-generated traces from the perplexity-based selection, as they are shorter and contain fewer semantic explanations. As shown in Table 1, the resulting CHR remains nearly unchanged, confirming that semantic explanation plays a crucial role. We report the following key takeaway.

Takeaway for Perplexity: *Perplexity is not a suitable selection criterion for ECI; length has a greater influence, with semantically richer CoT traces more effective at mitigating causal hallucination.*

Setting	PPL	Mean Token	CHR
Vanilla	-	-	84.55
Qwen	2.77	242	60.30
Llama	2.35	317	34.12
Perplexity	2.28	302	39.26
LongOnly	2.28	-	39.84

Table 1: Results validating whether the perplexity-based selection strategy is effective. “PPL" denotes perplexity, and “Mean Token" represents the average token length of each CoT trace. “Vanilla" denotes Qwen2.5-1.5B without CoT fine-tuning. “Qwen" and “Llama" use CoT traces generated by Qwen2.5-7B and Llama3.1-8B, respectively. “Perplexity" selects the lowest-perplexity trace across models, while “LongOnly" removes shorter Qwen traces from the perplexity-based selection set.

CoT Trace Length Previous study suggests that smaller models struggle to learn from longer CoT traces. However, our perplexity-based selection analysis contradicts this observation. Motivated by this observation, we further analyze the effect of trace length by fine-tuning Qwen2.5-1.5B using CoT traces of three different lengths. From Table 2, we observe that CHR consistently decreases as CoT trace length increases, indicating that smaller models can effectively learn from longer reasoning sequences. In particular, traces containing rich semantic explanations and reasoning steps achieve the lowest CHR. This indicates that reasoning steps are essential for reducing causal hallucination, alongside semantic explanations. We summarize this observation as the following key takeaway.

Takeaway for CoT Trace Length: *In ECI, smaller models benefit more from longer CoT traces; beyond richer semantic explanations, more reasoning steps are also crucial for mitigating causal hallucination.*

Distribution Gap Although rewriting CoT traces using the target model has been shown to reduce the distribution gap and improve fine-tuning effectiveness, it remains unclear whether this strategy is beneficial for ECI, and at which CoT trace length rewriting yields the best performance. To investigate this, we evaluate the rewriting strategy, adopting the rewriting prompt template from prior work, across CoT traces of different lengths, as summarized in Table 3. We observe that rewriting improves performance only for the shortest and longest CoT traces, while it increases CHR for

Setting	Mean Token	CHR
Vanilla	-	84.55
Qwen	242	59.79
Llama	317	34.68
LlamaThinking	482	30.60

Table 2: Comparison of CoT traces with different lengths. “Mean Token" represents the average token length of each CoT trace. “Vanilla" denotes Qwen2.5-1.5B without CoT fine-tuning. “Qwen” uses concise traces from Qwen2.5-7B, “Llama” uses longer traces from Llama3.1-8B, and “LlamaThinking” uses even longer CoT traces generated by Llama3.1-8B, guided by few-shot examples produced by Qwen3-235B-A22B (Thinking) (Yang et al., 2025).

Setting	PPL	CHR
Vanilla	-	84.55
Qwen	2.76	59.79
QwenR	2.75	55.98
Llama	2.35	34.68
LlamaR	2.36	42.81
LlamaR (Our prompt)	2.34	33.77
LlamaThinking	3.43	30.60
LlamaThinkingR	3.42	24.42

Table 3: Effect of applying rewriting on CoT traces of different lengths. “Qwen”, “Llama”, and “LlamaThinking” follow the same definitions as in Table 2. “QwenR”, “LlamaR”, and “LlamaThinkingR” denote applying the prior study of rewriting prompt, while “LlamaR (our prompt)” applies our revised rewriting prompt.

medium-length traces. Further analysis shows that rewriting reduces perplexity for both short and long traces but increases it for medium-length traces, suggesting that the rewriting strategy is effective only when it does not increase perplexity. To verify this, we design a revised rewriting prompt, which successfully reduces perplexity and lowers CHR. Notably, rewriting the longest traces with richer semantic explanations and reasoning steps still yields the lowest CHR.

Takeaway for Distribution Gap: *Rewriting strategy to reduce the distribution gap mitigates causal hallucination only when perplexity does not increase; the best performance occurs when it is applied to CoT traces enriched with semantic explanations and reasoning steps.*

Based on the findings and takeaways discussed above, we formulate the following two essential criteria that effective CoT traces must meet to mitigate causal hallucination in smaller models.

- **Criterion I:** *CoT traces must contain sufficient semantic explanations and reasoning steps.*
- **Criterion II:** *The distribution gap between CoT traces and the target models should be reduced without increasing perplexity.*

In all analyses reported above, the training set is drawn from the first 16 topics of EventStoryLine, while the test set comprises the remaining four topics. The topics are processed in ascending order. To ensure fairness, the training samples are held identical across the scope of each factor analysis. In addition, the rewriting prompt template follows that of prior work, and our designed prompt is presented below.

Official instruction: Below is an instruction that describes a task along with a reference answer. Using the reference answer as a guide, write your own response.

Our designed instruction: Rewrite the following response based on the given instruction and reference answer. Using the reference answer as a guide, write your own response.

Instruction: `{instruction}`

Reference: `{original CoT traces}`

Response:

3.3 CoT Trace Generation Pipeline

Based on the two criteria above, we design a two-step CoT trace generation pipeline tailored for mitigating causal hallucination in smaller models, as illustrated in Figure 2, while maintaining low cost, as it primarily relies on Llama3.1-8B with few-shot demonstrations. In the first step, our goal is to produce CoT traces that meet **Criterion I** by providing rich semantic explanations and reasoning steps. To achieve this, we employ Qwen3-235B-A22B (Thinking) to construct two few-shot demonstrations, one for a causal event pair and the other for a non-causal event pair. These demonstrations are then used to prompt Llama3.1-8B to generate the desired CoT traces, from which we retain only

those that produce correct answers. In the second step, the goal is to reduce the distribution gap between the CoT traces generated in the first step and the target model. To achieve this, we rewrite the CoT traces using the target model itself and verify that the rewritten traces do not increase perplexity, thereby ensuring that **Criterion II** is satisfied. Since our earlier analysis shows that the rewriting prompt from prior work does not increase perplexity for the targeted traces with rich semantic explanations and reasoning steps, we adopt this prompt in our pipeline. If a rewritten trace produces an incorrect answer, we retain the original trace from the first step instead of using the rewritten trace.

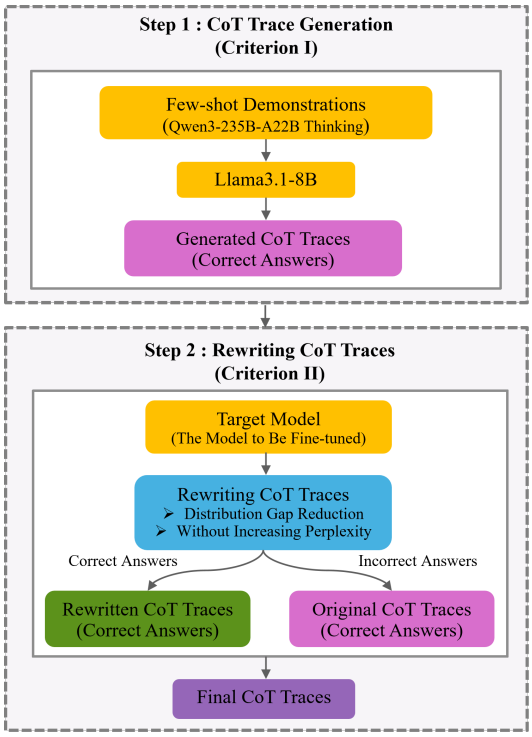

Figure 2: Overview of the proposed CoT trace generation pipeline.

4 Experiments

In this section, we first introduce the datasets, implementation details, and evaluation metrics. We then present experimental results, including comparative analysis, ablation studies, and investigations of generalization and robustness.

4.1 Datasets

Our evaluation primarily utilizes three datasets: EventStoryLine, Causal-TimeBank, and MAVEN-ERE, employing different configurations for distinct experimental goals.

For comparison, ablation, and robustness analyses, we consistently use sentence-level samples from the EventStoryLine dataset. Sentence-level means the input is strictly a single sentence. Across these three experiments, we use the same setup: we first sort the 20 topics in ascending order by topic ID, and subsequently perform 5-fold cross-validation for training and testing. It is important to note that, although the test set (the holdout set for each fold) remains complete, due to the generative capability limitations of models such as Llama 3.1-8B, the sample set used for model training includes only the subset for which the model successfully generated valid outputs.

For cross-dataset generalization testing, we use sentence-level samples from the Causal-TimeBank and MAVEN-ERE datasets as our test sets. Furthermore, in the cross-difficulty generalization experiment, we adopt document-level samples from EventStoryLine as our test set. Document-level means the input is a full passage, and we focus exclusively on inter-sentence event pairs, i.e., events that occur in different sentences. For the causal event pairs, we use all available samples; for the non-causal event pairs, we randomly sample 10,000 pairs. More detailed dataset statistics are provided in the Appendix A.

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.

Intra-sentence event pairs: (flattening, killing)

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 seismological center said the quake measured magnitude-5.9, but the U.S. Geological Survey in Golden, Colo., said it was a magnitude-6.1 temblor...

Inter-sentence event pairs: (flattening, quake), (killing, quake)

Additionally, the two examples above illustrate the distinction between sentence-level and document-level ECI settings. In the sentence-level setting, the input is restricted to a single sentence,

and the model evaluates only intra-sentence event pairs in which both events occur within the same sentence. In contrast, the document-level setting takes a passage as input and focuses on inter-sentence event pairs, where the two events appear in different sentences.

4.2 Implementation Details and Metrics

Implementation Details All of our training is based on the SFTTrainer implementation within the Transformer Reinforcement Learning (TRL) framework. The specific training hyperparameters are configured as follows: the per-device batch size is 1, the gradient accumulation steps are 8, and the total training is conducted for 1 epoch. We set the learning rate to $2e-4$ and use a cosine-annealed scheduler. To enhance training efficiency and conserve computational resources, we adopt the Low-Rank Adaptation (LoRA) fine-tuning technique. The key parameters for the LoRA configuration are: the LoRA rank is 8, the scaling factor is 16, and the dropout rate is 0.05. More comprehensive training and implementation details are presented in the Appendix B. Moreover, in all experiments, we fix the decoding temperature to 0. This deterministic setting eliminates sampling randomness and ensures that all inference results are fully reproducible across runs. Moreover, in this paper, we fix the decoding temperature to 0 for all experiments.

Metrics We use our proposed CHR to evaluate causal hallucination. Additionally, we use mean accuracy (mAcc) to assess the models' overall performance. The mAcc is defined as the average of the accuracy on causal event pairs and the accuracy on non-causal event pairs.

4.3 Results

Main Results To verify whether our proposed CoT trace generation pipeline can effectively reduce the causal hallucination in smaller language models, we test the effectiveness of our pipeline on two models: Qwen2.5-1.5B and Llama3.2-1B. Table 4 presents the comparative results on EventStoryLine, contrasting the model fine-tuned on CoT trace data generated by our pipeline against various LLMs and existing baseline methods. For a fairer comparison with existing approaches, we re-implement these methods by replacing their original base LLMs with Qwen2.5-1.5B, which we use in our experiments, thereby allowing us to assess whether these methods can effectively reduce causal hallucination in smaller models.

From the Table 4, we can observe that existing methods such as Dr.ECI, MuTQA, and MRBalance do not reduce causal hallucination of Qwen2.5-1.5B. In particular, Dr. ECI with Qwen2.5-1.5B yields a CHR of 100.00%, indicating that the method predicts a causal relation for every sample and increases causal hallucination. We also observe that prompt-based strategies, including zero-shot CoT prompting and ICL prompting, as well as fine-tuning with binary labels, fail to substantially mitigate hallucination. In contrast, fine-tuning the models on the CoT trace data generated by our pipeline yields a substantial reduction in hallucination for both models. Qwen2.5-1.5B achieves a CHR of 6.26% and Llama3.2-1B reaches 9.14%, corresponding to absolute reductions of 77.28% and 67.29% compared with the original models. After applying our method, the hallucination degree also becomes lower than that of much larger LLMs, including GPT-4, Llama3.1-8B and Qwen3-30B-A3B. Moreover, our pipeline not only suppresses causal hallucination but also improves overall reasoning accuracy, with mAcc rising from 52.97% to 66.00% for Qwen2.5-1.5B and from 55.58% to 63.44% for Llama3.2-1B. Further comparisons between model outputs before and after fine-tuning can be found in Appendix C. To provide a more comprehensive evaluation, we also report additional metrics, including FPR, TNR, and MCC, with results presented in Appendix D.

Ablation Results We further conduct an ablation study to examine the contribution of the rewriting step in our pipeline. We compare models fine-tuned on the original CoT traces, which contain rich semantic explanations and reasoning steps directly generated by Llama3.1-8B, against models fine-tuned on rewritten CoT traces produced by the target model itself. As shown in Table 5, removing the rewriting step substantially weakens the ability to reduce causal hallucination. For Qwen2.5-1.5B, CHR decreases from 83.54% to 30.39% without rewriting, but further to 6.26% with rewriting. A similar pattern is observed for Llama3.2-1B, where CHR decreases from 76.43% to 17.13% without rewriting, and to 9.14% with rewriting. Moreover, rewriting yields notable improvements in mAcc for both models. Importantly, as shown in Table 6, the rewriting step does not increase perplexity. These findings demonstrate that reducing the data distribution gap between the CoT traces and the target model, without increasing perplexity, is crucial for effectively suppressing causal hallucination.

Methods	CHR	mAcc
GPT3.5-turbo	43.43	57.36
GPT4	53.30	51.40
Llama3.1-8B	60.59	58.97
Qwen3-30B-A3B	60.27	58.76
Dr.ECI	100.00	50.00
MuTQA	46.03	28.82
MRBalance	82.45	59.10
Qwen2.5-1.5B (Vanilla)	83.54	52.97
Qwen2.5-1.5B (CoT)	69.77	51.48
Qwen2.5-1.5B (ICL)	69.57	55.32
Qwen2.5-1.5B (Binary)	66.67	56.74
Qwen2.5-1.5B (Our)	**6.26**	**66.00**
Llama3.2-1B (Vanilla)	76.43	55.58
Llama3.2-1B (CoT)	60.43	50.74
Llama3.2-1B (ICL)	58.42	46.43
Llama3.2-1B (Binary)	56.55	55.68
Llama3.2-1B (Our)	**9.14**	**63.44**

Table 4: Comparison results on EventStoryLine. "Vanilla" refers to the original model; "CoT" denotes zero-shot chain-of-thought prompting; "ICL" denotes in-context learning with the same exemplars as those used to elicit long CoT traces from Llama3.1-8B in our pipeline; "Binary" refers to fine-tuning with binary labels; and "Our" indicates fine-tuning using CoT trace data generated by our proposed pipeline.

Methods	CHR	mAcc
Qwen2.5-1.5B (Vanilla)	83.54	52.97
Qwen2.5-1.5B (w/o rewriting)	23.39	56.51
Qwen2.5-1.5B (w rewriting)	**6.26**	**66.00**
Llama3.2-1B (Vanilla)	76.43	55.58
Llama3.2-1B (w/o rewriting)	17.13	55.51
Llama3.2-1B (w rewriting)	**9.14**	**63.44**

Table 5: Ablation results on EventStoryLine. "Vanilla" refers to the original model; "w/o rewriting" denotes fine-tuning without rewriting the initial CoT traces that contain rich semantic explanations and reasoning steps. "w rewriting" denotes fine-tuning after rewriting these CoT traces using the target model.

Methods	PPL
Qwen2.5-1.5B (w/o rewriting)	3.23
Qwen2.5-1.5B (w rewriting)	**3.18**
Llama3.2-1B (w/o rewriting)	3.99
Llama3.2-1B (w rewriting)	**3.65**

Table 6: Perplexity comparison with and without CoT trace rewriting.

4.4 Generalization

Cross-dataset To evaluate the generalizability of the models fine-tuned, we further conduct cross-dataset experiments by applying the models fine-tuned on EventStoryLine to two external benchmarks: Causal-TimeBank and MAVEN-ERE. As shown in Table 7, the models fine-tuned achieve substantial reductions in causal hallucination alongside notable gains in predictive accuracy across both datasets. For Causal-TimeBank, the CHR of Qwen2.5-1.5B drops significantly from 84.55% to 11.37%, and its mAcc increases from 51.53% to 66.79%. Similarly, the CHR of Llama3.2-1B decreases from 67.79% to 4.26%, while mAcc improves from 45.85% to 61.51%. For MAVEN-ERE, Qwen2.5-1.5B reduces CHR from 84.69% to 11.13% and raises mAcc from 52.50% to 64.92%. Llama3.2-1B shows consistent improvements, with CHR dropping from 73.82% to 7.84% and mAcc rising from 50.79% to 62.86%. These results demonstrate that using our pipeline can generalize effectively to unseen corpora.

Cross-difficulty To further evaluate whether the models fine-tuned using the CoT trace data generated by our pipeline can generalize to more challenging scenarios, we additionally test them on the document-level ECI, where the input is a full passage, and the target event pairs occur across different sentences. This setting is substantially more difficult than the training-distribution sentence-level task due to longer contexts and more complex discourse structures. Using the models fine-tuned only on the first fold of EventStoryLine, we observe strong generalization improvements on this harder setting, as shown in Table 8. For Qwen2.5-1.5B, CHR decreases significantly from 54.52% to 1.41%, while mAcc increases from 59.76% to 69.85%. For Llama3.2-1B, CHR decreases from 93.94% to 4.51%, while mAcc increases from 51.62% to 59.45%. These results demonstrate that the proposed pipeline not only suppresses causal hallucination within the training distribution but also generalizes effectively to substantially more challenging document-level ECI tasks.

4.5 Robustness

To assess whether these models fine-tuned, truly learn causal relationships in ECI, we also conduct a robustness evaluation. After training, we inject incorrect intervention into the prompt to mislead these models: for each causal event pair, we tell them that no causal relationship exists, and for each non-causal pair, we tell them that a causal relationship does exist. As shown in Table 9, the mAcc of both models remains largely unchanged compared with the non-intervention setting, indi-

Methods	CHR	mAcc
Causal-TimeBank		
Qwen2.5-1.5B (Vanilla)	84.55	51.53
Qwen2.5-1.5B (Our)	**11.37**	**66.79**
Llama3.2-1B (Vanilla)	67.79	45.85
Llama3.2-1B (Our)	**4.26**	**61.51**
MAVEN-ERE		
Qwen2.5-1.5B (Vanilla)	84.69	52.50
Qwen2.5-1.5B (Our)	**11.13**	**64.92**
Llama3.2-1B (Vanilla)	73.82	50.79
Llama3.2-1B (Our)	**7.84**	**62.86**

Table 7: Cross-dataset results on Causal-TimeBank and MAVEN-ERE. "Vanilla" refers to the original model without any fine-tuning. "Our" denotes the model fine-tuned using the CoT trace data from the first fold of the five-fold cross-validation on EventStoryLine.

Methods	CHR	mAcc
Qwen2.5-1.5B (Vanilla)	54.52	59.76
Qwen2.5-1.5B (Our)	**1.41**	**69.85**
Llama3.2-1B (Vanilla)	93.94	51.62
Llama3.2-1B (Our)	**4.51**	**59.45**

Table 8: Document-level results on inter-sentence event pairs of EventStoryLine. "Vanilla" refers to the original model without any fine-tuning. "Our" denotes the model fine-tuned using the CoT trace data from the first fold of the five-fold cross-validation on EventStoryLine.

cating that the fine-tuned models are robust and do not simply follow misleading instructions. The slight decrease in CHR arises because incorrect intervention prompts disrupt correct predictions on causal pairs and flip some incorrect predictions on non-causal pairs, leading to a small net shift in the metric. The specific non-intervention and incorrect-intervention prompts used for causal and non-causal event pairs are shown below.

Instruction (Non-intervention for all event pairs): Is there a causal relationship between `<event A>` and `<event B>`?

Instruction (Incorrect intervention for causal event pairs): Is there a causal relationship...? *You may refer to the provided information that there is no causal relationship...*

Instruction (Incorrect intervention for non-causal event pairs): Is there a causal relationship...? *You may refer to the provided information that there is a causal relationship...*

Methods	CHR	mAcc
Qwen2.5-1.5B (Vanilla)	83.54	52.97
Qwen2.5-1.5B (Our)	6.26	66.00
Qwen2.5-1.5B (Interv.)	2.26	63.08
Llama3.2-1B (Vanilla)	76.43	55.58
Llama3.2-1B (Our)	9.14	63.44
Llama3.2-1B (Interv.)	4.34	60.34

Table 9: Robustness results on EventStoryLine. "Vanilla" refers to the original model without fine-tuning. "Our" denotes the model fine-tuned with our pipeline and evaluated without any intervention prompts. "Interv." represents the same fine-tuned model evaluated under incorrect intervention prompts during inference.

5 Conclusions

In this paper, we investigate how to mitigate causal hallucination in smaller LLMs for ECI. We first introduce the CHR, a simple yet informative metric that quantifies causal hallucination and guides both the analysis and evaluation of our methods. Through a systematic study of perplexity, CoT trace length, and distribution gap, we identify two key criteria for constructing effective CoT traces. Based on these insights, we develop a CoT trace generation pipeline that produces effective fine-tuning traces for smaller models, thereby substantially reducing their causal hallucination. Experiments on EventStoryLine show that fine-tuning with the CoT traces generated by our pipeline substantially reduces causal hallucination in smaller LLMs while also improving mean accuracy. The fine-tuned models further exhibit strong cross-dataset generalization to Causal-TimeBank and MAVEN-ERE, cross-difficulty generalization from sentence-level to document-level ECI, and robustness to misleading intervention prompts.

Limitations

Despite the effectiveness of our approach, several limitations remain. First, our pipeline is primarily designed for smaller models ($\leq$1.5B parameters). While we demonstrate substantial gains in this regime, larger models are also known to exhibit causal hallucination, and extending our pipeline to mitigate hallucination in large-scale models remains an important direction for future work. Second, our evaluation of generalization focuses on cross-dataset and cross-difficulty settings within ECI. However, causal hallucination is also prevalent in other causal reasoning tasks, such as causal discovery, which are not explored in this work. Fi-

nally, although our robustness analysis shows that the fine-tuned models are resistant to deliberately misleading intervention prompts, the intervention design is relatively simple and does not cover more adversarial scenarios that may arise in real-world applications.

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*.

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, Vancouver, Canada. Association for Computational Linguistics.

Qing Cheng, Zefan Zeng, Xingchen Hu, Yuehang Si, and Zhong Liu. 2025. A survey of event causality identification: Taxonomy, challenges, assessment, and prospects. *ACM Computing Surveys*, 58(3):1–37.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

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

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*.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, and 1 others. 2024. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*.

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.

Bo Li, Mingda Wang, Shikun Zhang, and Wei Ye. 2026a. Instruction data selection via answer divergence. *Preprint*, arXiv:2604.10448.

Bo Li, Shikun Zhang, and Wei Ye. 2026b. Data selection for multi-turn dialogue instruction tuning. *Preprint*, arXiv:2604.07892.

Yuetai Li, Xiang Yue, Zhangchen Xu, Fengqing Jiang, Luyao Niu, Bill Yuchen Lin, Bhaskar Ramasubramanian, and Radha Poovendran. 2025. Small models struggle to learn from strong reasoners. *arXiv preprint arXiv:2502.12143*.

Paramita Mirza and Sara Tonelli. 2016. Catena: Causal and temporal relation extraction from natural language texts. In *The 26th international conference on computational linguistics*, pages 64–75. ACL.

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

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zhaorui Yang, Tianyu Pang, Haozhe Feng, Han Wang, Wei Chen, Minfeng Zhu, and Qian Liu. 2024. Self-distillation bridges distribution gap in language model fine-tuning. *arXiv preprint arXiv:2402.13669*.

Zefan Zeng, Xingchen Hu, Qing Cheng, Weiping Ding, Wentao Li, and Zhong Liu. 2025. Zero-shot event causality identification via multi-source evidence fuzzy aggregation with large language models. *arXiv preprint arXiv:2506.05675*.

Dylan Zhang, Qirun Dai, and Hao Peng. 2025. The best instruction-tuning data are those that fit. *arXiv preprint arXiv:2502.04194*.

Yiheng Zhao and Jun Yan. 2025. Can llms identify event causality more accurately through debate? a systematic assessment of llms' factuality and reasoning. In *2025 IEEE Symposium on Trustworthy, Explainable and Responsible Computational Intelligence (CITREx)*, pages 1–8. IEEE.

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

Training Set Statistics			
Fold	**Total**	**Pos**	**Neg**
1	7021	2179	4842
2	7491	2122	5369
3	7378	2128	5250
4	7629	2213	5416
5	6697	2094	4603
Test Set Statistics			
Fold	**Total**	**Pos**	**Neg**
1	4108	630	3478
2	3244	648	2596
3	3326	696	2630
4	2404	528	1876
5	4916	682	4234

Table 10: Statistics of the five-fold split on EventStoryLine for both the training and test sets. "Pos" denotes causal samples, while "Neg" denotes non-causal samples.

Dataset	**Total**	**Pos**	**Neg**
Causal-TimeBank	7656	298	7358
MAVEN-ERE	19642	3359	16283

Table 11: Statistics of the Causal-TimeBank and MAVEN-ERE datasets. "Pos" denotes causal event pairs, and "Neg" denotes non-causal event pairs.

A Data Statistics

This appendix section summarizes the dataset statistics used across our experiments. Table 10 reports the five-fold split of the EventStoryLine dataset. Table 11 presents the test-set distributions of two widely used benchmarks, Causal-TimeBank and MAVEN-ERE. Finally, Table 12 provides the statistics of the document-level (inter-sentence) subset of EventStoryLine.

B Training Hyperparameters

This appendix section provides the training details used in our experiments. Table 13 summarizes the training hyperparameters and LoRA configuration used in our training process.

C Visualization of Model Outputs

This appendix section aims to qualitatively compare model outputs before and after fine-tuning. Figure 3 presents an example for Qwen2.5-1.5B, while Figure 4 shows a corresponding example for LLaMA-3.2-1B, both with non-causal ground-truth labels. As shown in the following, the original models incorrectly infer a causal link, whereas the fine-tuned models correctly identify the absence of causality and provides a more faithful explanation.

Dataset	**Total**	**Pos**	**Neg**
EventStoryLine	13000	3000	10000

Table 12: Document-level (inter-sentence) subset of the EventStoryLine. "Pos" denotes causal event pairs, and "Neg" denotes non-causal event pairs.

General Training Hyperparameters	
Training Framework	TRL
Epochs	1
Per-device Batch Size	1
Gradient Accumulation Steps	8
Effective Batch Size	8
Learning Rate	2×10^{-4}
Scheduler	Cosine Annealing
FP16 Training	Yes
Gradient Checkpointing	Enabled
Packing (Sequence Packing)	Enabled
LoRA Configuration	
LoRA Rank (r)	8
Scaling Factor (α)	16
LoRA Dropout	0.05

Table 13: Training hyperparameters and LoRA configuration for fine-tuning.

D Additional Evaluation Results with Extended Metrics

To provide a more comprehensive evaluation, we additionally report FPR, TNR, and MCC in this appendix. From Table 14, we observe that fine-tuning with CoT traces generated by our pipeline leads to significantly lower FPR and higher TNR and MCC, indicating reduced causal hallucination and improved overall prediction quality.

Qwen2.5-1.5B (Vanilla)	
FPR	88.80
TNR	11.20
MCC	0.075
Qwen2.5-1.5B (Ours)	
FPR	**37.13** (↓)
TNR	**62.87** (↑)
MCC (	**0.247** (↑)

Table 14: Additional evaluation results with FPR, TNR, and MCC on EventStoryLine. Lower FPR is better, while higher TNR and MCC are better. "Vanilla" refers to the original model and "Our" indicates fine-tuning using CoT trace data generated by our proposed pipeline.

Text: `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.`

Question: Is there a causal relationship between <argued> and <approves>? Ensuring your complete explanation comes first, then followed by the final answer in the strict format: [Final Answer: Yes/No].

Output before fine-tuning: The passage indicates that Mark Jay Heller argued that Lohan should be allowed to stay at Morningside Recovery until a judge approves her placement. This implies that there is a causal relationship between the two actions: Heller's argument and the judge's approval. **[Final Answer: Yes]**

Output after fine-tuning: The passage states that Mark Jay Heller argued that the actress should be allowed to stay until a judge approves her placement. While the argument expresses a request, it does not cause the judicial approval. The approval is a separate, future event not shown to be caused by the argument. **[Final Answer: No]**

Figure 3: Qualitative comparison of Qwen2.5-1.5B outputs before and after fine-tuning on a non-causal event pair.

Text: `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.`

Question: Is there a causal relationship between <argued> and <approves>? Ensuring your complete explanation comes first, then followed by the final answer in the strict format: [Final Answer: Yes/No].

Output before fine-tuning: To answer the question accurately, we need to analyze the relationship between the words argued and approves. The word argued is a verb that means to express or state an opinion or point of view, often in a persuasive or argumentative manner. It implies a level of intensity or conviction in one's stance or position. On the other hand, approves is a verb that means to give consent, approval, or acceptance of something, often in a formal or official capacity. Given the context of the passage, Mark Jay Heller is arguing that Lohan should be allowed to stay at Morningside Recovery, which implies that he is advocating for her to be approved or permitted to stay. Therefore, there is a causal relationship between argued and approves, as the former is a means of expressing an opinion or point of view, and the latter is a result of that expression. **[Final Answer: Yes]**

Output after fine-tuning: The passage states that Mark Jay Heller "argued that the actress should be allowed to stay until a judge approves her placement." Here, "argued" refers to Heller's action of presenting a case to the judge, while "approves" describes a future action by the judge that is conditional on the stay. The phrase "until a judge approves" establishes a temporal condition (i.e., the stay is permitted only up to the point of approval), not a causal link. Heller's argument is a request for temporary permission pending the judge's independent decision. The approval is portrayed as a separate, not as a consequence of Heller's argument. **[Final Answer: No]**

Figure 4: Qualitative comparison of Llama3.2-1B outputs before and after fine-tuning on a non-causal event pair.