PIPER: Benchmarking and Prompting Event Reasoning Boundary of LLMs via Debiasing-Distillation Enhanced Tuning

Zhicong Lu[1,2*], **Changyuan Tian** [1,2*], **Peiguang Li**[3†], **Li Jin**[1†], **Sirui Wang**[3,4]
Wei Jia[1,2], **Ying Shen**[5], **Guangluan Xu**[1]

[1] Key Laboratory of Target Cognition and Application Technology (TCAT), Aerospace Information Research Institute, Chinese Academy of Sciences
[2] School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences
[3] Meituan [4] Department of Automation, Tsinghua University
[5] Department of Finance, China University of Geosciences
{nazaritelzc, changyuanmails, jinlimails}@gmail.com

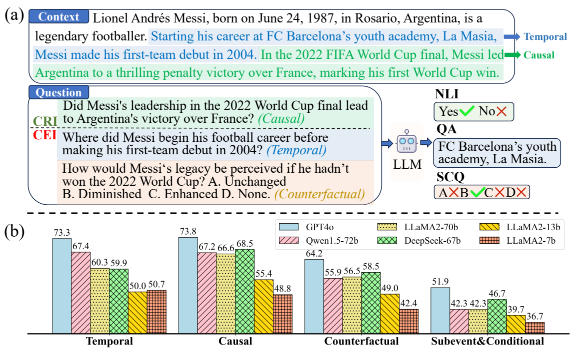

Figure 1: (a) shows examples of partial relational types and reasoning formats in the Contextual inter-event Relation Inference (CRI) and Contextual logical Event Inference (CEI) tasks within PIPER. (b) Overview of the evaluation results for mainstream aligned LLMs.

Abstract

While Large Language Models (LLMs) excel in diverse domains, their validity in event reasoning remains underexplored. Most existing works merely stagnate at assessing LLMs' event reasoning with a single event relational type or reasoning format, failing to conduct a complete evaluation and provide a practical solution for capability enhancement. In this paper, we propose **PIPER**, the first comprehensive benchmark for **P**robing **I**nto the **P**erformance boundary of LLMs in **E**vent **R**easoning. Motivated by our evaluation observations and error patterns analysis, we meticulously craft 10K diverse instruction-tuning demonstrations to alleviate event reasoning-oriented data scarcity. Additionally, a novel **D**ebiasing and **D**istillation-**E**nhanced **S**upervised **F**ine-**T**uning (**D^{2}E-SFT**) strategy is presented, which facilitates adhering to context and fixating significant contextual event information to elevate the event reasoning capability. Specifically, D^{2}E-SFT removes the given sample's context to construct an imagined sample, subtracting its logits to mitigate the bias of neglecting context and improve contextual faithfulness. To guide the model in emphasizing significant contextual event information, D^{2}E-SFT employs a context-refined sample to achieve self-distillation with the alignment of logits. Extensive experimental results demonstrate the effectiveness of our data and strategy in expanding the performance boundary of event reasoning.

1 Introduction

Event reasoning involves accurately comprehending inter-event relations or inferring logical events based on certain event relations (Zhou et al., 2022a,b; Tao et al., 2024a; Lu et al., 2024b). While Large Language Models (LLMs) have demonstrated exceptional performance across diverse domains, their validity in event reasoning remains underexplored. This capability is fundamental for LLMs to tackle a broad spectrum of complex tasks, such as explanation generation (Krishna et al., 2024), and task planning (Zhao et al., 2024)). Therefore, it is crucial to rigorously assess LLMs' event reasoning capability and explore effective solutions for enhancement.

Recently, some preliminary explorations have been made to investigate the performance of LLMs in specific-relation event reasoning. For instance, Tan et al. (2023) and Su et al. (2024) advocate for event order-related contextual question answering to examine LLMs' event temporal reasoning capability. Similarly, Yang et al. (2022) and Yu et al. (2023a) assess causal and counterfactual reasoning by requiring models to infer answers to questions prefixed with "*why*" or "*what if*" from the context. However, these cutting-edge works merely focus on a single event relational type (e.g., temporal) or reasoning format (e.g., question answering), making it challenging to conduct a complete evaluation regarding the event reasoning of LLMs. Moreover, they fail to seek a practical solution for enhancing this capability (Chu et al., 2024; Tao et al., 2024b;

*Equal Contribution
†Corresponding author.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 28591–28613

Wei et al., 2023), stagnating at the stage of providing evaluation observations.

To address the limitations of existing evaluations, our work is initiated by presenting the first comprehensive benchmark (**PIPER**) for **P**robing **I**nto the **P**erformance boundary of LLMs in **E**vent **R**easoning. It incorporates 14 curated datasets characterized by instance-level event reasoning, covering a wide range of event relational types (i.e., Temporal (T), Causal (Ca), Counterfactual (Co), Subevent, and Conditional (S&C)) and diverse reasoning formats, such as question answering (QA), natural language inference (NLI), single and multi-choice(s) selection (SCQ, MCQ). PIPER primarily adopts the Contextual inter-event Relation Inference (CRI) and Contextual logical Event Inference (CEI) tasks to examine the model's ability to perform event reasoning in the context. The specific patterns of CRI and CEI are depicted in Figure 1a.

Then, to quantify the event reasoning capability of contemporary LLMs, we thoroughly evaluate the mainstream proprietary and open-source models (e.g., GPT4o (Hurst et al., 2024), LLaMA (Touvron et al., 2023)). The brief overview shown in Figure 1b illustrates that across various relational types, even the best aligned model is far from satisfactory and there is a notable gap between open-source and proprietary models. To investigate the underlying reasons for this phenomenon, a statistical analysis of the evaluation results is conducted and we find that typical error cases stem from the difficulties in being faithful to the context or focusing on significant event information therein.

Motivated by our evaluation observations and subsequent analysis of error patterns, we deem that the inferior performance of LLMs in event reasoning results from the scarcity of event reasoning-oriented instruction tuning data and strategy. To conquer these gaps, in this paper, we engage GPT4o to cleanse and reorganize existing available datasets, crafting 10K diverse instruction-tuning demonstrations tailored for event reasoning. Additionally, we propose a novel **D**ebiasing and **D**istillation-**E**nhanced **S**upervised **F**ine-**T**uning (**D^{2}E-SFT**) strategy, which facilitates adhering to context and fixating significant contextual event information to elevate the event reasoning capability. Specifically, to extract the bias of neglecting the context, D^{2}E-SFT removes the given sample's context to construct an imagined sample. The logits of both samples are subsequently subtracted to mitigate the bias, thereby improving contextual faithfulness when performing event reasoning. Besides, to guide the model in emphasizing significant contextual event information, D^{2}E-SFT employs a context-refined sample to achieve self-distillation with the alignment of logits, further enhancing the event reasoning capability.

To validate the effectiveness of our data and strategy, we carry out extensive experiments on the PIPER benchmark with mainstream LLMs. Compared to traditional SFT, D^{2}E-SFT exhibits better performance across different types of event reasoning. In summary, our main contributions include:

1) We propose **PIPRE**, the first comprehensive benchmark for **P**robing **I**nto the **P**erformance boundary of Large Language Models in **E**vent **R**easoning, followed by a thorough evaluation on contemporary LLMs and error pattern analysis.

2) We cleanse and reorganize existing available datasets, crafting 10K diverse instruction-tuning demonstrations tailored for event reasoning to alleviate the scarcity of event reasoning-oriented data.

3) We present a novel **D**ebiasing and **D**istillation-**E**nhanced **S**upervised **F**ine-**T**uning (**D^{2}E-SFT**) strategy, which facilitates adhering to the context and fixating the significant contextual event information to elevate the event reasoning capability.

4) Extensive experimental results demonstrate the effectiveness of our crafted data and strategy. The PIPER benchmark and code will be released soon to foster future research in event reasoning.

2 PIPER Benchmark

In this section, we first introduce the process of benchmark construction. Then, a series of thorough evaluations are conducted to quantify the event rea-

Benchmark	**Total**	**M-RT**	**M-RF**
Mctaco (Zhou et al., 2019)	263	×	×
Cosmos (Huang et al., 2019)	2,985	✓	×
Cider (Ghosal et al., 2021)	1,988	✓	×
Tracie (Zhou et al., 2021)	2,124	×	×
Ester (Han et al., 2021)	1,030	✓	×
Ecare (Du et al., 2022)	2,132	×	✓
CQA (Yang et al., 2022)	2,449	×	×
TempReason (Tan et al., 2023)	4,426	×	×
IfQA (Yu et al., 2023a)	700	×	×
TimeBench (Chu et al., 2024)	1,537	×	×
EV^2 (Tao et al., 2024b)	1,226	✓	×
PIPER(ours)	18,366	✓	✓

Table 1: Comparison with existing event reasoning benchmarks. **Total** indicates the number of samples. **M-RT** and **M-RF** represent whether it incorporate multiple relational types and reasoning formats, respectively.

Models	Temporal						Causal					Counterfactual					Subevent&Conditional		
	Tem.	Tim.	Tra.	Mct.	Cid.	*Avg.*	Cst.	Cos.	Eca.	Cid.	*Avg.*	Cst.	Cos.	Ifq.	Wiq.	*Avg.*	Est.	Cid.	*Avg.*
GPT4o$_{FS}$	**73.4**	62.5	**71.4**	**71.8**	**89.7**	***73.8***	55.8	85.2	**86.5**	**77.7**	***76.3***	54.0	**94.0**	**78.8**	45.0	*67.9*	28.3	**69.9**	***49.1***
Qwen1.5$_{FS}$-72b	66.0	58.4	57.9	71.6	79.4	*66.7*	61.4	82.4	84.7	58.9	*71.9*	61.0	91.9	76.5	42.9	*68.1*	**34.4**	56.7	*45.5*
LLaMA2$_{FS}$-70b	70.3	61.9	58.5	62.5	74.3	*65.5*	**61.6**	**85.4**	83.7	70.7	*75.4*	**62.7**	87.0	77.1	**55.5**	***70.6***	32.4	57.0	*44.7*
DeepSeek$_{FS}$-67b	67.0	**63.1**	57.5	58.5	76.5	*64.5*	60.6	83.7	82.9	67.1	*73.6*	60.9	85.2	78.0	54.8	*69.7*	31.0	61.5	*46.2*
GPT4o$_{ZS}$	**74.8**	**62.2**	**72.2**	69.1	**88.2**	***73.3***	**46.2**	**85.7**	**85.5**	**77.9**	***73.8***	**43.0**	**94.0**	**76.5**	43.2	***64.2***	27.1	**76.7**	***51.9***
Qwen1.5$^{\dagger}_{ZS}$-72b	63.8	50.1	67.5	**71.6**	83.8	*67.4*	34.3	80.5	83.3	70.5	*67.2*	33.9	87.0	65.1	37.6	*55.9*	21.7	62.9	*42.3*
LLaMA2$^{\dagger}_{ZS}$-70b	63.4	58.1	56.4	55.3	68.4	*60.3*	39.7	78.0	74.9	73.8	*66.6*	32.4	73.9	69.9	**49.8**	*56.5*	23.9	60.7	*42.3*
DeepSeek$^{\dagger}_{ZS}$-67b	64.7	56.2	60.8	54.1	64.0	*59.9*	43.1	78.7	82.2	70.0	*68.5*	40.9	84.5	72.3	36.4	*58.5*	**29.3**	64.0	*46.7*
GPT4o$_{CoT}$	68.6	59.3	**74.9**	**66.6**	**89.7**	***71.8***	21.6	**87.1**	**86.2**	80.9	***68.9***	16.7	**94.0**	**68.9**	41.3	*55.2*	19.8	72.5	*46.1*
QwQ$_{CoT}$	65.4	57.4	73.8	52.1	84.6	*66.7*	25.8	76.3	83.4	82.9	*67.1*	21.2	87.3	67.7	43.5	*54.9*	19.2	73.6	*46.4*
DeepSeek − R1$_{CoT}$	**72.6**	**62.5**	74.8	56.2	88.2	*70.9*	28.8	66.7	85.8	**87.0**	*67.1*	**30.3**	82.0	**69.0**	**47.8**	***57.3***	22.3	**76.7**	***49.5***
Qwen1.5$^{\dagger}_{CoT}$-72b	59.3	45.0	59.6	53.7	85.3	*60.6*	17.3	77.6	76.7	69.2	*60.2*	13.6	83.5	61.3	36.2	*48.7*	14.1	60.1	*37.1*
LLaMA2$^{\dagger}_{CoT}$-70b	56.5	45.1	50.7	47.1	64.0	*52.7*	**33.7**	69.7	51.2	62.8	*54.3*	24.3	70.4	57.9	45.2	*49.5*	**25.8**	53.1	*39.4*
DeepSeek$^{\dagger}_{CoT}$-67b	57.0	48.3	58.6	61.4	75.7	*60.2*	30.0	76.6	78.8	71.8	*64.3*	24.7	82.8	65.9	39.7	*53.3*	21.7	62.9	*42.3*
Qwen1.5$_{FS}$-14b	**60.2**	56.8	**54.3**	**65.5**	**73.5**	***62.1***	**60.6**	**79.0**	**83.5**	**68.8**	***73.0***	**62.2**	**84.2**	**73.0**	44.8	***66.0***	**33.2**	**67.4**	***50.3***
LLaMA2$_{FS}$-13b	57.4	**61.8**	53.3	43.1	54.4	*54.0*	52.5	71.2	68.6	56.8	*62.3*	48.0	59.9	70.1	**52.1**	*57.5*	30.9	57.6	*44.2*
Qwen1.5$^{\dagger}_{ZS}$-14b	**58.3**	**54.0**	**55.1**	**55.6**	**74.3**	***59.5***	**32.4**	**75.5**	**79.6**	**63.8**	***62.8***	**30.4**	**81.7**	**69.2**	38.4	***54.9***	**24.3**	**62.4**	***43.3***
LLaMA2$^{\dagger}_{ZS}$-13b	54.1	52.8	52.3	39.1	51.5	*50.0*	28.9	60.6	72.2	59.6	*55.4*	21.1	57.4	62.9	**54.6**	*49.0*	21.5	57.9	*39.7*
Qwen1.5$^{\dagger}_{CoT}$-14b	**52.8**	**48.6**	**51.0**	**51.8**	**61.0**	***53.1***	**17.2**	**72.2**	**78.3**	**70.9**	***59.7***	12.6	**78.2**	**64.1**	36.9	***47.9***	**15.8**	**69.1**	***42.5***
LLaMA2$^{\dagger}_{CoT}$-13b	47.3	48.2	46.2	36.6	58.8	*47.4*	**17.2**	56.8	47.1	53.2	*43.6*	**12.9**	59.5	49.1	**43.6**	*41.3*	14.9	50.8	*32.9*
Qwen1.5$_{FS}$-7b	50.1	52.6	**55.1**	**53.0**	**70.6**	***56.3***	**56.3**	**77.6**	**79.0**	**60.5**	***68.4***	**52.8**	**78.9**	**63.3**	47.5	***60.6***	**30.8**	**68.0**	***49.4***
LLaMA2$_{FS}$-7b	49.6	58.7	50.7	39.3	52.9	*50.3*	44.5	53.2	62.4	50.0	*52.5*	43.3	45.4	61.8	**51.5**	*50.5*	27.9	52.0	*39.9*
DeepSeek$_{FS}$-7b	**53.5**	**59.5**	51.5	41.5	61.0	*53.4*	47.9	59.9	67.1	51.3	*56.5*	40.1	47.5	62.2	49.0	*49.7*	29.6	51.4	*40.5*
Qwen1.5$^{\dagger}_{ZS}$-7b	**48.1**	51.4	**55.1**	**48.9**	55.2	***51.7***	30.9	**76.0**	**80.0**	**62.6**	***62.4***	25.3	**75.4**	**57.2**	40.0	***49.5***	22.7	**57.0**	***39.8***
LLaMA2$^{\dagger}_{ZS}$-7b	46.4	**54.9**	52.4	39.5	**60.3**	*50.7*	20.4	52.1	67.6	55.1	*48.8*	19.5	50.7	55.4	**44.1**	*42.4*	20.3	53.1	*36.7*
DeepSeek$^{\dagger}_{ZS}$-7b	44.7	44.9	52.8	44.9	52.2	*47.9*	**32.5**	61.9	77.1	59.2	*57.7*	**29.5**	57.4	53.0	39.8	*44.9*	**25.3**	53.9	*39.6*
Qwen1.5$^{\dagger}_{CoT}$-7b	**49.4**	**46.4**	**53.6**	42.7	**60.3**	***50.5***	16.9	**63.0**	**74.6**	**64.1**	***54.6***	11.8	**60.2**	**60.1**	36.1	*42.0*	14.4	**55.3**	*34.9*
LLaMA2$^{\dagger}_{CoT}$-7b	33.2	37.0	46.1	37.0	52.9	*41.3*	14.9	43.5	46.5	46.3	*37.8*	12.1	45.8	39.7	39.6	*34.3*	13.4	47.8	*30.6*
DeepSeek$^{\dagger}_{CoT}$-7b	37.9	38.8	50.8	**45.9**	50.7	*44.8*	**26.6**	59.1	67.9	60.0	*53.4*	**21.6**	54.9	42.3	**53.1**	***43.0***	**22.3**	**55.3**	***38.8***

Table 2: Evaluation results of contemporary LLMs on PIPER benchmark. The subscript FS, ZS, CoT indicate few-shot (shot nums = 3), zero-shot, and chain-of-thought prompting, respectively. *Avg.* indicates the average scores of all datasets in a certain relational type. † indicates the model is aligned. Best in bold, second with an underline.

soning capability of contemporary LLMs. Finally, we conclude evaluation observations and perform a subsequent statistical analysis of error patterns.

2.1 Benchmark Construction

To validly evaluate the performance of event reasoning, we draw on previous works, (Han et al., 2021; Lu et al., 2023a) adopting the representative tasks: Contextual inter-event Relation Inference (CRI) and Contextual logical Event Inference (CEI) as the skeleton of PIPER. In contrast to existing evaluations that merely focus on a single event relational type (e.g., temporal) or reasoning format (e.g., qa), PIPER includes a wide range of event relational categories and reasoning styles, aiming to serve as a comprehensive benchmark.

Specifically, existing available datasets are first selected according to the task requirements of CRI and CEI. For each dataset, we filter out samples where the ground truth is absent from the context, as these samples require external knowledge to complete event reasoning. This ensures that the evaluation results mainly depend on the event reasoning capability rather than factors such as knowledge recall, thereby maintaining evaluation fairness. Then, we categorize the remaining samples by their relational types and reasoning formats, converting them into the evaluation format (e.g., zero-shot) to establish the PIPER benchmark. To the end, PIPER incorporates 14 curated datasets, totaling 18,366 evaluation samples, spanning 5 relational types and 4 reasoning formats. The comparison between PIPER and existing event reasoning benchmarks are illustrated in Table 1, with more detailed information about the benchmark construction provided in Appendix A.

2.2 Benchmark Evaluation and Analysis

Evaluation Setup. To quantify the event reasoning capability of LLMs, we conduct a thorough evaluation of mainstream proprietary and open-source models on the PIPER benchmark in three ways: zero-shot, few-shots (shot nums = 3), and chain-of-thought (CoT) (Wei et al., 2022) prompting. Following Kojima et al. (2022) and Chu et al. (2024), we append the reasoning trigger *Let's think step by step* to the prompt for performing CoT. The Exact-match and Token-level F1 are utilized to measure the accuracy of generated answers. We report their average scores to provide a more robust assessment. All evaluation experiments are conducted with a temperature of 0.0 for greedy decoding. More details about benchmark evaluation (e.g., prompt) are shown in Appendix B.

Quantitative Results. The evaluation results are presented in Table 2. We can observe that: (i) While scaling up model size improves performance, it is commonly accompanied by massive resource loss and performance bottlenecks, indicating that relying on doubling parameters to solve event reasoning is not advisable; (ii) Aross various relational types, even the best-performing model is far from satisfactory and there is still a notable gap between open-source (e.g., LLaMA2) and proprietary (i.e., GPT4o) models; (iii) Among relational categories, temporal and causal reasoning achieve superior performance. This is consistent with the fact that temporal and causal data are prevalent and occupy a larger proportion of the pre-training corpus.

(iv) Compared to open-source base models, the aligned models exhibit a slight performance degradation in event reasoning. This may be attributed to the lack of event reasoning-oriented instruction-tuning data during the alignment process, which weakens the model's event-related knowledge and impairs its event reasoning capability; (v) It is noteworthy that CoT prompting fails to elevate the event reasoning in most cases, even yielding a performance drop, which is in accordance with findings in (Su et al., 2024; Jin et al., 2023; Zhang et al., 2024). We attribute this to two reasons. One is the incorrect event logical reasoning occurred in the reasoning chain, which leads to the error accumulation and wrong answer, as shown in Table 16 of Appendix C. The other is the characteristics of event reasoning. CRI and CEI in PIPER require the model to grasp event relational and commonsense knowledge for locating beneficial event information in the context. Nevertheless, despite its proficiency in complex reasoning tasks, CoT is unable to compensate for the lack of knowledge.

Analysis of Error Patterns. To further investigate the underlying reasons for the inferior performance of aligned models in event reasoning, we conduct a statistical analysis on the CoT evaluation results of relatively small LLMs (e.g., 7b, 13b models). Specifically, we demand GPT4o to combine input prompts and generated answers, identifying the following error patterns: (1) neglecting the context, which denotes the generated answers contain hallucinated contents or deviate from the given context;

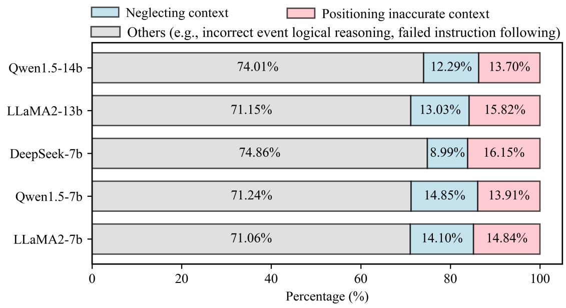

Figure 2: Detailed statistical results of the error patterns

(2) positioning inaccurate context, which denotes the model fails to emphasize the significant event information in the context, resulting in wrong answers; (3) others, such as incorrect event logical reasoning, failed instruction following and so on. Further details on the analysis of error patterns can be found in Appendix C. The statistical results are presented in Figure 2. We can discern that the former two error patterns account for nearly 30% of cases, indicating that the difficulties in following the context or focusing on the significant event information therein are underlying reasons contributing to poor event reasoning performance. Motivated by the above evaluation observations and error patterns analysis, we craft 10K diverse event reasoning-oriented instruction-tuning data and present **D^{2}E-SFT** strategy.

3 Methodology

In this section, the details of producing 10K instruction-tuning demonstrations are first described. Then, we elaborate on our proposed D^{2}E-SFT strategy, which facilitates adhering to context and fixating significant event information to elevate the event reasoning capability. As shown in the top right of Figure 3, D^{2}E-SFT creates an imagined sample by removing the original sample's context,

Dataset	**Type**	**Format**	**Ratio**(%)
Mctaco (Zhou et al., 2019)	T	MCQ	0.85
IfQA (Yu et al., 2023a)	Co	QA	4.20
Tracie (Zhou et al., 2021)	T	NLI	5.08
Ecare (Du et al., 2022)	Ca	SCQ	6.45
TempReason (Tan et al., 2023)	T	QA	7.00
TimeQA (Chen et al., 2021)	T	QA	7.00
WiQA (Yu et al., 2023a)	Co	QA	12.80
Cosmos(Huang et al., 2019)	Ca, Co	SCQ	12.80
Ester (Han et al., 2021)	Ca, Co, S&C	QA	17.51
CQA (Yang et al., 2022)	Ca, Co	QA	26.31

Table 3: Distribution of 10K event reasoning-oriented instruction-tuning data. Type and Format denotes event relational type and reasoning format, respectively.

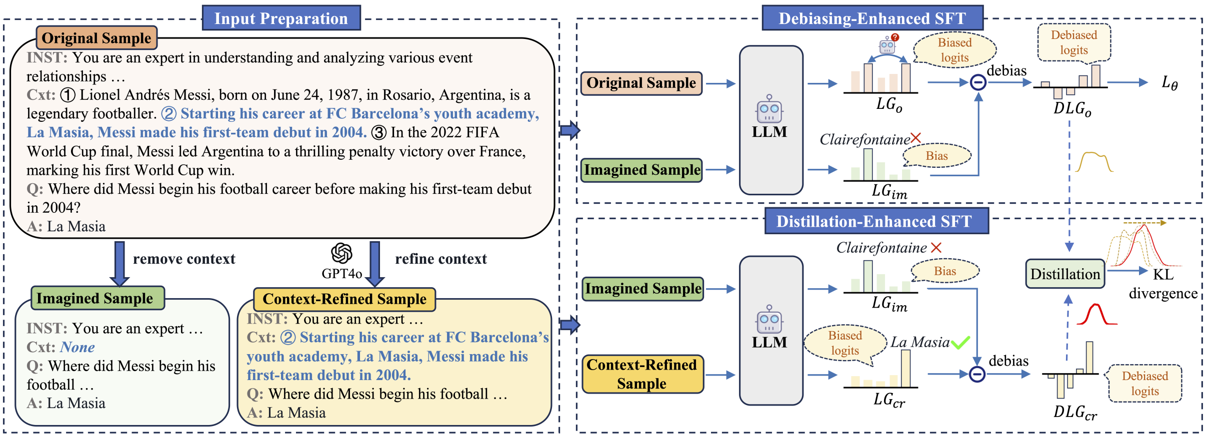

Figure 3: An illustration of the debiasing- and distillation-enhanced SFT strategy. The left depicts the construction of an imagined sample and a context-refined sample. The right presents the implementation of each strategy.

subtracting its logits to mitigate the bias of neglecting context. To guide the model in emphasizing significant contextual event information, D^{2}E-SFT employs a context-refined sample to achieve self-distillation with the alignment of logits, as depicted in the bottom right of Figure 3.

3.1 Instruction Tuning Data Production

According to our analysis of evaluation observations, we guess that the performance degradation of aligned models compared to base models primarily results from the lack of event reasoning-oriented instruction-tuning data during the alignment process. This deficiency weakens the model's event-related knowledge and impairs its event reasoning capability. To overcome this deficiency, we meticulously craft 10,000 diverse instruction-tuning demonstrations tailored for event reasoning. Considering the time consumption of annotating data from scratch, we follow previous works (Chu et al., 2024; Lu et al., 2024b) to cleanse and reorganize the existing available datasets.

Specifically, we start with collecting event reasoning-related training samples from the dataset involved in the PIPER benchmark. Note that we skip the datasets without providing training samples (e.g., Cider). To ensure data quality, we implement the following steps. (1) Context Filtering: given that our initial is to enhance the model's capability of performing event reasoning in the context, we filter out samples with excessively redundant or overly brief contexts to improve data relevance. (2) Quality Filtering: we engage GPT4o to exclude samples whose context fails to explicitly or implicitly support the ground truth of the sample's question to improve data rationality. Next, we apply a heuristic ratio to randomly select samples, integrating them into 10K raw data. To adapt to the alignment stage, the raw data is converted into an instruction-tuning formulation, acquiring 10K event reasoning-oriented demonstrations. The specific ratios for each dataset are presented in Table 3, with further details on the data production provided in Appendix D.

3.2 D^{2}E-SFT Strategy

Based on our statistical analysis of error patterns, we can confirm that neglecting context and positioning inaccurate context are underlying reasons contributing to weak event reasoning capability. To address these issues, we present a novel debiasing- and distilling-enhanced SFT strategy centered on adhering to context and fixating significant contextual event information, as shown in Figure 3.

Debiasing-Enhanced SFT Strategy. Inspired by prior research (Zhang et al., 2024; Shi et al., 2024), neglecting context can be regarded as a bias where the model overly relies on its internal parametric knowledge to answer questions. Hence, the core is to reasonably identify this bias and mitigate it effectively. To achieve this goal, we borrow the idea from counterfactual reasoning (Pearl, 2009; Chen et al., 2023) which infers outcomes under hypothetical conditions differing from the factual world to estimate the bias, and then adjust outcomes to eliminate it. Concretely, as depicted in the top for Figure 3, given the j-th original sample $Y_{j,o} = \{INST_j, Cxt_j, Q_j, A_j\}$, where $INST_j$, Cxt_j, Q_j, and A_j refers to the instruction, context, question and answer of the sample, respectively, we construct an imagined sample $Y_{j,im} = \{INST_j, Q_j, A_j\}$ by removing the context in the hypothetical world. Then, we simultaneously compute their distribution of logits $(LG) \in R^{l\times d}$ over the vocabulary as defined in (1), (2). Here, l and d represent the length of A and the vocabulary size.

$$LG_{j,0} = \sum_{k=1}^{l} F(A_{j,k}|INST_j, Cxt_j, Q_j, A_{j,<k}), \quad (1)$$

$$LG_{j,im} = \sum_{k=1}^{l} F(A_{j,k}|INST_j, Q_j, A_{j,<k}), \quad (2)$$

where F denotes the forward propagation function of the model. Notably, $LG_{j,o}$ is essentially the extracted bias, which reflects the model's internal parametric knowledge preference for answering questions without relying on the context. Last, to mitigate the bias, the logits of the imagined sample are subtracted with a ratio α, improving contextual faithfulness. The model can be trained by minimizing the cross entropy loss L_θ as (3):

$$L_\theta = -\sum_{k=1}^{l} log(softmax(LG_{j,o} - \alpha LG_{j,im})). \quad (3)$$

Distillation-Enhanced SFT Strategy. This strategy implements the knowledge self-distillation (Zhang et al., 2019) with the alignment of logits to guide the model in paying more attention to the significant contextual event information in the context. Particularly, as illustrated in the bottom of Figure 3, we start with the j-th original sample $Y_{j,o}$ = $\{INST_j, Cxt_j, Q_j, A_j\}$ and create a context-refined sample $Y_{j,cr}$ = $\{INST_j, Cxt_{j,r}, Q_j, A_j\}$, where $Cxt_{j,r}$ denotes the refined context. To obtain $Cxt_{j,r}$, we exploit GPT4o in advance to remove irrelevant contextual event information that does not contribute to answering the question. Then, we align their logits distribution $(LG) \in R^{l\times d}$ over the vocabulary through Kullback Leibler (KL) (Kullback and Leibler, 1951) divergence regularization as represented in (4), (5).

$$LG_{j,cr} = \sum_{k=1}^{l} F(A_{j,k}|INST_j, Cxt_j, Q_j, A_{j,<k}), \quad (4)$$

$$KL(LG_{j,cr}||LG_{j,o}) = \sum_{k=1}^{l} p(lg^k_{j,cr}) \cdot log\frac{lg^k_{j,cr}}{lg^k_{j,o}}, \quad (5)$$

where $lg^k \in R^d$ denote the distribution of the k-th token. By doing so, the model is driven to emphasize significant contextual event information, further enhancing the event reasoning capability.

To integrate the debiasing- and distillation-enhanced strategies, for each sample, we first obtain the debiased logits, denoted as $DLG_{j,o}$ for the original sample and $DLG_{j,cr}$ for the context-refined sample, as specified in (6), (7). Then, we combine the cross entropy loss L_θ in (3) with the debiased KL regularization to jointly guide the optimization of the model as defined in (8).

$$DLG_{j,o} = LG_{j,o} - \alpha LG_{j,im}, \quad (6)$$

$$DLG_{j,cr} = LG_{j,cr} - \alpha LG_{j,im}, \quad (7)$$

$$L = L_\theta + KL(DLG_{j,o}||DLG_{j,cr}), \quad (8)$$

4 Experiments

In this section, we introduce the datasets, baselines, and experimental settings. Then, we present experimental results and provide a detailed analysis.

4.1 Datasets

To verify the effectiveness of our data and strategy, we conduct experiments on our proposed benchmark, PIPER, to evaluate the event reasoning capability of diverse relational types. We ensure the test samples are completely non-overlapping with the crafted 10K instruction tuning demonstrations.

4.2 Baselines

We first introduce mainstream proprietary and open-source aligned models, namely GPT4o (Hurst et al., 2024), Qwen (Bai et al., 2023), LLaMA2 (Touvron et al., 2023), DeepSeek (Bi et al., 2024) as baselines to validate the efficacy of our crafted data. To further assess the effectiveness of our proposed D^2E-SFT strategy, we fine-tune some of the aforementioned models with the standard SFT strategy as additional baselines.

4.3 Experimental Settings

For the original aligned models, we evaluate them in a zero-shot way, consistent with the settings used in section 2.2. For the aligned models, we utilize LLaMA-Factory (Zheng et al., 2024) to fine-tune them with 8 A100 GPUs under mixed-precision training (fp16). During training, the hyper-parameters are set as follows: learning rate: 1.0e-5, batch size: 64, epochs: 3, warmup ratio: 0.1, and accumulation steps: 1 for all models. During inference, we adopt the greedy decoding with a temperature of 0.0 across all experiments. α is set to 0.5 for both training and inference. We directly choose the last checkpoint to evluate the fine-tuned

Models	Temporal						Causal					Counterfactual					Subevent&Conditional		
	Tem.	Tim.	Tra.	Mct.	Cid.	*Avg.*	Cst.	Cos.	Eca.	Cid.	Avg.	Cst.	Cos.	Ifq.	Wiq.	Avg.	Est.	Cid.	*Avg.*
GPT4o$_{ZS}$	74.8	62.2	72.2	69.1	88.2	73.3	46.2	85.7	85.5	77.9	73.8	43.0	94.0	76.5	43.2	64.2	27.1	76.7	51.9
Qwen1.5-72b❄	63.8	50.1	67.5	71.6	83.8	67.4	34.3	80.5	83.3	70.5	67.2	33.9	87.0	65.1	37.6	55.9	21.7	62.9	42.3
LLaMA2-70b❄	63.4	58.1	56.4	55.3	68.4	60.3	39.7	78.0	74.9	73.8	66.6	32.4	73.9	69.9	49.8	56.5	23.9	60.7	42.3
DeepSeek−-67b❄	64.7	56.2	60.8	54.1	64.0	59.9	43.1	78.7	82.2	70.0	68.5	40.9	84.5	72.3	36.4	58.5	29.3	64.0	46.7
Qwen1.5−7b❄	48.1	51.4	55.1	48.9	55.2	51.7	30.9	76.0	80.0	62.6	62.4	25.3	75.4	57.2	40.0	49.5	22.7	57.0	39.8
LLaMA2-7b❄	46.4	54.9	52.4	39.5	60.3	50.7	20.4	52.1	67.6	55.1	48.8	19.5	50.7	55.4	44.1	42.4	20.3	53.1	36.7
DeepSeek-7b❄	44.7	44.9	52.8	44.9	52.2	47.9	32.5	61.9	77.1	59.2	57.7	29.5	57.4	53.0	39.8	44.9	25.3	53.9	39.6
Qwen1.5-7b (SFT)🔥	64.0	66.3	50.0	67.2	50.7	59.7	77.6	77.1	82.0	50.0	71.7	77.5	80.3	68.0	58.9	71.2	33.3	50.0	41.7
LLaMA2-7b (SFT)🔥	64.1	68.2	50.0	63.1	50.7	59.2	77.1	73.6	80.2	50.0	70.2	76.4	73.7	69.0	59.6	69.7	36.2	50.0	43.1
DeepSeek-7b (SFT)🔥	64.8	66.8	50.0	68.2	50.7	60.1	75.3	75.5	82.0	50.0	70.7	74.9	73.9	70.6	59.2	69.7	33.9	50.0	41.9
Qwen1.5-7b (D^{2}E-SFT)🔥	65.6*	68.5*	76.0*	69.2*	72.1*	70.3*	79.2*	79.3*	83.7*	73.5*	78.9*	79.3*	82.8*	68.9*	60.6*	72.9*	36.6*	60.1*	48.4*
LLaMA2-7b (D^{2}E-SFT)🔥	66.5*	71.5*	74.3*	62.5	74.3*	69.8*	79.0*	75.5*	81.8*	56.3*	73.1*	77.9*	73.9	70.8*	64.6*	71.8*	37.9*	55.3*	46.6*
Deekseek-7b (D^{2}E-SFT)🔥	65.9*	71.1*	74.2*	69.6*	72.8*	70.7*	76.8*	77.4*	82.9*	55.6*	73.2*	76.4*	76.8*	70.4	62.2*	71.5*	35.3*	54.8*	45.0*

Table 4: Experimental results on the PIPER. *Avg.* indicates the average scores of all datasets in a certain relational type. ❄ and 🔥 denote model freezing and fine-tuning, respectively. Bold in the frozen model indicates the best performance under the same model size. Bold in the fine-tuned model indicates superior performance to the original frozen model. * represent its performance is better than the same model fine-tuned on the standard SFT strategy.

aligned models. Following Chu et al. (2024), the Exact-match and Token-level F1 are used to measure the accuracy of the generated answers.

4.4 Experimental Results

The experimental results on the PIPER benchmark are reported in Table 4. We can observe that: (1) In frozen models, the performance of the model with small size (i.e., 7b) in various relational event reasoning is far inferior to the large scale model (e.g., 72b). However, when fine-tuned on the crafted 10K event reasoning-oriented instruction tuning demonstrations with standard SFT strategy, their performance substantially surges in most cases, even surpassing the larger model. We believe this is because the crafted data contains a wealth of specialized knowledge relevant to the event and relation, enhancing the model's capability to understand and reason about the event. (2) In the scenario of natural language inference (NLI) (i.e., Tracie, Cider), fine-tuning on our crated 10K data with standard SFT strategy loses efficacy. We further scrutinize the generated answers of fine-tuned models and find that they suffer from the bias of neglecting the context, which induces the model to generate the same answer "positive" regardless of the context of the test sample, resulting in an accuracy of approximately 50%, akin to binary random guessing.

(3) On the contrary, when equipped with our proposed D^{2}E-SFT strategy, this deficiency disappears and significant performance improvements are observed for NLI tasks. Furthermore, compared to the standard SFT strategy, D^{2}E-SFT achieves superior performance in almost various relational categories and reasoning formats. We believe the reason is that D^{2}E-SFT separately designs imagined and context-refined samples to facilitate mitigating the bias of neglecting the context and emphasizing significant contextual event information, elevating the event reasoning capability. In the end, the huge performance gain demonstrates the effectiveness of our crafted data and D^{2}E-SFT strategy in expanding the performance boundary of event reasoning.

4.5 Ablation Study

To explicitly illustrate the efficacy of our proposed D^{2}E-SFT strategy, we conduct ablation studies to validate its core design on the PIPER benchmark. As shown in table 5, we define the following three ablation variants: **(1) -w/o DIE** removes the distillation-enhanced SFT strategy, retaining only the imagined sample to mitigate the bias of neglecting the context. **(1) -w/o DEE** removes the debiasing-enhanced SFT strategy, retaining only the context-refined sample to guide the model in emphasizing the significant contextual event information. **(3) -w/o BOTH** removes both the debiasing- and distillation-enahnces strategies.

Particularly, we can draw the following inferences based on the results in Table 5: (1) Removing the distillation-enhanced strategy (-w/o DIE) leads to performance drop, especially for the datasets (i.e., TempReason, TimeQA) with longer context length, which reflects the important role of the distillation-enhanced strategy in guiding model in emphasizing the significant contextual event information. (2) Removing the debiasing-enhanced strategy (-w/o DEE) also results in performance degradation, especially in the NLI tasks (i.e., Tracie, Cider), which underscores the critical role of debiasing-enhanced strategy in mitigating the bias of neglecting the context. Note that the distillation-enhanced strategy exerts the analogous slighter

Models	Temporal						Causal					Counterfactual					Subevent&Conditional		
	Tem.	Tim.	Tra.	Mct.	Cid.	*Avg.*	Cst.	Cos.	Eca.	Cid.	*Avg.*	Cst.	Cos.	Ifq.	Wiq.	*Avg.*	Est.	Cid.	*Avg.*
Qwen1.5-7b (D^{2}E-SFT)🔥	**65.6**	**68.5**	76.0	**69.2**	72.1	***70.3***	**79.2**	79.3	**83.7**	73.5	***78.9***	**79.3**	**82.8**	**68.9**	60.6	***72.9***	**36.6**	60.1	*48.4*
- w/o DIE 🔥	62.7	67.3	**77.5**	68.6	**73.1**	*69.8*	77.8	79.2	82.4	**74.1**	*78.4*	78.4	81.7	67.9	59.1	*71.8*	36.2	**61.0**	***48.6***
- w/o DEE 🔥	63.3	68.4	75.1	66.4	71.7	*69.0*	78.2	**79.9**	83.6	71.4	*78.2*	78.8	82.0	**68.9**	**61.9**	***72.9***	**36.6**	59.9	*48.3*
- w/o BOTH 🔥	64.0	66.3	50.0	67.2	50.7	*59.7*	77.6	77.1	82.0	50.0	*71.7*	77.5	80.3	68.0	58.9	*71.2*	33.3	50.0	*41.7*
LLaMA2-7b (D^{2}E-SFT)🔥	**66.5**	**71.5**	74.3	62.5	**74.3**	***69.8***	**79.0**	**75.5**	**81.8**	**56.3**	***73.1***	**77.9**	**73.9**	**70.8**	64.6	***71.8***	37.9	**55.3**	***46.6***
- w/o DIE 🔥	64.4	69.9	**75.2**	**63.1**	71.3	*68.8*	77.8	73.8	80.3	56.2	*72.0*	77.2	69.4	70.7	**65.0**	*70.6*	**39.0**	53.1	*46.1*
- w/o DEE 🔥	65.1	71.1	73.2	62.3	68.0	*67.9*	78.7	71.9	80.7	55.6	*71.7*	77.6	70.1	70.6	62.7	*70.2*	36.0	53.5	*44.7*
- w/o BOTH 🔥	64.1	68.2	50.0	63.1	50.7	*59.2*	77.1	73.6	80.2	50.0	*70.2*	76.4	73.7	69.0	59.6	*69.7*	36.2	50.0	*43.1*

Table 5: Ablation study results on the PIPER. *Avg.* indicates the average scores of all datasets in a certain relational type. 🔥 denotes model fine-tuning. DIE and DEE denote distillation- and debiasing-enhanced SFT strategies.

role. It is probably because the KL divergence in (8) regulates the model to focus on the contextual information, alleviating the bias of neglecting the context. (3) Removing both distillation- and debiasing-enhanced strategies (-w/o BOTH) causes the worst performance, indicating the necessity of each core design. (4) D^{2}E-SFT generally outperforms the variants with a single strategy, suggesting that different strategies not only serve the unique roles but also supplement each other. Overall, these inferences validate the efficacy of debiasing and distillation strategy in D^{2}E-SFT.

4.6 Hyperparamter Sensitivity Analysis

For the unique hyper-parameter "ratio α", we find in the experiments that as long as it is not set extremely unreasonable (e.g., too small 0.1 or too large 5.0), the model will gain the constantly good performance. This is because the extremely unreasonable ratio causes the debiasing effect of the model to be too weak or too strong, which is not conducive to model learning. The experimental results of hyperparameter sensitivity analysis on the ratio are reported in Table 6 (T, Ca, Co, and S&C represent the Temporal, Causal, Counterfactual, Subevent&Conditional respectively, we present the average performance for each event relational type). For the regular hyper-parameters (e.g., learning rate, batch size), we set them based on experience to ensure that the model can converge normally, and do not perform specific parameter adjustments.

Methods	T	Ca	Co	S&C
LLaMA2-7b (α=0.1)	67.4	72.6	69.7	44.8
LLaMA2-7b (α=0.2)	66.4	72.8	69.3	44.7
LLaMA2-7b (α=0.3)	69.5	72.8	69.9	44.6
LLaMA2-7b (α=0.5)	**69.8**	73.1	**71.8**	46.6
LLaMA2-7b (α=0.7)	69.4	72.4	70.7	**47.8**
LLaMA2-7b (α=0.9)	69.1	74.4	70.8	44.5
LLaMA2-7b (α=1.0)	69.5	**74.6**	70.6	45.4
LLaMA2-7b (α=2.0)	68.4	73.4	70.3	45.6
LLaMA2-7b (α=5.0)	66.9	72.8	70.4	44.5
Qwen1.5-7b (α=0.1)	68.8	77.1	71.5	46.7
Qwen1.5-7b (α=0.2)	69.7	78.3	71.1	47.9
Qwen1.5-7b (α=0.3)	**70.9**	78.5	72.6	49.7
Qwen1.5-7b (α=0.5)	70.3	**78.9**	72.9	48.4
Qwen1.5-7b (α=0.7)	68.6	78.8	**73.4**	49.9
Qwen1.5-7b (α=0.9)	67.3	77.8	72.0	49.5
Qwen1.5-7b (α=1.0)	69.1	77.4	72.3	**51.6**
Qwen1.5-7b (α=2.0)	68.2	77.8	72.4	48.8
Qwen1.5-7b (α=5.0)	67.4	76.6	71.5	48.3

Table 6: Experimental results of hyperparameter sensitivity analysis on the ratio.

5 Related Work

Event reasoning. It has been proven useful in event-centric tasks. Early works (Zhou et al., 2022b,a; Huang et al., 2019; Qin et al., 2019; Lu et al., 2023b) focused on enhancing the event reasoning capability of Pre-trained Language Models (PLMs). With the development of LLMs, researchers begin to explore their validity in event reasoning. Chu et al. (2024) and Su et al. (2024) adopt time-sensitive question answering to assess the event temporal reasoning capability. Yang et al. (2022) and Yu et al. (2023a) demand the model to deduce the answer to the question prefixed with "*why*" or "*what if*", evaluating their event causal and counterfactual reasoning capability. However, these works merely focus on a single event relational type or reasoning format, failing to conduct a complete evaluation. Moreover, they stagnate at the stage of providing evaluation observations (Chu et al., 2024; Tao et al., 2024b; Wei et al., 2023; Lu et al., 2024a), failing to seek a practical solution for enhancing this capability. In this paper, we propose PIPER, the first comprehensive benchmark for Probing Into the Performance boundary of LLMs in Event Reasoning. We also present a novel D^{2}E-SFT strategy for capability enhancement.

Instruction Tuning. It (Zhong et al., 2021; Wei et al.; Ouyang et al., 2022; Tian et al., 2025) has emerged as a critical stage for boosting the performance of LLMs on a wide range of downstream tasks by fine-tuning them on datasets composed of diverse task-specific instructions. For instance, Yu et al. (2024), Wen et al. (2024) and Xu et al. (2025) devote to producing diverse domain-specific SFT data, such as role-playing and mathematical reasoning. To overcome the catastrophic forgetting problem of SFT process, the self-distillation (Zhang et al., 2019; Yang et al., 2024) is used to learn the dataset generated by the model itself. Sun et al. (2024) leverages LLMs to identify informative biased samples, followed by a causal-guided approach to debias LLMs. In this paper, we meticulously craft 10K event reasoning-oriented instruction tuning data and propose a novel debiasing- and distillation-enhanced SFT strategy to improve event reasoning capability.

6 Conclusions

In this paper, we focus on exploring and improving the validity of Large Language Models (LLMs) in event reasoning. Concretely, we propose PIPER, the first comprehensive benchmark for Probing Into the Performance boundary of LLMs in Event Reasoning, totaling 18,366 samples, spanning 5 relational types and 4 reasoning formats. Then, a thorough evaluation of contemporary LLMs on PIPER and analysis of error patterns is conducted. We find that typical error cases stem from the difficulty in being faithful to the context or precisely focusing on significant event information therein. Motivated by our evaluation observations and analysis, we craft 10K diverse instruction tuning demonstrations tailored for event reasoning to alleviate the event reasoning-oriented data scarcity. Additionally, we present a novel D^2E-SFT strategy, which facilitates adhering to the context and fixating the significant contextual event information to elevate the event reasoning capability. Extensive experimental results demonstrate the effectiveness of our crafted data and strategy in expanding the performance boundary of event reasoning. We aspire for the PIPER benchmark, and 10K instruction tuning data to foster future research in this field.

Limitations

Despite the impressive results of our method, we have to admit our work has the following limitations: (1) during the production of 10K event reasoning-oriented instruction tuning data, we apply a random sampling strategy to obtain samples from datasets, possibly affecting the final quality. If the data ratio and sampling strategy can be meticulously adjusted, the data quality and final performance can be further improved. (2) During the statistical analysis of error patterns in the evaluation results, in addition to neglecting the context and positioning inaccurate context, incorrect event logical reasoning also accounts for a large proportion. Our work focuses on presenting **$\mathbf{D^2}$E-SFT** strategy to overcome the former two error patterns, with more explorations on enhancing the event logical reasoning left in the future.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 62206267, 62262040. This work was also supported by Jiangxi Natural Science Foundation under grant $20242BAB25069$ and Jiangxi Provincial Key Laboratory of Intelligent Systems and Human-Machine Interaction ($2024SSY03121$). We sincerely thank Kaiwen Wei, Nayu Liu for their constructive collaboration.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. arXiv preprint arXiv:2309.16609.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. 2024. Deepseek llm: Scaling open-source language models with longtermism. arXiv preprint arXiv:2401.02954.

Wenhu Chen, Xinyi Wang, and William Yang Wang. 2021. A dataset for answering time-sensitive questions. arXiv preprint arXiv:2108.06314.

Ziwei Chen, Linmei Hu, Weixin Li, Yingxia Shao, and Liqiang Nie. 2023. Causal intervention and counterfactual reasoning for multi-modal fake news detection. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 627–638, Toronto, Canada. Association for Computational Linguistics.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Haotian Wang, Ming Liu, and Bing Qin. 2024. TimeBench: A comprehensive evaluation of temporal reasoning abilities in large language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume

1: Long Papers), pages 1204–1228, Bangkok, Thailand. Association for Computational Linguistics.

Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2022. e-CARE: a new dataset for exploring explainable causal reasoning. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 432–446, Dublin, Ireland. Association for Computational Linguistics.

Deepanway Ghosal, Pengfei Hong, Siqi Shen, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2021. CIDER: Commonsense inference for dialogue explanation and reasoning. In Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 301–313, Singapore and Online. Association for Computational Linguistics.

Rujun Han, I-Hung Hsu, Jiao Sun, Julia Baylon, Qiang Ning, Dan Roth, and Nanyun Peng. 2021. ESTER: A machine reading comprehension dataset for reasoning about event semantic relations. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 7543–7559, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Stephen J Hoch. 1985. Counterfactual reasoning and accuracy in predicting personal events. Journal of Experimental Psychology: Learning, Memory, and Cognition, 11(4):719.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. 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 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. arXiv preprint arXiv:2410.21276.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez, Max Kleiman-Weiner, Mrinmaya Sachan, et al. 2023. Cladder: assessing causal reasoning in language models. In Proceedings of the 37th International Conference on Neural Information Processing Systems, pages 31038–31065.

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

Satyapriya Krishna, Jiaqi Ma, Dylan Slack, Asma Ghandeharioun, Sameer Singh, and Himabindu Lakkaraju. 2024. Post hoc explanations of language models can improve language models. Advances in Neural Information Processing Systems, 36.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. The annals of mathematical statistics, 22(1):79–86.

Zhicong Lu, Li Jin, Ziwei Chen, Changyuan Tian, Xian Sun, Xiaoyu Li, Yi Zhang, Qi Li, and Guangluan Xu. 2024a. Relation-aware multi-pass comparison deconfounded network for change captioning. IEEE Transactions on Circuits and Systems for Video Technology, 34(12):13349–13363.

Zhicong Lu, Li Jin, Peiguang Li, Yu Tian, Linhao Zhang, Sirui Wang, Guangluan Xu, Changyuan Tian, and Xunliang Cai. 2024b. Rethinking the reversal curse of LLMs: a prescription from human knowledge reversal. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 7518–7530, Miami, Florida, USA. Association for Computational Linguistics.

Zhicong Lu, Li Jin, Guangluan Xu, Linmei Hu, Nayu Liu, Xiaoyu Li, Xian Sun, Zequn Zhang, and Kaiwen Wei. 2023a. Narrative order aware story generation via bidirectional pretraining model with optimal transport reward. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 6274–6287.

Zhicong Lu, Li Jin, Guangluan Xu, Linmei Hu, Nayu Liu, Xiaoyu Li, Xian Sun, Zequn Zhang, and Kaiwen Wei. 2023b. Narrative order aware story generation via bidirectional pretraining model with optimal transport reward. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 6274–6287, Singapore. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Judea Pearl. 2009. Causal inference in statistics: An overview.

Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. 2019. Counterfactual story reasoning and generation. 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 5043–5053, Hong Kong, China. Association for Computational Linguistics.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024. Trusting your evidence: Hallucinate less with context-aware decoding. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 783–791, Mexico City, Mexico. Association for Computational Linguistics.

Zhaochen Su, Juntao Li, Jun Zhang, Tong Zhu, Xiaoye Qu, Pan Zhou, Yan Bowen, Yu Cheng, and Min Zhang. 2024. Living in the moment: Can large language models grasp co-temporal reasoning? In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13014–13033, Bangkok, Thailand. Association for Computational Linguistics.

Zhouhao Sun, Li Du, Xiao Ding, Yixuan Ma, Yang Zhao, Kaitao Qiu, Ting Liu, and Bing Qin. 2024. Causal-guided active learning for debiasing large language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024, pages 14455–14469. Association for Computational Linguistics.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2023. Towards benchmarking and improving the temporal reasoning capability of large language models. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14820–14835.

Niket Tandon, Bhavana Dalvi, Keisuke Sakaguchi, Peter Clark, and Antoine Bosselut. 2019. WIQA: A dataset for "what if..." reasoning over procedural text. 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 6076–6085, Hong Kong, China. Association for Computational Linguistics.

Zhengwei Tao, Xiancai Chen, Zhi Jin, Xiaoying Bai, Haiyan Zhao, and Yiwei Lou. 2024a. EVIT: Event-oriented instruction tuning for event reasoning. In Findings of the Association for Computational Linguistics: ACL 2024, pages 8966–8979, Bangkok, Thailand. Association for Computational Linguistics.

Zhengwei Tao, Zhi Jin, Yifan Zhang, Xiancai Chen, Haiyan Zhao, Jia Li, Bing Liang, Chongyang Tao, Qun Liu, and Kam-Fai Wong. 2024b. A comprehensive evaluation on event reasoning of large language models. arXiv preprint arXiv:2404.17513.

Changyuan Tian, Zhicong Lu, Zequn Zhang, Heming Yang, Wei Cao, Zhi Guo, Xian Sun, and Li Jin. 2025. Hypermixer: Specializable hypergraph channel mixing for long-term multivariate time series forecasting. Proceedings of the AAAI Conference on Artificial Intelligence, 39(19):20885–20893.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In International Conference on Learning Representations.

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

Yifan Wei, Yisong Su, Huanhuan Ma, Xiaoyan Yu, Fangyu Lei, Yuanzhe Zhang, Jun Zhao, and Kang Liu. 2023. MenatQA: A new dataset for testing the temporal comprehension and reasoning abilities of large language models. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 1434–1447, Singapore. Association for Computational Linguistics.

Bosi Wen, Pei Ke, Xiaotao Gu, Lindong Wu, Hao Huang, Jinfeng Zhou, Wenchuang Li, Binxin Hu, Wendy Gao, Jiaxing Xu, Yiming Liu, Jie Tang, Hongning Wang, and Minlie Huang. 2024. Benchmarking complex instruction-following with multiple constraints composition. In Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. 2025. Magpie: Alignment data synthesis from scratch by prompting aligned LLMs with nothing. In The Thirteenth International Conference on Learning Representations.

Linyi Yang, Zhen Wang, Yuxiang Wu, Jie Yang, and Yue Zhang. 2022. Towards fine-grained causal reasoning and qa. arXiv preprint arXiv:2204.07408.

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. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024, pages 1028–1043. Association for Computational Linguistics.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2024. Metamath: Bootstrap your own mathematical questions for large language models. In

The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net.

Wenhao Yu, Meng Jiang, Peter Clark, and Ashish Sabharwal. 2023a. IfQA: A dataset for open-domain question answering under counterfactual presuppositions. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 8276–8288, Singapore. Association for Computational Linguistics.

Wenhao Yu, Meng Jiang, Peter Clark, and Ashish Sabharwal. 2023b. Ifqa: A dataset for open-domain question answering under counterfactual presuppositions. arXiv preprint arXiv:2305.14010.

Letian Zhang, Xiaotong Zhai, Zhongkai Zhao, Yongshuo Zong, Xin Wen, and Bingchen Zhao. 2024. What if the tv was off? examining counterfactual reasoning abilities of multi-modal language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 21853–21862.

Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. 2019. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In Proceedings of the IEEE/CVF international conference on computer vision, pages 3713–3722.

Zirui Zhao, Wee Sun Lee, and David Hsu. 2024. Large language models as commonsense knowledge for large-scale task planning. Advances in Neural Information Processing Systems, 36.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), pages 400–410, Bangkok, Thailand. Association for Computational Linguistics.

Ruiqi Zhong, Kristy Lee, Zheng Zhang, and Dan Klein. 2021. Adapting language models for zero-shot learning by meta-tuning on dataset and prompt collections. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2856–2878, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. 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 3363–3369, Hong Kong, China. Association for Computational Linguistics.

Ben Zhou, Kyle Richardson, Qiang Ning, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2021. Temporal reasoning on implicit events from distant supervision. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1361–1371, Online. Association for Computational Linguistics.

Yucheng Zhou, Xiubo Geng, Tao Shen, Guodong Long, and Daxin Jiang. 2022a. Eventbert: A pre-trained model for event correlation reasoning. In Proceedings of the ACM Web Conference 2022, pages 850–859.

Yucheng Zhou, Tao Shen, Xiubo Geng, Guodong Long, and Daxin Jiang. 2022b. Claret: Pre-training a correlation-aware context-to-event transformer for event-centric generation and classification. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2559–2575.

A PIPER Benchmark Construction

PIPER is the first comprehensive benchmark for event reasoning, which incorporates 14 curated datasets, totaling 18,366 samples, spanning 5 relational types and 4 reasoning formats. PIPER primarily adopts the Contextual inter-event Relation Inference (CRI) and Contextual logical Event Inference (CEI) tasks as the skeleton. To elaborate on the benchmark construction, we first introduce the event semantic relations, and the meanings of the CRI and CEI tasks. Then, we describe the processing operations for each dataset involved in the PIPER benchmark. Last, we report the statistical results of the PIPER benchmark.

Regarding the event semantic relations.

(1) Temporal (Zhou et al., 2021): A pair of events (e_i, e_j) exhibits a TEMPORAL relation if e_i occurs *before* e_j or e_i occurs *after* e_j, or e_i and e_j occurs at the same time according to the context. Note that the above temporal relation is derived by comparing the start time of paired events rather than the end time.

(2) Causal (Han et al., 2021): A pair of events (e_i, e_j) exhibits a CAUSAL relation if e_i happens then e_j will definitely happen according to the context.

(3) Counterfactual (Hoch, 1985): A pair of events (e_i, e_j) exhibits a COUNTERFACTUAL relation if e_i happens in hypothetical conditions differing from the factual world then e_j imagines the consequences of something that is contrary to what actually happened or is factually true according to the context.

(4) Subevent (Han et al., 2021): There is a semantic hierarchy where a complex event e_k consists of a set of subevents $\{e_{k,1}, ..., e_{k,j}$, ..., $e_{k,n}\}$ according to the context.

(5) Conditional (Han et al., 2021): A pair of events (e_i, e_j) exhibits a CONDITIONAL relation if e_i facilitates, but may not necessarily leads to e_j according to the context. Note that Sub-event and Conditional are relatively small event semantic relations, so we merge their evaluation samples into one large category Subevent&Conditional.

Regarding the CRI and CEI tasks.

Formally, given a context C containing a pair of events (e_1, e_2) and a set of sub-relationships (e.g., *before*, *after*) in a certain event semantic relation (e.g., temporal), the CRI task aims to determine whether (e_1, e_2) exhibits a specific sub-relationships. Take an example shown in Figure 1a Question "Did Messi's leadership in the 2022 World Cup final lead to Argentina's victory over France?" is a typical CRI example, which requires the model to judge whether it constitutes CAUSAL relationship according to the context.

While contextual logical event inference (CEI) task aims to infer the logical event with the specific event semantic relation (e.g., temporal). For example in Figure 1a, question "Where did Messi begin his football career before making his first-team debut in 2004?" requires the model to infer the logical event that happens before the mentioned event according to the context.

Regarding the processing operations of each dataset in the PIPER benchmark.

To construct the PIPER benchmark, we select existing available datasets according to the task requirements of CRI and CEI mentioned above, followed by processing each dataset specifically. The specific processing operations are detailed as follows:

(1) TempReason (Temp.) (Tan et al., 2023): It is a temporal contextual question answering dataset, containing 5,397 entries for l2 (event-time reasoning) and 4,426 entries for l3 (event-event reasoning). Here, we choose the l3 entries and keep those whose context length does not exceed 600 words. Last, we achieve 1,010 test samples and convert them into the instruction-tuning format of question answering.

(2) Timeqa (Tim.) (Chen et al., 2021): It is a temporal contextual question answering dataset, including two splits, $Easy$ and $Hard$, which are mainly distinguished based on whether the temporal clues involved in the question are strictly aligned in the context. In this work, we choose the hard split, achieving 630 test samples and converting them into the instruction-tuning format of question answering.

(3) Tracie (Tra.) (Zhou et al., 2021): It is a temporal contextual natural language inference dataset, consisting of 4,248 test samples with the answers yes or no to judge the event order. In this work, we exclude the samples with the repeated context, achieving 2,039 test samples and converting them into the instruction-tuning format of natural language inference. Note that we regard the hypothesis and the premise as the context and the question, respectively.

(4) Mctaco (Mct.) (Zhou et al., 2019): It is a temporal contextual multi-choices selections dataset, containing 1,332 questions with 9,442 options. We chose the test samples typed with "Event Ordering", achieving 234 test samples and converting them into the instruction-tuning format of multi-choices selections. Note that we exclude the test samples with the blank answer.

(5) Cider (Cid.) (Ghosal et al., 2021): It is a commonsense inference dataset in the conversational scenario. The conversation content contains rich event semantic relationships, including TEMPORAL, CAUSAL, SUBEVENT, and CONDITIONAL. Here, we classify samples according to the relationship tags in the samples and achieve 136 test samples in temporal type, 1469 in causal type, and 356 in Subevent&Conditional type. Last, we convert them into the instruction-tuning format of natural language inference. Note that the conversational contents are regarded as the context.

(6) Ester (Est.) (Han et al., 2021): It is a contextual question answering datasets, including the event semantics relations of CAUSAL, COUNTERFACTUAL, SUBEVENT&CONDITIONAL. Here, we classify samples according to the relationship tags in the samples and achieve 110 test samples in causal type, 36 test samples in counterfactual type, and 108 test samples in subevent&conditional type. Last, we convert them into the instruction-tuning format of question answering.

(7) Cster (Cst.) (Yang et al., 2022; Han et al., 2021): It is a merged version of the causal and counterfactual samples in the ESTER and CQA. CQA is also a contextual question answering dataset, including the event semantics of CAUSAL, and COUNTERFACTUAL. We exclude the samples with blank answers and classify samples according

to the relationship tags in the samples, achieving 1,617 test samples in causal type, and 1,921 test samples in counterfactual type. Last, we convert them into the instruction-tuning format of question answering.

(8) Cosmos (Cos.) (Huang et al., 2019): It is a single-choice selection reading comprehension dataset with contextual commonsense reasoning, including the event semantics of CAUSAL, COUNTERFACTUAL. We exclude the samples with blank answers and classify samples according to the relationship tags in the samples, achieving 2701 test samples in causal type, and 284 test samples in counterfactual type. Last, we convert them into the instruction-tuning format of single-choice selection.

(9) Ecare (Eca.) (Du et al., 2022): It is an explainable causal reasoning dataset, including the reasoning format of single-choice selection and question answering. Here, we choose the samples with the single-choice selection format, achieving 2,122 test samples. Last, we convert them into the instruction-tuning format of single-choice selection.

(10) IfQA (IFq.) (Yu et al., 2023b): It is a contextual counterfactual question answering dataset, characterized by the sample's question with the prefix "if". Here, we exclude samples with blank answers and achieve 700 test samples. Last, we convert them into the instruction-tuning format of question answering.

(11) WiQA (Wiq.) (Tandon et al., 2019): It is a contextual single-choice selection dataset, characterized by the sample's question with the format "what if". Here, we directly adopt its 3,003 test samples and convert them into the instruction-tuning format of single-choice selection.

The statistical results of the PIPER benchmark are reported in Table 7. Concretely, the number of samples with event temporal reasoning is 4,049, with causal temporal reasoning is 7,907, with counterfactual reasoning is 5,944, with subevent&conditional reasoning is 464. The number of samples with the format of contextual inter-event relational inference (CRI) (i.e., NLI) is 3,864, with the format of contextual logical event inference (CEI) (i.e., QA, MCQ, SCQ) being 14,502.

B Details on Benchmark Evaluation

In this section, as illustrated in Table 11, Table 12, and Table 13, we present prompt templates utilized for benchmark evaluation across three distinct settings: zero-shot, few-shot, and chain-of-thought (CoT). These templates encompass four reasoning formats: natural language inference (NLI), multiple-choice questions (MCQ), single-choice questions (SCQ), and question answering (QA). Concretely, in the few-shot setting, since the model can actively imitate the format of the demonstration's answer, we do not set the answer format in the prompt. Instead, in the CoT settings, to improve the accuracy of extracting answers from generated

Type-Dataset	Format	Nums
Temporal-Tempreason (Tem.)	QA	1,010
Temporal-Timeqa (Tim.)	QA	630
Temporal-Tracie (Tra.)	NLI	2,039
Temporal-Mctaco (Mct.)	MCQ	234
Temporal-Cider (Cid.)	NLI	136
Causal-Cster (Cst.)	QA	1,617
Causal-Cosmos (Cos.)	SCQ	2,701
Causal-Ecare (Eca.)	SCQ	2,122
Causal-Cider (Cid.)	NLI	1,469
Counterfactual-Cster (Cst.)	QA	1,957
Counterfactual-Cosmos (Cos.)	SCQ	284
Counterfactual-Ifqa (Ifq.)	QA	700
Counterfactual-Wiqa (Wiq.)	SCQ	3,003
Subevent&Conditional-Ester (Est.)	QA	108
Subevent&Conditional-Cider (Cid.)	NLI	356
Overall	QA, SCQ, MCQ, NLI	18,366

Table 7: The statistical results of the PIPER benchmark.

Method	Train	Infer/T	Infer/Ca	Infer/Co	Infer/S&C
SFT	61min	87.2ms	58.6ms	74.9ms	86.2ms
D^2E-SFT	86min	159.3ms	110.5ms	133.8ms	144.4ms

Table 8: Computational costs for LLaMA2-7b fine-tuning with D^2E-SFT strategy.

Method	T	Ca	Co	/S&C
Qwen1.5-7b (SFT)	59.7	71.7	71.2	41.7
Qwen1.5-7b (DIE) D^2E-SFT	69.0	78.2	72.9	48.3
Qwen1.5-7b (GPT4o-D^2E-SFT)	70.3	78.9	72.9	48.4
(Qwen1.5-7b-D^2E-SFT)	69.4	78.5	72.2	49.1

Table 9: Experimental results on refining the context by itself. $(\cdot)$ represents the model used to refine the context and the fine-tuning strategy

Method	CIDER	SCT	SocialIQA	ANLI
LLAMA2-7b (instruct)	56.2	82.5	59.5	63.2
Qwen1.5-7b (instruct)	46.8	95.6	72.7	75.6
Deepseek-7b (instruct)	45.6	91.9	66.2	73.0
LLAMA2-7b (D^2E-SFT)	62.0	96.2	69.5	75.3
Qwen1.5-7b (D^2E-SFT)	68.6	97.0	74.5	77.0
Deepseek-7b (D^2E-SFT)	61.1	97.6	72.5	78.8

Table 10: Experimental results on refining the context by itself. $(\cdot)$ represents the model used to refine the context and the fine-tuning strategy

contents, we prompt the model with some triggers to generate the answers with the specific format, such as "Therefore, the final answer is:". Note that the prompts stay consistent when evaluating the performance of fine-tuned models in section 4.4.

C Details on Error Pattern Analysis

To further investigate the underlying reasons for the inferior performance of aligned models in event reasoning. We demand GPT4o to combine input prompts and generated answers, identifying the following three error patterns: (1) neglecting the context; (2) positioning inaccurate context; (3) others, such as incorrect event logical reasoning, failed instruction following and so on. Here, we show how to utilize GPT4o to complete this task, the specific examples are shown in Table 14, Table 15 and Table 16.

D Details on Data Production

To overcome the scarcity of the event reasoning-oriented instruction-tuning data, we meticulously craft 10K diversions instruction-tuning demonstrations tailored for event reasoning. The steps are as follows: (1) We collect event reasoning-related training samples from the dataset involved in the PIPER benchmark. Then, we process each dataset consistent with the operations detailed in Appendix A to perform context filtering, improving data relevance. (2) We engage GPT4o to exclude samples whose context fails to explicitly or implicitly support the ground truth of the sample's question to improve the data rationality. Here, we show how to utilize GPT4o to complete this task, the specific examples are shown in Table 17, we introduce the prompt templates employed for data production across four reasoning formats: Natural Language Inference (NLI), Multiple-Choice Questions (MCQ), Single-Choice Questions (SCQ), and Question Answering (QA). Through the above two steps, we achieve a set of filtered training sets for each dataset. Then, we apply a heuristic ratio that is consistent with the dataset distribution in the PIPER benchmark to randomly select samples, integrating them into 10K raw data. We ensure the test samples in the PIPER benchmark are completely non-overlapping with the crafted 10K data.Note that we do not carefully adjust the sampling ratio and strategy for simplicity, which we leave for the future to further improve the data quality.

E Computational Costs

For the computational costs, since the data required for $\mathrm{D^2E-SFT}$ (e.g., imagined sample, context-refined sample) can be easily prepared in advance, no additional training costs will be introduced. During the training and inference stage, benefiting from the advantages of GPU parallel computing, we expand the batch size to calculate the logits of the imagined sample in real-time to reduce the time cost. The specific computational costs computed with 8 A100-40G GPUs are shown in the following tables (we report the whole training time cost, per sample inference time for each relational type, T, Ca, Co, and S&C represent the Temporal, Causal, Counterfactual, Subevent&Conditional in Table 8, respectively). Although the model's performance improves dramatically across various event relational types, we acknowledge that it takes more inference time due to the debiased operation. We will explore a more efficient reasoning strategy in the future.

F Bootstrap the Context-refined Sample

In the proposed distillation-enhanced SFT strategy, to ensure the accuracy of guiding the model in focusing on the significant contextual event information through KL divergence, we adopt the powerful GPT4o to generate context-refined samples. Note that GPT4o functions solely as a context purifier by filtering out irrelevant information from the original context, without adding any new context. The self-distillation operation is achieved by KL divergence distillation between the original and context-refined samples. This approach differs fundamentally from the direct distillation of GPT4o's responses, as the learning signal derives entirely from the original context and no knowledge transfer occurs from GPT4o. Additionally, we further explore practicality of utilizing LLM itself to refine the context and conduct a preliminary exploration on Qwen1.5-7b. The experimental results are shown in Table 10 (T, Ca, Co, S&C represent the Temporal, Causal, Counterfactual, Subevent&Conditional respectively, we present the average performance for each event relational type).

We can observe that utilizing LLM itself to refine the context leads to a slight performance drop compared to adopting GPT4o to refine the context. This is because the quality of the context-refined sample filtered by LLM itself has decreased, which is not conducive to adopting KL divergence regularization to guide the model focusing on significant contextual information. However, utilizing LLM itself to refine the context is still far superior to the standard SFT strategy and outperforms the ablation variant Qwen-7b (DIE) which only adopts the debiased-enhanced SFT strategy), demonstrating the feasibility and effectiveness of utilizing LLM itself to refine the context.

G Out of Distribution Evaluation

In fact, our proposed PIPER Benchmark itself includes out-of-distribution (OOD) evaluations (i.e., CIDER). Note that our crafted 10K instruction tuning data does not contain CIDER training samples. Consequently, the CIDER dataset is evaluated on the OOD settings. To further demonstrate the effectiveness of our data and strategy, we additionally adopt the representatively comprehensive datasets (SCT, SocialIQA, ANLI) in the field of event reasoning to conduct the OOD evaluation. The experimental results are shown in the following tables (we evaluate the frozen instruct model, and our proposed D^2E-SFT model, the performance on the CIDER dataset is calculated by averaging the values of CIDER in all event relational types).

Zero-Shot Prompt Templates for Four Reasoning Formats	
Format	**Prompt**
NLI	
System	`You are an expert in understanding and analyzing various event relationships, including temporal, causal, counterfactual, hierarchical, and conditional relations. Your task is to accurately comprehend both explicit and implicit event relationships in the premise, perform reasonable event reasoning, and determine whether the hypothesis is compatible with the premise. Answer the question concisely without additional explanations. Respond in the format: 'The answer is: positive' (if the hypothesis is compatible with the premise) or 'The answer is: negative' (if the hypothesis is incompatible with the premise).`

User	`Premise: {Premise}`
	`Hypothesis: {Hypothesis}`
	`Question: Is the hypothesis compatible with the premise?`
	`Answer:`

MCQ	
System	`You are an expert in understanding and analyzing various event relationships, including temporal, causal, counterfactual, hierarchical, and conditional relations. Your task is to accurately comprehend both explicit and implicit event relationships in the context, perform reasonable event reasoning, and determine which options are the solution to the question (there may be multiple correct options). List only the correct options to answer the question concisely. Respond in the format: 'The correct options are:______'. and do not provide additional explanations.`

User	`Context: {Context}`
	`Question: {Question}. {Options}`
	`Answer:`

SCQ	
System	`You are an expert in understanding and analyzing various event relationships, including temporal, causal, counterfactual, hierarchical, and conditional relations. Your task is to accurately comprehend both explicit and implicit event relationships in the context, perform reasonable event reasoning, and determine which option is the solution for the question (there is only one correct option). List only the correct option to answer the question concisely. Respond in the format: 'The correct option is:______'. and do not provide additional explanations.`

User	`Context: {Context}`
	`Question: {Question}. {Options}`
	`Answer:`

QA	
System	`You are an expert in understanding and analyzing various event relationships, including temporal, causal, counterfactual, hierarchical, and conditional relations. Your task is to accurately comprehend both explicit and implicit event relationships in the context, perform reasonable event reasoning, and answer the question concisely without additional explanations. Respond in the format: 'The answer is:______'`

User	`Context: {Context}`
	`Question: {Question}`
	`Answer:`

Table 11: Zero-shot prompt templates with a predefined response format. System denotes the system prompt, while User signifies the subsequent user input. **NLI**, **MCQ**, **SCQ**, and **QA** represent four distinct reasoning formats: Natural Language Inference, Multiple-Choice Question, Single-Choice Question, and Question Answering, respectively. The notation $\{\cdot\}$ represents slots intended to be filled by the corresponding field value of a sample.

Few-Shot Prompt Templates for Four Reasoning Formats	
Format	**Prompt**
NLI	
System	You are an expert in understanding and analyzing various event relationships, including temporal, causal, counterfactual, hierarchical, and conditional relations. You accurately comprehend both explicit and implicit event relationships in the premise, perform reasonable event reasoning, and determine whether the hypothesis is compatible with the premise. You answer questions concisely, using only the 'positive'(compatible) or 'negative'(incompatible).
User	[Premise: {Premise} \n Hypothesis: {Hypothesis} \n Question: Is the hypothesis compatible with the premise? \n Answer: {Answer}] ×3 Premise: {Premise} Hypothesis: {Hypothesis} Question: Is the hypothesis compatible with the premise? Answer:
MCQ	
System	You are an expert in understanding and analyzing various event relationships, including temporal, causal, counterfactual, hierarchical, and conditional relations. You accurately comprehend both explicit and implicit event relationships in the context, perform reasonable event reasoning, and determine which options are the solutions to the question (there may be multiple correct options). You answer questions concisely, listing only all the correct options.
User	[Context: {Context} \n Question: {Question}. {Options} \n Answer: {Answer}] ×3 Context: {Context} Question: {Question}. {Options} Answer:
SCQ	
System	You are an expert in understanding and analyzing various event relationships, including temporal, causal, counterfactual, hierarchical, and conditional relations. You accurately comprehend both explicit and implicit event relationships in the context, perform reasonable event reasoning, and determine which option is the solution to the question (there is only one correct option). You answer questions concisely, providing only the correct option.
User	[Context: {Context} \n Question: {Question}. {Options} \n Answer: {Answer}]×3 Context: {Context} Question: {Question}. {Options} Answer:
QA	
System	You are an expert in understanding and analyzing various event relationships, including temporal, causal, counterfactual, hierarchical, and conditional relations. You accurately comprehend both explicit and implicit event relationships in the context, perform reasonable event reasoning, and answer the corresponding questions concisely.
User	[Context: {Context} \n Question: {Question} \n Answer: {Answer}] ×3 Context: {Context} Question: {Question} Answer:

Table 12: Few-shot prompt templates with three demonstrations. System denotes the system prompt, while User signifies the subsequent user input. **NLI**, **MCQ**, **SCQ**, and **QA** represent four distinct reasoning formats: Natural Language Inference, Multiple-Choice Question, Single-Choice Question, and Question Answering, respectively. The notation $\{\cdot\}$ represents slots intended to be filled by the corresponding field value of a sample. Similarly, the notation $\{\cdot\}$ identifies slots to be filled by the corresponding field value of a demonstration sample. The notation $\times 3$ denotes a 3-shot prompt configuration.

CoT Prompt Templates for Four Reasoning Formats	
Format	**Prompt**
NLI	
System	You are an expert in understanding and analyzing various event relationships, including temporal, causal, counterfactual, hierarchical, and conditional relations. Your task is to accurately comprehend both explicit and implicit event relationships in the premise, perform reasonable event reasoning, and determine with evidence whether the hypothesis is compatible with the premise.
User	Premise: {Premise} Hypothesis: {Hypothesis} Question: Is the hypothesis compatible with the premise? Answer: Let's think step by step and end with the format: "Therefore, the final answer is: positive (if the hypothesis is compatible with the premise)" or "Therefore, the final answer is: negative (if the hypothesis is incompatible with the premise)"
MCQ	
System	You are an expert in understanding and analyzing various event relationships, including temporal, causal, counterfactual, hierarchical, and conditional relations. Your task is to accurately comprehend both explicit and implicit event relationships in the context, perform reasonable event reasoning, and determine with evidence which options are the solution to the question (there may be multiple correct options).
User	Context: {Context} Question: {Question}. {Options} Answer: Let's think step by step and end with the format by listing only the correct options: "Therefore, the final correct options are:______".
SCQ	
System	You are an expert in understanding and analyzing various event relationships, including temporal, causal, counterfactual, hierarchical, and conditional relations. Your task is to accurately comprehend both explicit and implicit event relationships in the context, perform reasonable event reasoning, and determine with evidence which option is the solution for the question (there is only one correct option).
User	Context: {Context} Question: {Question}. {Options} Answer: Let's think step by step and end with the format by listing only the correct option: "Therefore, the final correct option is:______".
QA	
System	You are an expert in understanding and analyzing various event relationships, including temporal, causal, counterfactual, hierarchical, and conditional relations. Your task is to accurately comprehend both explicit and implicit event relationships in the context, perform reasonable event reasoning and provide evidence-supported answer to the question (the evidence should at least appear in the context).
User	Context: {Context} Question: {Question} Answer: Let's think step by step and end with the format: "Therefore, the final answer is:______".

Table 13: CoT prompt templates accompanied by "Let's think step by step." and a predefined end format. System denotes the system prompt, while User signifies the subsequent user input. **NLI**, **MCQ**, **SCQ**, and **QA** represent four distinct reasoning formats: Natural Language Inference, Multiple-Choice Question, Single-Choice Question, and Question Answering, respectively. The notation $\{\cdot\}$ represents slots intended to be filled by the corresponding field value of a sample.

Neglect the context
You are an expert in understanding and analyzing various event relationships, including temporal, causal, counterfactual, hierarchical, and conditional relations. Given the premise, the hypothesis, the question and the generated answer that contains the intermediate reasoning analysis, and the groundtruth answer (Note that in the given generated answer and groundtruth answer, 'positive' means that the hypothesis is compatible with the premise, and 'negative' means the hypothesis is incompatible with the premise), your task is to accurately comprehend both explicit and implicit event relationships in the premise, perform reasonable event reasoning, and precisely determine the error patterns of the generated answers. The error patterns include: (1) neglect the context, which intermediate reasoning analysis contains hallucinated context or the final generated answer is not included in the context, thus deviating the given context; (2) positioning inaccurate context, which denotes the model fails to emphasize the significant event information in the context, resulting in wrong answers; (3) others, such as incorrect event logical reasoning and so on. **Premise:** A: Why, Ann, I haven't seen you for ages. How have you been? B: Bob, it has been a long time, hasn't it? It must be at least a year. A: No. We talked at the Johnson's Christmas party last December. Don't you remember? B: That's right. Well, how are you? Still working for the Germen company? A: No. I changed my job about 4 months ago. I'm with the World Bank now. How about you? B: I'm still working at the University, but I transferred from the German Department to the Spanish Department. A: Don't you tell me you speak Spanish, too? B: Sure. I grew up in Spain, you know. German was what I studied in college. A: All I have ever managed to learn is a little Italian. B: Oh. excuse me, someone is knocking at the door. I'll talk to you again later, Bob. A: It was nice talking to you, Ann. **Hypothesis:** 'someone is knocking at the door' causes 'talk to you again later'. **Question:** Is the hypothesis compatible with the premise? **Generated Answer:** Great! Let's analyze the event relationships in the premise and determine whether the hypothesis is compatible with it. 1. Temporal relationship: The conversation between A and B takes place at some point in the past, as A mentions that it has been a long time since they last saw each other. 2. Causal relationship: There is no direct causal relationship between "someone is knocking at the door" and "talk to you again later". However, it is possible that the person knocking at the door is a mutual friend or acquaintance who will introduce A and B again, leading to another conversation. 3. Counterfactual relationship: There is no counterfactual relationship in the premise. 4. Hierarchical relationship: There is no hierarchical relationship between the events in the premise. 5. Conditional relationship: There is no conditional relationship in the premise. Now, let's evaluate the compatibility of the hypothesis with the premise: Hypothesis: 'someone is knocking at the door' causes 'talk to you again later'. Analysis: The hypothesis suggests that the person knocking at the door will lead to another conversation between A and B. However, there is no direct evidence in the premise to support this claim. While it is possible that the person knocking at the door will introduce A and B again, there is no guarantee that they will talk again. Therefore, the hypothesis is not compatible with the premise. Therefore, the final answer is: negative. The hypothesis is incompatible with the premise. **Groundtruth Answer:** positive Let's think step by step and end with the format: "Therefore, the error patten is one" (if you deem that the error pattern is: (1) neglect the context) or "Therefore, the error patten is two" (if you deem that the error pattern is: (2) positioning inaccurate context) or "Therefore, the error patten is three" (if you deem that the error pattern is: (3) others, such as incorrect event logical reasoning and so on.)
GPT-4o Output: To determine the error pattern in the generated answer, let's analyze the premise and the hypothesis step by step: 1. **Premise Context**: The conversation between A and B ends with B saying, "Oh, excuse me, someone is knocking at the door. I'll talk to you again later, Bob." This indicates that the knocking at the door is the reason B ends the conversation and suggests they will continue it later. 2. **Hypothesis**: The hypothesis states that 'someone is knocking at the door' causes 'talk to you again later'. This implies a causal relationship where the knocking at the door is the reason for postponing the conversation to a later time. 3. **Generated Answer Analysis**: - The generated answer incorrectly states that there is no direct causal relationship between "someone is knocking at the door" and "talk to you again later". However, the premise clearly shows that the knocking is the reason B gives for ending the conversation and planning to continue it later. - The generated answer introduces an unnecessary and speculative scenario about a mutual friend or acquaintance,which is not present in the premise. This is an example of neglecting the context by introducing hallucinated context. 4. **Groundtruth Answer**: The groundtruth answer is positive, indicating that the hypothesis is indeed compatible with the premise, as the knocking at the door is the reason for the conversation being postponed. Therefore, the error pattern in the generated answer is due to neglecting the context by introducing speculative reasoning not supported by the premise. The correct understanding of the premise should have recognized the direct causal relationship as stated in the hypothesis. Therefore, the error pattern is one.

Table 14: The first type of error identified by GPT-4o, originating from the analysis results for LLaMA2-7b's generated answers. In the **Context** field, gold-highlighted text denotes ground truth-related content. In the **Generated Answer** field, red-highlighted text indicates the location of major errors. In the **GPT-4o Output** field, blue-highlighted text signifies the key explanation of why the generated answer is incorrect.

Position inaccurate context

You are an expert in understanding and analyzing various event relationships, including temporal, causal, counterfactual, hierarchical, and conditional relations. Given the context, the question, the generated answer that contains the intermediate reasoning analysis, and the groundtruth answer, your task is to accurately comprehend both explicit and implicit event relationships in the context, perform reasonable event reasoning and precisely determine the error patterns of the generated answers. The error patterns include: (1) neglect the context, which denotes intermediate reasoning analysis contains hallucinated context or the final generated answer is not included in the context, thus deviating the given context; (2) positioning inaccurate context, which denotes the model fails to emphasize the significant event information in the context, resulting in wrong answers; (3) others, such as incorrect event logical reasoning and so on.

Context: In addition, the potential for cross-selling of services between Teladoc and Livongo is substantial as the two companies had less than a 25% overlap in client base before the merger. We build \$100 million of cost synergies and \$500 million of revenue synergies into our forecast by 2025, as sales and operations between the two companies consolidate and clinicians in the Teladoc network can directly refer members to Livongo. With Livongo, Teladoc also expands its data set (12.5 million-plus annual visits, 2 million weekly blood glucose readings) and in the long term can potentially use that scale to its advantage and integrate data products and informatics with Teladoc's offerings. The company's addressable market likely exceeds our estimates of core telehealth services, but we have excluded significant cross-selling opportunities due to the nascent size of the virtual health market. For the core telehealth services, we believe the company is likely to retain its first-mover advantage and scale by penetrating members of existing customers in addition to adding members of new clients under new logos. These market estimates do not include the potential to add US government payers under the recent government approval to employ telehealth services.

Question: What will happen if the nascent size of the virtual health market?

Generated Answer: Great, let's analyze the given context and answer the question. The context mentions that the potential for cross-selling of services between Teladoc and Livongo is substantial, and the companies had less than a 25% overlap in their client base before the merger. This suggests that there is a significant opportunity for growth through the consolidation of sales and operations between the two companies. Furthermore, the context states that Teladoc expands its data set with Livongo, which includes 12.5 million-plus annual visits and 2 million weekly blood glucose readings. This increase in data scale can potentially be used to the company's advantage in the long term, such as integrating data products and informatics with Teladoc's offerings. However, the context also mentions that the market size for virtual health services is nascent, which could potentially limit the company's growth.
Therefore, the final answer is: If the nascent size of the virtual health market, Teladoc's growth may be limited by the market's small size, which could hinder the company's ability to scale and achieve its full potential.

Groundtruth Answer: excluded significant cross-selling opportunities

Let's think step by step and end with the format: "Therefore, the error patten is one" (if you deem that the error pattern is: (1) neglect the context) or "Therefore, the error patten is two" (if you deem that the error pattern is: (2) positioning inaccurate context) or "Therefore, the error patten is three" (if you deem that the error pattern is: (3) others, such as incorrect event logical reasoning and so on.)

GPT-4o Output: Let's analyze the context and the generated answer step by step:
1. **Context Analysis**: The context discusses the potential for cross-selling services between Teladoc and Livongo, highlighting that there was less than a 25% overlap in their client base before the merger. It mentions the forecasted cost and revenue synergies by 2025 due to the consolidation of sales and operations. The context also notes that the nascent size of the virtual health market leads to the exclusion of significant cross-selling opportunities.
2. **Question Understanding**: The question asks about the implications of the nascent size of the virtual health market.
3. **Generated Answer Analysis**: The generated answer suggests that the nascent size of the virtual health market could limit Teladoc's growth, hindering its ability to scale and achieve its full potential. However, this interpretation does not directly address the specific point in the context about excluding significant cross-selling opportunities due to the nascent market size.
4. **Groundtruth Answer**: The groundtruth answer directly states that significant cross-selling opportunities are excluded due to the nascent size of the virtual health market.
5. **Error Pattern Identification**: The generated answer fails to emphasize the specific event information in the context regarding the exclusion of significant cross-selling opportunities. Instead, it provides a more general interpretation about growth limitations, which deviates from the specific context provided.
Therefore, the error pattern is two.

Table 15: The second type of error identified by GPT-4o, originating from the analysis results for LLaMA2-7b's generated answers. In the **Context** field, gold-highlighted text denotes ground truth-related content. In the **Generated Answer** field, red-highlighted text indicates the location of major errors. In the **GPT-4o Output** field, blue-highlighted text signifies the key explanation of why the generated answer is incorrect.

Others, such as incorrect event logical reasoning
You are an expert in understanding and analyzing various event relationships, including temporal, causal, counterfactual, hierarchical, and conditional relations. Given the context, the question, the generated answer that contains the intermediate reasoning analysis, and the groundtruth answer, your task is to accurately comprehend both explicit and implicit event relationships in the context, perform reasonable event reasoning and precisely determine the error patterns of the generated answers. The error patterns include: (1) neglect the context, which denotes intermediate reasoning analysis contains hallucinated context or the final generated answer is not included in the context, thus deviating the given context; (2) positioning inaccurate context, which denotes the model fails to emphasize the significant event information in the context, resulting in wrong answers; (3) others, such as incorrect event logical reasoning and so on. **Context:** Context: Edmund Rüdiger Stoiber (born 28 September 1941) is a German politician who served as the 16th Minister President of the state of Bavaria between 1993 and 2007 and chairman of the Christian Social Union (CSU) between 1999 and 2007. In 2002 , he ran for the office of Chancellor of Germany in the federal election , but in one of the narrowest elections in German history lost against Gerhard Schröder. On 18 January 2007 , he announced his decision to step down from the posts of minister-president and party chairman by 30 September. The election was one of modern Germanys closest votes. During his time as Minister-President of Bavaria , Stoiber pushed for the construction of a roughly 40-kilometer high-speed magnetic-levitation link from Munichs main station to its airport, to be built by Transrapid International , a consortium including ThyssenKrupp and Munich-based Siemens. After he left office , the German federal government abandoned the plans in 2008 because of spiraling costs of as much as €3.4 billion. **Question:** What position did Edmund Stoiber take after Mar 2001? **Generated Answer:** Great, let's analyze the given text to answer the question. Edmund Stoiber served as the Minister-President of Bavaria between 1993 and 2007, and as the chairman of the Christian Social Union (CSU) between 1999 and 2007. In 2002, he ran for the office of Chancellor of Germany in the federal election, but lost against Gerhard Schröder. On January 18, 2007, he announced his decision to step down from the posts of minister-president and party chairman by September 30. Now, let's think about the question: What position did Edmund Stoiber take after March 2001? March 2001 is a significant date in the text because it is mentioned that Stoiber pushed for the construction of a high-speed magnetic-levitation link from Munich's main station to its airport during his time as Minister-President of Bavaria. Therefore, the final answer is: After March 2001, Edmund Stoiber took the position of pushing for the construction of a high-speed magnetic-levitation link from Munich's main station to its airport. **Groundtruth Answer:** Minister-President of Bavaria Let's think step by step and end with the format: "Therefore, the error patten is one" (if you deem that the error pattern is: (1) neglect the context) or "Therefore, the error patten is two" (if you deem that the error pattern is: (2) positioning inaccurate context) or "Therefore, the error patten is three" (if you deem that the error pattern is: (3) others, such as incorrect event logical reasoning and so on.)
GPT-4o Output: To determine the error pattern in the generated answer, let's analyze the context and the reasoning process: 1. **Contextual Information**: The context provides information about Edmund Stoiber's political career, specifically his roles as Minister-President of Bavaria from 1993 to 2007 and chairman of the Christian Social Union (CSU) from 1999 to 2007. It also mentions his candidacy for Chancellor of Germany in 2002 and his announcement in January 2007 to step down from his positions by September 30, 2007. 2. **Question Analysis**: The question asks for the position Edmund Stoiber took after March 2001. The context clearly states that he was the Minister-President of Bavaria during this period, which is the correct answer. 3. **Generated Answer Analysis**: The generated answer incorrectly identifies Stoiber's role after March 2001 as "pushing for the construction of a high-speed magnetic-levitation link." This is an activity he was involved in during his tenure as Minister-President, not a new position he took after March 2001. The generated answer fails torecognize that his official position remained Minister-President of Bavaria. 4. **Error Pattern Identification**: The error in the generated answer arises from incorrect event logical reasoning. The model misinterprets an activity associated with Stoiber's role as a new position, rather than recognizing that his official position remained unchanged as Minister-President of Bavaria. Therefore, the error pattern is three.

Table 16: The third type of error identified by GPT-4o, originating from the analysis results for LLaMA2-7b's generated answers. In the **Context** field, gold-highlighted text denotes ground truth-related content. In the **Generated Answer** field, red-highlighted text indicates the location of major errors. In the **GPT-4o Output** field, blue-highlighted text signifies the key explanation of why the generated answer is incorrect.

GPT-4o Filter Prompt Templates for Four Reasoning Formats	
Format	**Prompt**
NLI	You are an expert in understanding and analyzing various event relationships, including temporal, causal, counterfactual, hierarchical, and conditional relations. Your task is to accurately comprehend event relationships in the premise, perform reasonable event reasoning, and determine with evidence whether the given answer is the correct solution to the given question (if the answer is the correct solution to the question, it should be explicitly or implicitly inferred from the context). Note that the given answer 'positive' means that the hypothesis is compatible with the premise, the given answer 'negative' means the hypothesis is incompatible with the premise. Premise: {Premise} Question: {Question} Answer: {Answer} Let's think step by step and end with the format: "Therefore, the answer is correct" (if you deem that the answer is the correct solution to the question) or "Therefore, the answer is wrong" (if you deem that the answer is the wrong solution to the question).
MCQ	You are an expert in understanding and analyzing various event relationships, including temporal, causal, counterfactual, hierarchical, and conditional relations. Your task is to accurately comprehend event relationships in the context, perform reasonable event reasoning, and determine with evidence whether the given answer is the correct solution to the given question (if the answer is the correct solution to the question, it should be explicitly or implicitly inferred from the context). Note that there may be multiple correct options. Context: {Context} Question: {Question}. {Options} Answer: {Answer}. Let's think step by step and end with the format: "Therefore, the answer is yes" (if you deem that the answer is the correct solution to the question) or "Therefore, the answer is no" (if you deem that the answer is the wrong solution to the question).
SCQ	You are an expert in understanding and analyzing various event relationships, including temporal, causal, counterfactual, hierarchical, and conditional relations. Your task is to accurately comprehend event relationships in the context, perform reasonable event reasoning, and determine with evidence whether the given answer is the correct solution to the given question (if the answer is the correct solution to the question, it should be explicitly or implicitly inferred from the context). Note that there is only one correct option. Context: {Context} Question: {Question}. {Options} Answer: {Answer} Let's think step by step and end with the format: "Therefore, the answer is yes" (if you deem that the answer is the correct solution to the question) or "Therefore, the answer is no" (if you deem that the answer is the wrong solution to the question).
QA	You are an expert in understanding and analyzing various event relationships, including temporal, causal, counterfactual, hierarchical, and conditional relations. Your task is to accurately comprehend event relationships in the context, perform reasonable event reasoning, and determine with evidence whether the given answer is the correct solution to the given question (if the answer is the correct solution to the question, it should be explicitly or implicitly inferred from the context). Context: {Context} Question: {Question} Answer: {Answer} Let's think step by step and end with the format: "Therefore, the answer is yes" (if you deem that the answer is the correct solution to the question) or "Therefore, the answer is no" (if you deem that the answer is the wrong solution to the question).

Table 17: GPT-4o Filter Prompt Templates. **NLI**, **MCQ**, **SCQ**, and **QA** represent four distinct reasoning formats: Natural Language Inference, Multiple-Choice Question, Single-Choice Question, and Question Answering, respectively. The notation $\{\cdot\}$ represents slots intended to be filled by the corresponding field value of a sample.