WizardEvent: Empowering Event Reasoning by Hybrid Event-Aware Instruction Tuning

Anonymous ACL submission

Abstract

Events refer to specific occurrences, incidents, or happenings that take place under a particular background. Event reasoning aims to reason according to certain relations. The cutting-edge techniques for event reasoning play crucial and fundamental abilities underlying various natural language processing applications. Large language models (LLMs) have made significant advancements in event reasoning owing to their wealth of knowledge and reasoning capabilities. However, open-source LLMs currently in use do not consistently demonstrate exceptional proficiency in managing event reasoning. This discrepancy arises from insufficient learning of knowledge of event relations and incomplete reasoning paradigms. In this paper, we propose WIZARDEVENT, the hybrid event-aware instruction tuning leading to better event reasoning abilities. Specifically, we first represent the events and relation of the event relational knowledge in a novel structure. We then mine the knowledge from raw text. Second, we introduce the prototypical event reasoning paradigms which include four reasoning formats. Lastly, we wrap our constructed event relational knowledge with our reasoning paradigms to create the instruction tuning dataset. We fine-tune to obtain WIZARDEVENT using this enriched dataset, significantly improving their event reasoning. The performance of WIZARDEVENT is rigorously evaluated through a series of extensive experiments across 10 event reasoning tasks. We also annotate a new dataset for event relational knowledge evaluation. The results from these evaluations demonstrate that WIZARDEVENT substantially outperforms other instruction-tuned models, indicating the success of our approach in enhancing LLMs' proficiency in event reasoning.

1 Introduction

Events are instances or occurrences that form the basic semantic building units in natural language. Event Reasoning (ER) is the ability to process and analyze the complex interconnections between events. This involves training models to understand the dynamics of event progression in the real world (Tao et al., 2023a). As a fundamental competency within Large Language Models (LLMs), ER supports a multitude of Natural Language Processing (NLP) tasks, including recommendation engines (Yang et al., 2020), interactive question-answer systems (Souza Costa et al., 2020), and AI Agents (Liu et al., 2023). Therefore, ER is essential for the advancement of LLMs.

Unlocking the full potential of ER hinges on the mastery of various reasoning abilities and event relational knowledge which reflects the logic of events and their relations. Humans excel in ER due to their extensive acquisition of such relational knowledge and proficient reasoning of comprehensive paradigms. In contrast, the proficiency of current open-source LLMs for ER is lacking (Tao et al., 2023a). There are primarily two reasons. First, the skill set of present-day open-source LLMs is predominantly stimulated by instruction-tuning (Taori et al., 2023; Chiang et al., 2023; Xu et al., 2023). However, datasets used for instruction-tuning often contain sparse critical event relational knowledge, resulting in an underdeveloped understanding of this essential information. Second, the realm of ER encompasses a variety of reasoning patterns, which includes both speculating events (Du et al., 2022; Sap et al., 2019) and complex inter-event relationship reasoning (Ning et al., 2018; Caselli and Vossen, 2017). Current methods of generating instructions face challenges in capturing this broad spectrum of reasoning paradigms (Wang et al., 2022b). It also incurs unbalanced abilities of different relations and paradigms.

In an effort to overcome the challenges outlined, we introduce WIZARDEVENT , which is obtained through our innovative **Ev**ent-oriented **I**nstruction

Tuning approach. To counteract the scarcity of event relational knowledge in training data, we design a novel formulation to represent the event relational knowledge. Furthermore, we present a technique for mining event relational knowledge from unprocessed datasets. Second, we aim to thoroughly encompass the various paradigms of ER by introducing what we term the hybrid event reasoning. We design four reasoning formats that collectively establish the foundational competencies necessary for effective ER. Lastly, we craft templates for each reasoning format and the corresponding inter-relations. These templates are subsequently integrated within our instruction tuning process, paired with the ER paradigms. Utilizing the resulting dataset, we fine-tune open-source LLMs, thereby enhancing their abilities to execute ER informed by event relational knowledge.

We conduct extensive experiments to testify to the effectiveness of WIZARDEVENT. We first evaluate the performance of WIZARDEVENT across 10 tasks of ER. We then annotate a new dataset for the evaluation of event relational knowledge. Results of automatic and human evaluations show that WIZARDEVENT outperforms other instruction-tuned models.

We summarize our contributions:

- We propose to enhance the basic ER ability for developing better LLMs. We present a novel formulation of event relational knowledge and introduce a method to construct it.
- We design the hybrid event reasoning. We then encapsulate event relational knowledge into the training dataset with the proposed paradigm. We then finetune open-source LLMs to obtain our model. The novel event-oriented formulation and reasoning paradigm may also shed light on other event-oriented methods.
- We conduct extensive experiments on 10 test sets for testing the abilities of ER and an annotated test set for event relational knowledge. Results show the effectiveness of WIZARDEVENT.

2 WIZARDEVENT Methodology

2.1 Overview

Our primary aim is to achieve an enhanced comprehension of event relational knowledge and various reasoning formats. An overview of the WIZARDEVENT training and evaluation process is illustrated in Figure 1. To accomplish this objective, we begin by introducing our formulation of event relational knowledge in Section 2.2. We also design a method to mine the knowledge from the raw dataset in Section 2.2. Then we propose the hybrid event reasoning paradigm which consists of four types of reasoning formats in Section 2.3. Last, we constructed instruction-tuning dataset with our event relational knowledge and reasoning paradigm.

2.2 Event Relational Knowledge

Formulation Existing LLMs fall short in learning event relational knowledge since it is sparse in finetuning datasets. These models tend to stimulate the abilities of event reasoning insufficiently and imbalancedly. In an endeavor to mitigate the limitation, we enhance the models with event relational knowledge. We initially introduce the formulation of event relational knowledge, which encompasses events and their inter-relations.

Given the $\mathbb{K}$, a set of event relational knowledge, any item in it is represented as $\mathcal{K} = (\mathcal{C}, \mathcal{E}^h, \mathcal{R}, \mathcal{E}^t)$, $\mathcal{K} \in \mathbb{K}$. $\mathcal{E}^h$ is the head event, $\mathcal{E}^t$ is the tail event, and $\mathcal{R}$ is the relation between them. $\mathcal{C}$ is the context describing the background information of both events. $\mathcal{K}$ entails rich semantic information of events. $\mathcal{K}$ is also rich in event relational and structural knowledge since it captures event inter-relations. Besides, $\mathcal{K}$ captures the necessary information for the events by including the context. Contextual information is important for an accurate understanding of events, as in the absence of contextual information, the understanding of the event is often prone to ambiguity. In summary, using $\mathcal{K}$ to capture different aspects of events may reduce the risk of event misunderstanding and enhance the conceptions of structure and semantics of the events, thereby improving the accuracy of achieving event reasoning.

Construction This section details the construction of $\mathcal{K}$, extracted from BookCorpus (Zhu et al., 2015). Initially, tail events $\mathcal{E}^t$ are located by identifying connectives indicative of relations, similar to the approach in Zhou et al. (2022a) using PDTB (Prasad et al., 2008). A child node with a VERB part-of-speech tag from these connectives is considered a tail event trigger. This VERB triggers a dependency tree traversal, capturing a verb-rooted subset of words forming $\mathcal{E}^t$.

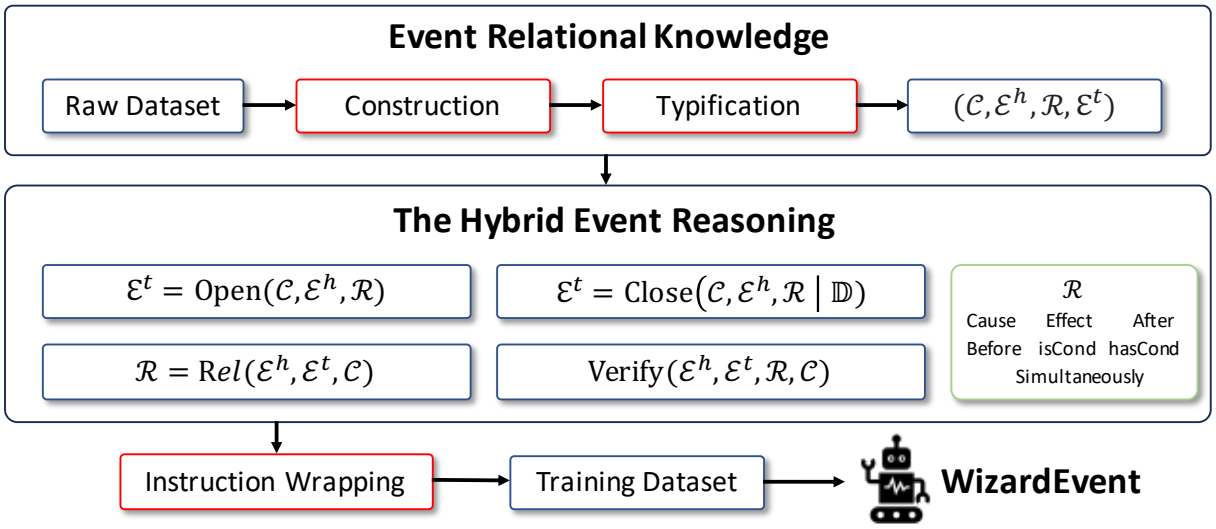

Figure 1: Overview of training process and evaluation of WIZARDEVENT. The training process encompasses Event-Oriented Instruction Tuning and Construction of event relational knowledge.

The extraction of the head event $\mathcal{E}^h$, relation $\mathcal{R}$, and contextual information $\mathcal{C}$ for $\mathcal{K}$ follows. Extracting $\mathcal{E}^h$ proves more complex due to the indirect linkage between its trigger and the relational connective. Instead of linguistic rules, an end-to-end relation parser like ASER's (Zhang et al., 2020) is used to analyze the text containing $\mathcal{E}^t$ and extract $\mathcal{E}^h$ and the connecting relations [1]. The parser outputs the relation $\mathcal{R}$. We focus on a predefined set of relations:

$$\mathcal{R} \in \mathbb{S}^{\mathcal{R}} = \{\texttt{Cause}, \texttt{Effect}, \texttt{After}, \texttt{Before}, \texttt{isCond}, \texttt{hasCond}, \texttt{Simultaneously}\}.$$

We concatenate sentences before the sentence of $\mathcal{E}^h$ as the context $\mathcal{C}$. Thus far, we have accomplished the construction of $\mathcal{K}$.

Typification After the construction, $\mathcal{K}$ could be atypical event relational knowledge. However, $\mathcal{K}$ should be salient patterns. Therefore, we conduct a further typification process.

For each $\mathcal{K} \in \mathbb{K}$, we extract the verbs of the head and tail events via dependency parsing. We obtain the pattern $(v_h, \mathcal{R}, v_t)$ of $\mathcal{K}$. We count the frequencies of all patterns for all $\mathcal{K} \in \mathbb{K}$. For each top frequent pattern, we sample 5 $\mathcal{K}$ of it. Then we collect our final salient event relational knowledge.

2.3 The Hybrid Event Reasoning

With event relational knowledge, existing LLMs are still limited in learning various event reasoning formats. To overcome this deficiency, we induce hybrid event reasoning paradigm. We focus on four types of event reasoning formats covering practical needs. We introduce them in the following.

Open Event Reasoning This is the most common ability that requires the model to generate tail event $\mathcal{E}^t$ based on head event $\mathcal{E}^h$, and context $\mathcal{C}$ according to relation $\mathcal{R}$:

$$\mathcal{E}^t = \text{Open}(\mathcal{E}^h, \mathcal{R}, \mathcal{C}). \quad (1)$$

Through learning to generate events, the model's comprehension of the event semantics is stimulated, enabling it to accomplish the event reasoning tasks in a manner more aligned with human understanding. Moreover, the model learns to draw proper information from the context to answer the event reasoning questions more precisely.

Close Event Reasoning Models should be able to discriminate the wrong events. Similar to DPO training (Rafailov et al., 2023), to enhance the model's event discrimination ability, we incorporate close event reasoning:

$$\mathcal{E}^t = \text{Close}(\mathcal{E}^h, \mathcal{R}, \mathcal{C} \mid \mathbb{D}). \quad (2)$$

$\mathbb{D}$ is the set of the candidate events including ground-truth tail event $\mathcal{E}^t$ and several negative candidates. Thus, close event reasoning is a multiple-choice formulation. This learning process further reinforced the model's comprehension of events and their interrelationships, enhancing the model's discriminative capabilities of event relational knowledge .

Relation Reasoning Determining the relations between events is another basic ability that has significant applications. We include it into our

[1] Only $\mathcal{E}^h$ preceding the tail event are considered.

Instructions:
Give me 5 instructions which are questions of the question type based on all input data.

Question type:
[$\mathcal{T}$]

Input data:
[context]: the context information.
[event1]: the first event.
[event2]: the second event.

Requirements:
1. The instructions should strictly be the question type asked.
2. If [context] in input data, ensure the instructions must include the placeholder name [context]. If [event1] in input data, ensure the instructions must include placeholder name [event1]. If [event2] in input data, ensure the instructions must include placeholder name [event2].
3. The instructions should be diversified.

Generate:

Generation Examples:
Cause – Close
Choose the event that is the direct cause of [event1] in the provided [context].
After - Open
In the scenario described by [context], what is the subsequent event after [event1]?
Before – Rel
Based on the [context], what is the sequence of occurrences between [event1] and [event2]?
isCond – Verify
Within the [context] provided, is [event2] an essential condition that must be met for [event1] to occur?

Figure 2: The above is the prompt for template generation. Question type $\mathcal{T}$ describes the relation and paradigm of the instruction we plan to generate. [context] is the placeholder for context information. [event1] and [event2] are placeholders for the head and tail events. The below is the generated template examples.

paradigm as:

$$\mathcal{R} = \text{Rel}(\mathcal{E}^h, \mathcal{E}^t, \mathcal{C}). \quad (3)$$

It requires the model to reason relations between two events given the context. This reasoning format further strengthens the model's event relational knowledge with improved relation understanding.

Event Fact Verification Given event relational fact $(\mathcal{E}^h, \mathcal{R}, \mathcal{E}^t, \mathcal{C})$, we require the model to determine whether the fact is true or not. This is also a wildly-used ability:

$$\text{Verify}(\mathcal{E}^h, \mathcal{E}^t, \mathcal{R}, \mathcal{C}). \quad (4)$$

This paradigm not only enhances the event understanding but also the relation between two events.

Training with these reasoning formats can effectively improve the event reasoning performances leading to better downstream applications.

	Close	Open	Rel	Verify
Cause	1,293	1,269	823	814
Effect	1,338	1,344	835	833
None	0	0	3,713	3,839
After	2,598	2,609	1,712	1,687
Before	2,644	2,649	1,671	1,678
Simul	996	987	690	663
IsCond	989	1,024	668	624
HasCond	1,007	1,025	695	674
Total	10,865	10,907	10,807	10,812

Table 1: Instruction tuning dataset statistics. Simul is short for simultaneously.

2.4 Instruction-Tuning Dataset

We then incorporate our constructed event relational knowledge into instruction tuning datasets with our reasoning paradigm.

We derive instruction templates by querying GPT4. Our method encompasses $|\mathbb{S}^{\mathcal{R}}|$ relations, coupled with four reasoning formats. Furthermore, we account for situations in which context $\mathcal{C}$ might be absent. Consequently, we require total amounts to $|\mathbb{S}^{\mathcal{R}}| \times 4 \times 2$ variations of instruction templates. For each kind, we ask GPT4 to list 20 prompts with the query. The prompt and the generation examples are shown in Figure 2. More examples are in the Appendix C.

After that, for each $\mathcal{K}$, we sample a reasoning format. We then wrap the $\mathcal{K}$ with a certain template according to the relation and the reasoning format. We replace the placeholder [event1] and [event2] with head event $\mathcal{E}^h$ and tail event $\mathcal{E}^t$, and the placeholder [context] by context $\mathcal{C}$ (if exists).

For Close, Rel, and Verify reasoning, we need negative candidates. We follow Zhou et al. (2022a) to retrieve the negative events to create candidate event set $\mathbb{D}$. We build a pool of events from the whole corpus and then retrieve the negative events using three heuristic rules. Specifically, given tail event $\mathcal{E}^t$, we build its negative events, in light of lexicon-based, PoS-based, or in-domain retrieval. For Close, we sample two events from all negative events and form candidate event set $\mathbb{D}$ with gold tail event $\mathcal{E}^t$. For Rel and Verify, we sample 1 event from $\mathbb{D}$ to form a new data with label None.

We balance the amount of different tasks and relation data and report the instruction tuning set statistics in Table 1. To maintain the general ability, we also mix with a general instruction tuning dataset GPT4Alpaca (Peng et al., 2023). Then we finetune the backbone LLM with the construct instruction tuning dataset to obtain WIZARDEVENT.

Model	ACC					F1				
	ECARE	COPA	MCTACO	SocialIQA	SCT	MATRES	ESL	TRACIE	ESTER	CQA
Close Source Models										
GPT3.5	80.30	94.00	92.25	71.03	95.03	44.98	61.41	59.00	23.88	19.95
Open Source Models										
Alpaca-7B (Taori et al., 2023)	69.42	68.00	84.41	54.40	86.16	4.23	45.83	59.41	20.64	19.35
WizardLM-7B (Xu et al., 2023)	64.96	68.00	80.28	46.21	82.31	34.70	32.97	53.07	17.61	12.00
Vicuna-7B (Chiang et al., 2023)	53.38	68.00	51.51	32.91	52.70	50.42	69.02	0.00	20.59	12.46
Baichuan2-7B (Yang et al., 2023a)	75.19	68.00	87.32	62.90	86.37	33.98	62.68	47.38	15.48	11.79
Llama2-7B	73.31	83.00	83.40	55.89	78.41	38.45	52.17	63.64	18.61	10.59
Llama2-7B-ALP	65.90	74.00	82.19	49.03	87.28	49.21	57.43	38.92	14.90	11.84
WizardEvent (Ours)	80.11	**94.00**	**89.84**	62.23	**92.68**	51.39	**76.09**	**66.67**	28.48	34.85
WizardEvent-13B (Ours)	**81.85**	**94.00**	87.22	**64.89**	92.36	**56.83**	68.30	**66.67**	**36.36**	**45.09**

Table 2: Main results on event reasoning. Bold number stands for best performances among all 7B models. Blue blank stands for outperforming GPT-3.5.

3 Experiments

3.1 Datasets

We incorporate ECARE (Du et al., 2022), MCTACO (Zhou et al., 2019), SocialIQA (Sap et al., 2019), SCT (Mostafazadeh et al., 2016), MATRES (Ning et al., 2018), ESL (Caselli and Vossen, 2017), TRACIE (Zhou et al., 2020), ESTER (Han et al., 2021), and CQA (Bondarenko et al., 2022) for test. These datasets can be used to assess the abilities of causal, temporal, intentional event reasoning, and event prediction respectively.

3.2 Baselines

We include Llama1 based Alpaca-7B (Taori et al., 2023). WizardLM-7B (Xu et al., 2023), Vicuna-7B (Chiang et al., 2023), Llama2-7B-chat (Touvron et al., 2023) are based on Llama2. We also include Baichuan2 (Yang et al., 2023a) and GPT3.5 as our baselines. For open-source models, we use the chat version of the open-source models for evaluation and use the model names for short. Llama2-ALP is Llama2-base only finetuned on GPT4Alpaca.

3.3 Implementation Settings

WizardEvent undergoes fine-tuning using academic resources. Precisely, we utilize 4 $\times$ NVIDIA A800 80G GPUs to train both 7B and 13B Llama2-base for 3 epochs. We use the DeepSpeed training framework[2], and ZERO-2 strategy (no offload) along with mixed-precision training (bf16). We use a standard AdamW optimizer and a linear warmup scheduler. The initial learning rate for AdamW is set to $2e^{-5}$, and the ratio for warmup is set to 0.03. The maximum sequence length for the model training is 512, and the batch size is configured as 64 per device. The entire fine-tuning process is completed within 4 hours.

We use Spacy[3] for all linguistic extraction. In our pilot experiments, we test multiple input prompts for each model to search for the optimum prompt for evaluation tasks. We observe minimal fluctuations in the results despite prompt variations. To mitigate the impact of other variables, we ensure consistency by employing the same prompt for all models when they undertake the same task. We turn the Close, Rel, and Verify into multiple-choice questions and require the model to answer by the label of choice. All prompts for evaluation can be found in the Appendix B.

3.4 Evaluation Metrics

We follow Tao et al. (2023a) to evaluate all models on automatic metrics. For ECARE, COPA, MCTACO, SocialIQA, SCT, MATRES, ESL we use accuracy. For TRACIE, ESTER, and CQA we use F1-score. For tasks of multiple-choice style, some models won't directly generate the label as the answer. We adopt a decoding protocol to parse the output answers and obtain the final prediction for all models. We find this protocol effective. We show this protocol in Appendix A.

3.5 Main Results

We show the main results in Table 2. We find both 7B and 13B WizardEvent significantly outperform Llama2-7B-ALP. In COPA, TRACIE, and CQA, WizardEvent improves larger than 20 percent. The results suggest our method effectively increases the event reasoning abilities. Enhancing event relational knowledge improves reasoning

[2]https://www.deepspeed.ai

[3]https://spacy.io

Model	ECARE	COPA	MCTACO	SocialIQA	SCT	MATRES	ESL	TRACIE	ESTER	CQA	AVG
REASONING PARADIGM EVALUATION											
Ours	80.11	94.00	89.84	62.23	92.68	51.39	76.09	66.67	28.48	34.85	67.63
- Close	74.34	88.00	91.25	61.36	90.91	49.94	79.89	59.03	24.18	26.39	$64.52^{-3.10}$
- Open	78.05	86.00	88.03	61.67	91.07	42.32	66.85	66.67	16.03	11.50	$60.81^{-6.81}$
- Rel	78.71	91.00	81.99	58.90	91.66	50.42	66.67	66.67	29.59	35.48	$65.10^{-2.52}$
- Verify	79.08	91.00	88.03	62.33	89.26	50.79	70.47	62.32	27.90	35.99	$65.71^{-1.91}$
TRAINING CURRICULUMS											
Serial	76.69	92.00	83.20	55.83	85.09	22.85	78.44	66.67	27.74	45.25	$63.37^{-4.25}$
Pipeline	78.94	91.00	90.04	61.87	91.50	50.42	73.91	61.27	26.11	32.94	$65.80^{-1.83}$

Table 3: Evaluation of reasoning paradigm and exploration of training curriculums. AVG stands for the average scores of all datasets. Red numbers are the decrease scores.

abilities on various relations and reasoning formats.

Compared with other instruction-tuning methods, WIZARDEVENT achieves the best performances. It indicating our method can effectively mitigate the sparsity of event relational knowledge and reasoning formats in general instruction-tuning datasets. WIZARDEVENT even excels GPT3.5 in 6 among all 10 test sets which further demonstrates the effectiveness of WIZARDEVENT.

We also find WIZARDEVENT performs particularly well on event relation extraction test sets such as ESL, MATRES, and fact verification test sets as TRACIE. This indicates that WIZARDEVENT effectively solves the imbalance of abilities on various reasoning formats.

3.6 Paradigm Evaluation

In this section, we evaluate the proposed reasoning paradigm. We conduct ablation experiments. We testify WIZARDEVENT four times each with one reasoning format ablated. We show the results in Table 3. As the average scores on all datasets, we find all reasoning formats work. Ablating any reasoning format would incur a drop in performances. The results demonstrate the effectiveness of our prototypical event reasoning abilities.

Among all reasoning formats, we find the Open reasoning format works most where WIZARDEVENT would drop average 6.81 scores on average. It could be the most basic ability for event reasoning. We also find there are sometimes exceptions where ablating some reasoning format increases performances on a few datasets. We believe that it may be attributed to different learning progress and curriculums of different abilities. We would probe it more in Section 3.10.

3.7 Generalization of LLMs

In this section, we evaluate the generalization of WIZARDEVENT on other backbone models. We conduct two experiments with backbone replacement to Qwen-7B (Bai et al., 2023) and Mistral-7B (Jiang et al., 2023). We compare WIZARDEVENT on different backbones to their models on only GPT4Alpaca. The results are shown in Table 4. We find WIZARDEVENT excels GPT4Alpaca on all backbones. The results indicate the generalization of our methods. Further, findings are consistent on different backbones. WIZARDEVENT can effectively boost relation extraction and verification datasets.

3.8 Event Relational Knowledge Evaluation

In this section, we evaluate whether our motivation holds that WIZARDEVENT can enhance the event relational knowledge of LLMs. However, it is hard to evaluate since there are no available datasets. To fulfill this goal, we annotate a novel test set to evaluate the event relational knowledge. Our dataset mainly testifies to Close and Rel reasoning formats. The construction process is:

Schema Graph Construction. We utilize the Event-Event Concept Knowledge Graph (EECKG) (Wang et al., 2022a), which is derived from ConceptNet through a combination of rule-based reasoning and crowdsourcing. Since we aim to evaluate event relational knowledge, we ensure events we use are abstract event types by filtering out concrete events. This is done by including nodes of fewer than two words and employing GPT4 to assess the abstractness of events. We then eliminate high-frequency events to avoid overly generic nodes. We split EECKG into smaller components, each representing a scenario of related events, by random walk on EECKG.

Model	ECARE	COPA	MCTACO	SocialIQA	SCT	MATRES	ESL	TRACIE	ESTER	CQA
Llama2+ALP	65.9	74.00	82.19	49.03	87.28	49.21	57.43	38.92	14.90	11.84
Ours (Llama2)	80.11$^{14.21}$	94.00$^{20.00}$	89.84$^{7.65}$	62.23$^{13.20}$	92.68$^{5.40}$	51.39$^{2.18}$	76.09$^{18.66}$	66.67$^{27.75}$	28.48$^{13.58}$	34.85$^{23.01}$
Qwen+ALP	76.17	90.00	88.83	64.74	94.71	43.17	37.68	49.14	10.84	10.90
Ours (Qwen)	80.3$^{4.13}$	88.00$^{-2.00}$	91.45$^{2.62}$	67.40$^{2.66}$	93.85$^{-0.86}$	55.74$^{12.57}$	54.71$^{17.03}$	66.67$^{17.53}$	27.49$^{16.65}$	34.08$^{23.18}$
Mistral+ALP	69.75	80.00	82.80	48.82	86.26	10.64	51.45	65.64	10.18	8.36
Ours (Mistral)	76.45$^{6.70}$	91.00$^{11.00}$	84.00$^{1.20}$	54.81$^{5.99}$	85.62$^{-0.64}$	50.42$^{39.78}$	68.12$^{16.67}$	66.67$^{1.03}$	16.04$^{5.86}$	20.21$^{11.85}$

Table 4: Generalization of LLMs. ALP stands for finetuned on GPT4Alpaca. Green numbers indicate improvements while red numbers stand for dropped performances.

Model	CLOSE	REL
Llama2+ALP	32.62	40.28
WIZARDEVENT (Llama2)	43.19$^{10.57}$	41.32$^{1.04}$
Qwen+ALP	43.73	33.53
WIZARDEVENT (Qwen)	44.62$^{0.89}$	39.28$^{5.75}$
Mistral+ALP	33.51	37.25
WIZARDEVENT (Mistral)	50.18$^{16.67}$	42.04$^{4.79}$

Table 5: Event relational knowledge evaluation. Green numbers indicate improvements.

Convert the components into DAGs by removing cycles and creating backward components with reversed edges and relations. We totally consider 6 relations: `Cause`, `Before`, `HasSubevent`, `Effect`, `After`, and `IsSubevent`[4].

Task Construction. For each component, we sample two abstract events to form questions. We calculate their relation according to their connecting path. For the Close task, we use GPT4 to generate 15 negative candidate events for each question. For the Rel reasoning task, we regard two events as input and relation as the answer.

Human Anotation. We recruit human annotators to fix and further filter the questions we constructed as the final answers.

We finally construct 558 Close reasoning data and 835 Rel reasoning data. As in the previous section, we train WIZARDEVENT on different backbones and compare them to GPT4Alpaca. We report the results on our event relational knowledge dataset in Table 5. We find WIZARDEVENT is indeed able to enhance the event relational knowledge on various LLMs. WIZARDEVENT improves on all backbones and reasoning paradigms which demonstrates our motivation holds. We show the evaluation prompts in Appendix B.

[4] `HasSubevent` and `IsSubevent` are not in our training relations. We use them for held-out relation evaluation.

3.9 Data Scaling

We conduct experiments on various numbers of training data. We vary the number from 1k to full. We show the results in Figure 3. We find the overall performance increases with the scaling of data. We discover a pattern that appeared across multiple test sets. That is, the model will first achieve better results at 3k data, then drop slightly at 5k, and finally continue to rise. We think it also results from different training curriculums of abilities.

3.10 Training curriculums

In this part, we probe how the training curricula affect WIZARDEVENT. Our WIZARDEVENT is trained with uniformly shuffling our data and GPT4Alpaca. We explore two extra training curriculums:

Serial. We first train GPT4Alpaca to endow the model with the general abilities for 3 epochs. Then train it with our data for another 3 epochs.

Pipline. We split GPT4Alpaca into 4 chunks and denote the i_{th} chunk as ALP_i. We train in the pipeline order as [Open, ALP_1, Close, ALP_2, Rel, ALP_3, Verify, ALP_4, ...]. We train until 3 epochs.

Results are in Table 3. We find both training curriculums drop. The drop of *Serial* may indicate that abilities of event reasoning are fundamental which don't deeply rely upon other basic abilities. *Pipline* also falls behind the uniform mix of all data showing a challenging problem in designing better training curriculums. We leave it for future work.

4 Related Work

Event Reasoning Event relational reasoning infers events of certain inter-relations. Du et al. (2022) aims to select the accurate cause or effect event from candidates. Zhou et al. (2019) serves as a dataset for event temporal reasoning. Existing

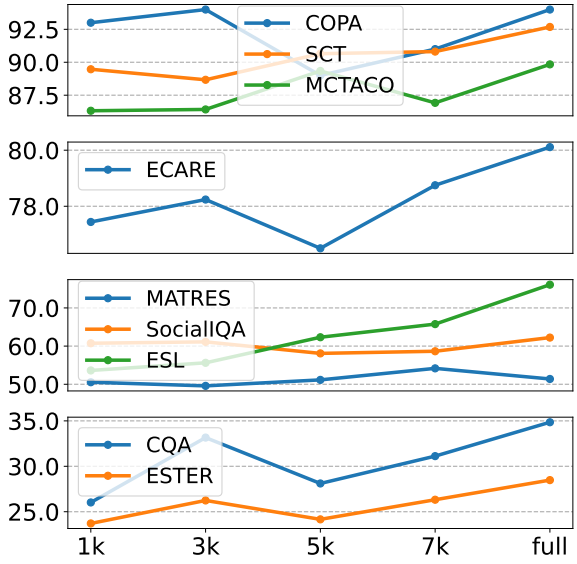

Figure 3: Performances of WIZARDEVENT with data scaling.

works present a scenario of counterfactual reasoning (Qin et al., 2019, 2020). In addition to single-event relation reasoning, existing works also reason events according to diversified event relations (Poria et al., 2021; Han et al., 2021; Yang et al., 2022). Tao et al. (2023b) further unifies datasets of several event-inter relations to transfer event relational knowledge to unseen tasks.

Predicting events necessitates the model to anticipate forthcoming occurrences grounded in the present context (Zhao, 2021). Mostafazadeh et al. (2016) employs a multiple-choice framework to predict future events by encompassing a diverse range of common-sense connections among events. Guan et al. (2019) establish a dataset oriented towards capturing event logic, enabling the generative prediction of future incidents.

Tao et al. (2023a) present the Event Semantic Processing including the event understanding, reasoning, and prediction of event semantics.

Instruction Tuning Instruction tuning refers to the process of fine-tuning a large language model based on specific instructions or guidance provided during training. Chung et al. (2022) finetunes on T5 with a scaling number of datasets which achieves strong few-shot performance even compared to much larger models. Taori et al. (2023) is trained by fine-tuning the LLaMA (Touvron et al., 2023) model using a dataset consisting instructions generated by text-davinci-003. Chiang et al. (2023) is an open-source chatbot created by fine-tuning LLaMA using user-shared conversations gathered from ShareGPT. Xu et al. (2023) extends the previous model by evolve-instruct algorithms to improve the model. Conover et al. (2023) leverages data on the Databricks platform.

In another line of research, instruction tuning is used to make a language model more focused and specialized in certain abilities or domains. Zhang et al. (2023a) trains a medical conversation model with different sources of datasets with instructions. Cui et al. (2023) propose a legal LLM named ChatLaw by legal domain dataset and mitigate hallucination of the model. Zhang et al. (2023b) train an LLM specialized for information extraction with data adapted from a knowledge graph. Yang et al. (2023b) design an automatic data curation pipeline and in building financial open-source LLM. Tang et al. (2023) propose a dataset to improve the tool manipulating ability of LLMs. Our work lies in this ability enhancement line of research.

Event-Aware Continuous Pretraining Considering both the pre-training and fine-tuning strategies, researchers are dedicated to improving event processing through fine-tuning techniques that incorporate events. In their study, Yu et al. (2020) inject intricate commonsense knowledge about events into pre-trained language models. Similarly, Zhou et al. (2022a,b) enhance language models by focusing on event-related tasks through event masking prediction and generation. However, these works struggle to effectively perform zero-shot reasoning. Our work is mainly different from theirs. We are methods of instruction tuning. Our focus is to stimulate the various abilities of event reasoning with the deliberate dataset.

5 Conclusion

In this study, we introduce Event-Oriented Instruction Tuning to enhance event reasoning capabilities and train our model WIZARDEVENT. We enhance the event relational knowledge and reasoning abilities of various formats. We first represent the event relational knowledge. Building upon this, we mine the knowledge through our method. We secondly introduce our hybrid event reasoning paradigm. Last, we create an instruction-tuning dataset based on the knowledge and reasoning paradigm. We fine-tune Llama2 to get our WIZARDEVENT model. We conduct experiments on 10 test sets and a new test set for event relational knowledge. Experiments show the effectiveness of our method.

Limitations

In this paper, we explore the training curriculum and find the uniform mix of data is the best. It has great potential to design better training curriculums and further investigate the dependencies of these abilities. We leave it to future work.

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*.

Alexander Bondarenko, Magdalena Wolska, Stefan Heindorf, Lukas Blübaum, Axel-Cyrille Ngonga Ngomo, Benno Stein, Pavel Braslavski, Matthias Hagen, and Martin Potthast. 2022. CausalQA: A benchmark for causal question answering. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3296–3308, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Tommaso Caselli and Piek Vossen. 2017. The event storyline corpus: A new benchmark for causal and temporal relation extraction. In *Proceedings of the Events and Stories in the News Workshop*, pages 77–86.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *arXiv preprint arXiv:2306.16092*.

Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2022. e-care: a new dataset for exploring explainable causal reasoning. *arXiv preprint arXiv:2205.05849*.

Jian Guan, Yansen Wang, and Minlie Huang. 2019. Story ending generation with incremental encoding and commonsense knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6473–6480.

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.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. 2023. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849.

Qiang Ning, Hao Wu, and Dan Roth. 2018. A multi-axis annotation scheme for event temporal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1318–1328.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Pengfei Hong, Romila Ghosh, Abhinaba Roy, Niyati Chhaya, et al. 2021. Recognizing emotion cause in conversations. *Cognitive Computation*, 13:1317–1332.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. 2008. The penn discourse treebank 2.0. In *LREC*.

Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. 2019. Counterfactual story reasoning and generation. *arXiv preprint arXiv:1909.04076*.

Lianhui Qin, Vered Shwartz, Peter West, Chandra Bhagavatula, Jena Hwang, Ronan Le Bras, Antoine Bosselut, and Yejin Choi. 2020. Back to the future: Unsupervised backprop-based decoding for counterfactual and abductive commonsense reasoning. *arXiv preprint arXiv:2010.05906*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*.

Tarcísio Souza Costa, Simon Gottschalk, and Elena Demidova. 2020. Event-qa: A dataset for event-centric question answering over knowledge graphs. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 3157–3164.

Qiaoyu Tang, Ziliang Deng, Hongyu Lin, Xianpei Han, Qiao Liang, and Le Sun. 2023. Toolalpaca: Generalized tool learning for language models with 3000 simulated cases. *arXiv preprint arXiv:2306.05301*.

Zhengwei Tao, Zhi Jin, Xiaoying Bai, Haiyan Zhao, Yanlin Feng, Jia Li, and Wenpeng Hu. 2023a. Eveval: A comprehensive evaluation of event semantics for large language models. *arXiv preprint arXiv:2305.15268*.

Zhengwei Tao, Zhi Jin, Haiyan Zhao, Chengfeng Dou, Yongqiang Zhao, Tao Shen, and Chongyang Tao. 2023b. Unievent: Unified generative model with multi-dimensional prefix for zero-shot event-relational reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7088–7102.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ya Wang, Cungen Cao, Zhiwen Chen, and Shi Wang. 2022a. Ecckg: An eventuality-centric commonsense knowledge graph. In *International Conference on Knowledge Science, Engineering and Management*, pages 568–584. Springer.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022b. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023a. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Chengbiao Yang, Weizhuo Li, Xiaoping Zhang, Runshun Zhang, and Guilin Qi. 2020. A temporal semantic search system for traditional chinese medicine based on temporal knowledge graphs. In *Semantic Technology: 9th Joint International Conference, JIST 2019, Hangzhou, China, November 25–27, 2019, Revised Selected Papers 9*, pages 13–20. Springer.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023b. Fingpt: Open-source financial large language models. *arXiv preprint arXiv:2306.06031*.

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

Changlong Yu, Hongming Zhang, Yangqiu Song, and Wilfred Ng. 2020. Cocolm: Complex commonsense enhanced language model with discourse relations. *arXiv preprint arXiv:2012.15643*.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, et al. 2023a. Huatuogpt, towards taming language model to be a doctor. *arXiv preprint arXiv:2305.15075*.

Hongming Zhang, Xin Liu, Haojie Pan, Yangqiu Song, and Cane Wing-Ki Leung. 2020. Aser: A large-scale eventuality knowledge graph. In *Proceedings of the web conference 2020*, pages 201–211.

Ningyu Zhang, Jintian Zhang, Xiaohan Wang, Honghao Gui, Yinuo Jiang, Xiang Chen, Shengyu Mao, Shuofei Qiao, Zhen Bi, Jing Chen, Xiaozhuan Liang, Yixin Ou, Ruinan Fang, Zekun Xi, Xin Xu, Liankuan Tao, Lei Li, Peng Wang, Zhoubo Li, Guozhou Zheng, and Huajun Chen. 2023b. Deepke-llm: A large language model based knowledge extraction toolkit.

Liang Zhao. 2021. Event prediction in the big data era: A systematic survey. *ACM Computing Surveys (CSUR)*, 54(5):1–37.

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. *arXiv preprint arXiv:1909.03065*.

Ben Zhou, Kyle Richardson, Qiang Ning, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2020. Temporal reasoning on implicit events from distant supervision. *arXiv preprint arXiv:2010.12753*.

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.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

A Decoding Protocol

We show our decoding protocol for extracting answers of CLOSE tasks as follows:

---

pattern = "the(?: correct)? (?:option|answer) should be[\s:]+([ABCDEFGH])"

if *Output* starts with an alphabetical number **then**
  Set *prediction* as the alphabetical number
else if re.match(*pattern*, *Output*) **then**
  Extract the *prediction* follow the *pattern*.
else
  prediction=$\underset{c \in \mathbb{D}}{argmax}(\text{WordOverlap}(c,$ *Ouput*)

B Event Reasoning Evaluation Prompts

We show prompts for evaluation on all tasks for all models in Figure 4 and Figure 5.

C Examples of Instruction Templates

We showcase examples of instruction templates in Figure 6 to 9.

Instruction:
Answer the question by selecting A, B.

Question:
What is the cause of "He got some rum."?
Choices:
A. The worker fremented some sugar cane with yeast.
B. Tom went out and want to hunt some cottontails.
The answer is:
Output:
A

(a) ECARE

Instruction:
Answer the question by selecting A, B

Question:
What is the effect of "The man turned on the faucet."?
Choices:
A. The toilet filled with water.
B. Water flowed from the spout.
The answer is:
Output:
B

(b) COPA

Instruction:
Answer the question by selecting A, B.

Context:
Durer's father died in 1502, and his mother died in 1513.
Question:
What happened after Durer's father died?
Choices:
A. Durer took care of his mother.
B. He got a new job.
The answer is:
Output:
A

(c) MCTACO

Instruction:
Answer the question by returning A, B or C.

Context:
Due to his car breaking down, Robin decided to ride with Jan's friends to school.
Question:
What will Robin want to do?
Choices:
A. Fix his car.
B. Avoid missing class.
C. Arrive on time to school.
The answer is:
Output:
A

(d) SocialIQA

Instruction:
Answer the question by returning A or B.

Context:
John was writing lyrics for his new album. He started experiencing writer's block. He tried to force himself to write but it wouldn't do anything. He took a walk, hung out with some friends, and looked at nature.
Question:
What is the next event?
Choices:
A. He felt inspiration and then went back home to write.
B. John then got an idea for his painting.
The answer is:
Output:
A

(e) SCT

Instruction:
Determine the type of temporal relationship between events by returning A, B, C or D.
Context:
But the tie – up with rosneft will keep bp in russia , allowing it to continue to explore and exploit the country 's vast energy resources , including in the arctic region.
Question:
What is the temporal relationship between "keep" and "explore"?
Choices:
A. "keep" happened before "explore".
B. "keep" and "explore" happened simultaneously.
C. "keep" happened after "explore".
D. Can't decide.
Answer:
Output:
C

(f) MATRES

Figure 4: Evaluation prompts on ECARE, COPA, MCTACO, SocialIQA, SCT and MATRES for all models.

Instruction:
Determine the type of causal relationship between events by returning A, B or C.
Context:
Navy foils Somali pirate attack off Gulf of Aden The Indian Navy patrolling the Gulf of Aden on Thursday thwarted a In their five operations so far, ⋯ Indian Navy is part of the international effort to ensure the safety and freedom of seaborne trade in this high – risk stronghold of modern – day piracy .
Question:
What is the causal relation between "foils" and "action"?
Choices:
A. "action" caused "foils".
B. "foils" caused "action".
C. There is no causal relationship between them.
Answer:
Output:
A

(a) ESL

Instruction:
Answer the question by returning A or B.

Context: Margaret was walking through town. She noticed a store window with an ad for a family shelter. The pictures of kids really struck a chord in her heart. She decided to donate. She walked inside and began the process.
Question: Is it true that she felt guilty starts after she saw an ad for a family shelter?
A. False
B. True
Only output A or B.
Answer:
Output:
B

(b) TRACIE

Instruction:
Context:
Wang is the acting head of Wangbolin District and Ning is Wang's deputy. ⋯ Preliminary investigations show that the mine was operating illegally because its production safety and coal mining certificates expired at the end of December last year.

Question:
What caused Wang, Ning, Gong and Feng to be removed from their posts in the Party?

Answer:
Output:
regulations that stipulate officials in charge should be punished for fatal accidents

(c) ESTER

Instruction:
Context:
However, over time we do expect Banco de Chile to accrue benefits from these investments as well as increased operating leverage as its revenue grows. ⋯ While we anticipate the base rate returning to a more normalized level of 3.00% by the end of 2025, persistently low rates can pressure profitability.

Question:
Why over time is useful?

Answer:
Output:
we do expect Banco de Chile to accrue benefits from these investments

(d) CQA

Instruction:
Answer the question by selecting A, B, C, D.

Question:
Which is the cause of "argue"?
Choices:
A. legislative_action
B. competing
C. emergency_response
D. reconciliation
The answer is :
Output:
B

(e) CLOSE of event knowledge evaluation.

Instruction:
Answer the question by selecting A, B or C.

Question:
Which is the temporal relationship between "trust" and "additional_challenges"?
Choices:
A. "trust" is before "additional_challenges".
B. "trust" is after "additional_challenges".
C. There is no obvious temporal relationship between "trust" and "additional_challenges".
The answer is :
Output:
A

(f) REL of event knowledge evaluation.

Figure 5: Evaluation prompts on ESL, TRACIE, ESTER, CQA, CLOSE and REL of event knowledge evaluation for all models.

Cause w/ ctx
Considering the [context], what can be considered as the initiating event that caused [event1]?
Cause w/o ctx
From the given options, which event is primarily responsible for causing [event1]?

Effect w/ ctx
Determine which event emerged as a result of [event1] within the specified [context].
Effect w/o ctx
Determine which event was directly caused by [event1] from the options provided.

After w/ ctx
From the options below, identify the event that happens immediately after [event1] within the [context].
After w/o ctx
Choose an event that happens after [event1] from the choices.

Before w/ ctx
Referring to [context], which of these events was the immediate precursor to [event1]?
Before w/o ctx
Identify the event that took place right before [event1].

isCond w/ ctx
Within the [context] provided, what event is deemed necessary for the completion of [event1]?
isCond w/o ctx
From the following events, which one is a prerequisite for [event1]?

hasCond w/ ctx
Considering the [context], which alternative represents the situation with the condition described as [event1]?
hasCond w/o ctx
From the list provided, which event is characterized by the condition of [event1]?

Simul w/ ctx
Identify the event that coincides with [event1] within the described [context].
Simul w/o ctx
Which of the following events occurred at the same time as [event1]?

Figure 6: Examples of instruction templates for **Close Event Reasoning** generated by GPT4.

Cause w/ ctx
Within [context], what specific action or decision directly resulted in [event1]?
Cause w/o ctx
What technological advancement allowed [event1] to take place?

Effect w/ ctx
What are the direct consequences of [event1] within the [context]?
Effect w/o ctx
How will the environment be altered as the effect of [event1]?

After w/ ctx
What event directly follows [event1] within the specified context of [context]?
After w/o ctx
Can you describe an event that occurs directly following [event1]?

Before w/ ctx
What event directly precedes [event1] in the context of [context]?
Before w/o ctx
Could you describe an occurrence that takes place right before [event1]?

isCond w/ ctx
What prior event sets the stage for [event1] in the context of [context]?
isCond w/o ctx
What specific action or event triggers [event1] as a prerequisite?

hasCond w/ ctx
In light of [event1], what subsequent event could logically occur in the provided [context]?
hasCond w/o ctx
Can you visualize a scenario that unfolds when the situation is identified as [event1]?

Simul w/ ctx
What concurrent event can be observed alongside [event1] in the described [context]?
Simul w/o ctx
What is a significant occurrence that takes place at the same time as [event1]?

Figure 7: Examples of instruction templates for **Open Event Reasoning** generated by GPT4.

Cause w/ ctx
In what ways is [event1] a direct cause of [event2] in the context of [context]?
Cause w/o ctx
Can you analyze the causal relation between [event1] and [event2]?

Effect w/ ctx
Can you explain the causal link between [event1] and [event2] within [context]?
Effect w/o ctx
What effect does [event1] have on bringing about [event2]?

After w/ ctx
Based on [context], how does [event1] temporally relate to [event2]?
After w/o ctx
What chronological relationship exists between [event1] and [event2]?

Before w/ ctx
Based on the [context], what is the sequence of occurrences between [event1] and [event2]?
Before w/o ctx
How does [event1] temporally relate to [event2] in terms of sequencing?

isCond w/ ctx
In what way does [event1] influence the occurrence of [event2] within [context]?
isCond w/o ctx
If [event1] did not occur, could [event2] still take place?

hasCond w/ ctx
How does [event1] influence [event2] within the given [context]?
hasCond w/o ctx
If [event1] takes place, under what conditions does [event2] follow?

Simul w/ ctx
Did [event1] occur before, after, or simultaneously with [event2] in the given [context]?
Simul w/o ctx
In the timeline of events, where does [event1] stand in relation to [event2]?

Figure 8: Examples of instruction templates for **Relation Reasoning** generated by GPT4.

Cause w/ ctx
Can you determine if [event2] led to [event1] in the given [context]?
Cause w/o ctx
In what way did [event2] precipitate the events that resulted in [event1]?

Effect w/ ctx
Can it be confirmed that [event2] occurs as a direct consequence of [event1] within the [context]?
Effect w/o ctx
Upon the happening of [event1], is [event2] a predictable consequence?

After w/ ctx
Based on [context], does [event2] occur following the completion of [event1]?
After w/o ctx
Can you verify if [event2] occurs following [event1]?

Before w/ ctx
In the given [context], does [event2] occur before [event1] becomes true?
Before w/o ctx
Can one confirm that the sequence puts [event2] happening before [event1]?

isCond w/ ctx
Does [event2] have to occur for [event1] to happen within [context]?
isCond w/o ctx
Does [event1] depend on [event2] happening first?

hasCond w/ ctx
Does the occurrence of [event1] directly condition the possibility of [event2] in the given [context]?
hasCond w/o ctx
Does [event2] only occur if the condition [event1] is met?

Simul w/ ctx
Given the [context], does [event2] happen simultaneously with [event1]?
Simul w/o ctx
Could you confirm the simultaneous occurrence of [event1] and [event2]?

Figure 9: Examples of instruction templates for **Fact Verification** generated by GPT4.