A Comprehensive Evaluation on Event Reasoning of Large Language Models

Anonymous ACL submission

Abstract

Event reasoning is a fundamental ability that underlies many applications. It requires event schema knowledge to perform global reasoning and needs to deal with the diversity of the inter-event relations and the reasoning paradigms. How well LLMs accomplish event reasoning in terms of competence and knowledge remains unknown. To mitigate this disparity, we comprehensively evaluate the abilities of event reasoning of LLMs. We introduce a novel benchmark EV^2 for EValuation of EVent reasoning. EV^2consists of two levels of evaluation of schema and instance and is comprehensive in relations and reasoning paradigms. We conduct extensive experiments on EV^2. We find that LLMs have abilities to accomplish event reasoning but their performances are far from satisfactory. We also notice the imbalance of event reasoning abilities in LLMs. Besides, LLMs have event schema knowledge, however, they're not aligned with humans on how to utilize the knowledge. Based on these findings, we introduce two methods to guide the LLMs to utilize the event schema knowledge. Both methods achieve improvements.

1 Introduction

Events are instances or occurrences that form the basic semantic building units encompassing the meanings of Activities, Accomplishments, Achievements, and States (Vendler, 1957). Event Reasoning is the ability to process and analyze events and their complex interconnections. Compared with other abilities, event reasoning is unique in some aspects. Firstly, it requires knowledge in the form of event schemas, capturing the progress of event evolution in scenarios, then performing global reasoning (Li et al., 2021a; Mao et al., 2021). As shown in Figure 1, each event instance is associated with an event type. All event types and their relations form the event schema knowledge which reflects the logic and mechanism of event evolution. Knowing "Memory" would often happen after "Learn" can help answer the reasoning question. Second, the inter-event relations and reasoning paradigms are various. Event reasoning incorporates reasoning events according to a certain relation (Du et al., 2022; Sap et al., 2019b) and reasoning inter-event relations (Ning et al., 2018; Caselli and Vossen, 2017). The queried relations are diversified such as causality (Roemmele et al., 2011), temporality (Zhou et al., 2019), and hierachy (Glavaš et al., 2014). There are various paradigms such as reasoning the event or the inter-relation.

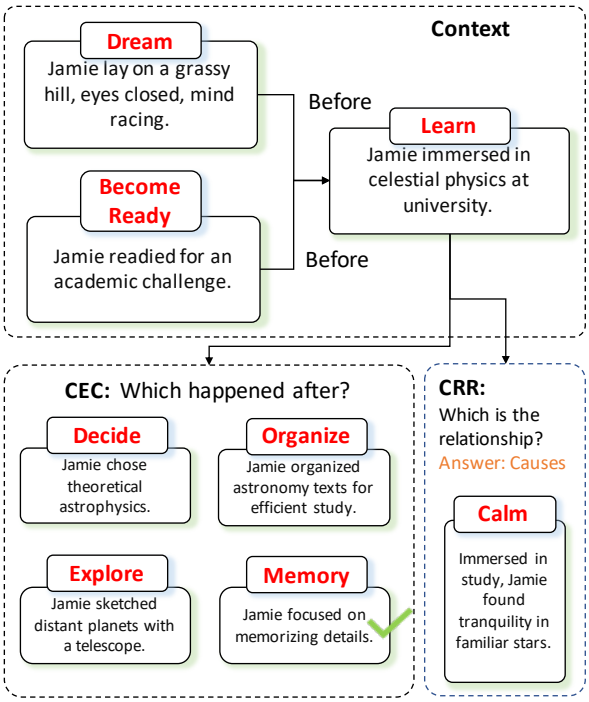

Figure 1: An example of event reasoning. The red words are event schema knowledge. The sentences below are event instances. In event reasoning, there are various paradigms such as Contextual Event Classification (CEC) and Contextual Relation Reasoning (CRR), and diverse inter-event relations.

As a fundamental competency within Large Language Models (LLMs), event reasoning sup-

ports 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, the enhancement of event reasoning abilities is essential for the advancement of LLMs.

LLMs like LLAMA (Touvron et al., 2023) series and GPT series (Brown et al., 2020) have demonstrated exceptional accomplishments in various natural language reasoning (Bang et al., 2023). Existing research has evaluated a broad spectrum of reasoning abilities of LLMs such as commonsense (Bian et al., 2023), sentence relations (Chan et al., 2023), and math (Arora et al., 2023). However, studies on the comprehensive evaluation of event reasoning of LLMs are scarce. Current works only focus on instance-level events, resulting in unclearness of how LLMs understand and utilize the event schema knowledge (Chan et al., 2023). Besides, they ignore the diversity of relations and paradigms (Yuan et al., 2023). These disparities hinge on the development of such crucial abilities of LLMs.

In this paper, we comprehensively evaluate event reasoning in knowledge and abilities. Since there are existing datasets that are comprehensive in relations and paradigms, and can cover both levels of schema and instance, we introduce a novel benchmark EV^2 for the **EV**aluation of **EV**ent reasoning. EV^2 is featured in evaluating both aligned schema-level and instance-level. The schema-level evaluation investigates the event schema knowledge of LLMs while the instance-level testifies the event reasoning abilities. Besides, to evaluate event reasoning in various types of relation and reasoning paradigms, EV^2 includes two event reasoning tasks, namely Contextual Event Classification (CEC) and Contextual Relation Reasoning (CRR) as shown in Figure 1. EV^2 is constructed from both GPT generation and human annotation. Utilizing EV^2, we comprehensively evaluate how well LLMs do event reasoning in terms of abilities and knowledge. Specifically, we mainly explore four research questions: 1) How proficient abilities of event reasoning do LLMs have? 2) To what extent do LLMs have the event schema knowledge? 3) Are LLMs aligned with humans in leveraging event schema knowledge? 4) Can LLMs perform better event reasoning with explicit guidance of leveraging event schema knowledge?

We conduct extensive experiments on EV^2 to answer these questions. The results provide insights into event reasoning that: 1) LLMs have the abilities of event reasoning, but are far from satisfactory and are imbalanced in different relations and reasoning paradigms. 2) LLMs have event schema knowledge. They can answer the schema-level questions with similar accuracy to the instance-level questions. However, the development of schema-level abilities falls behind those of instance-level. 3) LLMs are not aligned with humans in the aspect of leveraging event schema knowledge. 4) Based on the findings, we design two mentoring methods to guide the LLMs to utilize event schema knowledge. One is to directly add event schema knowledge to the prompt. The second is guiding in a chain-of-thought format. With the designed guidances for utilizing event schema knowledge, LLMs can perform better event reasoning. Especially with direct guidance, LLMs get significant improvements.

We summarize our contributions as follows:

- We evaluate event reasoning in both levels of schema and instance, and various relations and paradigms.
- We construct a novel benchmark EV^2 which features two levels of evaluation and comprehensive in relations and reasoning paradigms. We conduct extensive experiments to probe how LLMs perform event reasoning.
- We conclude several insights. Based on our findings, we design mentoring methods to guide LLMs to utilize event schema knowledge which achieves improvements in event reasoning.

2 Problem Formulation

Event reasoning is to anticipate the occurrences of certain relations or deduce interrelated correlations (Tao et al., 2023a). The relations encompass causality (Du et al., 2022), temporality (Zhou et al., 2019), and hierarchy (Glavaš et al., 2014).

Event reasoning requires comprehension of event schema knowledge. An event schema of a scenario is a schema-level graph $\mathcal{G}^s = (\mathbb{V}^s, \mathbb{E}^s)$[1], where $\mathbb{V}^s$ is the set of event types and $\mathbb{E}^s$ is the set of relations between events. Each edge in $\mathbb{E}^s$ is a relation triplet $(\mathcal{E}_i^s, \mathcal{R}, \mathcal{E}_j^s)$ standing for that there is the relation $\mathcal{R}$ between $\mathcal{E}_i^s$ and $\mathcal{E}_j^s$. With instantiation, we have the instance-level event

[1]Superscript s represents schema level.

graph $\mathcal{G}^i = (\mathbb{V}^i, \mathbb{E}^i)$[2]. An instance event $\mathcal{E}^i$ has an event type $\mathcal{E}^s$ but with detailed event arguments and context (Mitchell, 2005). The nodes and edges of these two graphs are corresponding, namely, each triplet in $\mathcal{G}^s$ has a corresponding triplet in $\mathcal{G}^i$ with the same inter-relation. In both levels, we consider totally six relation types, namely $\mathcal{R} \in$ {`Causes`, `IsResult`, `Before`, `After`, `IsSubevent`, `HasSubevent`}.

EV^2 consists of two event reasoning paradigms for both levels of schema and instance. The first is Contextual Event Classification (CEC) and the second is Contextualized Relation Reasoning (CRR).

CEC Given graph $\mathcal{G}$, either schema- or instance-level, queried event $\mathcal{E} \in \mathcal{G}$, and target relation $\mathcal{R}$, CEC requires the model to answer an event $\mathcal{E}_a$:

$$\mathcal{E}_a = \mathrm{M}(\mathcal{E}, \mathcal{R}, \mathcal{G}, \mathbb{C}). \quad (1)$$

M is the model, $\mathbb{C}$ is the candidate event set. CEC evaluates the model's comprehension of event semantics and structure.

CRR Given graph $\mathcal{G}$, either schema- or instance-level, two queried events $\mathcal{E}_i, \mathcal{E}_j \in \mathcal{G}$, CRR requires to determine the relation $\mathcal{R}$ between them:

$$\mathcal{R} = \mathrm{M}(\mathcal{E}_i, \mathcal{E}_j, \mathcal{G}). \quad (2)$$

CRR evaluates the understanding of event relations.

In both schema and instance levels, EV^2 has CEC and CRR tasks. Schema-level tasks require models to be rich in knowledge while tasks for instance need models to process detailed information.

3 Benchmark Construction

To create the EV^2 benchmark, we curate a comprehensive dataset through a three-stage process. Initially, the schema graph $\mathcal{G}^s$ is established. Then, GPT4 is employed to generate the instance graph $\mathcal{G}^i$. Lastly, human annotators are tasked with creating questions from $\mathcal{G}^s$ and $\mathcal{G}^i$.

3.1 Schema Graph Construction

We leverage EECKG (Wang et al., 2022b) to ensure a diverse range of event types in our schema. EECKG combines rule-based reasoning with crowdsourced insights, built on ConceptNet's structure. Nodes in EECKG represent verb phrases as events, and edges denote inter-event relations, focusing on `Causes`[3], `Before`, and `HasSubevent`.

[2]Superscript i represents instance level.

[3]The direction is that the head event causes the tail event. Other relations are the same.

Our objective mandates that the nodes within $\mathcal{G}^s$ should represent event types. Therefore, we filter EECKG nodes, removing concrete event instances. Preference is given to nodes with at most two words, as longer descriptions tend to include specific details. For events with fewer than two words, we use GPT4 to enhance our selection, ensuring the appropriate abstraction level for our schema graph with the following prompt:

> **### Instructions:**
> Determine which of the following candidate phrases are abstract and conceptual event types.

We identify a subset of remaining events that are too generic. To refine the event selection, we also exclude the most frequent events from our subset to avoid generic events.

We then dissect the interconnected EECKG into separate components, each representing a distinct scenario. To prevent semantic drift, we carefully control the size of each component. Starting from a node, we conduct a random walk until the number of nodes surpasses a set threshold, thus defining a component. This process is executed for all nodes to gather all components, as detailed in Algorithm 1. Post-extraction, we eliminate cycles to convert these structures into DAGs.

EECKG only contains forward event evolution relations such as `Causes`. We further include components of backward relations. We generate a reversed version for each component by inverting edge directions and replacing relations with their opposites: `IsResult`, `After`, and `IsSubevent`. This creates the backward components.

In preparation for constructing tasks for CEC and CRR, we label two events for each component. We sample three event pairs $(\mathcal{E}_h, \mathcal{E}_t)$ per component with a maximum inter-path length of four, utilizing their predecessors as background events. These pairs and background events form a schema graph. When the path length between $\mathcal{E}_h$ and $\mathcal{E}_t$ is two, the direct relation serves as the queried relation; for longer paths, we deduce the relation using Table 1. We construct a schema graph, queried event pair, and their relation $(\mathcal{E}_h, \mathcal{E}_t, \mathcal{R}, \mathcal{G}^s)$.

3.2 Instance Graph Construction

We next harvest instance graph $\mathcal{G}^i$ for each schema graph $\mathcal{G}^s$. For each node $\mathcal{E}^s \in \mathcal{G}^s$, we ask GPT4 to generate $\mathcal{E}^i$ using the following prompt:

Algorithm 1: Components Construction

Input : EECKG $\mathcal{G}, \mathcal{N}$
Output : A list of components $\mathbb{O}$.

$\mathbb{O}$ = [ ]
Function RandomWalk($start$, c):
  l = RandomInt($\mathcal{N}$, $\mathcal{N}$+2)
  if $l \leq len(c)$ **then**
    c.Append ($start$)
    return c
  n = Sample(*start*.Neighbors)
  if $n \notin c$ **then**
    $c \leftarrow$ RandomWalk$(n, c \cup \{n\})$
    return c
  return null
foreach $node \in G$ **do**
  $component$ = RandomWalk ($node$, [ ])
  $\mathbb{O}$.Append ($component$)
return $\mathbb{O}$

Instruction:
Generate an instance event for each abstract event. The abstract event is the event type of the instance event. All the instance events form a coherent story which maintain the relations of each abstract event. The integrated story should have explicit roles, location and time. The whole story should be detailed, diverse in topic and scenarios, and rich in knowledge.

We inherit the relations of $\mathcal{G}^s$ and obtain $\mathcal{G}^i$. We naturally obtain the instances of $\mathcal{E}_h$ and $\mathcal{E}_t$.

3.3 Question Construction

The last step is to construct questions of CEC and CRR in both schema and instance levels. For CEC, regarding schema and instance head events as the query and the tail as an answer, we ask GPT4 to generate 15 possible candidate instance events with their event types.

We then recruit 8 well-educated human annotators. Their missions are:

1) Revise or discard $\mathcal{G}^s$ if not valid. Ensure the events are abstract, the relations are correct, there's no scenario shifting in $\mathcal{G}^s$.
2) Revise or discard $\mathcal{G}^i$ if not valid. Ensure the events are concrete, the relations are correct, and the whole scenario of $\mathcal{G}^i$ is coherent and has no shifting.
3) Choose three proper negative candidate events with their event types. Ensure answering the question should rely on the context events.

We use the schema part of annotation as the schema-level questions and the instance part as instance-level questions. Then we complete CEC.

RULE	INDUCTION
$(\texttt{Before})^+$	Before
$(\texttt{After})^+$	After
$(\texttt{Before})^\star(\texttt{Causes})^+(\texttt{Before})^\star$	Before
$(\texttt{After})^\star(\texttt{IsResult})^+(\texttt{After})^\star$	After
$(\texttt{Before})^\star(\texttt{HasSubevent})^+(\texttt{Before})^\star$	Before
$(\texttt{Causes})^\star(\texttt{HasSubevent})^+(\texttt{Causes})^\star$	Causes
$(\texttt{After})^\star(\texttt{IsSubevent})^+(\texttt{After})^\star$	After
$(\texttt{IsResult})^\star(\texttt{IsSubevent})^+(\texttt{IsResult})^\star$	IsResult

Table 1: Relation induction rules. * denotes there exists zero or more. $^+$ means there is at least one.

S-CEC	I-CEC	S-CRR	S-CRR	AVG N	AVG E
492	558	767	835	3.62	2.78

Table 2: Statistic of EV^2. S and I are schema and instance. AVG N and AVG E stand for the average number of nodes and edges per graph respectively.

For CRR, we regard $\mathcal{E}_h^s$ and $\mathcal{E}_t^s$ as queried events and use the relation between them as the answer to form the schema-level question. For instance part, we adopt a similar way.

Our CEC task is a 4-way multiple-choice task. The CRR is a 3-way multiple-choice task. In CRR, the choices for temporal, causal, and hierarchy relations are [Before, After, Vague], [Causes, IsResult, None], and [IsSubevent, HasSubevent, None] respectively. We show examples of both tasks in Figure 1. We report the number of each task and the average nodes and edges of EV^2 in Figure 2.

3.4 Quality Inspection

We recruit other human annotators to inspect the quality of EV^2. We sample 100 data for all tasks. We ask them to give two scores for each sample:

Correct: Rate 1 if correct, otherwise rate 0.
Contextualized: Rate 1 if the answer relies on the context events, otherwise rate 0.

Finally, we get 91% for *Correct* and 92% for *Contextualized*. Human examination testifies that EV^2 is qualified. Besides, context events count.

3.5 Existing Dataset Comparison

We compare our benchmark to existing related datasets. We show detailed comparison in Table 3. Our benchmark is the only one that is for contextualized event reasoning of various relations and paradigms on both schema and instance levels.

DATASET	L	C	M-R	M-P
ALTLEX(Hidey, 2016)	*I*	✗	✗	✗
ASER(Zhang et al., 2020)	*S*	✗	✓	✗
ATOMIC(Sap et al., 2019a)	*S*	✗	✓	✗
COPA(Roemmele et al., 2011)	*I*	✗	✗	✗
CQA(Bondarenko et al., 2022)	*I*	✓	✓	✗
ECARE(Du et al., 2022)	*I*	✗	✗	✗
ESL(Caselli and Vossen, 2017)	*I*	✓	✗	✗
ESTER(Han et al., 2021)	*I*	✓	✓	✗
HIEVE(Glavaš et al., 2014)	*I*	✓	✗	✗
KAIROS(Li et al., 2021a)	*S*	✓	✗	✗
LDC2020E25(Li et al., 2021a)	*S*	✓	✗	✗
MATRES(Ning et al., 2018)	*I*	✓	✗	✗
MAVEN-ERE(Wang et al., 2022a)	*I*	✓	✗	✗
MCNC(Granroth-Wilding, 2016)	*I*	✓	✗	✗
MCTACO(Zhou et al., 2019)	*I*	✓	✗	✗
RED(O'Gorman et al., 2016)	*I*	✓	✓	✗
SCITE(Li et al., 2021b)	*I*	✓	✗	✗
SCT(Mostafazadeh et al., 2016)	*I*	✓	✗	✗
SocialIQA(Sap et al., 2019b)	*I*	✓	✓	✗
TB-Dense(Cassidy et al., 2014)	*I*	✓	✗	✗
TRACIE(Zhou et al., 2020)	*I*	✓	✗	✗
EV^2	*S I*	✓	✓	✓

Table 3: Comparison with existing event reasoning datasets. L stands for the included levels. C represents whether it's contextualized. M-R and M-P means if it has multi-relations and paradigms. *S* and *I* stand for schema and instance level.

4 Experiments

4.1 Evaluated LLMs

We evaluate 9 LLMs on event reasoning. For the open-source models, we evaluate their chat-version. For the closed-source models, we utilize their official APIs to conduct performance evaluations. Specifically, we employ the gpt-4-0125-preview version as the GPT4 and the gpt-3.5-turbo-1106 version as GPT3.5 in our experiments. For the open-source models, we include Qwen1.5-7B (Bai et al., 2023), Mistral-7B (Jiang et al., 2023), Baichuan-2-7B (Yang et al., 2023), Llama2-7B (Touvron et al., 2023), WizardLM-7B (Xu et al., 2023), Vicuna-7B (Chiang et al., 2023), and Alpaca-7B (Taori et al., 2023). Without loss of generosity, we use the model names to refer to the chat versions in the rest of our paper. For all evaluated LLMs, we use the same prompt. We show prompts in Figure 5-8 in the Appendix.

5 Results and Findings

5.1 How proficient abilities of event reasoning do LLMs have?

In this part, we mainly probe the abilities of how existing LLMs complete the event reasoning of the instance level.

LLMs have the abilities of event reasoning, but even the strongest GPT-4 is far from satisfactory. We evaluate CEC and CRR at the instance level. We show the results of different relations in Figure 2. For CEC, GPT4 performs the best. Models like Qwen1.5-7B, Mistral-7B, and GPT3.5 are in the second tier. Qwen1.5-7B and Mistral-7B are both better than GPT3.5. Qwen1.5-7B can even excel GPT4 in the temporal and causal relations. The other models such as WizardLM-7B almost fail, obtaining lower than 40% accuracy. For CRR, GPT4 excels all other models as well. However, unlike CEC, there is no obvious difference in the performance of other models for CRR.

We show the average performance of instance-level CEC and CRR in columns I-CEC and I-CRR in Table 4. We only show models that have basic abilities, namely CEC accuracy above 50.00 or CRR above 40.00 in Table 4, while other models may lack analytical significance. Overall, existing LLMs such as GPT4, and Qwen1.5-7B have CEC-tain event reasoning abilities. However, even the strongest GPT4 can only achieve 63.80 and 61.20 accuracy in each task showing there's much room for improvements of event reasoning.

The abilities of LLMs to deal with different relations and reasoning paradigms are unbalanced. Comparing CEC to CRR, as relation-wise results shown in Figure 2 and average performances in columns I-CEC and I-CRR in Table 4, LLMs perform better for CEC than CRR. We compute the average scores of four listed models in Table 4. We find I-CEC is much higher than I-CRR, with 58.91 to 46.18. The results significantly suggest that CRR is harder than CEC. Existing pretraining and SFT datasets may be biased in paradigms.

We then analyze performances on different relations. As shown in Figure 2, LLMs perform best in causality relation. Then, temporal, and hierarchy relations are tied. That further indicates the imbalance training of different relations. Methods and datasets of balanced abilities on relations are needed. Transferring abilities of different relations could also be feasible (Tao et al., 2023b).

This is a crucial finding. Chan et al. (2023) conduct causal event classification such as ECARE (Du et al., 2022), and relation reasoning such as MATRES (Ning et al., 2018). They directly compare these two groups of results and conclude the gaps are merely from differences in relations. However, they ignore the difference in reasoning

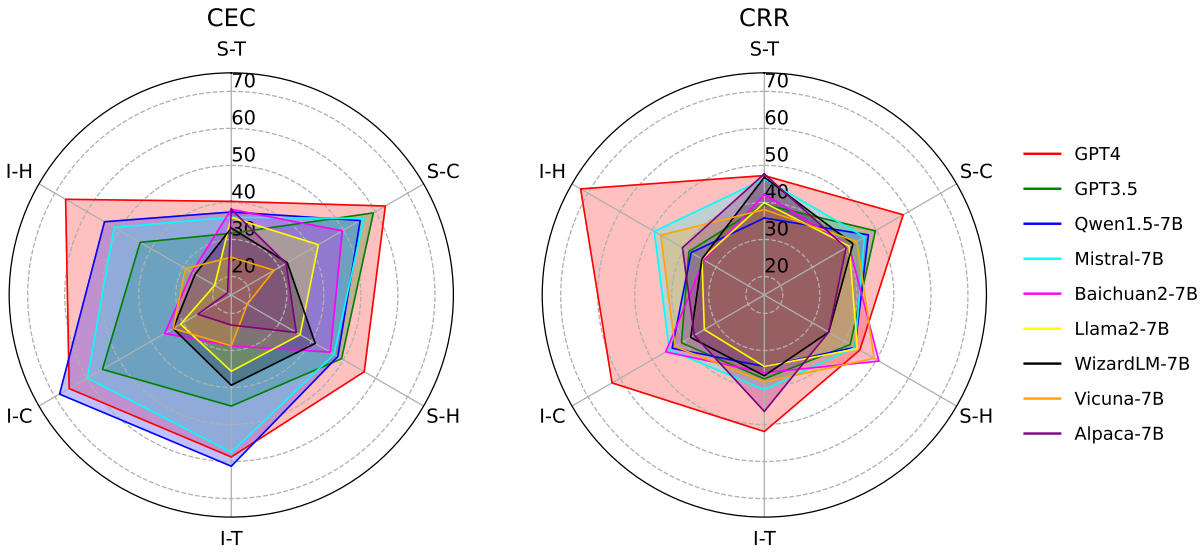

Figure 2: Results of CEC and CRR. S and I stand for schema- and instance-level. Relation types of Causality, Temporality, and Hierarchy are denoted as C, T, and H.

paradigms. Leveraging EV^2, with disentangling relations and formulations, we investigate event reasoning with less bias.

LLMs excel in forward CEC compared with backward. We calculate the average scores of forward relation (`After`, `IsResult`) and backward relation (`Before`, `Causes`). The results are shown in columns I-F, and I-B in Table 4. We find that the average of I-F is significantly better than I-B. It also suggests that the training dataset is unbalanced in relations. Less of the training data is used for backward relations, resulting in poorer performances on those. However, backward relations are important in abduction scenarios. Methods should be designed to enhance such abilities.

CEC improves faster than CRR with model development. We investigate the improvement trends of CEC and CRR. In Figure 3. When models have poor event reasoning abilities, their performances lie around the balanced line showing no significant differences in tasks. With the development, the CEC improves much faster than CRR such models as GPT3.5, Mistral-7B, and Qwen1.5-7B. This investigation appeals to the need for training in comprehensive event reasoning abilities.

5.2 To what extent do LLMs have the event schema knowledge?

In the previous section, we acknowledge that LLMs can complete event reasoning to some extent. However, whether they are endowed with event schema knowledge remains unknown. In this part, we mainly explore to what extent LLMs have the event schema knowledge, i.e. of the schema level.

Model	S-F	S-B	I-F	I-B	S-CEC	I-CEC	S-CRR	I-CRR
GPT4	54.22	55.84	65.34	61.84	55.48	63.80	52.80	61.20
GPT3.5	47.88	50.94	55.68	49.12	49.79	50.18	45.37	39.52
Qwen	42.96	52.45	67.05	65.37	48.98	63.98	43.00	40.00
Mistral	44.37	52.08	67.61	53.71	48.98	57.71	46.00	44.00
AVG	47.35	52.82	63.92	57.51	50.8	58.91	46.79	46.18

Table 4: Average performances. AVG stands for the average scores of all models on that column. S and I stand for schema- and instance-level. F and B are forward and backward relations.

LLMs have event schema knowledge. We evaluate CEC and CRR on the schema level. The results are shown in Figure 2, and the average scores are reported in Table 4. We find LLMs already have event schema knowledge and can complete both CEC and CRR tasks at the schema level to some extent. However, in Table 4, we observe that S-CEC lags I-CEC, suggesting that LLMs are more adept at reasoning at the instance level.

Event schema knowledge increases falling behind reasoning at the instance level. We probe how event schema knowledge increases with the development of LLMs. We depict CEC performance comparisons of LLMs on instance- and schema-level in Figure 4. When the models initially can reason about events, they also have event schema knowledge. At this time, models can perform comparatively or even better in schema-level event reasoning. With the development, models perform instance-level reasoning better than schema-level. It indicates that the accumulation of event schema

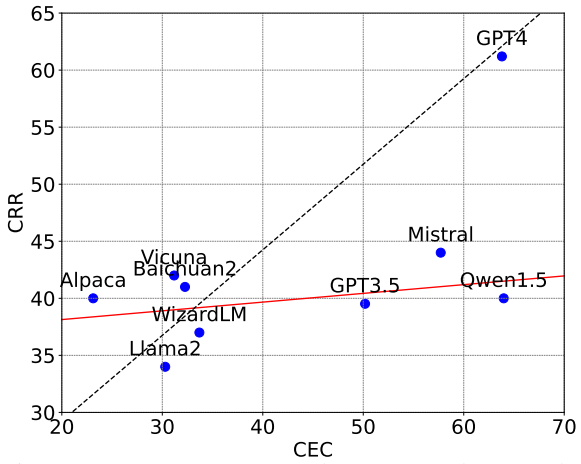

Figure 3: Improvements trend on CEC and CRR. The dashline represents the balanced improvement with slope 3/4 considering the CEC is a 4-way multiple-choice task while CRR has three choices. The red line is the regression line of models except GPT4.

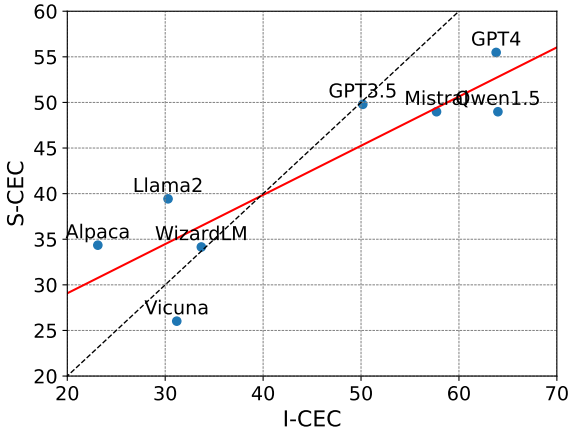

Figure 4: Comparisons between CEC performances on instance- and schema-level. The dashed line represents the balanced improvement with slope 1. The red line is the regression line of all models

knowledge falls behind the reasoning at the instance level. This finding demonstrates that enhancing event schema knowledge may further improve these abilities to obtain better general LLMs.

5.3 Are LLMs aligned with humans in the aspect of leveraging event schema knowledge

In this section, we investigate how LLMs leverage event schema knowledge to complete event reasoning. We first provide the instance-level question for the models and then ask them to generate the required event schema knowledge to solve the task. Then we evaluate the accuracy of the generated event schema knowledge.

Since we have the ground truth event schema knowledge for each question, the only challenge is to guide the LLMs to generate in a similar format for calculating accuracy. The instruction of our prompt first asks LLMs to generate the event types of each instance event in data. Based on the event types, it requires the LLMs to further generate relation triplets needed for the question.

However, we find the LLMs would generate event types of different words but correct contents. To mitigate this problem, we prepare a list of candidate event types for each data to make it a classification setting. To keep the task difficult, we first conduct KMeans clustering on all event types in our dataset[4]. We obtain 1000 clusters. For each data, we assign 20 random candidates in total including the correct ones. The negative event types are chosen from different clusters.

[4]We use `all-mpnet-base-v2` for encoding.

	CEC		CRR	
	ET	REL	ET	REL
GPT4	70.71	37.30	70.66	49.41
GPT3.5	13.43	15.78	18.55	21.14
Mistral-7B	11.15	9.00	11.88	15.15

Table 5: Event schema knowledge Alignment. ET is the event type accuracy. REL is relation triplet F1-score.

After the generation, we calculate the accuracy of event types and F1-scores of relation triplets respectively comparing with the human-labeled event schema. We regard a correct triplet if all the head and tail event types and the inter-relation align with the human labels. We show detailed examples in Figures 9-10 in Appendix.

The results are in Table 5. We find only GPT4 can generate correct event types while other models all fail. For relation triplet generation, even GPT4 can not output proper event schemas[5]. It significantly suggests that LLMs may not leverage event schema knowledge as humans when solving event reasoning tasks. Alignment of using such knowledge could further improve the performances.

5.4 Can LLMs perform better event reasoning with explicit guidance of leveraging event schema knowledge?

In the previous section, we find LLMs may not leverage event schema knowledge as human does. It raises an interesting question how well LLMs perform if we guide them to explicitly use such knowledge? In this section, we probe this question.

We design two guiding methods:

[5]GPT4 excels other may be attributed to 1) its better alignment. 2) The dataset is originally generated by GPT4.

	CEC		CRR	
	W.O.S	W.T.S	W.O.S	W.T.S
GPT4	63.80	69.89 (6.09↑)	61.2	63.11 (1.91↑)
GPT3.5	50.18	60.92 (10.74↑)	39.52	45.99 (6.47↑)
Mistral-7B	57.71	63.26 (5.55↑)	44.00	47.07 (3.07↑)
Llama2-7B	30.29	38.17 (7.88↑)	34.00	43.35 (9.35↑)
WizardLM	33.69	29.93 (3.76↓)	37.00	44.91 (7.91↑)
Vicuna-7B	31.18	34.41 (3.23↑)	42.00	42.40 (0.40↑)

Table 6: Direct guidance with schema knowledge. W.T.S and W.O.S stands for with and without event knowledge guidance. We also report the difference between them.

CEC	W.O.S	W.T.S
GPT4	63.80	67.92 (4.12↑)

Table 7: CoT guidance with schema knowledge.

Direct: Directly add the event type of each instance event into the prompt.

CoT: Guide the LLMs in a CoT-style to 1) generate the event types of each instance event. 2) reason with the event types. This is a more practical method since we would not know the event types in advance in real scenarios.

We show the performances of direct guidance in Table 6. We find incorporating event schema knowledge significantly improves event reasoning. It shows great potential to solve event reasoning with the fusion of event schema knowledge.

We report the results of the CoT guidance in Table 7. We only report results of GPT-4 since we find other models are unable to follow this instruction. We find in CEC, CoT guidance can improve performance. However, the improvement of CoT lags those of Direct, indicating great space for better methods. Developing advanced guidance for all LLMs remains a challenging research problem. We show the example of prompt and GPT4 generation in Figures 11 and 12 in Appendix.

6 Related Work

Event Reasoning 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. Current works present a scenario of incorporating 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.

6.1 Evaluations for LLMs

Evaluating the capacities of LLMs is the foundation of using and improving them. One group of research evaluates the general abilities of LLMs (Hendrycks et al., 2020; Zheng et al., 2023; Zhong et al., 2023; Bang et al., 2023) Besides, existing works evaluate LLMs in specific tasks (Bang et al., 2023; Bian et al., 2023; Gao et al., 2023; Wei et al., 2023) Related to event reasoning, Yuan et al. (2023) evaluated the ability to solve event relation extraction. Tao et al. (2023a) present the Event Semantic Processing including the event understanding, reasoning, and prediction of event semantics. Chan et al. (2023) investigates relation reasoning between sentences. Compared with them, we are the first to introduce the evaluation for both schema- and instance-level event reasoning. Moreover, we comprehensively evaluate the performances of various relations and reasoning paradigms.

7 Conclusion

In this paper, we evaluate the event reasoning of LLMs. We introduce a novel benchmark EV^2 which features both levels of schema and instance. It evaluates event schema knowledge and reasoning abilities. Besides, EV^2 can be used to comprehensively evaluate the event reasoning in various relations and reasoning paradigms. We conduct extensive experiments on EV^2. We obtain many insights such as: 1) LLMs have the abilities of event reasoning, but are far from satisfactory and are unbalanced in different relations and reasoning paradigms. 2) LLMs have event schema knowledge. However, with the development of LLMs, this knowledge increases slowly compared with the increase of abilities of event instance reasoning. 3) LLMs are not aligned with human to leverage event schema knowledge in event reasoning. 4) Based on the findings, we design two methods, namely Direct and CoT, to guide the LLMs to utilize event schema knowledge. With our designed guidances for utilizing event schema knowledge, LLMs can perform better event reasoning.

Limitations

We guide the LLMs to utilize the event schema knowledge in two ways. The *Direct* effects most. However, the more practical way *CoT* falls behind *Direct* indicating the potential of a better method of guidance. We leave it to future work.

References

Daman Arora, Himanshu Gaurav Singh, et al. 2023. Have llms advanced enough? a challenging problem solving benchmark for large language models. *arXiv preprint arXiv:2305.15074*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Ning Bian, Xianpei Han, Le Sun, Hongyu Lin, Yaojie Lu, and Ben He. 2023. Chatgpt is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models. *arXiv preprint arXiv:2303.16421*.

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.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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.

Taylor Cassidy, Bill McDowell, Nathanael Chambers, and Steven Bethard. 2014. An annotation framework for dense event ordering. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 501–506.

Chunkit Chan, Jiayang Cheng, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. 2023. Chatgpt evaluation on sentence level relations: A focus on temporal, causal, and discourse relations. *arXiv preprint arXiv:2304.14827*.

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.

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*.

Jun Gao, Huan Zhao, Changlong Yu, and Ruifeng Xu. 2023. Exploring the feasibility of chatgpt for event extraction. *arXiv preprint arXiv:2303.03836*.

Goran Glavaš, Jan Šnajder, Parisa Kordjamshidi, and Marie-Francine Moens. 2014. Hieve: A corpus for extracting event hierarchies from news stories. In *Proceedings of 9th language resources and evaluation conference*, pages 3678–3683. ELRA.

Mark Granroth-Wilding. 2016. What happens next? event prediction using a compositional neural network model. In *AAAI Conference on Artificial Intelligence*.

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.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

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

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*.

Manling Li, Sha Li, Zhenhailong Wang, Lifu Huang, Kyunghyun Cho, Heng Ji, Jiawei Han, and Clare Voss. 2021a. The future is not one-dimensional: Complex event schema induction by graph modeling for event prediction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5203–5215.

Zhaoning Li, Qi Li, Xiaotian Zou, and Jiangtao Ren. 2021b. Causality extraction based on self-attentive bilstm-crf with transferred embeddings. *Neurocomputing*, 423:207–219.

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*.

Qianren Mao, Xi Li, Hao Peng, Jianxin Li, Dongxiao He, Shu Guo, Min He, and Lihong Wang. 2021. Event prediction based on evolutionary event ontology knowledge. *Future Generation Computer Systems*, 115:76–89.

Alexis Mitchell. 2005. The automatic content extraction (ace) program-tasks, data, and evaluation.

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.

Tim O'Gorman, Kristin Wright-Bettner, and Martha Palmer. 2016. Richer event description: Integrating event coreference with temporal, causal and bridging annotation. In *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*, pages 47–56, Austin, Texas. Association for Computational Linguistics.

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.

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*.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019a. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3027–3035.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019b. 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.

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*.

Zeno Vendler. 1957. Verbs and times. *The philosophical review*, pages 143–160.

Xiaozhi Wang, Yulin Chen, Ning Ding, Hao Peng, Zimu Wang, Yankai Lin, Xu Han, Lei Hou, Juanzi

Li, Zhiyuan Liu, et al. 2022a. Maven-ere: A unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction. *arXiv preprint arXiv:2211.07342*.

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

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, et al. 2023. Zero-shot information extraction via chatting with chatgpt. *arXiv preprint arXiv:2302.10205*.

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. 2023. 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.

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

Chenhan Yuan, Qianqian Xie, and Sophia Ananiadou. 2023. Zero-shot temporal relation extraction with chatgpt. *arXiv preprint arXiv:2304.05454*.

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.

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*.

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*.

Appendix

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

Context:
"study" is a subevent of "analyse". "analyse" is after "think". "pass_class" is after "study".

Question:
Which event has the subevent of "think"? Choices:
A. research
B. attend_conference
C. plan_project
D. talk_to The answer is

The answer is

Figure 5: Prompt of schema-level CEC.

Instructions:
Answer the question by selecting A, B, C, D. Note that all events appearing in "Context", "Question", and "Choices" refer to the specific events described in "Instances".

Instances:
event9:
In anticipation of the challenging final exams, she had dedicated countless evenings in the library, poring over textbooks and academic papers on climate change.
event65:
These insights were further refined through discussions with her mentor, Professor Ramirez, who provided valuable feedback and perspectives.
event30:
One pivotal moment was when she deciphered the complex data on global warming trends, developing a comprehensive presentation.
event99:
After months of preparation, Alice finally received her diploma in Environmental Science from the university.
event86:
This breakthrough came after she spent a weekend in solitude at a cabin in the woods, reflecting on the interconnectedness of natural systems.
event32:
After her deep reflections, she crafted an ambitious project plan aiming to initiate a community-driven reforestation program, outlining steps for local engagement and environmental restoration efforts.
event5:
Alice attended an international conference on sustainable development, where she presented her findings on the effectiveness of renewable energy sources in reducing carbon emissions.
event90:
She conducted an in-depth analysis of historical environmental policy reforms to understand their impact on current climate advocacy strategies.

Context:
"event9" is a subevent of "event30". "event30" is after "event86". "event99" is after "event9".

Question:
Which event has the subevent of "event86"?
Choices:
A. event32
B. event65
C. event90
D. event5

The answer is

Figure 6: Prompt of instance-level CEC.

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

Context:
"fall" is a subevent of "exercise". "miss" causes "fall". "miss" causes "exercise".

Question:
Which is the causal relationship between "fall" and "lack_energy"?
Choices:
A. "fall" causes "lack_energy".
B. "fall" is result of "lack_energy".
C. There is no obvious causal relationship between "fall" and "lack_energy".

The answer is

Figure 7: Prompt of schema-level CRR.

Instructions:
Answer the question by selecting A, B or C. Note that all events appearing in "Context", "Question", and "Choices" refer to the specific events described in "Instances".

Instances:
event66:
Sitting in the second row, the jurors leaned forward, focusing intently on every word spoken by the witness, understanding the gravity of the details being shared.
event88:
The court case of John Doe for alleged embezzlement commenced on a rainy Monday morning at the downtown courthouse.
event90:
During the proceedings, a key witness was called to the stand to provide a detailed account of the financial transactions in question.

Context:
"event88" causes "event90".

Question:
Which is the subordinate relationship between "event90" and "event66"?
Choices:
A. "event66" is subevent of "event90".
B. "event90" is subevent of "event66".
C. There is no obvious subordinate relationship between "event90" and "event66".

The answer is

Figure 8: Prompt of instance-level CRR.

Instructions:
In a scenario explained by the "Context", the "Question" ask about selecting one event that has a certain event relationship with a particular event, from all possible tail events provided by the "Choices", and the "Instances" explain all the events in detail with a few sentences. Event semantic knowledge refers to the abstract event types to which specific events belong, and the relationships between these abstract event types. Please output the event semantic knowledge used in solving the following problem. Note that all possible abstract event types in the "Schema", and the relationships between abstract events include HasSubevent, IsSubevent, Before, After, Causes, and IsResult. For the tuple [event0, relation, event1], HasSubevent indicates that event1 is a subevent of event0, IsSubevent indicates that event0 is a subevent of event1, Before indicates that event0 occurs before event1, After indicates that event0 occurs after event1, Causes indicates that event0 causes event1, and IsResult indicates that event0 is the result of event1. Output in JSON format, don't generate other sentences.

Requirements:
Abstract event types can only be chosen from "Schema", and the relationships of abstract event types can only be selected from HasSubevent, IsSubevent, Before, After, Causes, and IsResult. Follow the format in examples, output in JSONL format. The key "event type" should correspond to a value that is a dictionary with events as keys and their abstracted categories as values. The key "event relation" should correspond to a value that is a list of tuples [event0, relation, event1]. The relationships between events include HasSubevent, IsSubevent, Before, After, Causes, and IsResult.

Schema:
artistic_innovation, relocate, pass_class, experience_emotional_distress, seek_guidance, think, drinking_coffee, answer, attend_conference, tell_lies, feeling_homesick, bored, review_notes, research, talk_to, study, plan_project, getting_exercise, analyse, gaining_recognition

Here are some examples:

Instances:
event53:
This revelation ultimately prompted an individual in the courtroom audience to discretely exit the room and later that evening, the same individual, driven by fear of exposure, went on to commit a fatal assault against a witness who could connect him to the crime.
event67:
While Mr. Smith was providing his account, he mentioned a key detail that was previously overlooked—a unique tattoo that he glimpsed on the perpetrator's arm.
event91:
Two weeks later, during the heated court proceedings at the downtown courthouse, the homeowner, Mr. Smith, was called to testify before the jury regarding the night of the incident.
event64:
In a quiet suburban neighborhood, a burglary occurred at the Smith residence, where an unknown assailant broke in and stole valuable heirlooms late at night.
event51:
Upon hearing the new testimony about the tattoo, a juror who happened to have an interest in body art quietly made a mental note to research the design's origins, intrigued by its possible cultural significance.
event13:
Upon hearing Mr. Smith's testimony, one juror with claustrophobia experienced a severe panic attack, which led to the court session being temporarily adjourned as the juror was rushed to the hospital for medical attention.
event90:
In a surprising turn of events during the tea break, the court stenographer, overwhelmed by guilt, approached the judge and admitted to tampering with court transcripts in a previous unrelated case, sparking an investigation into judicial misconduct.

Context:
"event64" is before "event91". "event67" is a subevent of "event91".

Question:
Which is caused by "event67"?
Choices:
A. event13
B. event90
C. event53
D. event51

Event type and event relation:
{"event_type": {"event67": "talk", "event53": "kill", "event64": "commit_crime", "event91": "take_stand", "event51": "hide_evidence", "event13": "escape", "event90": "confess"}, "event_relation": [["commit_crime", "Before", "take_stand"], ["take_stand", "HasSubevent", "talk"], ["talk", "Causes", "kill"]]}

Now, based on the above, please output the event semantic knowledge used in solving the following problem.

Instances:
event9:
In anticipation of the challenging final exams, she had dedicated countless evenings in the library, poring over textbooks and academic papers on climate change.
event65:
These insights were further refined through discussions with her mentor, Professor Ramirez, who provided valuable feedback and perspectives.
event30:
One pivotal moment was when she deciphered the complex data on global warming trends, developing a comprehensive presentation.
event99:
After months of preparation, Alice finally received her diploma in Environmental Science from the university.
event86:
This breakthrough came after she spent a weekend in solitude at a cabin in the woods, reflecting on the interconnectedness of natural systems.
event32:
After her deep reflections, she crafted an ambitious project plan aiming to initiate a community-driven reforestation program, outlining steps for local engagement and environmental restoration efforts.
event5:
Alice attended an international conference on sustainable development, where she presented her findings on the effectiveness of renewable energy sources in reducing carbon emissions.
event90:
She conducted an in-depth analysis of historical environmental policy reforms to understand their impact on current climate advocacy strategies.

Context:
"event9" is a subevent of "event30". "event30" is after "event86". "event99" is after "event9".
Question:
Which event has the subevent of "event86"?
Choices:
A. event32
B. event65
C. event90
D. event5

Event type and event relation:

Ground Truth:
{"event_type": {"event86": "think", "event65": "talk_to", "event9": "study", "event30": "analyse", "event99": "pass_class", "event32": "plan_project", "event5": "attend_conference", "event90": "research"}, "event_relation": [["study", "IsSubevent", "analyse"], ["analyse", "After", "think"], ["pass_class", "After", "study"], ["think", "IsSubevent", "talk_to"]]}

Figure 9: A example of detailed Prompt and ground truth for alignment evaluation on CEC.

Instructions:
In a scenario explained by the "Context", the "Question" ask about selecting one relationship between two events, from all possible relationships provided by the "Choices", and the "Instances" explain all the events in detail with a few sentences. Event semantic knowledge refers to the abstract event types to which specific events belong, and the relationships between these abstract event types. Please output the event semantic knowledge used in solving the following problem. Note that all possible abstract event categories in the "Schema", and the relationships between abstract events include HasSubevent, IsSubevent, Before, After, Causes, and IsResult. For the tuple [event0, relation, event1], HasSubevent indicates that event1 is a subevent of event0, IsSubevent indicates that event0 is a subevent of event1, Before indicates that event0 occurs before event1, After indicates that event0 occurs after event1, Causes indicates that event0 causes event1, and IsResult indicates that event0 is the result of event1.

Requirements:
Abstract event types can only be chosen from "Schema", and the relationships of abstract event types can only be selected from HasSubevent, IsSubevent, Before, After, Causes, and IsResult. Follow the format in examples, output in JSONL format. The key "event type" should correspond to a value that is a dictionary with events as keys and their abstracted categories as values. The key "event relation" should correspond to a value that is a list of tuples [event0, relation, event1]. The relationships between events include HasSubevent, IsSubevent, Before, After, Causes, and IsResult.
Schema:
injury, write_story, review_appeal, competing, improving_skill, delivering_verdict, experience_illness, competing_against, discuss_results, engage_physical_activity, trial, plan, gaining_weight, hear_testimony, conduct_research, develop_immunity, decorate_venue, organize_thoughts, educate_child, paying_attention

Here are some examples:

Instances:
event53:
This revelation ultimately prompted an individual in the courtroom audience to discretely exit the room and later that evening, the same individual, driven by fear of exposure, went on to commit a fatal assault against a witness who could connect him to the crime.
event67:
While Mr. Smith was providing his account, he mentioned a key detail that was previously overlooked—a unique tattoo that he glimpsed on the perpetrator's arm.
event91:
Two weeks later, during the heated court proceedings at the downtown courthouse, the homeowner, Mr. Smith, was called to testify before the jury regarding the night of the incident.
event64:
In a quiet suburban neighborhood, a burglary occurred at the Smith residence, where an unknown assailant broke in and stole valuable heirlooms late at night.

Context:
"event64" is before "event91". "event67" is a subevent of "event91".

Question:
Which is the causal relationship between "event67" and "event53"?
Choices:
A. "event67" causes "event53".
B. "event67" is result of "event53".
C. There is no obvious causal relationship between "event67" and "event53".
Event type and event relation:
{"event_type": {"event67": "talk", "event53": "kill", "event64": "commit_crime", "event91": "take_stand"}, "event_relation": [["commit_crime", "Before", "take_stand"], ["take_stand", "HasSubevent", "talk"], ["talk", "Causes", "kill"]]}

Now, based on the above, please output the event semantic knowledge used in solving the following problem.

Instances:
event66:
Sitting in the second row, the jurors leaned forward, focusing intently on every word spoken by the witness, understanding the gravity of the details being shared.
event88:
The court case of John Doe for alleged embezzlement commenced on a rainy Monday morning at the downtown courthouse.
event90:
During the proceedings, a key witness was called to the stand to provide a detailed account of the financial transactions in question.

Context:
"event88" causes "event90".

Question:
Which is the subordinate relationship between "event90" and "event66"?
Choices:
A. "event66" is subevent of "event90".
B. "event90" is subevent of "event66".
C. There is no obvious subordinate relationship between "event90" and "event66".

Event type and event relation:

Ground Truth:
{"event_type": {"event90": "hear_testimony", "event88": "trial", "event66": "paying_attention"}, "event_relation": [["trial", "Causes", "hear_testimony"], ["hear_testimony", "HasSubevent", "paying_attention"]]}

Figure 10: A example of detailed Prompt and ground truth for alignment evaluation on CRC.

Instructions:
Answer the question by selecting A, B, C, D. Note that all events appearing in "Context", "Question", and "Choices" refer to the specific events described in "Instances".

Instances:
event52:
Prior to the physical confrontation, a vigorous debate took place, with individuals from both sides passionately voicing their opinions on the matter at hand.
event61:
The root of the evening's events can be traced back to an organized competition earlier that day, which brought together activists from various factions to discuss their differences on policy issues.
event23:
This unfortunate incident occurred when tensions escalated between opposing groups, leading to a physical altercation among participants.
event39:
During the late hours of a cold Friday night, a local journalist suffered a broken arm while covering a heated protest outside a government building.
event60:
Efforts were made to mediate between the opposing factions through a series of workshops aimed at fostering mutual understanding and respect.
event95:
First responders were dispatched to the scene of a minor road accident involving two vehicles, where they provided medical assistance and traffic control.
event46:
Legislators engaged in lengthy discussions to amend existing laws, aiming to improve public safety and reduce the incidents of violence at public events.

Context:
"event39" is a result of "event23". "event23" is a result of "event52".
Question:Which is the cause of "event52"?
Choices:
A. event60
B. event46
C. event95
D. event61

Event Type: Refers to common knowledge about events, which has certain universal rules and does not change with time or location.
Instance Event: The instantiation of an abstract event in a specific scenario. The relationships and development of instance events are related to the background.
Type Relationship: Refers to the relationship between two events.
Please think as this way:
Firstly, think the event type for each event in the question and choices, then infer the relationship between the question and choices and find out the answer. In addition, Context should also be taken into account, as it will provide a specific background for this question.

For example:
Event0: A captivating billboard about a new smartphone catches Sarah's attention on her way to work. This can be categorized as the event type [good_advertisement].
Event1: Feeling intrigued, Sarah decides to allocate part of her savings towards purchasing the gadget advertised. This can be categorized as the event type [spend_money].
From the event types of Event0[good_advertisement] and Event1[spend_money], because it is a consensus that good advertisements can prompt people to spend money, we can infer that [good_advertisement] causes [spend_money].
So when the question is: which event is caused by Event0, the answer is Event1.

Note: When there is no completely correct answer, please select the closest correct answer.

Figure 11: An example of prompt for CoT guidance.

Given the context and the instances described, we need to determine the cause of "event52," which is described as a vigorous debate that took place before a physical confrontation ("event23"), which itself led to a journalist suffering a broken arm during a protest ("event39").

Let us analyze the choices based on their event types:

A. event60 - This event involves efforts to mediate between opposing factions through workshops aimed at fostering mutual understanding and respect. The event type here can be categorized as [conflict_resolution].

B. event46 - This event involves legislators engaging in discussions to amend laws for improving public safety and reducing violence at public events. The event type can be categorized as [legislative_action].

C. event95 - This event is about first responders attending to a minor road accident. The event type here can be categorized as [emergency_response].

D. event61 - This event is about an organized competition earlier that day, which brought together activists from various factions to discuss their differences on policy issues. The event type can be categorized as [organized_discussion].

Given the sequence of events leading to "event52," which is a vigorous debate before a physical confrontation, we need to find an event that could logically precede and cause a debate among individuals with opposing views.

Comparing the event types:

- [conflict_resolution] (A) typically follows conflicts or debates, not precedes them.
- [legislative_action] (B) is more of a response to ongoing issues rather than a direct cause of a specific debate.
- [emergency_response] (C) is unrelated to the context of debates or confrontations over policy issues.
- [organized_discussion] (D) is a setting that naturally precedes and can cause vigorous debates, as it involves bringing together individuals with differing views in a structured manner, which can escalate into more heated discussions.

Therefore, the cause of "event52" is most closely related to the event type [organized_discussion], making the correct answer:

D. event61

Figure 12: An example of GPT4 generation of CoT guidance.