A Comprehensive Evaluation on Event Reasoning of Large Language Models

Zhengwei Tao[1−4], Zhi Jin[1,2]✉, Yifan Zhang[1,2], Xiancai Chen[1,2], Haiyan Zhao[1,2], Jia Li[1,2]
Bin Liang[3,4], Chongyang Tao[5], Qun Liu[6], Kam-Fai Wong[3,4]

[1]School of Computer Science, Peking University,
[2]MoE Key Lab. of High Confidence Software Technologies(PKU), China
[3]Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong
[4]MoE Key Lab. of High Confidence Software Technologies(Hong Kong), China
[5] Beihang University,
[6] Huawei Noah's Ark Lab
{tttzw, xiancaich, yifanzhang}@stu.pku.edu.cn, {zhijin, zhhy.sei, lijiaa}@pku.edu.cn, chongyang@buaa.edu.cn,
bin.liang@cuhk.edu.hk, qun.liu@huawei.com, kfwong@se.cuhk.edu.hk

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. The extent to which LLMs excel in event reasoning across various relations and reasoning paradigms has not been thoroughly investigated. Additionally, it is still unclear whether LLMs utilize event knowledge in the same way humans do. To mitigate this disparity, we comprehensively evaluate the abilities of event reasoning of LLMs on different relations, paradigms, and levels of abstraction. We introduce a novel benchmark EV^2 for EValuation of EVent reasoning. EV^2consists of two levels of evaluation on schema and instance and is comprehensive in relations and reasoning paradigms. We conduct extensive experiments on EV^2. We find that 1) LLMs have abilities to accomplish event reasoning but their performances are far from satisfactory. 2) There are imbalances of event reasoning abilities on different relations and paradigms. 3) LLMs have event schema knowledge, however, they're not aligned with humans on how to utilize the knowledge. Based on these findings, we guide the LLMs in utilizing the event schema knowledge as memory for improvements in event reasoning.

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 of event evolution. Knowing the event occurrence

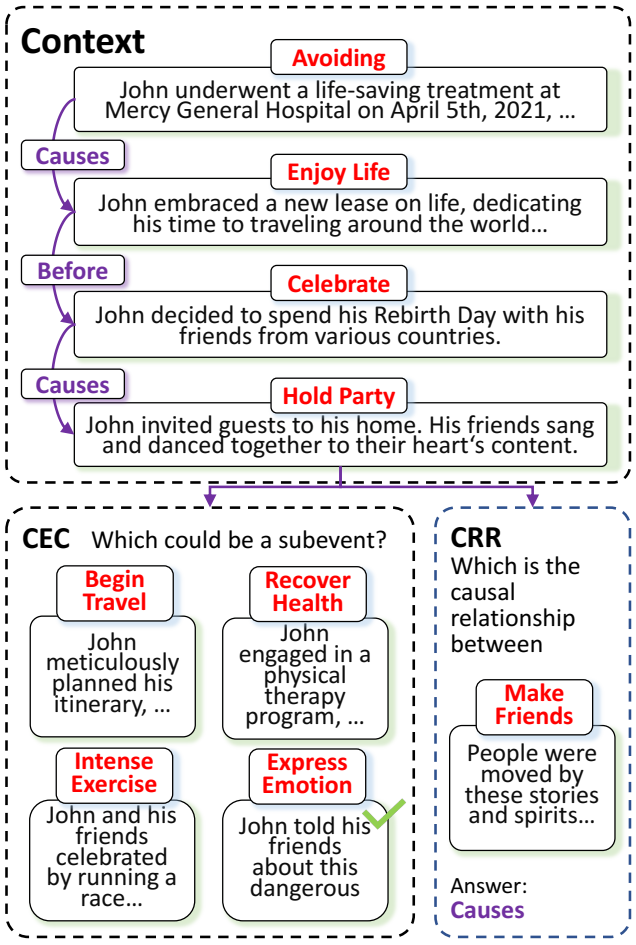

Figure 1: The red words are event schema knowledge where event instances are below. There are various paradigms such as Contextual Event Classification (CEC) and Contextual Relation Reasoning (CRR), and diverse relations.

✉ Corresponding author.

chain as "Avoiding", "Enjoy life", "Celebrate", and "Hold Party" would result in the subevent "Express Emotion".

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, Wu, and Roth 2018; Caselli and Vossen 2017). The queried relations are diversified such as causality (Roemmele, Bejan, and Gordon 2011), temporality (Zhou et al. 2019), and hierarchy (Glavaš et al. 2014). There are various paradigms such as reasoning the event or the inter-relation. As a fundamental competency within LLMs, event reasoning supports a multitude of Natural Language Processing (NLP) tasks, including recommendation engines (Yang et al. 2020), interactive question-answer systems (Souza Costa, Gottschalk, and Demidova 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; Xu 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, Singh et al. 2023). However, studies on the comprehensive evaluation of event reasoning of LLMs are scarce. The incompleteness of current event reasoning evaluation is reflected in two aspects. First, 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). Investigating the event knowledge of LLMs and how they employ them underlines applications such as event-based memory systems. Besides, they ignore the diversity of relations and paradigms (Yuan, Xie, and Ananiadou 2023). Such findings could be biased since they neglect discrepancies brought by different aspects. 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 no existing datasets that are comprehensive in relations and paradigms, and can cover both levels of schema and instance, we introduce a 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, EV^2 evaluates 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, and three types of relations of causality, temporality, and hierarchy. Utilizing EV^2, we evaluate how well LLMs do event reasoning in abilities and knowledge. 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 embed imbalanced abilities in different relations and reasoning paradigms. 3) LLMs have event schema knowledge. However, LLMs are not aligned with humans in the aspect of leveraging event schema knowledge. Based on the findings, we investigate guiding the LLMs to utilize event schema knowledge. With the guidance, LLMs can perform better event reasoning which sheds light on modeling event knowledge as memory of LLMs to enhance event reasoning.

We summarize our contributions as follows:

- We first comprehensively evaluate event reasoning in both abstraction levels of schema and instance, and various relations and paradigms.
- We construct a 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 guide LLMs to utilize event schema knowledge as memory achieving improvements in event reasoning.

Problem Formulation

Event reasoning is to anticipate the events by certain relations or deduce interrelated correlations (Tao et al. 2023a). 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 $(\mathcal{E}_t^s, \mathcal{R}, \mathcal{E}_j^s)$ standing for that there is the relation $\mathcal{R}$ between $\mathcal{E}_t^s$ and $\mathcal{E}_j^s$. With instantiation, we have the instance-level event 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 description (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`$\}$[3]. `Causes` and `IsResult` are Causal relations, `Before` and `After` belong to Temporal type while `IsSubevent` and `HasSubevent` are Hierarchical type. We consider two event reasoning paradigms for both the schema and instance levels: Contextual Event Classification (CEC) and 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

[1]Superscript s represents schema level.

[2]Superscript i represents instance level.

[3]In this work, we denote the direction of the relation to that the former event have relation to the latter event. For example, ($\mathcal{E}_t$, `Causes`, $\mathcal{E}_j$) stands for $\mathcal{E}_i$ causes $\mathcal{E}_j$.

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 $\mathcal{E}_t, \mathcal{E}_j \in \mathcal{G}$, CRR requires to determine the relation $\mathcal{R}$ between them:

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

CRR evaluates the understanding of event relations.

Both schema and instance levels have CEC and CRR tasks. Schema-level tasks require models to be rich in knowledge while instance need models to process details.

Benchmark Construction

Constructing the EV^2 benchmark is challenging since events and their relations are semantically abstract concepts compared with entity concepts. The occurrence of events and their relations not only follow objective natural laws but are also influenced by social and humanistic factors. Therefore, annotating such data is extremely label-intensive leading to a lack of evaluation dataset of the task. Previous works mainly construct such datasets by extracting events and relations from some unlabeled corpus such as news reports (Caselli and Vossen 2017; Ning, Wu, and Roth 2018; O'Gorman, Wright-Bettner, and Palmer 2016). However, such a method suffers from limited event relational patterns and domain-specific language expression. To mitigate such problems, in this work, we construct our comprehensive event reasoning evaluation from scratch.

In EV^2 , we evaluate event reasoning in various relation and reasoning paradigms of both schema and instance levels. In our pilot trial, we directly annotate each question and answer by human annotators. We find it extremely time-consuming since human annotators must imagine scenarios and guarantee their correctness. Besides, the diversity of the events and their relations are poor. Annotators only write the most common events. Therefore, we propose a five-stage construction process. Generally, we first synthesize both the schema and corresponding instance event prompting-graphs automatically as prompts for later annotations. Due to the large size of the synthesis graphs, before formal annotating, we then recruit annotators to remove incorrect graphs which are hard to modify. After that, the annotators curate schema and instance graphs based on the generated prompts. Then, we adapt the graphs into questions and answers. Finally, we recruit human validators to examine the quality of the data. We celebrate this process in the following sections.

Prompting-Graph Synthesizing

To cover more scenarios, we synthesize event graphs as prompts for annotations instead of annotating from scratch. Specifically, we first establish the schema graph $\mathcal{G}^s$. Then we employ GPT4 to generate the instance graph $\mathcal{G}^i$.

Schema Graph The schema graph represents event occurrence knowledge. Harvesting such knowledge is a research point (Du 2022). We here 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`, `Before`, and `HasSubevent`.

Our objective mandates that the nodes within $\mathcal{G}^s$ should represent event types. Therefore, we filter EECKG nodes, removing concrete event instances. We keep 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[4].

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 threshold, thus defining a component. This process is executed for all nodes to gather components, as Algorithm 1 in the Appendix. Then 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 then sample three event pairs $(\mathcal{E}_h, \mathcal{E}_t)$ per component with a maximum inter-path length of four with their predecessors as background events. These pairs and background events form a schema graph. If the path length between $\mathcal{E}_h$ and $\mathcal{E}_t$ is two, their relation serves as the queried relation; for longer paths, we deduce the relation using predefined rules as shown in Appendix Table 6. We construct a schema graph, queried event pair, and their relation $(\mathcal{E}_h, \mathcal{E}_t, \mathcal{R}, \mathcal{G}^s)$.

Instance Graph 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 $\mathcal{E}^i$.

We inherit the relations of $\mathcal{G}^s$ to obtain $\mathcal{G}^i$. We naturally obtain the instances of $\mathcal{E}_h$ and $\mathcal{E}_t$. We obtain 1,600 schema prompting graphs and 1,600 corresponding instances.

Manual Filtering

After curating the prompting graphs of both levels, the next is to annotate based on the prompting graphs. However, we find some of the prompting graphs are incorrect and hard to modify. Therefore, before formal annotation, we launch another manual filtering step to remove such graphs.

[4]See all prompts in extension version.

S-CEC	I-CEC	S-CRR	I-CRR	GRAPH PAIRS
492	491	730	735	491

Table 1: Number of EV^2. S and I are schema and instance.

We then recruit 8 well-educated human annotators where they each process 200 data. Their missions are to investigate the $\mathcal{G}^s$ and $\mathcal{G}^i$ by the following steps:

1) Check whether $\mathcal{G}^s$ can be modified, the events are abstract, the relations are correct, and $\mathcal{G}^s$ tells an entire story of a scenario.
2) Check whether $\mathcal{G}^i$ can be modified., the events are concrete, the relations are correct, and $\mathcal{G}^i$ tells an entire story of a scenario.
3) Check whether $\mathcal{G}^s$ and $\mathcal{G}^i$ are consistent, there are obvious schema-instance relations between them.

If any of these conditions are not met, we discard this datum. After this filtering process, we remain 491 graph pairs.

Annotation

In this stage, we formally annotate based on filtered prompting-graphs. The missions are to 1) rewrite the events and relations in both schema and instance graphs to make them strictly valid. 2) identify a query event $\mathcal{E}_h^s$ and an answer event $\mathcal{E}_t^s$ in the graph for later question adaptation. 3) write candidates as negative choices considering the answer event where each candidate event consists of a schema and an instance event. For the second mission, 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 schema events for prompting. We recruit another 10 annotators. The annotation is completed on our annotating system. We describe the detailed annotation process in the Appendix. Each annotator should rewrite correct data alongside the following standards:

1) Rewrite $\mathcal{G}^s$ and $\mathcal{G}^i$ making them correct as high-probability knowledge, and do not consider low-probability situations.
2) Rewrite $\mathcal{G}^s$ and $\mathcal{G}^i$ leading to the coherence of the whole graph, and there's no semantic drift.
3) Rewrite $\mathcal{G}^s$ and $\mathcal{G}^i$ making them consistent. Schema events and instance events require a clear distinction in hyper-hypo relation.
4) Rewrite the target event making it only can be answered when considering the whole graph.
5) Rewrite the instance events making them should be expressed independently without connective expressions such as "After $\mathcal{E}_1$" to avoid information leakage.

Question Adaptation

The last is to construct questions of CEC and CRR in both schema and instance levels based on the annotated graphs. The schema part of annotation is for the schema-level questions and the instance part is for instance-level.

For schema-level CEC, we naturally use the queried event $\mathcal{E}_h^s$ and other nodes except for the answer event $\mathcal{E}_t^s$ as context to form a question. Then we use the answer event $\mathcal{E}_t^s$ as the answer. We do similarly at the instance level. 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.

Finally, 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.

DATASET	L	C	M-R	M-P
ALTLEX(Hidey 2016)	*I*	✗	✗	✗
ASER(Zhang et al. 2020)	*S*	✗	✓	✗
ATOMIC(Sap et al. 2019a)	*S*	✗	✓	✗
COPA	*I*	✗	✗	✗
CQA(Bondarenko 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, Wu, and Roth 2018)	*I*	✓	✗	✗
MAVEN-ERE(Wang et al. 2022a)	*I*	✓	✓	✗
MCNC(Granroth-Wilding 2016)	*I*	✓	✗	✗
MCTACO(Zhou et al. 2019)	*I*	✓	✗	✗
RED	*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 2020)	*I*	✓	✗	✗
EV^2	*S I*	✓	✓	✓

Table 2: Comparison with existing event reasoning datasets. L stands for the included levels. C represents whether it's contextualized. M-R and M-P means multi-relations and paradigms. *S* and *I* stand for schema and instance level.

Quality Validation

After that, we recruit another three human validators to guarantee the quality of both tasks in EV^2. They delete the non-qualified data by the following criteria:

1) Delete the data if the logic of the graph is incorrect.
2) Delete the data if any of the negative candidates can also be the correct answer.
3) Delete the data if it can be answered without the context.

After the quality validation, we have our final EV^2 benchmark. We report the number of each task and the average nodes and edges in Table 1.

Existing Dataset Comparison

We compare our benchmark to existing related datasets. We show detailed comparison in Table 2. Our benchmark is the only one that is for contextualized event reasoning of various relations and paradigms on both schema and instance levels.

Model	S-CEC	I-CEC	S-CRR	I-CRR
GPT4o	68.93	66.60	62.05	62.04
GPT4	68.11	68.43	63.01	63.81
GPT3.5	65.43	64.77	54.52	43.95
Mistral-7B	52.47	54.18	51.64	55.24
Qwen2-7B	62.14	63.75	52.05	47.62
Baichuan2-7B	52.88	29.94	51.64	45.31
Orca2-7B	59.88	60.08	46.16	45.17
Chatglm2-6B	55.76	30.96	52.47	49.66
Interlm2-7B	65.84	62.12	48.63	57.28
Llama2-7B	45.06	29.74	46.58	41.22
Vicuna-7b	25.93	27.09	52.05	51.43

Table 3: Average performances with updated models. S and I stand for schema- and instance-level.

	CEC		CRR	
	ET	REL	ET	REL
GPT4o	65.53	40.23	69.84	51.24
GPT4	72.78	40.57	73.19	52.79
GPT3.5	10.05	15.58	16.19	28.63
Mistral-7B	22.93	16.28	25.00	23.87
Qwen2-7B	11.40	15.97	13.07	16.58

Table 4: Event schema knowledge Alignment. ET is the event type accuracy. REL is relation triplet F1-score.

Experiments Results and Findings

Evaluated LLMs We evaluate 11 LLMs on event reasoning. For the open-source models, we evaluate their chat-version. We evaluate GPT4o, GPT4, GPT3.5. For the closed-source models, we utilize their official APIs to conduct performance evaluations. For the open-source models, we include Mistral-7B (Jiang 2023), Qwen2-7B (Yang et al. 2024), Baichuan-2-7B (Yang et al. 2023), Orca2-7B (Mitra et al. 2023), Chatglm2-6B (GLM 2024), Internlm2-7B (Cai et al. 2024), Llama2-7B (Touvron et al. 2023), and Vicuna-7B (Chiang et al. 2023). Without loss of generosity, we use the model names to refer to the chat versions. We list the model and prompt details in the Appendix.

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 and detailed results in Tables 7 and 9 in the Appendix. For CEC, GPT4 performs the best. Among all open-source LLMs, Qwen2 is the best while Internlm2 holds the second. Early models such as Baichuan2-7B, Chatglm2-6B, Llama2-7B, and Vicuna-7B fall in this task where they score under 40%. For CRR, GPT4 excels all other models as well. Among all open-source LLMs, Internlm2-7B and Mistral-7B performs in the first tie.

We show the average performance of instance-level CEC and CRR in columns I-CEC and I-CRR in Table 3. Overall, existing LLMs such as GPT4, and Qwen2-7B have CEC event reasoning abilities. However, even the strongest GPT4 can only achieve 68.43 (4-way multiple choice) and 63.81 (3-way multiple choice) 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 imbalanced. Comparing CEC to CRR, as relation-wise results shown in Figure 2 and average performances in columns I-CEC and I-CRR in Table 3, LLMs perform better for CEC than CRR (note that CEC is a 4-way multiple choice task while CRR is of 3-way). To compare, we compute the average scores of I-CEC and I-CRR on models achieving above 40% and 30%[5]. We find I-CEC is much higher than I-CRR, with average scores 62.84 and 53.58. The results significantly suggest that CRR is harder and insifeciently solved than CEC. Existing pretraining and SFT datasets may be biased in paradigms.

We then analyze performances on different relations. We compute the average scores of relations on models achieving above 40% and 30% on average I-CEC and I-CRR. The I-CEC average scores of Temporal, Causal, Hierarchical are 50.11, 68.71, and 61.22 while in I-CRR the scores are 43.31, 58.36, and 52.17. With these results alongside scores shown in Figure 2, LLMs perform best in Causal relation. Then, Temporal relation is the worst. It indicates current training can enable LLMs to reason causality. It also trains the event hierarchy comprehension. However, temporal reasoning is the hardest. More methods should be established to handle this problem. 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, Wu, and Roth 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 paradigms. By EV^2, with disentangling relations and formulations, we investigate event reasoning with less bias.

CEC improves faster than CRR with model development. We investigate the improvement trends of CEC and CRR. The Figure 3 in the Appendix, when models have poor event reasoning abilities, their CEC performances lie around the balanced line showing no significant differences in tasks. With the development, the CEC improves much faster than CRR for all models. This investigation further appeals to the need for training in balanced event reasoning abilities.

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

[5]Models under these scores may lack statistic significance.

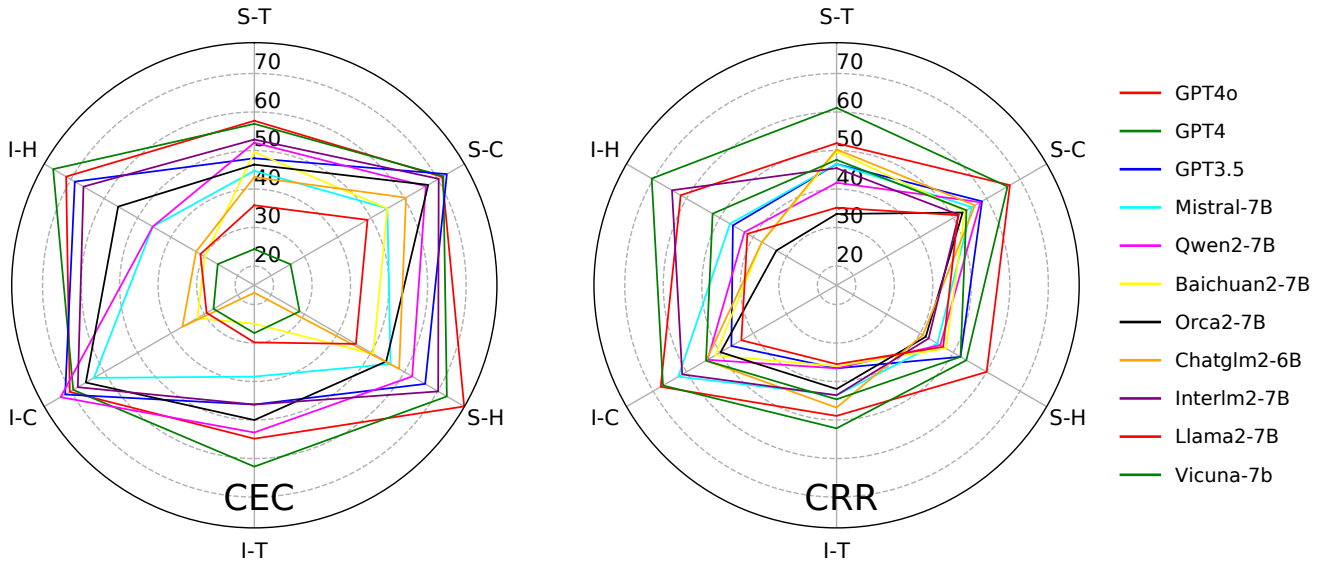

Figure 2: Results of CEC and CRR. S and I stand for schema- and instance-level. Relation types of Causality, Temporality, and Hierarchy are denoted as C, T, and H.

Model	CEC						CRR					
	Temporal	Causal	Hierarchical	W.T.S	W.O.S	Δ	Temporal	Causal	Hierarchical	W.T.S	W.O.S	Δ
GPT4o	58.06	73.45	85.71	71.49	66.60	4.89↑	56.45	73.04	75.65	69.25	62.04	7.21↑
GPT4	61.29	75.86	77.92	72.51	68.43	4.08↑	53.23	70.74	68.70	65.99	63.81	2.18↑
GPT3.5	52.42	77.59	77.92	71.28	64.77	6.51↑	44.09	59.22	47.83	53.61	43.95	9.66↑
Mistral-7B	45.16	70.34	74.03	64.56	54.18	10.38↑	48.92	63.36	50.43	57.69	55.24	2.45↑
Qwen2-7B	61.29	78.97	77.92	74.34	63.75	10.59↑	47.31	66.13	55.65	59.73	47.62	12.11↑
Baichuan2-7B	41.13	56.90	57.14	52.95	29.94	23.01↑	46.24	58.29	40.87	52.52	45.31	7.21↑
Orca2-7B	53.23	77.59	76.62	71.28	60.08	11.20↑	44.09	65.90	51.30	58.10	45.17	12.93↑
Chatglm2-6B	38.71	65.86	54.55	57.23	30.96	26.27↑	47.31	53.00	50.43	51.16	49.66	1.50↑
Interlm2-7B	57.26	80.69	81.82	74.95	62.12	12.83↑	53.23	70.28	72.17	66.26	57.28	8.98↑
Llama2-7B	36.29	45.86	49.35	43.99	29.74	14.25↑	42.47	51.61	53.91	49.66	41.22	8.44↑
Vicuna-7b	27.42	28.62	27.27	28.11	27.09	1.02↑	45.16	54.15	52.17	51.56	51.43	0.13↑

Table 5: Guidance with schema knowledge on instance level CEC and CRR tasks. W.T.S and W.O.S stands for average performances with and without event knowledge guidance. Δ is the difference between them.

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.

LLMs have event schema knowledge. We evaluate CEC and CRR on schema level. The results are shown in Figure 2 and detailed results in Tables 8 and 10 in the Appendix, and the average scores are reported in Table 3. 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 3, we observe that S-CEC lags I-CEC, suggesting advanced 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 and CRR performance comparisons of LLMs on instance- and schema-level in Figure 4 in the Appendix. 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 on par with schema-level. It indicates that the accumulation of event schema 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.

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 gen-

erated 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. The models need to select the correct event types for all event instances in the question and determine relations between each type. To keep the task difficult, we first conduct KMeans clustering on all event types in our dataset[6]. 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. Models need to find out the correct event types from given options. We calculate the accuracy of event types and F1-scores of relation triplets 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 the Appendix.

The results are in Table 4. We find only GPT4 and GPT4o can generate correct event types. GPT4 excels may be attributed to 1) its better alignment. 2) The dataset is prompted by GPT4. Therefore in this part, we mainly focus on other models rather than GPT4 series. However, we find rest models all fail to generate corresponding schema knowledge. For relation triplet generation, even GPT4 can not output proper event schemas. 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.

Can LLMs perform better on 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? Here we probe this question.

We conduct experiments in which we directly add the schema event of each instance event in the question into the prompt. We also add relations of these schema events.

We demonstrate the performances of this guidance in Table 5. Incorporating event schema knowledge significantly improves event reasoning. It shows great potential to solve event reasoning with the fusion of event schema knowledge. These results provide important insights that event schema knowledge could be used as the memory of LLMs to improve solving practical and domain-specific problems. Constructing and retrieving proper event schema knowledge is another challenge. We leave them to future works.

Related Work

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

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, Wang, and Huang (2019) establish a dataset oriented towards capturing event logic, enabling the generative prediction of future incidents.

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 2023; Zhong 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 2023; Li et al. 2024). Constructing benchmarks in the LLM-human interactive way is a regular strategy for complicated tasks (Huang et al. 2024).

Related to event reasoning, Yuan, Xie, and Ananiadou (2023) evaluated the event relation extraction. Tao et al. (2023a) present the Event Semantic Processing including the event understanding, reasoning, and prediction. 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.

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 a comprehension of event schema knowledge. 3) LLMs are not aligned with human to leaverage event schema knowledge in event reasoning. Based on the findings, we guide the LLMs to utilize event schema knowledge. With our guidance, LLMs can perform better on event reasoning.

Acknowledgements

This work is supported by National Natural Science Foundation of China under Grant No. 62436006. We thank all annotators for contributing to this work.

References

Arora, D.; Singh, H. G.; et al. 2023. Have LLMs Advanced Enough? A Challenging Problem Solving Benchmark For Large Language Models. *arXiv preprint arXiv:2305.15074*.

Bang, Y.; Cahyawijaya, S.; Lee, N.; Dai, W.; Su, D.; Wilie, B.; Lovenia, H.; Ji, Z.; Yu, T.; Chung, W.; et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Bian, N.; Han, X.; Sun, L.; Lin, H.; Lu, Y.; and He, B. 2023. Chatgpt is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models. *arXiv preprint arXiv:2303.16421*.

Bondarenko, A. e. a. 2022. CausalQA: A Benchmark for Causal Question Answering. In *Proceedings of the 29th International Conference on Computational Linguistics*, 3296–3308. Gyeongju, Republic of Korea: International Committee on Computational Linguistics.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Cai, Z.; Cao, M.; Chen, H.; Chen, K.; Chen, K.; Chen, X.; Chen, X.; Chen, Z.; Chen, Z.; Chu, P.; et al. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.

Caselli, T.; and Vossen, P. 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*, 77–86.

Cassidy, T.; McDowell, B.; Chambers, N.; and Bethard, S. 2014. An annotation framework for dense event ordering. In *Proceedings of the 52nd ACL*, 501–506.

Chan, C.; Cheng, J.; Wang, W.; Jiang, Y.; Fang, T.; Liu, X.; and Song, Y. 2023. Chatgpt evaluation on sentence level relations: A focus on temporal, causal, and discourse relations. *arXiv preprint arXiv:2304.14827*.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; Stoica, I.; and Xing, E. P. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

Du, L.; Ding, X.; Xiong, K.; Liu, T.; and Qin, B. 2022. e-CARE: a new dataset for exploring explainable causal reasoning. *arXiv preprint arXiv:2205.05849*.

Du, X. e. a. 2022. RESIN-11: Schema-guided Event Prediction for 11 Newsworthy Scenarios. In *Proceedings of the NAACL 2022*, 54–63. Hybrid: Seattle, Washington + Online: Association for Computational Linguistics.

Gao, J.; Zhao, H.; Yu, C.; and Xu, R. 2023. Exploring the feasibility of chatgpt for event extraction. *arXiv preprint arXiv:2303.03836*.

Glavaš, G.; Šnajder, J.; Kordjamshidi, P.; and Moens, M.-F. 2014. HiEve: A corpus for extracting event hierarchies from news stories. In *Proceedings of 9th language resources and evaluation conference*, 3678–3683. ELRA.

GLM, T. 2024. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. arXiv:2406.12793.

Granroth-Wilding, M. 2016. What Happens Next? Event Prediction Using a Compositional Neural Network Model. In *AAAI Conference on Artificial Intelligence*.

Guan, J.; Wang, Y.; and Huang, M. 2019. Story ending generation with incremental encoding and commonsense knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 6473–6480.

Han, R.; Hsu, I.-H.; Sun, J.; Baylon, J.; Ning, Q.; Roth, D.; and Peng, N. 2021. ESTER: A machine reading comprehension dataset for reasoning about event semantic relations. In *Proceedings of the EMNLP 2021.*, 7543–7559.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Hidey, C. 2016. Identifying Causal Relations Using Parallel Wikipedia Articles. In *Proceedings of the 54th ACL*, 1424–1433. Berlin, Germany: Association for Computational Linguistics.

Huang, S.; Zhong, W.; Lu, J.; Zhu, Q.; Gao, J.; Liu, W.; Hou, Y.; Zeng, X.; Wang, Y.; Shang, L.; Jiang, X.; Xu, R.; and Liu, Q. 2024. Planning, Creation, Usage: Benchmarking LLMs for Comprehensive Tool Utilization in Real-World Complex Scenarios. arXiv:2401.17167.

Jiang, A. Q. e. a. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Li, M.; Li, S.; Wang, Z.; Huang, L.; Cho, K.; Ji, H.; Han, J.; and Voss, C. 2021a. The Future is not One-dimensional: Complex Event Schema Induction by Graph Modeling for Event Prediction. In *Proceedings of the EMNLP 2021.*, 5203–5215.

Li, Z.; Li, Q.; Zou, X.; and Ren, J. 2021b. Causality extraction based on self-attentive BiLSTM-CRF with transferred embeddings. *Neurocomputing*, 423: 207–219.

Li, Z.; Xu, X.; Shen, T.; Xu, C.; Gu, J.-C.; and Tao, C. 2024. Leveraging large language models for nlg evaluation: A survey. *arXiv preprint arXiv:2401.07103*.

Liu, X.; Yu, H.; Zhang, H.; Xu, Y.; Lei, X.; Lai, H.; Gu, Y.; Ding, H.; Men, K.; Yang, K.; et al. 2023. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*.

Mao, Q.; Li, X.; Peng, H.; Li, J.; He, D.; Guo, S.; He, M.; and Wang, L. 2021. Event prediction based on evolutionary event ontology knowledge. *Future Generation Computer Systems*, 115: 76–89.

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

Mitra, A.; Del Corro, L.; Mahajan, S.; Codas, A.; Simoes, C.; Agarwal, S.; Chen, X.; Razdaibiedina, A.; Jones, E.; Aggarwal, K.; et al. 2023. Orca 2: Teaching small language models how to reason. *arXiv preprint arXiv:2311.11045*.

Mostafazadeh, N.; Chambers, N.; He, X.; Parikh, D.; Batra, D.; Vanderwende, L.; Kohli, P.; and Allen, J. 2016. A corpus

and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the NAACL 2016.*, 839–849.

Ning, Q.; Wu, H.; and Roth, D. 2018. A Multi-Axis Annotation Scheme for Event Temporal Relations. In *Proceedings of the 56th ACL*, 1318–1328.

O'Gorman, T.; Wright-Bettner, K.; and Palmer, M. 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)*, 47–56. Austin, Texas: Association for Computational Linguistics.

Poria, S.; Majumder, N.; Hazarika, D.; Ghosal, D.; Bhardwaj, R.; Jian, S. Y. B.; Hong, P.; Ghosh, R.; Roy, A.; Chhaya, N.; et al. 2021. Recognizing emotion cause in conversations. *Cognitive Computation*, 13: 1317–1332.

Qin, L.; Bosselut, A.; Holtzman, A.; Bhagavatula, C.; Clark, E.; and Choi, Y. 2019. Counterfactual story reasoning and generation. *arXiv preprint arXiv:1909.04076*.

Qin, L.; Shwartz, V.; West, P.; Bhagavatula, C.; Hwang, J.; Bras, R. L.; Bosselut, A.; and Choi, Y. 2020. Back to the future: Unsupervised backprop-based decoding for counterfactual and abductive commonsense reasoning. *arXiv preprint arXiv:2010.05906*.

Roemmele, M.; Bejan, C. A.; and Gordon, A. S. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Sap, M.; Le Bras, R.; Allaway, E.; Bhagavatula, C.; Lourie, N.; Rashkin, H.; Roof, B.; Smith, N. A.; and Choi, Y. 2019a. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 3027–3035.

Sap, M.; Rashkin, H.; Chen, D.; LeBras, R.; and Choi, Y. 2019b. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*.

Souza Costa, T.; Gottschalk, S.; and Demidova, E. 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*, 3157–3164.

Tao, Z.; Jin, Z.; Bai, X.; Zhao, H.; Feng, Y.; Li, J.; and Hu, W. 2023a. EvEval: A Comprehensive Evaluation of Event Semantics for Large Language Models. *arXiv preprint arXiv:2305.15268*.

Tao, Z.; Jin, Z.; Zhao, H.; Dou, C.; Zhao, Y.; Shen, T.; and Tao, C. 2023b. UniEvent: Unified Generative Model with Multi-Dimensional Prefix for Zero-Shot Event-Relational Reasoning. In *Proceedings of the 61st ACL*, 7088–7102.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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

Wang, X.; Chen, Y.; Ding, N.; Peng, H.; Wang, Z.; Lin, Y.; Han, X.; Hou, L.; Li, J.; Liu, Z.; et al. 2022a. Maven-ere: A unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction. *arXiv preprint arXiv:2211.07342*.

Wang, Y.; Cao, C.; Chen, Z.; and Wang, S. 2022b. ECCKG: An Eventuality-Centric Commonsense Knowledge Graph. In *International Conference on Knowledge Science, Engineering and Management*, 568–584. Springer.

Wei, X. e. a. 2023. Zero-shot information extraction via chatting with chatgpt. *arXiv preprint arXiv:2302.10205*.

Xu, X.; Tao, C.; Shen, T.; Xu, C.; Xu, H.; Long, G.; and Lou, J.-g. 2023. Re-reading improves reasoning in language models. *arXiv preprint arXiv:2309.06275*.

Yang, A.; Xiao, B.; Wang, B.; Zhang, B.; Bian, C.; Yin, C.; Lv, C.; Pan, D.; Wang, D.; Yan, D.; et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Yang, A.; Yang, B.; Hui, B.; Zheng, B.; Yu, B.; Zhou, C.; Li, C.; Li, C.; Liu, D.; Huang, F.; et al. 2024. Qwen2 Technical Report. *arXiv preprint arXiv:2407.10671*.

Yang, C.; Li, W.; Zhang, X.; Zhang, R.; and Qi, G. 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*, 13–20. Springer.

Yang, L.; Wang, Z.; Wu, Y.; Yang, J.; and Zhang, Y. 2022. Towards Fine-grained Causal Reasoning and QA. *arXiv preprint arXiv:2204.07408*.

Yuan, C.; Xie, Q.; and Ananiadou, S. 2023. Zero-shot Temporal Relation Extraction with ChatGPT. *arXiv preprint arXiv:2304.05454*.

Zhang, H.; Liu, X.; Pan, H.; Song, Y.; and Leung, C. W.-K. 2020. ASER: A large-scale eventuality knowledge graph. In *Proceedings of the web conference 2020*, 201–211.

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

Zheng, L. e. a. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685*.

Zhong, W. e. a. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*.

Zhou, B.; Khashabi, D.; Ning, Q.; and Roth, D. 2019. " Going on a vacation" takes longer than" Going for a walk": A Study of Temporal Commonsense Understanding. *arXiv preprint arXiv:1909.03065*.

Zhou, B. e. a. 2020. Temporal reasoning on implicit events from distant supervision. *arXiv preprint arXiv:2010.12753*.