EventRelBench: A Comprehensive Benchmark for Evaluating Event Relation Understanding in Large Language Models

Jie Gong[1]*, **Biaoshuai Zheng**[2], **Qiwang Hu**[3]
[1]School of Artificial Intelligence, Wuhan University
[2]School of Computer Science, South China Normal University
[3]School of Computer Engineering and Science, Shanghai University

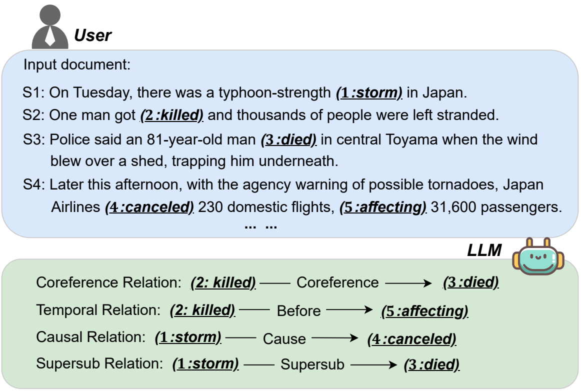
gongjie1130@gmail.com, 2021023327@m.scnu.edu.cn, hqw17@shu.edu.cn

Abstract

Understanding event relationships is critical for tasks such as narrative comprehension, information extraction, and reasoning in natural language processing. Despite the remarkable advancements of large language models (LLMs) across diverse NLP tasks, current studies have not systematically evaluated their ability to capture the complexities of event relations. To this end, we aim to assess LLMs on event relationship extraction (ERE) by designing the benchmark **EventRelBench**. **EventRelBench** comprises 35K diverse event relation questions covering four key categories—coreference, temporal, causal, and supersub relations. These questions are provided at two levels of granularity: document-level and sentence-level. Extensive experiments on different sizes and types of LLMs show that existing LLMs still fall short in accurately extracting and understanding event relationships. To address this gap, we introduce **EventRelInst**, a 48K instruction fine-tuning dataset in the event relation extraction domain. Experimental results not only highlight the shortcomings of current general-purpose LLMs in extracting event relationships but also demonstrate the effectiveness of **EventRelInst**. [1]

1 Introduction

Event relations are crucial for numerous downstream NLP applications, including narrative comprehension, information extraction, question answering, and commonsense reasoning (Chaturvedi et al., 2017; Zhou et al., 2019; Ning et al., 2020; Li et al., 2021; Zhang et al., 2021; Han et al., 2021; Zhuang et al., 2023). Event relation extraction (ERE) facilitates the construction of structured event graphs and the inference of implicit dependencies among events. While early work on event relation extraction relied on feature-engineered classifiers and structured prediction frameworks (Han et al., 2019; Wen and Ji, 2021; Hwang et al., 2022; Tan et al., 2023; Gong and Hu, 2025), the advent of large language models (LLMs) has opened new opportunities for capturing rich contextual cues and implicit knowledge. Recent LLMs (Ouyang et al., 2022; Touvron et al., 2023; Achiam et al., 2023; Zhao et al., 2023; Jiang et al., 2023; Fang et al., 2023; Grattafiori et al., 2024; Team et al., 2025) have demonstrated impressive zero-shot and few-shot performance across a wide spectrum of benchmarks. These benchmarks range from reasoning-oriented tasks such as GSM8K (Cobbe et al., 2021) to knowledge-intensive challenges like factual question answering (Pan et al., 2023; Hu et al., 2024).



Input document:
S1: On Tuesday, there was a typhoon-strength ***(1:storm)*** in Japan.
S2: One man got ***(2:killed)*** and thousands of people were left stranded.
S3: Police said an 81-year-old man ***(3:died)*** in central Toyama when the wind blew over a shed, trapping him underneath.
S4: Later this afternoon, with the agency warning of possible tornadoes, Japan Airlines ***(4:canceled)*** 230 domestic flights, ***(5:affecting)*** 31,600 passengers.
... ...

Coreference Relation: ***(2: killed)*** —— Coreference ⟶ ***(3:died)***
Temporal Relation: ***(2: killed)*** —— Before ⟶ ***(5:affecting)***
Causal Relation: ***(1:storm)*** —— Cause ⟶ ***(4:canceled)***
Supersub Relation: ***(1:storm)*** —— Supersub ⟶ ***(3:died)***

Figure 1: An Example of LLM predicting four event relationships (**Coreference**, **Temporal**, **Causal**, and **Supersub**) in the given document.

However, systematic evaluation of LLMs' capacity to understand and extract complex event relationships remains lacking. Existing LLM benchmarks typically focus on individual relation types in isolation (e.g., temporal relation) or operate at a single granularity (e.g., sentence-level). This leaves open questions about LLMs' capacity to comprehend event relations across diverse categories and text scopes. Without a unified evaluation benchmark, it is difficult to pinpoint specific weaknesses

[1] *Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 9084–9099
November 4-9, 2025 ©2025 Association for Computational Linguistics

of general-purpose LLMs or to develop targeted solutions.

To bridge this gap, we introduce **EventRelBench**, a comprehensive benchmark designed to evaluate LLMs on four crucial event relation categories—temporal, causal, coreference, and supersub. EventRelBench consists of 35k multiple-choice questions drawn from diverse domains, ensuring rich variety in linguistic constructions and knowledge demands. Two granularity levels are included: document-level and sentence-level.

As shown in Figure 1, LLM predicts four event relationships in the given document. Event(2: killed) has a coreference event relationship with event(3: died). Event(2: killed) and event(5: affecting) have a temporal relationship of *"before"*. Event(1: storm) has a causal relationship of *"cause"* with event(4: canceled). The subevent of event(1: storm) is event(3: died).

Through extensive experiments on **EventRelBench** with several open-source and leading limited-access LLMs, we have observed several key findings: **1)** Existing LLMs exhibit poor performance on event relation extraction tasks, with some models (e.g., AquilaChat-7B and Gemma-7B-IT (Team et al., 2024)) even underperforming random guessing, highlighting substantial potential for improvement. **2)** LLMs perform better on document-level tasks than on sentence-level tasks. This contrasts with the typical behavior observed in conventional models. We believe this may be attributed to the ability of LLMs to better capture and understand contextual information over long distances. **3)** Performance varies markedly across LLMs for different event relation extraction tasks, reflecting each model's distinct strengths and limitations.

Motivated by observed shortcomings, we further introduce **EventRelInst**, a 48K example instruction fine-tuning dataset tailored to the event relation extraction task. **EventRelInst** converts domain-specific data into explicit instruction–response pairs. This format enables models to better grasp the distinct demands of each relation type and the corresponding contextual granularity. Our experiments show that fine-tuning with **EventRelInst** can lead to significant performance gains compared to benchmark models. This outcome validates the effectiveness and potential of **EventRelInst**.

In summary, our main contributions are as follows:

- We introduce **EventRelBench**, an English large-scale, multi-category benchmark for event-relation extraction, covering temporal, causal, coreference, and supersub relations at both the sentence and document levels.
- We present **EventRelInst**, an instruction fine-tuning dataset of 48K examples specifically designed for the event-relation extraction task. It substantially enhances LLMs' ability to understand and distinguish complex event relationships.
- Extensive experimental results reveal that (a) general-purpose LLMs exhibit clear limitations on **EventRelBench**; and (b) instruction fine-tuning with **EventRelInst** produces substantial performance improvements in open-source LLMs.

2 Related Work

2.1 Event Relation Extraction

Event relation extraction constitutes a fundamental information extraction task that underpins various downstream applications (Zhang et al., 2020). Numerous studies have been conducted on event relation extraction tasks, covering a variety of event relation types, including coreference relations (Lu et al., 2022; Ahmed et al., 2024), temporal relations (Wang et al., 2020; Zhou et al., 2021; Yuan et al., 2023; Tan et al., 2024), causal relations (Chen et al., 2022, 2023; Man et al., 2024), and supersub relations (Wang et al., 2021; Wu et al., 2024).

Recent work has begun to explore how to leverage the capabilities of large language models for event relation extraction tasks (Wang et al., 2022b; Gao et al., 2023; Huang et al., 2023a; Ma et al., 2023; Qiu et al., 2023; Chen et al., 2024a; Yuan et al., 2024; Hu et al., 2025). These pioneering efforts have employed a range of strategies to adapt LLMs for the event relation extraction task.

2.2 Benchmarks for Large Language Models

The advent of LLMs has highlighted the critical need for systematic benchmarking frameworks to enable standardized capability evaluation. Existing benchmarks can be broadly categorized into two types. **1)** General-knowledge and reasoning benchmarks, such as the MMLU (Hendrycks et al., 2021a), employ multiple-choice questions drawn from real-world examinations and academic literature, covering a wide array of subject areas. **2)**

Task	Sources	Nums
Coreference Relation	EventStoryLine, ECB+	8,617
Temporal Relation	MATRES, TCR, Causal-TimeBank	8,953
Causal Relation	MAVEN-ERE, EventStoryLine	8,208
Supersub Relation	HiEve, MAVEN-ERE	8,960

Table 1: Statistics of EventRelBenchmark

There are specialized benchmarks designed for multilingual evaluation, including those targeting non-English languages (Huang et al., 2023b) as well as bilingual settings (Zhong et al., 2024). HELM (Liang et al., 2022) utilizes seven metrics across forty-two tasks to evaluate LLMs, examining dimensions ranging from accuracy to robustness. BIG-bench (Srivastava et al., 2022) evaluates large language models across 204 diverse tasks spanning domains such as linguistics, software engineering, and beyond. CELLO (Chen et al., 2024b) evaluates large vision-language models' (LVLMs) causal reasoning capabilities through 12 structured tasks grounded in visual scenes. MMLongBench-Doc (Ma et al., 2024) evaluates LVLMs' long-context document understanding through 1,062 expert-annotated questions requiring evidence from five modalities (text, image, chart, table, layout). In the domain of program synthesis, MBPP (Austin et al., 2021) and HumanEva (Chen et al., 2021) assess functional correctness by generating programs from natural language docstrings. Our benchmark mainly focuses on evaluating the capacity of LLMs to understand and extract complex event relationships.

3 Benchmark Construction

3.1 Tasks

To systematically evaluate the ability of LLMs to understand and extract event relations, we focus on four types of event relations and carefully select raw data from diverse sources, as summarized in Table 1.

Event Coreference Relation occurs when multiple event mentions within a text denote the same event instance. In essence, if different phrases or sentences describe the same underlying occurrence, those mentions are considered coreferential and grouped into an event cluster. We selected event coreference relations from the ECB+ (Cybulska and Vossen, 2014) and EventStoryLine (Caselli and Vossen, 2017) datasets. ECB+ comprises 982 news articles across 43 event topics. Human annotators identified event mentions within each article and labeled their actions, temporal expressions, locations, and participant roles. They then linked these annotated mentions into both intra-document and cross-document coreference chains. On average, each event topic has 11 article instances. EventStoryLine annotators identify coreferential chains of event mentions both within and across documents according to the ECB+ guidelines. Two event mentions are considered coreferential if they describe the same action component and share participants, temporal anchors, and locations. Detailed statistics of event coreference relationships in **EventRelBench** are shown in Appendix A.

Event Temporal Relation refers to the temporal ordering of events based on their occurrence in time. In this paper, we consider four key temporal relations: *BEFORE:* if event A happened before event B. *AFTER:* if event A happened after event B. *EQUAL:* if event A and event B happen at the same time. *VAGUE:* if event A and event B cannot determine the temporal ordering. We selected event temporal relations from the MATRES (Ning et al., 2018b), TCR (Ning et al., 2018a), and Causal-TimeBank (Mirza et al., 2014) datasets. MATRES comprises annotations over 36 TimeBank-Dense documents, collected via a two-step annotation pipeline, and encompasses approximately 1.8K start-point temporal relations. This dataset thus serves as a high-quality, reproducible resource for temporal relation extraction research. TCR employs the same annotation scheme as MATRES but is limited by a much smaller dataset. Causal-TimeBank comprises 184 documents and 6,813 annotated events. It was created by annotating event relations within the TempEval-3 corpus. Detailed statistics of event temporal relationships in **EventRelBench** are shown in Appendix A.

Event Causal Relation describes scenarios in which one event (the cause) influences the occurrence of another event (the effect). These relations can be categorized into two types: *CAUSE:* the effect event inevitably follows from the cause event. *CAUSED_BY:* the effect event is understood as resulting from the preceding cause event. We selected event causal relations from the MAVEN-ERE (Wang et al., 2022a) and EventStoryLine (Caselli and Vossen, 2017) datasets. MAVEN-ERE is a uniformly annotated English event relation dataset that covers 4,480 Wikipedia articles and comprises over one million event relations. These relations were produced through a multi-stage, refined annotation process and organized using an innovative timeline-sorting scheme. EventStoryLine is a dataset centered on event annotation within the news corpus. It contains pairs of causally related events, making it well-suited for modeling event causal chains and extracting meaningful causal relationships between events. Detailed statistics of event causal relationships in **EventRelBench** are shown in Appendix A.

Event Supersub Relation refers to instances where one event (the subevent) constitutes a component or a smaller part of another event (the superevent). Understanding such relations is crucial for revealing the hierarchical structure of events within the given text. We selected event supersub relations from the HiEve (Glavaš et al., 2014) and MAVEN-ERE (Wang et al., 2022a). HiEve is a hierarchically structured event dataset for news texts. It comprises 100 articles, totaling 1,354 sentences and 33,273 annotated tokens. On average, each document contains approximately 32 event mentions. The overall inter-annotator agreement, measured by F-score, is 69%. MAVEN-ERE contains 15,841 supersub relationships, providing a large-scale, high-quality resource for event hierarchy analysis and multi-level narrative understanding. Detailed statistics of event supersub relationships in **EventRelBench** are shown in Appendix A.

3.2 Annotation and Quality Control

Multiple-choice questions provide a practical means of evaluating the complex capabilities of LLMs. Key benchmarks, such as the ARC (Clark et al., 2018), HellaSwag (Zellers et al., 2019), MMLU (Hendrycks et al., 2021a), and TruthfulQA (Lin et al., 2022), all employ multiple-choice formats. These benchmarks target distinct aspects of model proficiency. For instance, ARC presents challenging science questions; HellaSwag evaluates commonsense reasoning; MMLU assesses the breadth and depth of LLM's factual and conceptual knowledge; and TruthfulQA gauges a model's tendency to replicate human falsehoods. Furthermore, universal metrics for assessing overall output quality are currently lacking (Sai et al., 2022), and multiple-choice benchmarks help mitigate this limitation by providing clear, accuracy-based assessments (Hendrycks et al., 2021b). Moreover, prior work has demonstrated that LLMs exhibit reliable calibration in multiple-choice settings (Kadavath et al., 2022). However, there are also studies (Zheng et al., 2023) showing that there are systematic biases in answer letter preferences and order effects. Consequently, we also adopt multiple-choice questions as a straightforward yet effective proxy for evaluating LLMs' performance.

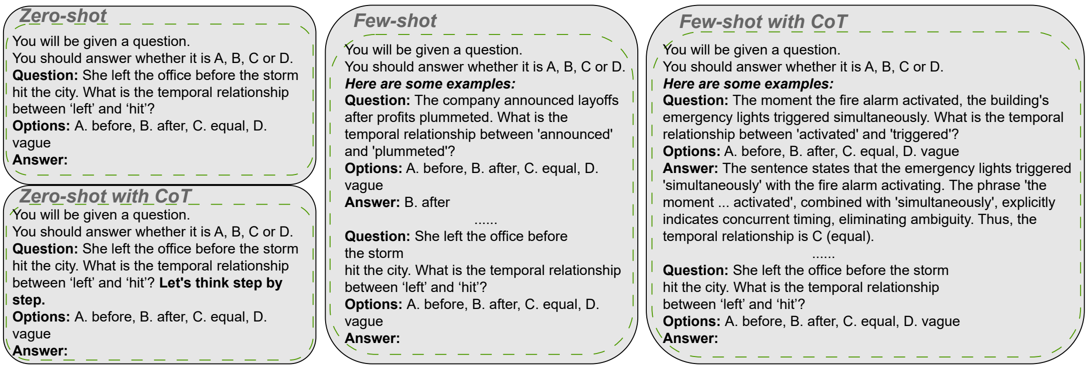
A total of eight annotators participated in the annotation process. We ask annotators to add questions after the original corpus without distorting the semantics and provide correct labels for the questions. Following the annotation of the correct answer for each question, we employed a systematic process to generate plausible yet incorrect options to complete the multiple-choice question format. This process was tailored to each relation type to ensure meaningful and challenging choices. The detailed annotation process is in the appendix B. By converting the original corpus into questions, the Question-Answering approach can more effectively evaluate the ability of LLMs to understand event relationships. Annotators were compensated based on both the quantity and quality of their annotations.

To ensure the quality of the annotated questions, we implemented a two-fold validation strategy. First, the two authors of this paper acted as meta-reviewers, randomly sampling 100 questions for each of the four event relation types in EventRelBench. They manually verified the correctness of the assigned labels, achieving an average label accuracy of 91.7% across the 800 sampled questions. Second, we assessed inter-annotator agreement (IAA) by dividing the eight annotators into two groups. Each group re-annotated a random set of 400 questions originally annotated by the other group. The resulting IAA was 82.8%. The IAA, measured by Cohen's Kappa, is 75.6%. These results collectively demonstrate that the labels in EventRelBench are of good quality.

4 Methodology

4.1 Models

To provide a comprehensive view of the status of LLMs' ability to understand event relationships, we evaluate 10 publicly accessible LLMs. These models have undergone the three distinct training phases of pretraining, instruction tuning, and reinforcement learning (Ouyang et al., 2022), covering multiple organizations and a diverse range of model sizes. The detailed description is provided in Ap-



Zero-shot

You will be given a question.
You should answer whether it is A, B, C or D.
Question: She left the office before the storm hit the city. What is the temporal relationship between 'left' and 'hit'?
Options: A. before, B. after, C. equal, D. vague
Answer:

Zero-shot with CoT

You will be given a question.
You should answer whether it is A, B, C or D.
Question: She left the office before the storm hit the city. What is the temporal relationship between 'left' and 'hit'? **Let's think step by step.**
Options: A. before, B. after, C. equal, D. vague
Answer:

Few-shot

You will be given a question.
You should answer whether it is A, B, C or D.
Here are some examples:
Question: The company announced layoffs after profits plummeted. What is the temporal relationship between 'announced' and 'plummeted'?
Options: A. before, B. after, C. equal, D. vague
Answer: B. after
......
Question: She left the office before the storm
hit the city. What is the temporal relationship between 'left' and 'hit'?
Options: A. before, B. after, C. equal, D. vague
Answer:

Few-shot with CoT

You will be given a question.
You should answer whether it is A, B, C or D.
Here are some examples:
Question: The moment the fire alarm activated, the building's emergency lights triggered simultaneously. What is the temporal relationship between 'activated' and 'triggered'?
Options: A. before, B. after, C. equal, D. vague
Answer: The sentence states that the emergency lights triggered 'simultaneously' with the fire alarm activating. The phrase 'the moment ... activated', combined with 'simultaneously', explicitly indicates concurrent timing, eliminating ambiguity. Thus, the temporal relationship is C (equal).
......
Question: She left the office before the storm
hit the city. What is the temporal relationship
between 'left' and 'hit'?
Options: A. before, B. after, C. equal, D. vague
Answer:

Figure 2: Illustration of prompts using four different settings. (Zero-shot, Zero-shot with CoT, Few-shot, Few-shot with CoT)

pendix C.

4.2 Prompt Strategy

As illustrated in Figure 2, we employ four types of prompts to elicit event relation judgments from LLMs: zero-shot, zero-shot with chain-of-thought (CoT) (Kojima et al., 2022), few-shot, and few-shot with CoT (Wei et al., 2022). First, we supply each model with a task instruction defining the classification schema, denoted as $\mathcal{M}$: "You will be given a question. You should answer whether it is A (before), B (after), C (equal), or D (vague)." This instruction establishes both the expected input format and the outputs. For any input $\mathcal{X}$, we then obtain a label $\mathcal{Y}$ from the LLMs $\mathcal{F}$: $\mathcal{Y} = \mathcal{F}\ (\mathcal{X}, \mathcal{M})$.

Zero-Shot Prompt In the zero-shot setting, the model generates Y based solely on $\mathcal{X}$ and $\mathcal{M}$. For example, given $\mathcal{X}$: "She left the office before the storm hit the city. What is the temporal relationship between 'left' and 'hit'?", the model should directly generate the answer "A (before)".

Zero-Shot with CoT To encourage explicit reasoning, we append "Let's think step by step" (Kojima et al., 2022) to the prompt. This two-stage formulation prompts the model to decompose its thought process.

Few-Shot Prompt In the few-shot setting, we precede $\mathcal{X}$ with four examples question answer pairs. This can help the model understand the input and output formats.

Few-Shot with CoT Here, each example includes a brief reasoning scaffold. As depicted in Figure 2, for $\mathcal{X}$: "The company announced layoffs after profits plummeted. What is the temporal relationship between 'announced' and 'plummeted'?" Our reasoning approach entails first identifying the key noun phrases in the events, "the company" (subject of "announced") and "profits" (subject of "plummeted"), to clarify who or what is involved in each event. We then focus on the modifying elements, such as the temporal connector "after," which directly governs the relationship between the two verbs. In this sentence, "after" modifies the verb "announced," specifying that the action of announcing happened subsequent to the action of plummeting. Therefore, the temporal sequence is unambiguous: "plummeted" occurred first, followed by "announced". Therefore, the answer is B (after).

Dataset Construction When constructing the instruction fine-tuning dataset, we introduced new news sources while keeping the samples from EventRelBenchmark non-overlapping, and combined strict deduplication and manual annotation. The specific process is as follows: 1) Data source and preliminary screening: In addition to the multiple datasets used in building EventRelBenchmark, a small amount of public news corpus was collected to enrich the field and style. All sentences that are exactly the same as those in Benchmark are removed to ensure that the samples are not repeated. 2) Semantic Deduplication: For the remaining candidate samples, use Sentence-BERT (Reimers and Gurevych, 2019) to encode the sentences into vectors, and calculate the average cosine similarity with all Benchmark samples respectively; Only news and public data samples with a cosine similarity of < 0.7 with any Benchmark sample are retained to prevent high similarity overlap. 3) Dataset size: Finally, about 48,000 high-quality instruction samples were obtained, and the detailed statistical data are shown in Appendix D. The class distribution in EventRelInst reflects the true frequencies in our source corpus and annotation process. We mitigate potential imbalance effects through macro F_1 evaluation.

Each instruction example follows a specific format. First, it provides a text explaining the task requirements, such as "You will be asked a question. You should answer it as A, B, C, or D" for a four-choice task, or "You should answer it as A or B" for a two-choice task. Next, the example presents a context sentence containing two events to be analyzed. It's important to note that the specific form of the question varies depending on the type of relationship being examined. Correspondingly, the options provided are also a specific set of relationships: for example, temporal relationships correspond to four options: A, B, C, and D, while other relationships correspond to two options: A and B. Finally, each example is accompanied by the correct "gold standard" answer, which also follows the format of A–D for temporal relationships and A–B for other relationships. A specific example is as follows:

Instruction: You will be given a question. You should answer whether it is A, B, C or D.

Context: "Another cousin Georgina Cid said Elian father had intended to flee Cuba himself but was being coerced by the Castro government to stay and make certain statements. What is the temporal relationship between 'said' and 'intended'?"

Options: A. before B. after C. equal D. vague

Gold Label: B

Fine-tuning on EventRelInst We conducted LoRA (Low-Rank Adaptation) (Hu et al., 2022) fine-tuning of five mainstream open-source LLMs: Llama-3-8B and the Qwen-2.5 series (7B, 3B, 1.5B, and 0.5B) on EventRelInst instruction-tuning dataset in order to systematically assess its generality and effectiveness across model scales. Specifically, we injected rank-8 adapters with a scaling factor of 16 into the query and key projection matrices of each model's self-attention layers, initializing all LoRA weights with He normal sampling. All experiments were executed under a unified DeepSpeed Zero-2 configuration. Core hyperparameters were held constant across models: an effective batch size of 128 (via gradient accumulation), a linear warmup of 500 steps to a peak learning rate of 1×10^{-5} followed by cosine decay to 1×10^{-6}, weight decay of 0.1, $\mathrm{Adam}\,\beta_2 = 0.95$, gradient clipping with a maximum norm of 1.0, a maximum sequence length of 600 tokens, and bf16 precision. Experimental results show that using the EventRelInst dataset for instruction tuning can significantly improve the event relation extraction performance of model configurations of different sizes.

5 Experiments

In this section, we present and analyze the performance of various LLMs on EventRelBench, which evaluates models' capabilities across four prompting paradigms: Zero-shot, Zero-shot with CoT, Few-shot, and Few-shot with CoT. We also include results for fine-tuned variants (FT) of several models using EventRelInst based on LoRA.

5.1 Main Results

Table 2 shows the average performance of 10 publicly available LLMs on the EventRelBench under various configurations, each evaluated over three independent runs. From these results, we can draw the following conclusions:

Overall Performance As shown in Table 2, there is a clear performance gap between base models and their fine-tuned counterparts. The best-performing model is Qwen2.5-7B-FT, achieving an accuracy of 60.0% and F1 score of 54.5%, significantly outperforming its base version (Qwen2.5-7B: 53.1% / 44.1%). This highlights the effectiveness of EventRelInst and fine-tuning strategy in improving event relation understanding.

Following Qwen2.5-7B-FT, the top-performing general-purpose models without fine-tuning are ERNIE-3.5-8K (Acc: 52.9%, F1: 49.9%) and DeepSeek-V3 (Acc: 55.6%, F1: 53.2%), both demonstrating strong performance under zero-shot and few-shot paradigms. Notably, DeepSeek-V3 shows remarkable consistency across prompting styles, suggesting a robust general understanding of event relations.

In contrast, models such as AquilaChat-7B and Gemma-7B-IT consistently underperform, with overall accuracy around 44–46% and F1 scores below 35%, indicating limited temporal and relational reasoning capabilities without fine-tuning.

Effect of Prompting Strategies Prompting methods significantly influence model performance:

Zero-shot vs. CoT prompting: Most models show only marginal improvements or even slight declines in performance when Chain-of-Thought reasoning is introduced. For instance, GPT-3.5-Turbo's accuracy remains nearly unchanged (47.7%

Model	Zero-shot		Zero-shot w/ CoT		Few-shot		Few-shot w/ CoT		Overall	
	Acc.	F_1	Acc.	F_1	Acc.	F_1	Acc.	F_1	Acc.	F_1
Bloomz-7B	46.0	41.6	45.4	41.9	45.4	41.6	47.1	42.4	46.0	41.9
ChatGLM2-6B-32k	46.0	36.6	46.4	34.7	47.4	37.7	49.4	38.1	47.3	36.8
AquilaChat-7B	46.6	36.1	46.2	34.8	46.8	34.5	46.3	30.4	46.5	34.0
Gemma-7B-IT	41.8	33.6	43.7	35.6	45.6	32.7	47.4	31.9	44.6	33.5
ERNIE-3.5-8K	56.7	53.2	56.1	52.3	50.6	48.5	48.2	45.7	52.9	49.9
DeepSeek-V3	56.3	53.2	56.0	52.5	56.2	53.9	54.0	53.3	55.6	53.2
Claude-3-Haiku	48.5	40.3	48.0	38.5	54.3	51.0	50.6	42.1	50.4	43.0
GPT-3.5-Turbo	47.7	39.4	47.9	38.5	51.9	47.2	50.8	46.3	49.6	42.9
GPT-4o-mini	56.1	52.9	56.1	52.3	56.2	52.7	55.8	52.5	56.1	52.6
GPT-4o	57.4	54.3	57.5	54.1	57.6	**54.4**	**57.2**	**53.9**	57.4	54.2
LLaMA-3-8B	46.4	31.7	48.5	32.5	48.5	36.3	49.6	39.1	48.3	34.9
LLaMA-3-8B-FT	51.6	44.6	51.6	43.1	51.1	40.4	49.7	42.7	51.0	42.7
Qwen2.5-0.5B	45.9	38.6	45.8	38.8	46.1	38.9	45.7	38.5	45.9	38.7
Qwen2.5-0.5B-FT	46.6	39.3	46.2	38.8	46.4	38.7	46.8	38.2	46.5	38.8
Qwen2.5-1.5B	46.1	38.2	46.2	37.5	46.0	38.4	45.6	37.6	46.0	37.9
Qwen2.5-1.5B-FT	47.8	39.4	47.3	38.8	47.9	38.7	47.5	38.2	47.6	38.8
Qwen2.5-3B	48.1	38.7	48.1	38.5	47.5	37.9	47.8	37.6	47.9	38.2
Qwen2.5-3B-FT	50.1	43.4	50.3	43.5	49.9	42.6	50.0	42.4	50.1	43.0
Qwen2.5-7B	53.8	44.3	53.9	44.3	53.6	45.0	51.1	42.9	53.1	44.1
Qwen2.5-7B-FT	**63.7**	**59.0**	**63.3**	**56.1**	**58.3**	51.8	54.5	51.0	**60.0**	**54.5**

Table 2: Results obtained by employing various prompt formats across 10 publicly accessible LLMs. FT means fine-tuning using Lora on the EventRelInst instruction dataset.

→ 47.9%), while Claude-3-Haiku's F1 score drops from 40.3 to 38.5. This suggests that naive CoT prompting may not benefit event relation understanding unless carefully engineered.

Few-shot prompting generally offers modest gains in performance. GPT-3.5-Turbo improves from 47.7% (Zero-shot) to 51.9% (Few-shot), and Claude-3-Haiku jumps from 48.5% to 54.3% in accuracy. This demonstrates that few-shot exemplars help models better align with the task format and logic.

Few-shot with CoT only benefits a few models. While GPT-3.5-Turbo and DeepSeek-V3 maintain their performance, models like AquilaChat and Gemma-7B-IT experience performance degradation, likely due to reasoning noise introduced by suboptimal CoT generation.

Fine-tuning Gains Instruction fine-tuning with EventRelInst yields consistent and sometimes substantial improvements across all model scales and variants: Qwen2.5-7B-FT outperforms its base model by +6.9 accuracy and +10.4 F1, showing the greatest relative improvement across all models. Even smaller models such as Qwen2.5-3B-FT and Qwen2.5-1.5B-FT benefit from fine-tuning, with improvements of 2–4 points in both metrics. These results affirm that even lightweight models can significantly benefit from domain-specific instruction tuning for structured relation extraction tasks.

Model Size vs. Performance Interestingly, model size does not always correlate with better performance. For instance, LLaMA-3-8B performs worse (Acc: 48.3%) than smaller models like Qwen2.5-3B-FT (50.1%), and its fine-tuned version (LLaMA-3-8B-FT) still lags behind top models. This suggests that architecture, training data, and fine-tuning strategy play a more crucial role than sheer parameter count.

Summary of Findings Instruction tuning is essential for high performance on structured event relation benchmarks. Few-shot prompting benefits strong base models like GPT-3.5-Turbo and Claude-3-Haiku more than weaker ones. CoT prompting provides inconsistent improvements and may introduce noise if not carefully guided. DeepSeek-V3 and ERNIE-3.5 emerge as strong out-of-the-box performers. Qwen2.5-7B-FT achieves state-of-the-art results on our benchmark, validating the effectiveness of EventRelInst.

In Table 3, we present the performance of LLMs in four event relations under the Few-shot CoT setting. This reveals the relative difficulty LLMs have in understanding and extracting event relations, providing insights for future training of event relation knowledge in LLMs. From Table 3, it is observed that LLMs exhibit relatively poorer performance

Model	Causal		Coreference		Supersub		Temporal	
	Acc.	F_1	Acc.	F_1	Acc.	F_1	Acc.	F_1
Bloomz-7B	52.7	52.1	48.4	48.0	**58.2**	48.1	29.3	22.4
ChatGLM2-6B-32k	54.6	48.1	55.0	48.7	50.5	42.0	30.8	21.6
AquilaChat-7B	50.6	36.2	53.7	36.9	49.4	34.4	32.0	15.0
Gemma-7B-IT	41.0	37.5	66.0	41.6	50.3	33.7	32.3	15.6
ERNIE-3.5-8K	69.1	69.3	67.1	64.2	20.4	20.1	**38.6**	31.8
DeepSeek-V3	74.2	**74.3**	69.1	65.5	38.9	32.3	35.8	**32.3**
Claude-3-Haiku	61.3	38.5	70.2	66.9	40.3	37.8	32.2	26.0
GPT-3.5-Turbo	61.2	61.3	55.1	53.1	57.0	**50.3**	30.8	22.0
LLaMA-3-8B	51.0	35.5	58.2	52.4	55.9	47.3	33.6	21.2
LLaMA-3-8B-FT	52.8	39.8	62.8	59.3	50.3	34.2	33.6	27.9
Qwen2.5-0.5B	48.5	48.3	53.5	45.8	49.2	48.2	32.8	21.0
Qwen2.5-0.5B-FT	51.5	51.0	53.5	37.9	49.5	48.2	33.4	21.1
Qwen2.5-1.5B	46.2	43.6	58.0	55.0	50.4	35.8	30.4	19.5
Qwen2.5-1.5B-FT	48.8	47.8	61.2	55.9	50.5	35.0	31.3	20.0
Qwen2.5-3B	55.2	55.2	50.3	37.5	51.0	47.4	36.7	16.2
Qwen2.5-3B-FT	55.7	55.2	60.0	54.3	51.3	49.7	34.0	15.7
Qwen2.5-7B	71.1	70.6	61.3	53.4	48.3	34.3	32.7	26.6
Qwen2.5-7B-FT	**73.8**	73.6	**71.8**	**69.4**	46.9	40.9	35.3	29.2

Table 3: Results of different LLMs using Few-shot w/ CoT prompts across four event relations.

on the temporal relation, significantly lower than the other three event relationships. This is because LLMs lack the ability to understand long-term contexts. The timing of events often depends on clues across sentences or even paragraphs. Insufficient long-term dependency will miss key time cues. We analyze the LLMs in understanding four event relations in Sec. 5.2.

5.2 Analysis

In this section, we explore LLMs' understanding capabilities focusing on four event relations coreference, temporal, causal, and supersub relation. We also examine their performance on different prompt strategies.

Event Causal Relation In the causal relation task, models that undergo LoRA-based instruction fine-tuning demonstrate a clear edge, with Qwen2.5-7B-FT attaining over 73% accuracy and F1. This outperforms even strong generalist models like DeepSeek-V3 (74.2%/74.3%) by leveraging dedicated causal reasoning examples during tuning. Notably, smaller models such as Qwen2.5-0.5B-FT still gain 3–4 points post-fine-tuning, illustrating the broad applicability of the approach. Meanwhile, underperformers like Gemma-7B-IT highlight that without specialized tuning, certain architectures struggle with causal logic extraction.

Event Coreference Relation LLMs show robust capability in coreference resolution, with several untuned models, including Claude-3-Haiku (70.2%/66.9%), DeepSeek-V3 (69.1%/65.5%), and ERNIE-3.5-8K (67.1%/64.2%) already surpassing 65% accuracy. Fine-tuning via our instruction dataset elevates top models to nearly 72% accuracy and almost 70% F1, yielding consistent 8–10 point improvements in accuracy for architectures like LLaMA-3-8B and Qwen2.5-3B. This underscores the value of high-quality coreference annotations in closing recall gaps and handling long-tail event mentions. A few cases of slight F1 declines suggest that tuning data must maintain diversity to avoid overfitting to narrow coreference patterns.

Event Supersub Relation Semantic hierarchy extraction remains more challenging: Bloomz-7B (58.2%/48.1%) and GPT-3.5-Turbo (57.0%/50.3%) lead the pack out of the box, indicating strong baseline knowledge of category relations. However, instruction fine-tuning yields only modest improvements (1–3 accuracy points), even for Qwen2.5-7B-FT, which rises to 48.3% F1. This suggests that our current fine-tuning examples lack sufficient coverage of diverse supersub patterns and complexities. Future work should incorporate recursive and cross-domain hierarchy instances.

Event Temporal Relation Time ordering remains the hardest relation type, with most

F1 scores lingering below 30%—for example, Qwen2.5-7B-FT achieves only 29.2% F1 despite fine-tuning. Models exhibit minimal accuracy gains (2–3 points) after LoRA tuning, and some even see minor F1 drops, indicating that our temporal examples may not fully capture multi-hop or intersecting timelines. Stronger performance by ERNIE-3.5-8K (38.6% accuracy) and DeepSeek-V3 (35.8% accuracy) suggests that certain pretraining strategies help, but overall, explicit temporal markers and richer time-sequence reasoning tasks are needed. Enhancing prompt design with timestamp normalization, interval calculation, and narrative chain tasks should be priorities to bolster temporal inference capabilities.

6 Conclusion

In this work, we present **EventRelBench**, a comprehensive evaluation suite comprising approximately 35K questions spanning four types of event relations. This benchmark is designed to assess whether LLMs can effectively understand, extract, and reason over event relations. Our empirical analysis reveals that, despite employing various prompting strategies, existing LLMs still face significant challenges in achieving optimal performance on event relation extraction tasks. To address this limitation, we introduce **EventRelInst**, a 48K instruction fine-tuning dataset specifically tailored for event relation extraction. Experimental results demonstrate that **EventRelInst** greatly enhances the performance of general-purpose LLMs on tasks involving event relation understanding. We hope that this novel benchmark and instruction fine-tuning dataset will advance research in this domain and serve as a foundation for improving LLMs' capabilities in understanding event relation knowledge.

Limitations

Despite its contributions, our work has several limitations that warrant consideration. First, although EventRelBenchmark encompasses four fundamental event relation categories, it does not capture the full diversity and granularity of interactions. Second, because both the EventRelBenchmark and EventRelInst were constructed primarily from English text, their applicability to non-English languages remains untested. Third, despite rigorous annotation guidelines and quality checks, subtle ambiguities in defining event boundaries and relation strength may have introduced inconsistencies, potentially capping the performance ceiling even for optimally tuned models. Fourth, while instruction fine-tuning with EventRelInst demonstrably improves extraction accuracy, it may also induce overfitting to our specific prompt formulations, thereby hindering zero-shot adaptation to novel instruction styles. Fifth, although we benchmarked a representative spectrum of LLMs, we did not exhaustively explore the wider space of model scales, specialized architectures, or diverse inference setups, which may exhibit different strengths and weaknesses.

Ethical Considerations

When developing and deploying EventRelBench and EventRelInst, we must guard against annotation bias and ensure diverse, de-identified data; mitigate privacy and sensitive-content risks through strict PII filtering; address hallucinations by integrating fact-checking and uncertainty estimates; prevent dual-use misuse via access controls and licensing; and minimize environmental impact by reporting energy use, adopting efficient fine-tuning methods, and sharing compressed models, thereby balancing technical advancement with ethical responsibility.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Shafiuddin Rehan Ahmed, George Arthur Baker, Evi Judge, Michael Regan, Kristin Wright-Bettner, Martha Palmer, and James H Martin. 2024. Linear cross-document event coreference resolution with x-amr. *arXiv preprint arXiv:2404.08656*.

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 1:1.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. 2021. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*.

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

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

Snigdha Chaturvedi, Haoruo Peng, and Dan Roth. 2017. Story comprehension for predicting what happens next. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1603–1614, Copenhagen, Denmark. Association for Computational Linguistics.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Meiqi Chen, Yixin Cao, Kunquan Deng, Mukai Li, Kun Wang, Jing Shao, and Yan Zhang. 2022. ERGO: Event relational graph transformer for document-level event causality identification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2118–2128, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Meiqi Chen, Yixin Cao, Yan Zhang, and Zhiwei Liu. 2023. CHEER: Centrality-aware high-order event reasoning network for document-level event causality identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10804–10816, Toronto, Canada. Association for Computational Linguistics.

Meiqi Chen, Yubo Ma, Kaitao Song, Yixin Cao, Yan Zhang, and Dongsheng Li. 2024a. Improving large language models in event relation logical prediction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9451–9478, Bangkok, Thailand. Association for Computational Linguistics.

Meiqi Chen, Bo Peng, Yan Zhang, and Chaochao Lu. 2024b. Cello: Causal evaluation of large vision-language models. *arXiv preprint arXiv:2406.19131*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Agata Cybulska and Piek Vossen. 2014. Using a sledgehammer to crack a nut? lexical diversity and event coreference resolution. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4545–4552, Reykjavik, Iceland. European Language Resources Association (ELRA).

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2021. Glm: General language model pretraining with autoregressive blank infilling. *arXiv preprint arXiv:2103.10360*.

Yin Fang, Qiang Zhang, Ningyu Zhang, Zhuo Chen, Xiang Zhuang, Xin Shao, Xiaohui Fan, and Huajun Chen. 2023. Knowledge graph-enhanced molecular contrastive learning with functional prompt. *Nature Machine Intelligence*, 5(5):542–553.

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

Goran Glavaš, Jan Šnajder, Marie-Francine Moens, and Parisa Kordjamshidi. 2014. HiEve: A corpus for extracting event hierarchies from news stories. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3678–3683, Reykjavik, Iceland. European Language Resources Association (ELRA).

Jie Gong and Qiwang Hu. 2025. Extracting military event temporal relations via relative event time prediction and virtual adversarial training. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3305–3317, Albuquerque, New Mexico. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Rujun Han, I Hsu, Jiao Sun, Julia Baylon, Qiang Ning, Dan Roth, Nanyun Peng, et al. 2021. Ester: A machine reading comprehension dataset for event semantic relation reasoning. *arXiv preprint arXiv:2104.08350*.

Rujun Han, I-Hung Hsu, Mu Yang, Aram Galstyan, Ralph Weischedel, and Nanyun Peng. 2019. Deep structured neural network for event temporal relation extraction. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 666–106, Hong Kong, China. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Xuming Hu, Junzhe Chen, Xiaochuan Li, Yufei Guo, Lijie Wen, Philip S Yu, and Zhijiang Guo. 2024. Towards understanding factual knowledge of large language models. In *The Twelfth International Conference on Learning Representations*.

Zhilei Hu, Zixuan Li, Xiaolong Jin, Long Bai, Jiafeng Guo, and Xueqi Cheng. 2025. Large language model-based event relation extraction with rationales. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7484–7496, Abu Dhabi, UAE. Association for Computational Linguistics.

Feng Huang, Qiang Huang, YueTong Zhao, ZhiXiao Qi, BingKun Wang, YongFeng Huang, and SongBin Li. 2023a. A three-stage framework for event-event relation extraction with large language model. In *International Conference on Neural Information Processing*, pages 434–446. Springer.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023b. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Advances in Neural Information Processing Systems*.

EunJeong Hwang, Jay-Yoon Lee, Tianyi Yang, Dhruvesh Patel, Dongxu Zhang, and Andrew McCallum. 2022. Event-event relation extraction using probabilistic box embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 235–244, Dublin, Ireland. Association for Computational Linguistics.

Jinhao Jiang, Kun Zhou, Zican Dong, Keming Ye, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Structgpt: A general framework for large language model to reason over structured data. *arXiv preprint arXiv:2305.09645*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *ArXiv preprint*, abs/2205.11916.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Manling Li, Sha Li, Zhenhailong Wang, Lifu Huang, Kyunghyun Cho, Heng Ji, Jiawei Han, and Clare Voss. 2021. 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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Yaojie Lu, Hongyu Lin, Jialong Tang, Xianpei Han, and Le Sun. 2022. End-to-end neural event coreference resolution. *Artificial Intelligence*, 303:103632.

Yubo Ma, Yixin Cao, YongChing Hong, and Aixin Sun. 2023. Large language model is not a good few-shot information extractor, but a good reranker for hard samples!

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, et al. 2024. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *arXiv preprint arXiv:2407.01523*.

Hieu Man, Franck Dernoncourt, and Thien Huu Nguyen. 2024. Mastering context-to-label representation transformation for event causality identification with diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18760–18768.

Paramita Mirza, Rachele Sprugnoli, Sara Tonelli, and Manuela Speranza. 2014. Annotating causality in the TempEval-3 corpus. In *Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL)*, pages 10–19, Gothenburg, Sweden. Association for Computational Linguistics.

Qiang Ning, Zhili Feng, Hao Wu, and Dan Roth. 2018a. Joint reasoning for temporal and causal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2278–2288, Melbourne, Australia. Association for Computational Linguistics.

Qiang Ning, Hao Wu, Rujun Han, Nanyun Peng, Matt Gardner, and Dan Roth. 2020. TORQUE: A reading comprehension dataset of temporal ordering questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1158–1172, Online. Association for Computational Linguistics.

Qiang Ning, Hao Wu, and Dan Roth. 2018b. 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, Melbourne, Australia. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6981–7004, Toronto, Canada. Association for Computational Linguistics.

Yifu Qiu, Zheng Zhao, Yftah Ziser, Anna Korhonen, Edoardo M Ponti, and Shay B Cohen. 2023. Are large language models temporally grounded? *ArXiv preprint*, abs/2311.08398.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Ananya B Sai, Akash Kumar Mohankumar, and Mitesh M Khapra. 2022. A survey of evaluation metrics used for nlg systems. *ACM Computing Surveys (CSUR)*, 55(2):1–39.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, et al. 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2107.02137*.

Xingwei Tan, Gabriele Pergola, and Yulan He. 2023. Event temporal relation extraction with Bayesian translational model. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1125–1138, Dubrovnik, Croatia. Association for Computational Linguistics.

Xingwei Tan, Yuxiang Zhou, Gabriele Pergola, and Yulan He. 2024. Set-aligning framework for auto-regressive event temporal graph generation. *arXiv preprint arXiv:2404.01532*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *ArXiv preprint*, abs/2302.13971.

Haoyu Wang, Muhao Chen, Hongming Zhang, and Dan Roth. 2020. Joint constrained learning for event-event relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 696–706, Online. Association for Computational Linguistics.

Haoyu Wang, Hongming Zhang, Muhao Chen, and Dan Roth. 2021. Learning constraints and descriptive segmentation for subevent detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5216–5226, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiaozhi Wang, Yulin Chen, Ning Ding, Hao Peng, Zimu Wang, Yankai Lin, Xu Han, Lei Hou, Juanzi Li, Zhiyuan Liu, Peng Li, and Jie Zhou. 2022a. MAVEN-ERE: A unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 926–941, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xingyao Wang, Sha Li, and Heng Ji. 2022b. Code4struct: Code generation for few-shot structured prediction from natural language.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *ArXiv preprint*, abs/2201.11903.

Haoyang Wen and Heng Ji. 2021. Utilizing relative event time to enhance event-event temporal relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10431–10437, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ting-Ting Wu, Xiao Ding, Li Du, Bing Qin, and Ting Liu. 2024. Reasoning subevent relation over heterogeneous event graph. *Knowledge and Information Systems*, 66(9):5311–5333.

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

Chenhan Yuan, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. Back to the future: Towards explainable temporal reasoning with large language models. In *Proceedings of the ACM on Web Conference 2024*, pages 1963–1974.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

Hongming Zhang, Daniel Khashabi, Yangqiu Song, and Dan Roth. 2020. Transomcs: From linguistic graphs to commonsense knowledge. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 4004–4010. ijcai.org.

Zixuan Zhang, Hongwei Wang, Han Zhao, Hanghang Tong, and Heng Ji. 2021. EventKE: Event-enhanced knowledge graph embedding. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1389–1400, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023. Large language models are not robust multiple choice selectors. *arXiv preprint arXiv:2309.03882*.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2024. AGIEval: A human-centric benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2299–2314, Mexico City, Mexico. Association for Computational Linguistics.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. " going on a vacation" takes longer than" going for a walk": A study of temporal commonsense understanding. *arXiv preprint arXiv:1909.03065*.

Yichao Zhou, Yu Yan, Rujun Han, J. Harry Caufield, Kai-Wei Chang, Yizhou Sun, Peipei Ping, and Wei Wang. 2021. Clinical temporal relation extraction with probabilistic soft logic regularization and global inference. In *AAAI 2021, IAAI 2021, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14647–14655. AAAI Press.

Ling Zhuang, Hao Fei, and Po Hu. 2023. Syntax-based dynamic latent graph for event relation extraction. *Information Processing Management*, 60(5):103469.

A Statistics of EventRelBench

Options	Yes	No	All
Sentence	3000	3000	6000
Doc	1506	1111	2617
All	4506	4111	8617

Table 4: Detailed statistics of event coreference relationships in EventRelBench.

Options	Before	After	Equal	Vague	All
Sentence	2300	2300	475	1534	6609
Doc	660	670	1014	0	2344
All	2960	2970	1489	1534	8953

Table 5: Detailed statistics of event temporal relationships in EventRelBench.

Options	Cause	Caused by	All
Sentence	4000	4000	8000
Doc	128	80	208
All	4128	4080	8208

Table 6: Detailed statistics of event causal relationships in EventRelBench.

Options	Cause	Caused by	All
Sentence	4000	4000	8000
Doc	510	450	960
All	4510	4450	8960

Table 7: Detailed statistics of event supersub relationships in EventRelBench.

B Distractor Generation for MCQs

1. **Relation-Specific Option Sets**

    - Temporal relations: (*before, after, equal, vague*)
    - Causal relations: (*causes, caused by*)
    - Coreference relations: (*yes, no*)
    - Super-Sub relations: (*Super-Sub, Sub-Super*)

2. **Candidate Pool Construction**

    For each gold event pair (E_1, E_2) with label L, we build:

    - A **positive pool** of other pairs labeled L (to yield plausible same-relation distractors).
    - A **negative pool** of pairs labeled ¬L (to provide contrastive distractors).

3. **Distractor Selection per Relation**

    - **Temporal (4-way)**: select the top **three** candidates (mix of positive and negative) to accompany the correct label, yielding four options.
    - **Causal (2-way)**: binary choice; select the single most confusable negative pool candidate as the sole distractor.
    - **Coreference (2-way)**: binary choice; one positive-pool candidate when gold = “no,” or one negative-pool candidate when gold = “yes.”
    - **Super-Sub (2-way)**: binary choice; one inverted-hierarchy candidate (e.g., Sub-Super if gold = Super-Sub).

4. **Quality Validation**

    - We randomly sampled **200 MCQs per relation type** (800 total) and had eight annotators judge each distractor on grammatical plausibility, semantic coherence, and clear incorrectness.
    - **Pass rate**: 89% of distractors passed on first submission; 11% were replaced and re-validated to ensure each MCQ retained the intended number of valid distractors.

5. **Illustrative MCQs** As shown in Table 8, we provide one fully detailed example per relation type.

C The Detailed Introduction to the LLMS

This section provides a detailed overview of each language model evaluated in EventRelBench. For each model, we describe the architecture, training data, parameter scale, notable features, and prior performance characteristics.

Bloomz-7B is a multilingual, instruction-tuned variant of the BLOOM family with approximately 7 billion parameters (Le Scao et al., 2023). It builds on a Transformer decoder-only architecture with 30 layers and 32 attention heads, supporting a 2048

Relation	Prompt & Events	Options	Gold
Temporal	Another cousin Georgina Cid said Elian father had intended to flee Cuba himself but was being coerced by the Castro government to stay and make certain statements. What is the temporal relationship between 'said' and 'intended'?	A.before B.after C.equal D.vague	B
Causal	A series of nine bombs exploded in which one person was killed and 20 injured. What is the causal relationship between 'exploded' and 'killed'?	A.causes B.caused by	A
Coreference	One man got killed and thousands of people were left stranded. Police said an 81-year-old man died in central Toyama when the wind blew over a shed, trapping him underneath. Do 'killed' and 'died' have an event coreference relationship?	A.yes B.no	A
Super-Sub	The Fandango! Tour was a concert tour by American rock band ZZ Top. Launched in support of their fourth studio album "Fandango!". What is the relationship between 'tour' and 'support'?	A.Super-Sub B.Sub-Super	A

Table 8: One fully detailed example per relation type

token context window. Pre-trained on a diverse corpus covering dozens of languages and then fine-tuned on the cross-lingual Task Mixture (xP3), it delivers strong zero-shot instruction following across languages and competitive performance on standard NLP benchmarks.

ChatGLM2-6B-32k is an open-source, decoder-only bilingual chat model from Tsinghua's group, with 6 billion parameters and a 32 K-token context window (Du et al., 2021). It extends ChatGLM2-6B using FlashAttention and positional interpolation for long-context understanding, and leverages Multi-Query Attention for faster, lower-memory inference. Pre-trained on 1.4 trillion Chinese–English token pairs with human-preference alignment, it excels at dialogue coherence and long-form reasoning.

AquilaChat-7B is an open-source conversational SFT model released by BAAI, built on the Aquila-7B. It was fine-tuned with supervised learning and reinforcement learning to improve dialogue quality and instruction following. While it benefits from Aquila's efficient low-level operators and parallel training methods, yielding faster inference compared to unoptimized implementations.

Gemma-7B-IT is Google's 7-billion-parameter instruct-tuned variant of the Gemma family (Team et al., 2024), released in late 2024. Built on a decoder-only Transformer architecture, it was fine-tuned on general conversational and instruction datasets (e.g., UltraChat) and supports up to 8K tokens of context. While well-suited for a broad range of text-generation tasks, including Q&A, summarization, and reasoning.

ERNIE-3.5-8K released on July 1, 2024, is Baidu's flagship large-scale model supporting an 8 K-token context window (Sun et al., 2021). Pre-trained on massive Chinese and English corpora using knowledge-enhanced techniques that integrate structured knowledge graphs, it excels at capturing complex semantic relations, making it particularly strong on tasks like event relation extraction.

DeepSeek-V3 is an advanced AI language model built on an innovative Mixture-of-Experts (MoE) architecture, featuring a total of 671 billion parameters with 37 billion activated per token to maximize efficiency and performance (Liu et al., 2024). DeepSeek-V3 is pre-trained on 14.8 trillion high-quality tokens. It delivers great results across complex reasoning, code generation, and multilingual understanding benchmarks while maintaining an extended 128 K context window for long-form inputs. Remarkably, DeepSeek-V3 was developed at a fraction of the typical cost, under $6 million in compute.

Claude-3-Haiku launched on March 13, 2024, is Anthropic's fastest and most cost-effective model in the Claude 3 family (Anthropic, 2024). It supports both text and image inputs, offers a massive 200K token context window, and delivers near-instantaneous responses, making it ideal for real-time enterprise applications, while still achieving strong results on industry-standard benchmarks.

GPT-3.5-Turbo is OpenAI's flagship chat-optimized, decoder-only model with a 4096-token context window, launched in March 2023 as the default Chat Completions endpoint (Achiam et al., 2023). On February 16, 2024, OpenAI automatically upgraded it to the gpt-3.5-turbo-0125 variant, adding bug fixes and improved format-following.

LLaMA-3-8B is Meta's 8-billion-parameter, decoder-only Transformer model released in April 2024 as part of the LLaMA 3 family (Grattafiori et al., 2024). It was pretrained on over 15 trillion tokens of multilingual text to improve reasoning and generation capabilities, and supports an 8 K-token context window for longer inputs. The model employs Grouped-Query Attention to enhance inference efficiency and scalability, and is offered in both base and instruction-tuned variants under the Meta Llama 3 Community License. LLaMA-3-8B delivers strong performance on general NLP benchmarks, coding tasks, and multilingual applications while remaining lightweight enough to run on a wide range of hardware setups.

Qwen2.5 is a family of dense, decoder-only language models released by Alibaba Cloud (Bai et al., 2023). We have selected 0.5B, 1.5B, 3B, and 7B parameter sizes. Each model is pretrained on up to 18 trillion tokens. For each size, both base and instruction-tuned versions are provided, and all open-weight variants are publicly available. These models outperform their Qwen2 predecessors on instruction following, long-text generation, structured-data understanding, and code tasks, making them versatile tools across NLP and developer workflows.

Event Relation	**# Nums**
Coreference	14749
Temporal	8523
Causal	23780
Supersub	864
All	47916

Table 9: Statistics of the EventRelInst.

D Statistics of the EventRelInst

As shown in Table 9, we provide the statistics of the EventRelInst.