ETRQA: A Comprehensive Benchmark for Evaluating Event Temporal Reasoning Abilities of Large Language Models

Sigang Luo[1]*, Yinan Liu[1]*, Dongying Lin[1], Yingying Zhai[1]†,
Bin Wang[1,3], Xiaochun Yang[1,3], Junpeng Liu[2]
[1]School of Computer Science and Engineering, Northeastern University, Shenyang, China
[2]The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China
[3]National Frontiers Science Center for Industrial Intelligence and Systems optimization,
Northeastern University, Shenyang, China
2301925@stu.neu.edu.cn, liuyinan@cse.neu.edu.cn, 2472128@stu.neu.edu.cn
{zyy, binwang, yangxc}@mail.neu.edu.cn, junpengliu@hkust-gz.edu.cn

Abstract

Event temporal reasoning (ETR) aims to model and reason about the relationships between events and time, as well as between events in the real world. Proficiency in ETR is a significant indicator that a large language model (LLM) truly understands the physical world. Previous question-answering datasets available for evaluating the ETR ability lack a systematic taxonomy and pay limited attention to compound questions. In this paper, we propose a unified taxonomy for event temporal questions and construct a comprehensive benchmark ETRQA, to evaluate the ETR abilities of LLMs based on this taxonomy. ETRQA not only inherits and expands the evaluation content of existing datasets but also contains multiple categories of compound questions. We evaluate two leading LLM series, Llama and Qwen, on ETRQA across various settings. Our experimental results indicate that large-scale LLMs exhibit certain ETR abilities. Yet they do not perform well in solving specific types of reasoning tasks, including reasoning involving time spans, reasoning for compound questions, and reasoning with fine temporal granularity. Additionally, we hope ETRQA can benefit the temporal reasoning research community for future studies.[1]

1 Introduction

Recently, large language models (LLMs) have demonstrated impressive performance in numerous reasoning tasks (Qiao et al., 2023; Huang and Chang, 2023; Zhao et al., 2024; Chang et al., 2024). However, it has been observed that they still underperform in a fundamental aspect of human cognition, namely temporal reasoning (Chu et al., 2024; Wang and Zhao, 2024). Temporal reasoning is typically categorized into three levels: symbolic, commonsense, and event temporal reasoning (Chu et al., 2024). Event temporal reasoning (ETR) is the most comprehensive of these levels, as it not only encompasses the understanding and application of the first two types of reasoning but also involves modeling and reasoning about the relationships between events and time, as well as between events in the real world (Tan et al., 2023). The proficient ETR ability is an important indicator that an LLM truly grasps the physical world. Therefore, we focus on the task of ETR.

Most question-answering datasets available for evaluating the ETR ability are constructed using temporal knowledge graphs. They generally generate questions based on facts that evolve over time (Saxena et al., 2021; Chen et al., 2021; Tan et al., 2023; Chen et al., 2023; Tan et al., 2024; Fatemi et al., 2024). For example, the question "What team did LeBron James play for in 2009?" is based on the facts that LeBron James has played for different teams over various periods. Each fact can be regarded as a temporal event, and we refer to such questions as event temporal questions in this paper. To answer such questions, LLMs need to perform ETR. However, there exist two issues in these previous datasets: (1) due to the lack of a systematic taxonomy, existing datasets are often designed with several empirically defined question types, which results in incomplete evaluations, as shown in Table 1; (2) existing datasets rarely focus on compound questions, which are critical for evaluating the ETR abilities of LLMs.

To address the above issues, we first establish a unified taxonomy based on previous studies. This taxonomy categorizes event temporal questions by question composition and answer type, providing a standardized reference that can be widely adopted. Based on this taxonomy, we propose a question-answering dataset for ETR, named ETRQA. ETRQA inherits the evaluation content of existing datasets while expanding upon them

* Equal contributions.

† Corresponding authors.

[1] https://github.com/sigangluo/ETRQA

Findings of the Association for Computational Linguistics: ACL 2025, pages 23321–23339
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

through a more comprehensive coverage and examination of question composition and answer type. Specifically, ETRQA contains multiple types of compound questions, which make the dataset more complex and challenging. For instance, to solve the compound question "What was the team that LeBron James played for the longest time from 2009 to 2015?", LLMs need to not only understand the temporal constraint "from 2009 to 2015" but also compare the durations of events to find "for the longest time" within that range. It is clear that such compound questions are well-suited for evaluating the comprehensive ETR abilities of LLMs.

To evaluate the ETR abilities of LLMs, we adopt two industry-leading LLM series: the Llama series (Grattafiori et al., 2024) and the Qwen series (Yang et al., 2025) on ETRQA. The latest models, Llama-3.3-70B-Instruct and Qwen2.5-72B-Instruct, have demonstrated competitive performance, matching or even surpassing GPT-4o (Hurst et al., 2024) on many benchmarks, making them highly representative. We show the effects of prompting strategies, instruction tuning, model sizes, and general reasoning distillation in the task of ETR via LLMs. Furthermore, we analyze and discuss the difficulties and challenges LLMs face. The experimental results show that: (1) large-scale LLMs demonstrate a certain level of ETR abilities compared with small-scale LLMs, but there is still room for improvement; (2) careful thought before answering and post-training techniques (e.g., instruction tuning and general reasoning distillation) are crucial for enhancing the models' ETR abilities; (3) LLMs struggle with reasoning involving time spans and show a decline in performance when reasoning for compound questions and reasoning with fine temporal granularity.

Our contributions can be summarized as follows:

(1) We propose a unified taxonomy for event temporal questions, providing a standardized reference that can be widely adopted.

(2) By expanding the evaluation content of existing datasets and systematically designing multiple types of compound questions, we construct a comprehensive and extensive benchmark for ETR evaluation.

(3) We conduct a quantitative evaluation of two industry-leading LLM series. Our analysis examines the effects of prompting strategies, instruction tuning, model sizes, and general reasoning distillation, offering valuable insights into the challenges and potential improvements in ETR.

2 A Taxonomy of Event Temporal Questions

The proposed taxonomy of event temporal questions, shown in Table 1, consists of two aspects: (1) question composition and (2) answer type.

2.1 Question Composition

An event temporal question can either be a non-compound question that contains only a temporal constraint or a temporal comparison, or a compound question that includes both a temporal constraint and a temporal comparison.

Temporal constraint. A temporal constraint limits the time range of a question. It consists of a temporal signal and a temporal expression. For example, in the temporal constraint "in 2009", the temporal signal is "in" and the temporal expression is "2009". Inspired by previous work (Jia et al., 2018; Chen et al., 2021), we identify four types of temporal signals: $DURING$, IN, $BEFORE$, and $AFTER$. The typical trigger words for these temporal signal types are as follows: (1) $DURING$: "during" and "from ... to ..."; (2) IN: "in"; (3) $BEFORE$: "before" and "prior to"; (4) $AFTER$: "after" and "following".

Temporal constraints can be divided into three categories based on the form of the temporal expression: (1) explicit temporal constraint denotes the time range that can be directly determined without additional context or computation (e.g., "in 2009" and "from 2009 to 2015"); (2) implicit temporal constraint requires determining the time range based on specific temporal events (e.g., "in the time when the Nuggets won the NBA championship" and "during the period when James was playing for the Heat"); (3) relative temporal constraint (Tan et al., 2024) involves calculating the target time based on a reference time and the specified time span to determine the time range (e.g., "in the second year before 2001" is equal to "in 1999" and "during the 2 years before 2002" is equal to "from 2000 to 2001").

Explicit and implicit temporal constraints involve all four temporal signal types, while relative temporal constraints involve only the temporal signal types $DURING$ and IN, as they require considering both the reference and target times.

Temporal comparison. Temporal events can be compared from two aspects (Xiong et al., 2024): (1) Order and (2) Duration. Order involves determining the chronological order of events. For example,

Dataset			CronQuestions (Saxena et al., 2021)	TimeQA (Chen et al., 2021)	TempReason (Tan et al., 2023)	MultiTQ (Chen et al., 2023)	Complex-TR (Tan et al., 2024)	ToT (Fatemi et al., 2024)	**ETRQA**
Question Composition									
Temporal constraint	Explicit	$DURING$		✓				✓	✓
		IN	✓	✓	✓	✓	✓	✓	✓
		$BEFORE/AFTER$		✓		✓			✓
	Implicit	$DURING$	✓				✓		✓
		IN				✓		✓	✓
		$BEFORE/AFTER$	✓		✓	✓	✓	✓	✓
	Relative	$DURING$							✓
		IN					✓		✓
Temporal comparison	Order		✓			✓		✓	✓
	Duration								✓
Compound	Explicit + Order					✓			✓
	Implicit + Order					✓			✓
	Relative + Order								✓
	Explicit + Duration								✓
	Implicit + Duration								✓
	Relative + Duration								✓
Answer Type									
Entity			✓	✓	✓	✓	✓	✓	✓
Time	Time point		✓			✓		✓	✓
	Time interval								✓
	Time span							✓	✓
Numeric								✓	✓

Table 1: Summary of event temporal question-answering datasets according to our proposed taxonomy.

in the question "What was the team that LeBron James played for the first time?", the ordinal term "first" indicates a comparison of order. Duration involves comparing the time span of events. For example, in the question "What was the team that LeBron James played for the longest time?", the term "longest" indicates a comparison of durations.

Compound. Temporal constraints and temporal comparisons can appear independently in questions or be combined to form more complex questions (Chen et al., 2023), an important category that is often overlooked by existing datasets. We introduce 14 types of compound questions by combining three kinds of temporal constraints excluding the temporal signal type IN and two kinds of temporal comparisons, as shown in Table 1. Note that IN confines the time range to a single point, making any comparison impossible. Therefore, compound questions involve performing a temporal comparison on temporal events that satisfy the temporal constraint, which are used to better evaluate the comprehensive ETR ability.

2.2 Answer Type

The type of answer is determined by the question word or phrase. We categorize the answers to event temporal questions into three types: (1) Entity involves identifying entities using question words like "Who". (2) Time is relevant for time-related questions, including time points (e.g., 2001) using question words like "When", time intervals (e.g., between 2000 and 2001) using question phrases like "During which period", and the time span (e.g., 2 years) using question phrases like "How long". (3) Numeric involves counting entities, time points, or time intervals using question phrases like "How many", and is relevant for statistical questions.

3 Dataset Construction

3.1 Temporal Event Preprocessing

We use the Wikidata (Vrandečić and Krötzsch, 2014) dump of September 2024 as the data source for extracting temporal events. Temporal events can be classified into two categories based on different annotating ways: (1) temporal interval events and (2) temporal point events. A temporal interval event can be represented as a quintuple $\langle s, r, o, t_s, t_e \rangle$, where s is a subject, r is a relation, o is an object, t_s is its start time and t_e is its end time. A temporal point event can be represented as a quadruple $\langle s, r, o, t_p \rangle$, where t_p is the event time. We select 16 relations that represent temporal interval events and 6 relations for temporal point events, as shown in Appendix A.2. We then extract all corresponding temporal events and group them by subject and relation to form temporal event groups. These include the temporal interval event group $\{\langle s, r, o_i, t_{s_i}, t_{e_i} \rangle \mid i = 1, 2, \ldots, N\}$ and the temporal point event group $\{\langle s, r, o_i, t_{p_i} \rangle \mid i = 1, 2, \ldots, N\}$, where N represents the number of events. Within these groups, we remove those containing only a single event or where the events involve the same object to avoid pseudo-temporal reasoning (Chen et al., 2022). Temporal event groups are the basis for constructing event temporal questions since they reflect the evolving process of facts over time. For example, the temporal event group representing LeBron James playing for different teams during different periods can lead to event temporal questions like “What team did LeBron James play for in 2009?”. Additionally, Wikidata annotates time with three levels of temporal granularity: year, month, and day. We preserve this feature, so our dataset also involves multi-granularity temporal reasoning (Chen et al., 2023). We note that within a temporal event group, the temporal granularity of different events may be inconsistent. Therefore, we standardize the granularity within each group to a unified level. For example, if a temporal event group contains both year and month granularities, we standardize it to the year granularity.

3.2 Event Temporal Question Context Construction

Considering the potential data leakage issue with LLMs (Fatemi et al., 2024; Xiong et al., 2024), where prior knowledge of constructed questions may be implicitly contained within model training, it is possible for LLMs to provide direct answers without any reasoning. For example, when asked “What was the team that LeBron James played for the longest time?”, if the training corpus includes the information “LeBron James played the longest for the Cavaliers.” the LLM may deliver the answer directly. Therefore, we evaluate LLMs in an anonymized open-book question answering setting, in which the LLM is provided an anonymized context to answer the anonymized question. This shifts the evaluation focus to the LLM’s ability to understand the context and perform the reasoning process itself. Specifically, we anonymize the temporal event group by mapping and replacing the specific entities in the subject and object positions of all temporal events with anonymized entities denoted as $E*$, where $*$ denotes a number. The anonymized temporal event group is textualized to serve as anonymized context. The questions are constructed based on anonymized temporal event groups, effectively avoiding data leakage issues. These contexts vary in length, time coverage, co-temporal types (Su et al., 2024), temporal event types, and relation composition (i.e., contexts involve two types of relations under implicit temporal constraints), thereby providing diversity in reasoning at the contextual level. Templates for textualizing the structured temporal events and the example of an anonymized context can be found in Appendix A.1 and Appendix B, respectively.

3.3 QA Pairs Construction

The automated construction of QA pairs is achieved through question template filling and rule-based answer acquisition. We have designed a total of 566 question templates for different relations. In designing these templates, we ensure comprehensive coverage of the taxonomy for event temporal questions. Details regarding the design of the question templates can be found in Appendix A.2. Each of these templates has a corresponding rule for obtaining answers, which can be executed on structured temporal event groups. For instance, given a temporal event group denoting the different positions E1 held during various periods, a compound question template like “What was the $\{ord\}$ position $\{s\}$ held $\{tc\}$?” could be filled. Here, the placeholder $\{ord\}$ is for ordinal terms indicating a comparison of order, $\{s\}$ is for the subject, and $\{tc\}$ is for the temporal constraint. An instance might be “What was the first position E1 held after 2002?”. The filled temporal constraint (i.e., after 2002) needs to ensure that there are multiple events satisfying the constraint, and then the answer is derived by performing a comparison of order (i.e., determining the “first” event) among the candidate events. For the same context, a single question template can generate multiple questions with similar reasoning processes (e.g., by sampling different times within temporal constraints). To prevent the dataset from containing redundant questions involving similar reasoning over the same context, we limit each template to produce only one question per context.

3.4 Dataset Statistics

Following Chen et al. (2023), we divide the Wikidata dump into training, development, and test splits, ensuring there is no overlap of entities among them. For each split, we conduct temporal event preprocessing, construct context and QA pairs, and perform sampling. As a result, we obtain a dataset with an approximate ratio of 8:1:1 for training, development, and test sets, totaling 160k questions. Detailed statistics are shown in Table 2.

3.5 Quality Check

Following Fatemi et al. (2024), we conduct multiple rounds of quality checks on the dataset. Specifically, considering the time cost, we sample 545 QA pairs and their corresponding contexts from the test set, covering various question templates without accounting for the different relations involved in these templates. We perform manual reviews, primarily focusing on: (1) answer accuracy; (2) potential format, grammatical, and semantic errors; (3) question ambiguity. This process repeats until no further issues are identified in the dataset.

		Train	Dev	Test
Question Composition				
Non-compound	Explicit	17,505	2,308	2,384
	Implicit	27,691	3,820	4,012
	Relative	7,812	1,024	1,073
	Order	4,916	648	658
	Duration	2,400	300	300
Compound	Explicit + Order	14,248	1,858	1,907
	Implicit + Order	24,311	3,304	3,425
	Relative + Order	4,916	648	658
	Explicit + Duration	7,200	900	900
	Implicit + Duration	12,820	1,737	1,759
	Relative + Duration	2,400	300	300
Answer Type				
Entity		34,161	4,496	4,621
Time	Time point	22,141	3,047	3,241
	Time interval	23,772	3,147	3,186
	Time span	19,657	2,581	2,593
Numeric		26,488	3,576	3,735
Total		126,219	16,847	17,376

Table 2: Dataset statistics of ETRQA

4 Experiments

4.1 Experimental Setting

Evaluation Metrics. Accuracy is used as the evaluation metric. For multi-answer cases, it is measured using set accuracy (Zhong et al., 2023), which considers a prediction correct only if the predicted set matches the ground truth answer set exactly.

LLMs for Evaluation. We evaluate ETR abilities of two industry-leading LLM series: the Llama series (Grattafiori et al., 2024) and the Qwen series (Yang et al., 2025), which include models of varying sizes, as well as both base and instruction-tuned versions. Additionally, DeepSeek releases general reasoning-enhanced versions of the Llama and Qwen models. These models are fine-tuned using 800k samples curated with DeepSeek-R1 (Guo et al., 2025), achieving significant improvements in reasoning performance. Therefore, we select them to evaluate whether general reasoning distillation contributes to enhancing ETR abilities. We run these models without quantization on four A800-80G GPUs, with the temperature parameter set to 0 to ensure reproducibility.

Setting Details. We evaluate the LLMs under four common prompting strategies: zero-shot, few-shot (Brown et al., 2020), zero-shot CoT (Kojima et al., 2022), and few-shot CoT (Wei et al., 2022). In the zero-shot setting, LLMs are required to provide the answer directly based on the context. In the few-shot setting, additional examples are provided to help LLMs understand the task. These examples are sampled from the training set. In the zero-shot CoT setting, we add the phrase “Let’s think step by step.” as a CoT trigger after the question and prompt the LLM to provide the final answer using “The final answer is:”. In the few-shot CoT setting, CoT examples are provided to guide step-by-step reasoning. These CoT examples are automatically annotated using a rule-based program. We evaluate the general reasoning enhanced Llama and Qwen models only under the zero-shot CoT setting, as they are fine-tuned with long CoT data and, in practice, do not follow the reasoning patterns of the examples we provide in the few-shot CoT setting. All prompt templates can be found in Appendix C.

4.2 Overall Performance

We conduct the evaluation based on question composition. The experimental results of the base models, instruction-tuned models, and general reasoning enhanced models on the ETRQA test set are presented in Appendix D, Table 3 and Table 4, respectively.

Effect Analysis of Prompting Strategy. To verify the effect of different prompting strategies, we evaluate the overall accuracy of the instruction-tuned models and base models under different prompting strategies. From the results shown in Figure 1, we can see all models achieve their best performance under the few-shot CoT setting. The best-performing model is Llama-3.3-Instruct, with an overall accuracy of 78.7%, demonstrating that LLMs possess a certain level of ETR ability, though there is still room for improvement. In the zero-shot CoT setting, the base and small-scale models show a significant decline in performance compared to the instruction-tuned large-scale models, revealing the inherent limitations of these models in ETR. Furthermore, we observe consistent performance improvements for base and small-scale models when moving from zero-shot CoT to few-shot CoT, as explicit reasoning examples help com-

Method	Non-Compound						Compound							Overall
	E.	I.	R.	O.	D.	Avg.	E. + O.	I. + O.	R. + O.	E. + D.	I. + D.	R. + D.	Avg.	
Llama-3.1-8B-Instruct														
ZS	36.6	27.4	24.9	46.8	15.3	30.8	35.4	26.0	26.3	16.0	14.0	12.0	24.2	27.4
FS	39.4	33.7	31.3	55.8	17.3	36.1	34.8	31.1	28.0	19.3	19.6	11.3	27.5	31.7
ZS CoT	27.1	17.6	16.1	52.4	32.0	23.3	30.2	21.8	23.3	21.6	17.5	16.7	22.7	23.0
FS CoT	58.6	56.0	50.2	66.0	35.0	56.0	52.5	48.5	41.6	34.3	31.4	15.3	43.0	49.3
Llama-3.1-70B-Instruct														
ZS	57.5	45.8	38.6	73.9	30.0	49.8	55.3	52.4	40.7	33.8	33.3	25.7	45.7	47.7
FS	59.7	54.4	42.3	80.9	33.3	55.7	58.2	55.9	40.6	32.9	34.8	24.7	47.7	51.6
ZS CoT	83.9	78.7	67.6	85.1	63.7	78.7	77.1	73.6	58.1	64.6	58.7	47.0	68.5	73.4
FS CoT	82.6	80.9	74.3	84.7	59.0	80.0	81.4	78.5	68.5	63.8	60.4	51.3	72.4	76.1
Llama-3.3-70B-Instruct														
ZS	58.6	47.9	36.6	75.8	30.3	51.1	55.2	52.6	40.6	31.1	28.2	19.0	44.2	47.5
FS	56.0	52.2	37.9	81.9	29.3	52.9	57.3	56.4	42.9	29.6	33.0	22.3	47.2	50.0
ZS CoT	**87.2**	**82.9**	72.5	86.8	68.0	**82.5**	81.0	80.0	63.7	**69.2**	**63.8**	46.0	73.6	77.9
FS CoT	82.9	82.3	75.6	**88.9**	67.0	81.6	**84.2**	**82.0**	**73.6**	67.9	**63.8**	**58.3**	**76.1**	**78.7**
Qwen2.5-7B-Instruct														
ZS	33.7	28.3	21.2	46.5	18.0	30.0	35.1	25.3	26.3	18.9	15.7	12.0	24.5	27.2
FS	36.2	32.7	24.0	53.8	23.3	33.9	37.5	28.9	30.7	19.6	20.6	13.3	27.8	30.7
ZS CoT	54.0	42.9	40.4	62.2	37.0	47.0	52.8	41.5	36.3	33.9	28.1	17.3	39.3	43.0
FS CoT	57.8	50.1	46.6	66.3	34.3	52.5	57.6	49.4	40.4	34.9	30.9	30.3	44.7	48.5
Qwen2.5-14B-Instruct														
ZS	46.4	37.6	31.3	64.1	23.3	40.9	41.1	36.2	34.5	25.8	21.9	19.0	32.7	36.6
FS	47.7	41.5	36.6	63.5	15.7	43.4	39.2	34.2	31.6	15.9	16.3	11.7	29.0	36.0
ZS CoT	59.0	55.0	53.5	75.4	53.3	57.5	63.0	59.0	49.8	50.1	45.2	36.0	54.8	56.1
FS CoT	58.8	58.1	49.8	76.1	53.7	58.5	68.1	64.3	52.1	52.8	45.3	35.3	58.3	58.4
Qwen2.5-32B-Instruct														
ZS	59.0	45.5	43.0	66.4	28.0	50.0	49.5	43.4	38.3	30.1	26.1	23.0	38.9	44.3
FS	59.2	47.7	40.2	73.9	24.3	51.2	49.1	42.5	34.5	27.8	25.9	18.0	37.8	44.3
ZS CoT	79.9	76.3	73.7	86.6	58.0	77.1	73.3	71.9	62.3	57.4	56.0	49.7	66.2	71.5
FS CoT	83.0	76.3	77.1	79.2	71.3	78.4	80.1	74.7	70.5	66.2	57.9	56.3	70.8	74.5
Qwen2.5-72B-Instruct														
ZS	56.0	47.3	40.2	73.1	23.0	50.0	49.8	44.7	39.4	28.4	24.6	18.3	38.9	44.3
FS	58.6	51.0	42.7	81.3	36.3	53.9	55.7	48.0	41.0	32.0	28.0	23.0	42.8	48.2
ZS CoT	84.1	81.7	76.2	88.1	66.0	81.6	74.6	73.5	67.3	59.9	58.7	50.7	68.3	74.7
FS CoT	84.7	80.7	**79.3**	88.4	**75.0**	82.0	80.4	76.7	**73.6**	65.9	56.6	54.7	71.5	76.6

Table 3: Experimental results of instruction-tuned models under four common prompting strategies. The abbreviations E., I., R., O., D. refer to Explicit, Implicit, Relative, Order, and Duration respectively. The results in bold (resp. underline) denote the best (resp. second) results.

pensate for their limited reasoning abilities. In contrast, instruction-tuned large-scale models already possess considerable reasoning abilities, so for simple, non-compound questions, few-shot CoT does not always outperform zero-shot CoT, since the effectiveness of few-shot CoT depends on factors such as the quality, complexity, order, and number of the provided examples. However, for more complex, compound questions, few-shot CoT still leads to better performance in most cases, indicating that these models still need to further learn the reasoning patterns required for solving compound questions.

Providing explicit step-by-step reasoning via CoT generally enhances model performance compared to direct input-output prompting strategies in zero-shot and few-shot settings. However, this influence can be inconsistent (Chu et al., 2024). For example, Llama-3.1-8B performs worse in the zero-shot CoT setting compared to the zero-shot and few-shot settings. We provide the error analysis in Appendix E.

Effect Analysis of Instruction Tuning. As shown in Figure 1, for each LLM, its instruction-tuned version generally outperforms the corresponding base model. Moreover, we observe that providing examples helps narrow the performance gap between the base and instruction-tuned versions of the same LLM. Specifically, the gap is smaller in the few-shot setting than in the zero-shot setting. A similar trend is observed in the CoT setting, where the performance gap is narrower under few-shot CoT compared to zero-shot CoT. These findings suggest that providing examples can partially mitigate

Method	Non-Compound						Compound							Overall
	E.	I.	R.	O.	D.	Avg.	E. + O.	I. + O.	R. + O.	E. + D.	I. + D.	R. + D.	Avg.	
Llama-3.1-8B†	76.1	63.8	55.5	73.9	29.3	65.8	61.6	55.5	51.5	32.0	31.8	21.7	48.4	56.8
Llama-3.3-70B-Instruct†	**94.0**	**90.3**	84.1	**87.7**	**59.3**	**89.2**	**85.5**	**82.2**	**76.7**	**61.3**	**65.0**	39.7	**75.6**	**82.2**
Qwen2.5-7B†	65.6	53.0	49.4	70.4	26.0	56.5	57.2	49.1	38.8	28.0	26.7	15.7	42.4	49.2
Qwen2.5-14B†	86.2	84.1	77.4	85.7	43.7	82.5	76.9	74.7	66.3	47.8	54.9	33.7	66.5	74.3
Qwen2.5-32B†	92.8	89.6	**86.8**	85.7	51.7	88.5	78.7	77.7	71.9	55.2	59.4	**40.0**	70.4	79.2

Table 4: Experimental results of general reasoning enhanced models in the zero-shot CoT setting. † denotes that the model has undergone general reasoning distillation. The abbreviations E., I., R., O., D. refer to Explicit, Implicit, Relative, Order, Duration. The results in bold (resp. underline) denote the best (resp. second) results.

the issue where base models, lacking instruction tuning, struggle with ETR.

Effect Analysis of Model Size. As shown in Figure 2, we evaluate how the scale of LLMs affects ETR performance using four model sizes (i.e., 7B, 14B, 32B, and 72B) of the Qwen instruction-tuned models under different prompting strategies. It can be observed that regardless of the prompting strategy used, performance improves as model size increases. However, compared to the performance improvement when scaling from 7B to 14B and from 14B to 32B, the improvement from 32B to 72B is much smaller, which demonstrates that further expansion will yield limited performance improvements of the model after its size increases to a certain extent. Therefore, it is necessary to explore other methods to further enhance the model's ETR ability.

Effect Analysis of General Reasoning Distillation. As shown in Table 4, general reasoning distillation significantly enhances the models' ETR abilities, especially for small-scale models. The best-performing model is the general reasoning enhanced Llama-3.3-70B-Instruct, with an overall accuracy of 82.2%, showing a 3.5% improvement over its underlying model, Llama-3.3-Instruct. However, these models tend to overthink (Chen et al., 2025) when dealing with questions involving duration comparisons, leading to poor performance on such questions. Specifically, even with the maximum output token limit set to $2,048$, these models fail to provide an answer within a reasonable number of reasoning steps.

4.3 Challenges in Event Temporal Reasoning

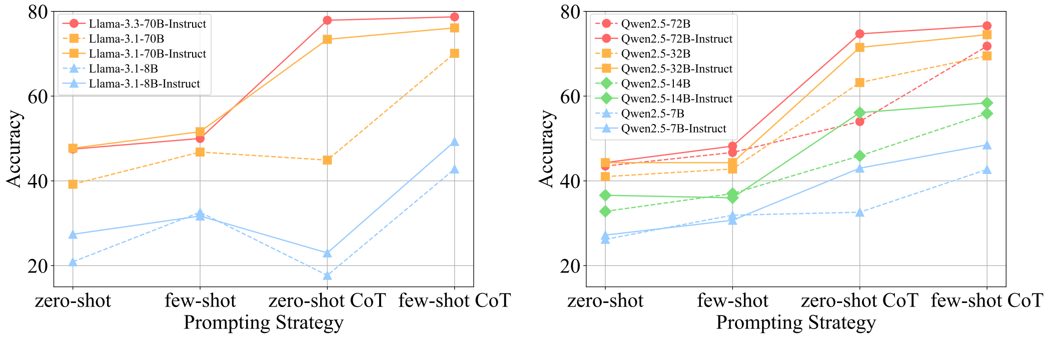
Compound questions are more challenging for existing LLMs. When handling compound and non-compound questions, the models exhibit a noticeable performance gap. The best performance on non-compound questions is achieved by the general reasoning enhanced Llama-3.3-70B-Instruct, with an average accuracy of 89.2%. In contrast, the best performance on compound questions is achieved by Llama-3.3-70B-Instruct, with an average accuracy of 76.1%, marking a difference of 13.1%, which highlights the greater challenge posed by compound questions and indicates that LLMs' multi-step ETR abilities still need to be enhanced.

Method	Entity	Time			Numeric
		Time point	Time interval	Time span	
Llama-3.3-70B-Instruct†	**88.6**	**93.4**	**78.7**	47.1	91.9
Qwen2.5-32B†	**88.6**	88.2	70.8	42.6	**92.3**
Llama-3.3-70B-Instruct	84.2	87.2	71.4	**56.5**	82.6
Qwen2.5-72B-Instruct	81.9	86.5	68.5	40.8	84.6

Table 5: Performance of models on different answer types.† denotes that the model has undergone general reasoning distillation. The the best results are in bold.

LLMs struggle with reasoning involving time spans. We observe that all models demonstrate weaker performance when handling questions involving relative temporal constraints or duration comparisons, particularly when both are compounded. In such compounded cases, Llama-3.3-70B-Instruct model performs the best, yet it only achieves an accuracy of 58.3%. Both relative temporal constraints and duration comparisons are related to the understanding and computation of time spans. Additionally, as shown in Table 5, we analyze the models' performance across different answer types. When questions require a precise time span as an answer, the models also show poor performance. Therefore, a major focus for improving ETR abilities in LLMs is enhancing their abilities to understand and accurately compute time spans.

LLMs' performance declines when reasoning with fine temporal granularity. We analyze the models' performance across different temporal granularities, as shown in Table 6. The models perform best at the year level, but their performance declines significantly at the month and day levels. The two instruction-tuned models exhibit



(a) Llama series.

(b) Qwen series.

Figure 1: Overall performance of base models and instruction-tuned models on ETRQA under four common prompting strategies.

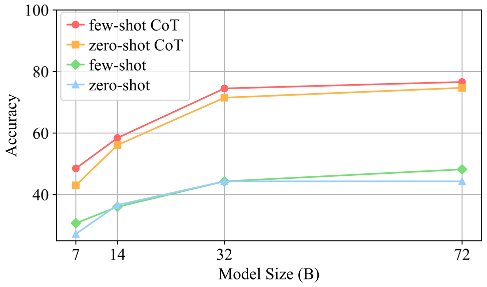


Figure 2: Overall performance of different model sizes in the Qwen series on ETRQA under four common prompting strategies.

Method	Year	Month	Day
Llama-3.3-70B-Instruct†	**89.8**	**80.2**	**76.3**
Qwen2.5-32B†	88.0	77.4	71.7
Llama-3.3-70B-Instruct	82.7	75.1	75.7
Qwen2.5-72B-Instruct	81.4	71.1	71.3

Table 6: Performance of models on different temporal granularities. † denotes that the model has undergone general reasoning distillation. The results in bold denote the best results.

similar performance at the month and day levels, while the two general reasoning enhanced models show a clear trend of decline. Additionally, it is observed that the general reasoning enhanced Llama-3.3-70B-Instruct improves its accuracy by 7.1% at the year level, 5.2% at the month level, and only 0.6% at the day level compared to its underlying model, Llama-3.3-70B-Instruct. This is because general reasoning distillation often focuses on enhancing mathematical abilities, which aids in time calculations and comparisons. However, at the day level, additional temporal commonsense, such as the different number of days in each month and the occurrence of leap years, also plays a role. Therefore, it is still necessary to strengthen the models' ETR abilities at finer temporal granularities.

5 Related Work

Early temporal NLP research primarily focuses on temporal information extraction tasks, leading to the development of early evaluation benchmarks such as TempEval (UzZaman et al., 2014). These efforts are mainly centered around extracting events, times, and temporal relations from text. However, with advancements in language models, particularly large language models, attention gradually shifts to the more challenging task of temporal reasoning. This shift leads to the emergence of numerous temporal reasoning datasets. Among them, comprehensive benchmarks such as TRAM (Wang and Zhao, 2024) and TimeBench (Chu et al., 2024) represent a major leap forward in temporal reasoning evaluation. These benchmarks consolidate various existing temporal reasoning datasets to facilitate a thorough evaluation of temporal reasoning abilities. TimeBench categorizes temporal reasoning into three levels: symbolic (Thukral et al., 2021; Tan et al., 2023), commonsense (Zhou et al., 2019; Qin et al., 2021; Virgo et al., 2022; Zhang and Wan, 2023), and event temporal reasoning. Our work focuses on the third category, which involves not only understanding and applying the first two types of reasoning but also modeling and reasoning about the relationships between events and time, as well as between events.

Many existing datasets can be used to evaluate event temporal reasoning abilities. Some of these datasets are specifically designed to evaluate the temporal reasoning abilities of language models. TimeQA (Chen et al., 2021) is an early dataset used to evaluate a model's ability to understand and reason about time-evolving facts. TempReason (Tan et al., 2023) evaluates temporal reasoning across three levels: time-time, time-event, and event-event relations. Its follow-up work, ComplexTR (Tan et al., 2024), emphasizes the importance of multi-answer and multi-hop (i.e., relative time constraints) temporal reasoning. These three datasets share a common limitation: the question types are limited, and the complexity of reasoning primarily arises from external contexts (e.g., Wikipedia). This makes it difficult to pinpoint the specific challenges that models face in event temporal reasoning. TGQA (Xiong et al., 2024) addresses potential data leakage in evaluation by anonymizing data through replacing entities with fictional ones. However, ToT (Fatemi et al., 2024) suggests that this method may introduce incorrect claims and adopts a simpler, non-semantic anonymization approach, while extending the context to more complex graph structures. Despite this, both datasets define question types empirically, lacking a systematic taxonomy. ComplexTempQA (Gruber et al., 2024) divides temporal questions into three categories: attribute, comparison, and counting questions. However, some of the question types it defines involve mathematical reasoning, which makes it unsuitable for independently evaluating temporal reasoning. Additionally, several time question-answering datasets built for temporal knowledge graph question answering (TKGQA) (Jia et al., 2018; Saxena et al., 2021; Jia et al., 2021; Neelam et al., 2022; Chen et al., 2023; Zhang et al., 2024) , table question answering (Gupta et al., 2023), or heterogeneous source question answering (Jia et al., 2024) can also be used for evaluation, but they share similar limitations as the aforementioned datasets.

To address the aforementioned limitations, we conduct an evaluation based on event temporal questions and establish a unified taxonomy. Furthermore, we introduce more challenging compound questions to further increase the complexity and diversity of reasoning. Based on this taxonomy, we propose the ETRQA benchmark to comprehensively evaluate event temporal reasoning abilities.

6 Potential impact

We believe that enhancing event temporal reasoning abilities is crucial for applying large language models (LLMs) to various real-world tasks related to temporal events, such as event timeline summarization (Song et al., 2025). Our work will have an impact on temporal reasoning in the following areas:

Dataset For the unified taxonomy of event temporal questions that we propose, we use it to guide the creation of more comprehensive question templates and the process of question generation. In practice, datasets can also be constructed through crowdsourcing or LLM synthesis, and the taxonomy we present will serve as a key guide for both of these approaches.

Model Training and Evaluation Our experimental results show that LLMs still do not perform well in solving certain types of reasoning tasks. Therefore, our dataset can be introduced during the model's pre-training or post-training phases to enhance its learning of event temporal reasoning. Furthermore, our dataset can not only be used for a comprehensive evaluation of LLMs' event temporal reasoning abilities but also to assess the impact of specific ability enhancements (e.g., mathematical reasoning) on the model's performance in event temporal reasoning tasks.

Retrieval-Enhanced Generation Applying event temporal reasoning to real-world scenarios often requires integrating context retrieval. Our paper presents a novel approach for retrieval-enhanced event temporal reasoning research, specifically decoupling reasoning from retrieval. Our experimental results provide the upper bound for the event temporal reasoning abilities of some mainstream models, and when these models are combined with retrieval systems, the key objective will be to optimize the retrieval system to help the models achieve this upper bound.

7 Conclusion

To comprehensively evaluate the event temporal reasoning abilities of LLMs, we develop a unified taxonomy for event temporal questions and propose a comprehensive question-answering benchmark named ETRQA, by inheriting and expanding the evaluation content of existing datasets, as well as designing various types of compound questions. We conduct a thorough evaluation of two industry-leading LLM series on ETRQA. The experimental results indicate that the ETR abilities of LLMs still have room for improvement, particularly in reasoning involving time spans, reasoning for compound questions, and reasoning with fine temporal granularity.

Limitations

Our work still has some limitations as follows: (1) The questions are constructed using manual templates, which may lack sufficient linguistic diversity. Therefore, we plan to incorporate human paraphrasing or utilize paraphrases generated by LLMs in the future. (2) The context of the questions is relatively simplistic. In real-world scenarios, contexts feature more diverse expressions of event and time relationships. For example, "LeBron James won the NBA MVP in 2008, and he won again a year later." In the future, we aim to generate more varied contexts based on rules or by using LLMs, which will further challenge the performance of LLMs. (3) Due to cost considerations, we have not yet evaluated closed-source models. However, considering the rapidly narrowing gap between open-source and closed-source models (Guo et al., 2025), evaluating industry-leading open-source models can reflect the current performance levels of LLMs.

Acknowledgments

The work is partially supported by the National Key Research and Development Program of China (2024YFF0617702), the National Natural Science Foundation of China (Nos.U22A2025, 62402097, 62232007, U23A20309), the Joint Funds of Natural Science Foundation of Liaoning Province (2023-BSBA-132), and 111 Project (No.B16009).

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. ACM Trans. Intell. Syst. Technol., 15(3).

Wenhu Chen, Xinyi Wang, and William Yang Wang. 2021. A dataset for answering time-sensitive questions. In Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. 2025. Do not think that much for 2+3=? on the overthinking of o1-like llms. Preprint, arXiv:2412.21187.

Ziyang Chen, Jinzhi Liao, and Xiang Zhao. 2023. Multi-granularity temporal question answering over knowledge graphs. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 11378–11392, Toronto, Canada. Association for Computational Linguistics.

Ziyang Chen, Xiang Zhao, Jinzhi Liao, Xinyi Li, and Evangelos Kanoulas. 2022. Temporal knowledge graph question answering via subgraph reasoning. Know.-Based Syst., 251(C).

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Haotian Wang, Ming Liu, and Bing Qin. 2024. TimeBench: A comprehensive evaluation of temporal reasoning abilities in large language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1204–1228, Bangkok, Thailand. Association for Computational Linguistics.

Bahare Fatemi, Mehran Kazemi, Anton Tsitsulin, Karishma Malkan, Jinyeong Yim, John Palowitch, Sungyong Seo, Jonathan Halcrow, and Bryan Perozzi. 2024. Test of time: A benchmark for evaluating llms on temporal reasoning. Preprint, arXiv:2406.09170.

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

Raphael Gruber, Abdelrahman Abdallah, Michael Färber, and Adam Jatowt. 2024. Complextempqa: A large-scale dataset for complex temporal question answering. Preprint, arXiv:2406.04866.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. Preprint, arXiv:2501.12948.

Vivek Gupta, Pranshu Kandoi, Mahek Vora, Shuo Zhang, Yujie He, Ridho Reinanda, and Vivek Srikumar. 2023. TempTabQA: Temporal question answering for semi-structured tables. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 2431–2453, Singapore. Association for Computational Linguistics.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In Findings of the Association for Computational Linguistics: ACL 2023, pages 1049–1065, Toronto, Canada. Association for Computational Linguistics.

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

Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. 2018. Tempquestions: A benchmark for temporal question answering. In Companion Proceedings of the The Web Conference 2018, WWW '18, page 1057–1062, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Zhen Jia, Philipp Christmann, and Gerhard Weikum. 2024. Tiq: A benchmark for temporal question answering with implicit time constraints. In Companion Proceedings of the ACM Web Conference 2024, WWW '24, page 1394–1399, New York, NY, USA. Association for Computing Machinery.

Zhen Jia, Soumajit Pramanik, Rishiraj Saha Roy, and Gerhard Weikum. 2021. Complex temporal question answering on knowledge graphs. CIKM '21, page 792–802, New York, NY, USA. Association for Computing Machinery.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022.

Sumit Neelam, Udit Sharma, Hima Karanam, Shajith Ikbal, Pavan Kapanipathi, Ibrahim Abdelaziz, Nandana Mihindukulasooriya, Young-Suk Lee, Santosh Srivastava, Cezar Pendus, Saswati Dana, Dinesh Garg, Achille Fokoue, G P Shrivatsa Bhargav, Dinesh Khandelwal, Srinivas Ravishankar, Sairam Gurajada, Maria Chang, Rosario Uceda-Sosa, Salim Roukos, Alexander Gray, Guilherme Lima, Ryan Riegel, Francois Luus, and L Venkata Subramaniam. 2022. A benchmark for generalizable and interpretable temporal question answering over knowledge bases. Preprint, arXiv:2201.05793.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2023. Reasoning with language model prompting: A survey. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5368–5393, Toronto, Canada. Association for Computational Linguistics.

Lianhui Qin, Aditya Gupta, Shyam Upadhyay, Luheng He, Yejin Choi, and Manaal Faruqui. 2021. TIMEDIAL: Temporal commonsense reasoning in dialog. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 7066–7076, Online. Association for Computational Linguistics.

Apoorv Saxena, Soumen Chakrabarti, and Partha Talukdar. 2021. Question answering over temporal knowledge graphs. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 6663–6676, Online. Association for Computational Linguistics.

Jiayu Song, Mahmud Akhter, Dana Atzil Slonim, and Maria Liakata. 2025. Temporal reasoning for timeline summarisation in social media. Preprint, arXiv:2501.00152.

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

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

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2024. Towards robust temporal reasoning of large language models via a multi-hop QA dataset and pseudo-instruction tuning. In Findings of the Association for Computational Linguistics: ACL 2024, pages 6272–6286, Bangkok, Thailand. Association for Computational Linguistics.

Shivin Thukral, Kunal Kukreja, and Christian Kavouras. 2021. Probing language models for understanding of temporal expressions. In Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pages 396–406, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Naushad UzZaman, Hector Llorens, James Allen, Leon Derczynski, Marc Verhagen, and James Pustejovsky. 2014. Tempeval-3: Evaluating events, time expressions, and temporal relations. Preprint, arXiv:1206.5333.

Felix Virgo, Fei Cheng, and Sadao Kurohashi. 2022. Improving event duration question answering by leveraging existing temporal information extraction data. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 4451–4457, Marseille, France. European Language Resources Association.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. Commun. ACM, 57(10):78–85.

Yuqing Wang and Yun Zhao. 2024. TRAM: Benchmarking temporal reasoning for large language models. In Findings of the Association for Computational Linguistics: ACL 2024, pages 6389–6415, Bangkok, Thailand. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022.

Siheng Xiong, Ali Payani, Ramana Kompella, and Faramarz Fekri. 2024. Large language models can learn temporal reasoning. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 10452–10470, Bangkok, Thailand. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, et al. 2025. Qwen2.5 technical report. Preprint, arXiv:2412.15115.

Tingyi Zhang, Jiaan Wang, Zhixu Li, Jianfeng Qu, An Liu, Zhigang Chen, and Hongping Zhi. 2024. MusTQ: A temporal knowledge graph question answering dataset for multi-step temporal reasoning. In Findings of the Association for Computational Linguistics: ACL 2024, pages 11688–11699, Bangkok, Thailand. Association for Computational Linguistics.

Yunxiang Zhang and Xiaojun Wan. 2023. Situatedgen: Incorporating geographical and temporal contexts into generative commonsense reasoning. In Advances in Neural Information Processing Systems, volume 36, pages 67355–67373. Curran Associates, Inc.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2024. A survey of large language models. Preprint, arXiv:2303.18223.

Victor Zhong, Weijia Shi, Wen-tau Yih, and Luke Zettlemoyer. 2023. RoMQA: A benchmark for robust, multi-evidence, multi-answer question answering. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 7055–7067, Singapore. Association for Computational Linguistics.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. “going on a vacation” takes longer than “going for a walk”: A study of temporal commonsense understanding. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3363–3369, Hong Kong, China. Association for Computational Linguistics.

A Templates

A.1 Context Templates

In Table 7, we show example templates for textualizing the structured temporal event. In the templates, the placeholders $\{s\}$ and $\{o\}$ are used to fill in the subject and object, respectively. The placeholders $\{t_s\}$ and $\{t_e\}$ are used to fill in the start and end times of a temporal interval event, respectively, while $\{t_p\}$ is used to fill in the event time of a temporal point event.

A.2 Question Templates

In designing question templates, besides ensuring comprehensive coverage of the taxonomy of event temporal questions, we also take the following into consideration:

- We create templates for both types of temporal event groups (i.e., temporal interval event groups and temporal point event groups) to cover the two common ways of annotating temporal events.
- To further enhance the diversity of questions, we consider the different focuses that questions might have on temporal event groups and design templates accordingly: (1) questions that focus on the entire temporal event group, such as “What was the position that E1 held for the longest time?”; (2) questions that focus on specific events within a temporal event group sharing the same object, such as “When did E1 receive the award of E2 for the first time?”; (3) for temporal interval event groups, questions can emphasize either the start or the end, thereby focusing reasoning solely on the start or end time of the events, such as “What distinct positions did E1 begin holding after 2002?”.

We design question templates for the representative relations in Wikidata shown in Table 8. In Table 10, we use the relation “position held” to present examples of question templates for questions based on temporal interval event groups. In Table 11, we use the relation "award received" to present examples of question templates for questions based on temporal point event groups. In the templates, the placeholders $\{s\}$ and $\{o\}$ are used to fill in the subject and object, respectively. The placeholder $\{tc\}$ is used to fill in the time constraint, the placeholder $\{ord\}$ is used to fill in ordinal terms, indicating a comparison of order, and the placeholder $\{dur\}$ is used to fill in terms like “longest”, indicating a comparison of durations.

For each designed template, we perform grammatical and semantic checks as well as ambiguity resolution. The primary strategy for resolving ambiguity is to add additional clarifications within the question template.

B Dataset Sample

Each sample in the dataset includes an anonymized context and a QA pair constructed based on the anonymized temporal event groups. This anonymized context is generated by textualizing anonymized temporal event groups. A sample is shown in Figure 3.

C Prompts

The prompt templates for four common prompting strategies used in the experiments are provided. Figure 4 is the zero-shot prompt template, Figure 5 is the few-shot prompt template, Figure 6 is the zero-shot CoT prompt template, and Figure 7 is the few-shot CoT prompt template. In the prompts, we include instructions indicating the answer format, such as "The answer is a time interval, format the interval as (start_time, end_time)," to ensure that the LLMs provide answers that can be interpreted by the evaluation program.

D Base Models Experimental Results

The experimental results of the base model on the ETRQA test set are presented in Table 12.

E Error Analysis

To understand why the base and small-scale models in the Llama series exhibit inconsistent behavior under the zero-shot CoT setting, particularly a significant performance drop, while other models tend to show steady improvement, we conduct an error analysis. We examine the responses of Llama-3.1-8B, Llama-3.1-70B, Qwen2.5-7B, and Qwen2.5-72B, including both base and instruction-tuned versions. We categorize the errors into four types: (1) Incorrect Answer: The response contains an incorrect answer; (2) Incomplete Answer: In cases where multiple answers are expected, the model fails to return the full set of correct answers; (3) Uncertainty Error: The model does not provide an answer, either because it believes there is none or it is uncertain; (4) Instruction-Following Error:

The model fails to follow the instruction to include the phrase “The final answer is:”, making it difficult to extract the final answer.

As shown in Table 9, the base and small-scale models in the Llama series tend to exhibit more instruction-following errors under the zero-shot CoT setting, while the Qwen series exhibits fewer such errors. This discrepancy may be attributed to differences in prompt robustness across models. These observations suggest that the ability to reliably follow instructions plays an important role in supporting the reasoning performance of language models.

Context:
E1 began working for the employer of E4 on 01 September 2000.
E1 finished working for the employer of E4 on 31 August 2001.
E1 began working for the employer of E5 on 01 September 2001.
E1 finished working for the employer of E5 on 31 August 2005.
E1 worked for the employer of E3 from 01 September 2005 to 30 May 2009.
E1 began working for the employer of E3 on 01 June 2009.
E1 finished working for the employer of E3 on 28 September 2010.
E1 worked for the employer of E3 from 01 September 2011 to 24 March 2012.
E1 began working for the employer of E2 on 01 May 2013.
E1 finished working for the employer of E2 on 26 May 2016.
E1 worked for the employer of E2 from 01 January 2017 to 31 May 2019.
E1 worked for the employer of E6 from 01 November 2021 to 30 July 2022.
Question:
What was the employer that E1 worked for for the second shortest time? Considering all periods during which E1 worked for it.
Answer:
E4

Figure 3: A sample from ETRQA dataset.

Answer the question based on the context:
`{context}`
`{answer format instruction}`
Please provide the answer directly, without explanation.
Question: `{question}`
Answer:

Figure 4: Zero-shot prompt template

Example:
Answer the question based on the context:
`{context}`
`{answer format instruction}`
Please provide the answer directly, without explanation.
Question: `{question}`
Answer: `{answer}`

...

Refer to the examples above.
Answer the question based on the context:
`{context}`
`{answer format instruction}`
Please provide the answer directly, without explanation.
Question: `{question}`
Answer:

Figure 5: Few-shot prompt template

Answer the question based on the context:
`{context}`
`{answer format instruction}`
Please think step by step, and at the end of your reasoning, use “The final answer is: ” followed by the answer in the correct format, then stop reasoning.
Question: `{question}`
Let’s think step by step.

Figure 6: Zero-shot CoT prompt template

Example:
Answer the question based on the context:
`{context}`
`{answer format instruction}`
Please think step by step, and at the end of your reasoning, use ’The final answer is:’ followed by the answer in the correct format, then stop reasoning.
Question: `{question}`
Let’s think step by step.
`{CoT}`
The final answer is: `{answer}`

...

Refer to the examples above.
Answer the question based on the context:
`{context}`
`{answer format instruction}`
Please think step by step, and at the end of your reasoning, use ’The final answer is:’ followed by the answer in the correct format, then stop reasoning.
Question: `{question}`
Let’s think step by step.

Figure 7: Few-shot CoT prompt template

Temporal Event Type	**Example Template**	**Example**
Temporal interval event	$\{s\}$ began holding the position of $\{o\}$ in/on $\{t_s\}$. $\{s\}$ finished holding the position of $\{o\}$ in/on $\{t_e\}$. $\{s\}$ held the position of $\{o\}$ from $\{t_s\}$ to $\{t_e\}$.	E1 began holding the position of E2 in 2002. E1 finished holding the position of E2 in 2005. E1 held the position of E2 from 2002 to 2005.
Temporal point event	$\{s\}$ received the award of $\{o\}$ in/on $\{t_p\}$.	E1 received the award of E2 in 2002.

Table 7: Example templates for textualizing the structured temporal event.

Temporal Event Type	Wikidata Id	Relation
Temporal interval event	P39	position held
	P108	employer
	P106	occupation
	P54	member of sports team
	P102	member of political party
	P463	member of
	P27	country of citizenship
	P937	work location
	P551	residence
	P69	educated at
	P6087	coach of sports team
	P1308	position holder
	P286	head coach
	P6	head of government
	P35	head of state
	P488	chairperson
Temporal point event	P166	award received
	P1411	nominated for
	P512	academic degree
	P2522	competition won
	P793	significant event
	P410	military or police rank

Table 8: Representative relations for temporal interval events and temporal point events in Wikidata

Method	Error Type				Total Errors
	Incorrect Answer	Incomplete Answer	Uncertainty Error	Instruction-Following Error	
Llama-3.1-8B	66.29%	2.66%	0.02%	31.03%	14,306
Llama-3.1-8B-Instruct	37.76%	3.03%	0.08%	59.13%	13,383
Llama-3.1-70B	71.89%	7.34%	1.41%	19.36%	9,577
Llama-3.1-70B-Instruct	78.91%	8.91%	0.54%	11.64%	4,614
Qwen2.5-7B	83.67%	6.80%	0.45%	9.08%	11,710
Qwen2.5-7B-Instruct	89.84%	8.91%	0.35%	0.90%	9,898
Qwen2.5-72B	85.65%	6.42%	0.46%	7.47%	7,994
Qwen2.5-72B-Instruct	91.11%	8.59%	0.18%	0.11%	4,388

Table 9: Error analysis of different models under the zero-shot CoT setting. We report the total number of errors for each model and the proportion of each error type.

Question Composition	**Example Template**	**Example Question**	**Answer Type**
Temporal constraint	What distinct positions did $\{s\}$ hold $\{tc\}$?	What distinct positions did E1 hold after 2002?	Entity (multiple)
Temporal constraint	How many distinct positions did $\{s\}$ hold $\{tc\}$?	How many distinct positions did E1 hold after 2002?	Numeric
Order	What was the $\{ord\}$ position $\{s\}$ held?	What was the first position E1 held?	Entity
Order	During which period did $\{s\}$ hold the position for the $\{ord\}$ time?	During which period did E1 hold the position for the first time?	Time interval
Order	How long did $\{s\}$ hold the position for the $\{ord\}$ time?	How long did E1 hold the position for the first time?	Time span
Duration	What was the position that $\{s\}$ held for the $\{dur\}$ time? Considering all periods during which $\{s\}$ held it.	What was the position that E1 held for the longest time? Considering all periods during which E1 held it.	Entity
Duration	During which period did $\{s\}$ hold the position for the $\{dur\}$ time? List all periods during which $\{s\}$ held it.	During which period did E1 hold the position for the longest time? List all periods during which E1 held it.	Time interval (multiple)
Duration	What was the total duration $\{s\}$ held the position for the $\{dur\}$ time? Summing all periods during which $\{s\}$ held it.	What was the total duration E1 held the position for the longest time? Summing all periods during which E1 held it.	Time span
Temporal constraint + Order	What was the $\{ord\}$ position $\{s\}$ held $\{tc\}$?	What was the first position E1 held after 2002?	Entity
Temporal constraint + Order	During which period did $\{s\}$ hold the position for the $\{ord\}$ time $\{tc\}$? only counting the portions $\{tc\}$.	During which period did E1 hold the position for the first time after 2002? Only counting the portions after 2002.	Time interval
Temporal constraint + Order	How long did $\{s\}$ hold the position for the $\{ord\}$ time $\{tc\}$? only counting the portions $\{tc\}$.	How long did E1 hold the position for the first time after 2002? Only counting the portions after 2002.	Time span
Temporal constraint + Duration	What was the position that $\{s\}$ held for the $\{dur\}$ time $\{tc\}$? Considering all periods during which $\{s\}$ held it but only counting the portions $\{tc\}$.	What was the position that E1 held for the longest time after 2002? Considering all periods during which E1 held it but only counting the portions after 2002.	Entity
Temporal constraint + Duration	During which period did $\{s\}$ hold the position for the $\{dur\}$ time $\{tc\}$? List all periods during which $\{s\}$ held it but only counting the portions $\{tc\}$.	During which period did E1 hold the position for the longest time after 2002? Considering all periods during which E1 held it but only counting the portions after 2002.	Time interval (multiple)
Temporal constraint + Duration	What was the total duration $\{s\}$ held the position for the $\{dur\}$ time $\{tc\}$? Summing all periods during which $\{s\}$ held it but only counting the portions $\{tc\}$.	What was the total duration E1 held the position for the longest time after 2002? Summing all periods during which E1 held it but only counting the portions after 2002.	Time span
Temporal constraint	What distinct positions did $\{s\}$ begin holding $\{tc\}$?	What distinct positions did E1 begin holding after 2002?	Entity (multiple)
Temporal constraint	How many distinct positions did $\{s\}$ begin holding $\{tc\}$?	How many distinct positions did E1 begin holding after 2002?	Numeric
Temporal constraint + Order	What was the $\{ord\}$ position $\{s\}$ began holding $\{tc\}$?	What was the first position E1 began holding after 2002?	Entity
Temporal constraint + Order	When did $\{s\}$ begin holding the position for the $\{ord\}$ time $\{tc\}$?	When did E1 begin holding the position for the first time after 2002?	Time point
Temporal constraint	During which period did $\{s\}$ hold the position of $\{o\}$ $\{tc\}$? Considering all periods during which $\{s\}$ held it but only counting the portions $\{tc\}$.	During which period did E1 hold the position of E2 after 2002? Considering all periods during which E1 held it but only counting the portions after 2002.	Time interval (multiple)
Temporal constraint	How many times did $\{s\}$ hold the position of $\{o\}$ $\{tc\}$?	How many times did E1 hold the position of E2 after 2002?	Numeric
Order	During which period did $\{s\}$ hold the position of $\{o\}$ for the $\{ord\}$ time?	During which period did E1 hold the position of E2 for the first time?	Time interval
Order	How long did $\{s\}$ hold the position of $\{o\}$ for the $\{ord\}$ time?	How long did E1 hold the position of E2 for the first time?	Time span
Duration	During which period did $\{s\}$ hold the position of $\{o\}$ for the $\{dur\}$ time?	During which period did E1 hold the position of E2 for the longest time?	Time interval
Duration	How long did $\{s\}$ hold the position of $\{o\}$ for the $\{dur\}$ time?	How long did E1 hold the position of E2 for the longest time?	Time span
Temporal constraint + Order	During which period did $\{s\}$ hold the position of $\{o\}$ for the $\{ord\}$ time $\{tc\}$? only counting the portions $\{tc\}$.	During which period did E1 hold the position of E2 for the first time after 2002? only counting the portions after 2002.	Time interval
Temporal constraint + Order	How long did $\{s\}$ hold the position of $\{o\}$ for the $\{ord\}$ time $\{tc\}$? only counting the portions $\{tc\}$.	How long did E1 hold the position of E2 for the first time after 2002? only counting the portions after 2002.	Time span
Temporal constraint + Duration	During which period did $\{s\}$ hold the position of $\{o\}$ for the $\{dur\}$ time $\{tc\}$? only counting the portions $\{tc\}$.	During which period did E1 hold the position of E2 for the longest time after 2002? only counting the portions after 2002.	Time interval
Temporal constraint + Duration	How long did $\{s\}$ hold the position of $\{o\}$ for the $\{dur\}$ time $\{tc\}$? only counting the portions $\{tc\}$.	How long did E1 hold the position of E2 for the longest time after 2002? only counting the portions after 2002.	Time span
Temporal constraint	When did $\{s\}$ begin holding the position of $\{o\}$ $\{tc\}$? List all if began multiple times.	When did E1 begin c E2 after 2002? List all if began multiple times.	Time point (multiple)
Temporal constraint	How many times did $\{s\}$ begin holding the position of $\{o\}$ $\{tc\}$?	How many times did E1 begin holding the position of E2 after 2002?	Numeric
Order	When did $\{s\}$ begin holding the position of $\{o\}$ for the $\{ord\}$ time?	When did E1 begin holding the position of E2 for the first time?	Time point
Temporal constraint + Order	When did $\{s\}$ begin holding the position of $\{o\}$ for the $\{ord\}$ time $\{tc\}$?	When did E1 begin holding the position of E2 for the first time after 2002?	Time point

Table 10: Example question templates for questions based on temporal interval event groups.

Question Composition	Example Template	Example Question	Answer Type
Temporal constraint	What awards did $\{s\}$ receive $\{tc\}$? Don't need deduplicate.	What awards did E1 receive after 2002? Don't need deduplicate.	Entity (multiple)
Temporal constraint	How many awards did $\{s\}$ receive $\{tc\}$? Don't need deduplicate.	How many awards did E1 receive after 2002? Don't need deduplicate.	Numeric
Temporal constraint + Order	What was the $\{ord\}$ award $\{s\}$ received $\{tc\}$?	What was the first award E1 received after 2002?	Entity
Order	When did $\{s\}$ receive the award for the $\{ord\}$ time?	When did E1 receive the award for the first time?	Time point
Temporal constraint + Order	When did $\{s\}$ receive the award for the $\{ord\}$ time $\{tc\}$?	When did E1 receive the award for the first time after 2002?	Time point
Temporal constraint	When did $\{s\}$ receive the award of $\{o\}$ $\{tc\}$? List all if received multiple times.	When did E1 receive the award of E2 after 2002? List all if received multiple times.	Time point (multiple)
Temporal constraint	How many times did $\{s\}$ receive the award of $\{o\}$ $\{tc\}$?	How many times did E1 receive the award of E2 after 2002?	Numeric
Order	When did $\{s\}$ receive the award of $\{o\}$ for the $\{ord\}$ time?	When did E1 receive the award of E2 for the first time?	Time point
Temporal constraint + Order	When did $\{s\}$ receive the award of $\{o\}$ for the $\{ord\}$ time $\{tc\}$?	When did E1 receive the award of E2 for the first time after 2002?	Time point

Table 11: Example question templates for questions based on temporal point event groups.

Method	Non-Compound						Compound							Overall
	E.	I.	R.	O.	D.	Avg.	E. + O.	I. + O.	R. + O.	E. + D.	I. + D.	R. + D.	Avg.	
Llama-3.1-8B														
ZS	21.6	17.8	14.6	36.3	11.3	19.7	29.6	23.0	22.8	17.8	15.4	11.3	22.0	20.9
FS	38.7	32.9	33.3	53.2	17.7	35.6	38.5	33.3	31.2	21.1	20.5	12.3	29.8	32.6
ZS CoT	18.6	12.7	12.8	33.0	17.0	16.1	25.3	20.0	16.7	15.2	14.7	12.7	19.1	17.7
FS CoT	48.1	46.8	46.0	65.7	30.7	48.0	45.4	44.1	39.7	26.8	26.8	18.0	38.0	42.8
Llama-3.1-70B														
ZS	51.8	41.6	36.7	61.4	19.7	44.7	43.7	38.5	32.1	24.3	23.6	15.0	34.0	39.2
FS	53.3	48.0	36.3	78.7	35.0	49.9	54.2	51.7	45.1	31.3	28.0	17.3	43.9	46.8
ZS CoT	46.5	48.4	42.7	70.4	30.7	48.2	51.5	48.8	35.4	28.2	30.3	19.7	41.7	44.9
FS CoT	78.6	73.7	70.7	80.2	60.3	74.7	76.6	72.0	65.0	53.2	52.0	43.3	65.7	70.1
Qwen2.5-7B														
ZS	33.4	27.0	25.3	44.4	13.3	29.5	32.9	23.5	24.6	16.8	16.1	11.7	23.0	26.2
FS	39.1	33.0	29.3	50.8	22.7	35.3	37.9	31.0	29.2	20.0	21.0	13.0	28.6	31.9
ZS CoT	38.0	29.0	33.0	52.0	28.0	33.8	40.8	34.0	26.9	25.9	23.5	16.7	31.5	32.6
FS CoT	52.9	43.1	43.3	59.0	29.0	46.7	50.9	42.8	36.2	30.1	26.6	24.0	39.0	42.7
Qwen2.5-14B														
ZS	42.7	32.7	30.1	51.7	18.3	36.2	39.3	31.8	27.7	24.0	21.3	16.3	29.7	32.8
FS	46.8	37.9	34.7	61.2	19.3	41.2	42.7	36.9	30.9	25.1	22.8	16.3	33.0	37.0
ZS CoT	54.7	42.1	42.9	72.3	44.0	48.2	56.1	46.4	35.4	37.2	34.5	28.3	43.8	45.9
FS CoT	64.5	58.6	55.3	66.9	43.7	60.0	64.1	57.6	48.3	42.8	38.7	29.3	52.2	55.9
Qwen2.5-32B														
ZS	54.6	38.8	38.6	63.2	23.3	44.6	47.4	41.7	38.1	30.8	25.5	20.7	37.7	41.0
FS	57.3	44.8	40.6	70.4	21.3	49.0	48.0	41.6	35.3	28.1	24.4	17.3	36.9	42.8
ZS CoT	73.4	63.6	61.5	**86.2**	50.7	67.4	69.4	63.6	52.3	49.9	50.0	41.7	59.2	63.2
FS CoT	76.0	72.3	71.8	77.5	57.0	73.1	77.2	71.0	64.6	**58.6**	52.1	44.3	66.0	69.5
Qwen2.5-72B														
ZS	57.7	46.0	43.2	67.0	22.0	49.8	48.9	42.5	35.4	27.4	25.1	19.0	37.6	43.5
FS	61.2	48.5	46.7	74.2	25.0	53.0	51.9	46.6	37.8	30.3	27.2	19.7	40.7	46.7
ZS CoT	64.4	55.1	43.2	76.7	45.3	57.6	56.3	57.4	41.3	40.7	43.6	29.7	50.6	54.0
FS CoT	**79.1**	**75.1**	**73.0**	81.3	**63.0**	**76.0**	**78.8**	**73.1**	**67.8**	58.3	**54.2**	**46.3**	**67.8**	**71.8**

Table 12: Experimental results of base models under four common prompting strategies. The abbreviations E., I., R., O., D. refer to Explicit, Implicit, Relative, Order, and Duration respectively. The results in bold (resp. underline) denote the best (resp. second) results.