Living in the Moment:
Can Large Language Models Grasp Co-Temporal Reasoning?

Anonymous ACL submission

Abstract

Temporal reasoning is fundamental for large language models (LLMs) to comprehend the world. Current temporal reasoning datasets are limited to questions about single or isolated events, falling short in mirroring the realistic temporal characteristics involving concurrent nature and intricate temporal interconnections. In this paper, we introduce CoTempQA, a comprehensive co-temporal Question Answering (QA) benchmark containing four co-temporal scenarios (Equal, Overlap, During, Mix) with 4,749 samples for evaluating the co-temporal comprehension and reasoning abilities of LLMs. Our extensive experiments reveal a significant gap between the performance of current LLMs and human-level reasoning on CoTempQA tasks. Even when enhanced with Chain of Thought (CoT) methodologies, models consistently struggle with our task. In our preliminary exploration, we discovered that mathematical reasoning plays a significant role in handling co-temporal events and proposed a strategy to boost LLMs' co-temporal reasoning from a mathematical perspective. We hope that our CoTempQA datasets will encourage further advancements in improving the co-temporal reasoning capabilities of LLMs.

1 Introduction

Recent advanced Large Language Models (LLMs) like GPT-4 (OpenAI, 2023) have shown impressive capabilities in understanding, generating, and reasoning about natural language (Wei et al., 2022a; Zhao et al., 2023; Chang et al., 2023). Despite their advancements, these models fall short in mastering temporal reasoning (Chu et al., 2023), which is fundamental for humans to comprehend the world and distinguish daily events (Chen et al., 2021; Tan et al., 2023), requiring a complex integration of capabilities, involving implicit arithmetic calculations (Zhu et al., 2023a), understanding logical implications (Wei et al., 2022c), and leveraging extensive world knowledge (Chu et al., 2023).

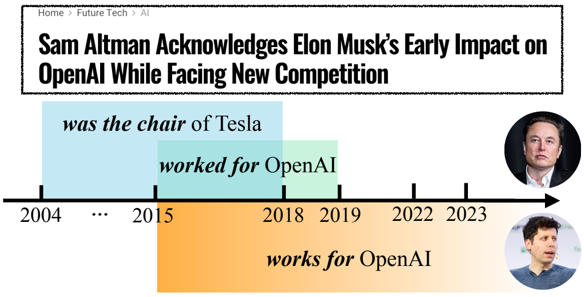


Figure 1: Understanding concurrent is crucial for us to understand how individuals navigate and influence diverse aspects of real-world scenarios. For instance, when *Elon Musk* was the chair of *Tesla*, he also worked for *OpenAI*. Concurrently, *Sam Altman* was working for *OpenAI*, too. Their simultaneous experiences greatly influenced subsequent decision-making at *OpenAI*.

Current studies in temporal reasoning mainly focus on time-sensitive question-answering (TSQA). Chen et al. (2021) first introduced the TimeQA dataset, constructing time-evolving facts for a given subject and formulating questions based on the specific timestamp within the evolutionary facts. TempLAMA (Dhingra et al., 2022) extracted structured facts from the Wikidata Knowledge Base (Vrandečić and Krötzsch, 2014) for closed-book TSQA. Furthermore, TempReason (Tan et al., 2023) translated explicit temporal expressions into the implicit event information within questions, offering a more comprehensive evaluation framework of TSQA. Given the fact *"Elon musk held the position of Tesla's chairman from 2004 to 2018"*, the models are tasked with accurately interpreting and responding to time specifiers in the questions, i.e., *"Which position did Elon Musk hold in 2005?"* in TimeQA (Chen et al., 2021) or *"Which position did Elon Musk held before he worked for OpenAI?"* in TempReason (Tan et al., 2023).

The datasets mentioned above provide a straightforward way to evaluate LLMs' capabilities in tem-

Datasets	Question	Answer
TimeQA (2021)	Which school did Sam Altman attended in 2005? Which position did Elon Musk hold in 2005?	Stanford University chairman of Tesla
TempLAMA (2022)	In 2005, Sam Altman attended _X_. In 2005, Elon Musk hold the position of _X_.	Stanford University chairman of Tesla
TempReason (2023)	Which school did Sam Altman attend before he held the position of president of Y Combinator? Which position did Elon Musk held before he worked for OpenAI?	Standford University chairman of Tesla
CoTempQA (ours)	When Elon Musk was working for OpenAI, where did he work for within the same time interval? (**Overlap**) While Elon Musk was working for OpenAI, where did Sam Altman work for concurrently? (**During**)	Tesla, SpaceX OpenAI

Table 1: Example questions of prior TSQA datasets and our CoTempQA datasets. Our question is divided into two components: the condition C is marked in blue, and the inquiry A is highlighted in red.

Interpretation	Relation
●——x——● ○——y——○	x is equal to y
●—x—● ○—y—○	x overlaps with y
●—x—● ○——y——○	x during y

Table 2: Interpretation of three co-temporal relations.

Mode	Questions	Subjects	#Facts	#Answers
Equal	436	401	11.65	1.17
Overlap	653	591	14.51	1.23
During	3,096	2,161	15.05	1.33
Mix	563	434	12.54	2.27
Total	4,748	3,587	14.45	1.41

Table 3: Statistics of our datasets. #Facts and #Answers represent the average number of facts and answers within the subject and question, respectively.

poral reasoning. However, as LLMs evolve, there is an urgent need to evaluate their proficiency in more realistic scenarios. As shown in Figure 1, the reality might present a more intricate and multifaceted nature, involving concurrent events and complex temporal interconnections over time (UzZaman et al., 2012). Current datasets mainly question single or isolated events and might not fully reflect the realistic temporal characteristics. Therefore, we create the Co-Temporal QA (CoTempQA) datasets to complement existing corpora by focusing on the concurrent nature of time and co-temporal relations in real-world situations.

Experiments conducted on both closed-book and open-book QA settings across 14 large language models reveal that even advanced models GPT-4 is well below a satisfactory co-temporal reasoning performance. Specifically, GPT-4 achieves an overall score of 54.7, and the best open-source LLM is 30.1, which significantly falls behind the human performance of 92.8. We also observe that the representative reasoning enhancement strategies, e.g., Chain-of-Thought (CoT) (Wei et al., 2022b), fail to consistently improve and even reduce the temporal reasoning capabilities of LLMs in some scenarios.

Throughout the investigation on our CoTempQA, we observed that mathematical reasoning plays a crucial role in handling co-temporal events. Building on this insight, we propose a simple but effective Math-reasoning CoT (Mr-CoT) strategy to boost the co-temporal reasoning capability of LLMs, achieving a remarkable 10.8 point improvement over existing baselines. However, it is important to note that there remains a nonnegligible gap between the performance of our proposed Mr-CoT and human-level reasoning in handling complex, concurrent temporal relations. We hope our research could inspire more great works to improve the co-temporal ability of LLMs.

2 The CoTempQA Datasets

2.1 The Taxonomy of Co-temporal Relations

Co-temporal relations are fundamental to understanding how events interconnect in time. These relationships highlight when facts or events happen simultaneously, which can be categorized into three distinct types (Pustejovsky et al., 2003), as shown in Table 2. Each of them represents a unique manner, whether events coincide with or overlap with each other in the temporal aspect. We divide these relations into four different scenarios below:

- **Equal:** Facts occur simultaneously, representing a strict co-temporal relationship. This is the simplest form of co-temporality but is essential for understanding facts that happen concurrently without any duration differences.
- **Overlap:** One fact is entirely contained within the timeline of another, reflecting a more complex interaction of timelines.
- **During:** Facts that partially coincide in time. This scenario is more common in real-world set-

Algorithm 1 Identifying Co-temporal Facts

```
1:  Input: Set of facts $F$, each fact as $(s, r, o, t_s, t_e)$
2:  Output: Set of co-temporal facts with their minimum temporal units
3:  function MINMAXTIME($f_i$, $f_j$)
4:      $(s_i, r_i, o_i, t_{s_i}, t_{e_i}) \leftarrow f_i$
5:      $(s_j, r_j, o_j, t_{s_j}, t_{e_j}) \leftarrow f_j$
6:      $start \leftarrow \max(t_{si}, t_{sj})$
7:      $end \leftarrow \min(t_{ei}, t_{ej})$
8:      $T_{\text{min}} \leftarrow (start, end)$
9:      if $start \leq end$ then return $T_{\text{min}}$
10:     else return None
11:     end if
12: end function
13: $R \leftarrow$ empty set                ▷ $R$ is the set of co-temporal facts
14: for each $f_i$ in $F$ do
15:     for each $f_j$ in $F$ where $f_i \neq f_j$ do
16:         $T_{\text{min}} \leftarrow$ MINMAXTIME($f_i$, $f_j$)
17:         if $T_{\text{min}}$ is not None then
18:             $R \leftarrow R \cup \{(f_i, f_j, T_{\text{min}})\}$
19:         end if
20:     end for
21: end for
22: return $R$
```

tings, where facts often intersect for a part of their duration without overlapping.

- **Mix:** A combination of the three types above. This category is particularly challenging as it involves the complexity and variability of real-world temporal relationships, necessitating a comprehensive level of co-temporal reasoning.

2.2 Structuring Temporal Facts

We utilize the Wikidata (Vrandečić and Krötzsch, 2014) dump of September 20, 2023 as our knowledge source for extracting time-dependent facts. Following Dhingra et al. (2022) and Tan et al. (2023), we focus on nine time-sensitive entity relations and keep a maximum of 2,000 subjects for each relation type. To structure the information, we transform the knowledge triples and qualifiers into a quintuplet format of (s, r, o, t_s, t_e), where s is the subject, r is the relation, o is the object, t_s and t_e are the start time and end time. We group all the temporal facts by subject, denoted as $S = \{(s, r_i, o_i, t_{s_i}, t_{e_i}) | i \in 1 \dots N\}$, where N is the number of facts within a group. We keep the groups that contain three or more temporal facts.

2.3 Extracting Co-temporal Facts

Building on our approach to structuring time-dependent facts from Wikidata, we compare the timestamps of different facts to find overlaps. Given a fact f_i, we distinguish co-temporal fact f_j in five scenarios based on the combinations where either the subject, relation, or object changes while the other elements remain constant or change, i.e., $(\mathcal{S}, \mathcal{R}, \overline{\mathcal{O}})$, $(\mathcal{S}, \overline{\mathcal{R}}, \overline{\mathcal{O}})$, $(\overline{\mathcal{S}}, \mathcal{R}, \mathcal{O})$, $(\overline{\mathcal{S}}, \mathcal{R}, \overline{\mathcal{O}})$, $(\overline{\mathcal{S}}, \overline{\mathcal{R}}, \overline{\mathcal{O}})$. An overline indicates a change in a specific element when comparing two related facts. For instance, $(\mathcal{S}, \mathcal{R}, \overline{\mathcal{O}})$ represents the scenario where the subject and relationship remain the same. Given the fact (*Elon Musk*, *employer*, *SpaceX*), the fact (*Elon Musk*, *employer*, *OpenAI*) fits this pattern. We exclude the scenarios $(\mathcal{S}, \overline{\mathcal{R}}, \mathcal{O})$ and $(\overline{\mathcal{S}}, \mathcal{R}, \overline{\mathcal{O}})$ since it is unrealistic for the same subject and object to have different relationships, or for the same object to have the same relationship with different subjects concurrently. The detailed illustrations are shown in Appendix B. Taken $(\overline{\mathcal{S}}, \overline{\mathcal{R}}, \overline{\mathcal{O}})$ as an example, we detail the extraction of co-temporal facts in the MINMAXTIME function (lines 3-11) from Algorithm 1. This framework identifies the complex co-temporal relations between events, allowing for a more intuitive understanding of how multiple events and states are interrelated in the temporal dimension.

2.4 QA Pairs Construction

Upon identifying co-temporal facts $(f_i, f_j, T_{\text{min}})$, we construct the query Q, which comprises two fundamental components: a condition C and an inquiry A. From the co-temporal facts, we select the intersection fact as the condition fact f_{cond} and the other as the query fact f_{query}. Given the facts f_{cond} and f_{query}, we construct questions for the object by manually-defined question templates, which can be found in Table 8. Based on the temporal relations identified through T_{min}, we categorize the tasks into four distinct classes: **Equal**, **Overlap**, **During**, and **Mix**. We acknowledge that in real life, multiple events can happen simultaneously, and therefore, a single temporal question might have multiple correct answers. To address this, we aggregate all valid answers for query Q to a set. As detailed in Table 3, the average number of our answers within the question is 1.42.

3 The Performance of LLMs on COTEMPQA

3.1 Experimental Setup

We investigate the co-temporal reasoning abilities of large language models within two problem settings: (1) **C**losed-**B**ook **QA** (**CBQA**) is widely recognized task format in time-sensitive QA research (Dhingra et al., 2022; Liska et al., 2022; Tan et al., 2023). In this setting, the language model is given only the question and tasked with generating the answer without relying on external natural language texts. The primary challenge here involves the retention and temporal reasoning of knowledge pertinent to the question. (2) In the **O**pen-**B**ook **QA** (**OBQA**) setting, we provide all the relevant temporal facts within the group $S = \{(s, r_i, o_i, t_{s_i}, t_{e_i}) | i \in 1...N\}$ in a structured format directly into the prompt, which is in contrast to previous studies (Chen et al., 2021; Wei et al., 2023) that utilized Wikipedia as the knowledge base. This process shifts the evaluation's emphasis towards the reasoning process itself, thereby minimizing the influence of the model's inherent factual extraction capabilities on the outcomes (Tan et al., 2023; Chu et al., 2023). Here, the language model is tasked with deriving answers by considering the time ranges associated with all potential answers.

3.2 LLMs for Evaluation

We perform comprehensive experiments on 14 representative large language models including (1) **ChatGPT** (Ouyang et al., 2022) ChatGPT is a chat model aligned through Supervised Fine-tuning (SFT) and Reinforcement Learning from Human Feedback (RLHF). GPT-4 is an upgraded version of ChatGPT with enhanced reasoning capabilities, making it the most powerful LLM. Since the model is constantly updated, we used the `gpt3.5-turbo-0613` and `gpt4-0613` for consistent evaluation. (2) **LLaMA2** (Touvron et al., 2023) LLaMA2 is one of the most popular open-source foundation models trained on 2T tokens with efficient group query attention (Ainslie et al., 2023). (3) **Code-LLaMA** (Roziere et al., 2023) Code-LLaMA models is a code generation model built on LLaMA2 and further trained on 500B tokens of code. (4) **WizardMath** (Luo et al., 2023a) WizardMath is also built on LLaMA2 and further trained on their proposed Reinforcement Learning from Evol-Instruct Feedback (RLEIF) (Xu et al., 2023) to enhance the mathematical reasoning abilities of LLaMA2. (5) **WizardCoder** (Luo et al., 2023b) WizardCoder, similar to WizardMath, adapts the RLEIF method to the domain of code. The implementation details of our experiments are shown in Appendix A.

3.3 Evaluation Metrics

Prior works followed the SQuAD benchmark's evaluation protocol (Rajpurkar et al., 2016), using exact match (EM) and token-level F_1 score. These metrics calculate the highest scores across all references, tending to overestimate performance in task settings involving questions with multiple possible answers. Following Zhong et al. (2022), we adopt a stricter **Acc.** score, where a prediction is correct only if it aligns with all the gold answers for a question. Additionally, we also evaluate our methods by answer-level F_1 score (**F_1**), which is a stricter metric compared to token-level F_1 score.

3.4 Results and Analysis

The main results are shown in Table 4. We report human performance to serve as an upper bound. From the results, we can observe:

LLMs partially grasp co-temporal reasoning Our analysis reveals that, despite GPT-4 exhibiting the best performance among all LLMs, there is still a considerable disparity compared to human performance (54.7 vs. 92.8), indicating significant potential for further improvement in co-temporal reasoning. Meanwhile, there is a general trend in improving co-temporal reasoning capabilities as the size of the models' parameters increases. We also discover that models exhibit different reasoning capabilities in different co-temporal scenarios. Take GPT-4 for further illustration, in the simple co-temporal reasoning task, i.e., the **Equal** scenario, GPT-4 demonstrates strong performance, achieving a 92.7 score overall. However, its performance significantly declines in more complex scenarios. Specifically, in the **Overlap** category, GPT-4's accuracy falls to 59.4, decreasing further to 50.1 in the **During** category. In the most challenging category, **Mix**, which combines various temporal relations, GPT-4's performance drops to 45.0. This phenomenon indicates that existing LLMs can effectively process and reason about straightforward concurrent events. However, they encounter difficulties in more complex tasks that require a deeper understanding and comprehension of co-temporal reasoning.

CBQA is more challenging for LLMs LLMs exhibit significantly weaker performance in the **CBQA** compared to the **OBQA**, as reflected in the GPT-4's performance (14.5 vs. 54.7). Interestingly, GPT-4 is outperformed by GPT-3.5 in **CBQA**. Our error analysis indicates that GPT-4 often responds with *"uncertain"* when unsure, unlike GPT-3.5, which tends to provide direct answers. This discovery is also found in previous works (OpenAI, 2023; Wei et al., 2023). This characteristic hin-

Model	Equal			Overlap			During			Mix			Overall
	Acc.	F_1	Avg.	Acc.	F_1	Avg.	Acc.	F_1	Avg.	Acc.	F_1	Avg.	
The Closed Book Question Answer (CBQA) setting													
GPT-3.5-Turbo	**13.8**	**14.8**	**14.3**	11.3	**14.3**	12.8	**15.0**	**22.9**	**18.9**	**0.0**	**15.5**	**7.7**	**16.3**
GPT-4	11.2	12.3	11.8	**11.5**	14.0	**12.7**	14.8	18.5	16.7	**0.0**	13.6	6.8	14.5
The Open Book Question Answer (OBQA) setting													
GPT-3.5-Turbo	59.4	66.3	62.8	40.1	48.5	44.3	31.5	42.9	37.2	0.7	46.1	23.4	38.9
GPT-4	**91.1**	**94.3**	**92.7**	**55.3**	**63.5**	**59.4**	**44.3**	**55.8**	**50.1**	**23.4**	**66.5**	**45.0**	**54.7**
CodeLLaMA-7B	6.4	**27.7**	**17.0**	3.1	14.6	8.8	3.1	15.8	9.5	**2.0**	**24.1**	**13.0**	10.5
WizardCoder-7B	9.2	21.1	15.1	4.7	14.8	9.8	6.3	15.9	11.1	0.5	20.4	10.5	11.2
LLaMA-7B	4.1	18.9	11.5	4.7	**19.5**	12.1	4.5	19.5	12.0	0.2	23.8	12.0	12.0
WizardMath-7B	**12.4**	16.5	14.4	**9.2**	15.2	**12.2**	**11.6**	**20.5**	**16.0**	0.4	22.0	11.2	**14.8**
CodeLLaMA-13B	7.6	28.3	18.0	4.1	17.0	10.6	3.3	19.3	11.3	**3.2**	**28.6**	**15.9**	12.4
WizardCoder-13B	8.3	16.6	12.4	7.0	17.8	12.4	9.5	19.7	**14.6**	1.1	24.1	12.6	13.9
LLaMA-13B	11.2	**31.2**	21.2	5.8	**21.6**	**13.7**	5.0	**20.6**	12.8	1.1	26.9	14.0	13.8
WizardMath-13B	**23.9**	29.0	**26.4**	**10.9**	15.1	13.0	**11.7**	17.1	14.4	0.0	13.2	6.6	**14.4**
CodeLLaMA-34B	16.1	46.5	31.3	9.8	27.0	18.4	8.1	28.4	18.3	4.4	40.3	22.4	20.0
WizardCoder-34B	19.5	26.3	22.9	15.2	22.4	18.8	15.9	23.9	19.9	0.9	25.9	13.4	19.2
LLaMA-70B	11.9	41.7	26.8	10.0	32.5	21.2	9.4	33.5	21.4	**5.2**	**42.5**	**23.8**	22.2
WizardMath-70B	**36.7**	**46.8**	**41.8**	**23.6**	**33.7**	**28.6**	**25.5**	**37.1**	**31.3**	0.4	32.9	16.6	**30.1**
Human	97.0	98.3	97.7	91.1	93.5	92.3	82.0	87.0	84.5	88.0	96.2	92.1	92.8

Table 4: Experimental results of each model in the **CBQA** and **OBQA** settings of our proposed CoTempQA. Notably, we only report the performance of GPT-3.5 and GPT-4 in **CBQA** setting as the open-source LLMs are almost negligible here, and closed-book human evaluations largely depend on individual knowledge, leading to significant variations between different individuals. The best performance of each model is **bold**.

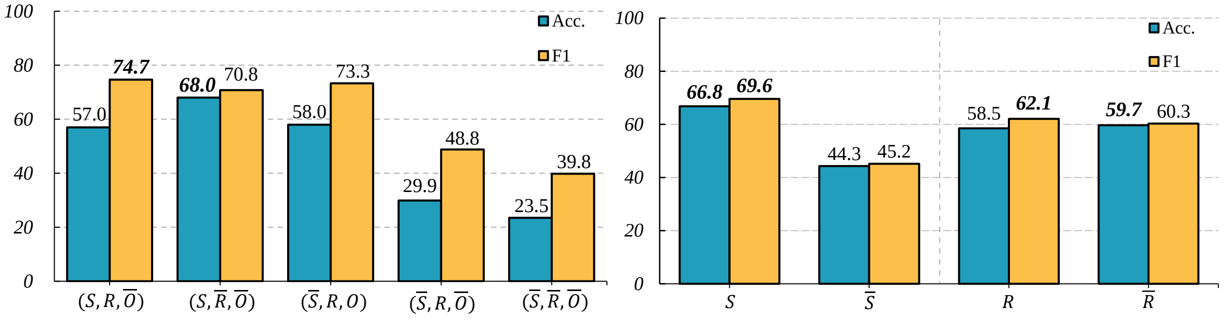
ders GPT-4's effectiveness in co-temporal **CBQA**, where precise answers are needed. While constructing our datasets, we concentrated on the top 2,000 subjects for each relationship type. These subjects are typically well-covered in pre-training stages, as Wikipedia is a significant part of their training data (Touvron et al., 2023). Despite this prior exposure, LLMs' reduced capability in **CBQA** underscores a challenging yet crucial situation, where LLMs frequently must respond to contemporaneous questions without supplementary context in the real-world application. Their ability to deliver precise answers relying solely on their pre-trained knowledge base is essential. This result underlines the need to enhance the co-temporal reasoning abilities of LLMs, empowering them to comprehend and reason about concurrent events.

Different aspects of capability benefit co-temporal reasoning differently We also explore the effects of augmenting the model's additional abilities using SFT on co-temporal reasoning. We observe that models specialized in mathematical reasoning, such as WizardMath-70B, show significant improvements in co-temporal reasoning, scoring 30.1, compared to the foundational LLaMA-70B model's 22.2, as detailed in Table 4. This enhanced capability suggests a strong correlation between the skills utilized in mathematical reasoning and those required for understanding and interpreting complex temporal relationships. However, LLMs with capabilities in code reasoning do not show consistent improvement in their performance. While these models are good at tasks involving sequential logic, their performance in co-temporal reasoning indicates that these skills may not directly translate to a robust understanding of simultaneous temporal relations.

3.5 Data Analysis

In Section 2.3, we categorize co-temporal facts into five scenarios. Building on this classification, this section delves into investigating how various types of fact elements influence LLMs' ability to perform co-temporal reasoning. To ensure fairness in our experiments, we excluded questions with multiple answers and standardized the number of questions across all co-temporal relations. Figure 2 illustrates GPT-4's performance with various element types. Additional results concerning different LLMs are presented in Table 7, and results consistently align with the findings shown below:

The influence of triple element types As observed in Figure 2a, the complexity of co-temporal reasoning for models increases with the number of changing elements. Among the scenarios, $(\overline{\mathcal{S}}, \mathcal{R}, \overline{\mathcal{O}})$, $(\overline{\mathcal{S}}, \overline{\mathcal{R}}, \overline{\mathcal{O}})$ are particularly challenging compared to others. It indicates that LLMs encounter significant challenges when dealing with scenarios of high complexity, where multiple elements undergo simultaneous changes. The analysis



(a) Triple Element Types

(b) Triplet Element Types

Figure 2: Performance of GPT-4 under different co-temporal element types in the **OBQA** setting of our CoTempQA. The overline indicates we changed the element in fact to others. The best performance of each element type is **bold**.

Model	Equal			Overlap			During			Mix			Overall
	Acc.	F_1	Avg.	Acc.	F_1	Avg.	Acc.	F_1	Avg.	Acc.	F_1	Avg.	
						The Closed Book Question Answer (CBQA) setting							
GPT-4	11.2	12.3	11.8	11.5	14.0	12.7	14.8	18.5	16.7	0.0	13.6	6.8	14.5
+ CoT	12.2	14.4	13.3	8.4	12.5	10.5	12.1	18.6	15.3	1.6	14.3	8.0	13.6
+ Fs	26.4	29.6	28.0	17.6	21.2	19.4	20.6	26.7	23.7	0.0	21.7	10.9	22.0
+ Fs&CoT	**32.1**	**35.2**	**33.6**	**19.9**	25.7	22.8	**23.2**	29.5	26.4	0.5	25.6	13.1	25.0
+ Fs&Mr-CoT	24.8	30.6	27.7	16.7	**29.9**	**23.3**	20.8	**35.7**	**28.2**	**3.9**	**31.7**	**17.8**	**26.3**
						The Open Book Question Answer (OBQA) setting							
GPT-4	91.1	94.3	92.7	55.3	63.5	59.4	44.3	55.8	50.1	23.4	66.5	45.0	54.7
+ CoT	87.8	90.0	88.9	46.2	58.7	52.5	43.5	57.0	50.2	29.5	71.6	50.5	54.1
+ Fs	87.4	91.4	89.4	62.6	72.5	67.6	55.9	68.6	62.2	30.6	71.9	51.2	64.2
+ Fs&CoT	**96.8**	**97.3**	**97.1**	61.3	71.4	66.3	55.7	69.4	62.5	32.1	73.2	52.7	65.0
+ Fs&Mr-CoT	95.9	97.2	96.5	**77.9**	**83.9**	**80.9**	**69.0**	**78.8**	**73.9**	**50.3**	**82.2**	**66.2**	**75.8**

Table 5: Performance of GPT-4 under Zero-shot CoT (CoT) prompting, Few-shot (Fs) prompting, Few-shot CoT (Fs&CoT) prompting and our proposed Few-Shot Mr-CoT (FS&Mr-CoT) prompting in **CBQA** and **OBQA**.

below further investigates which elements present the most significant challenges to co-temporal reasoning capabilities.

The influence of triplet element types In the left part of Figure 2b, we observe a notable decline in the model's performance, i.e., 22.5 point decrease in **Acc.** and 24.4 in $\mathbf{F_1}$ when it engages in reasoning involving multiple subjects. These findings highlight models' inherent difficulty when processing information from multiple concurrent subjects. While the reasoning process for handling multiple subjects shares similarities with single-subject scenarios, real-world situations are inherently more complex and variable. The model is required to integrate information across different subjects and understand complex relationships that extend beyond a single domain or context. On the other hand, in the right part of Figure 2b, we discover that the difference in the fact relation does not significantly impact the abilities of LLMs in co-temporal reasoning. This is attributed to the fact that relationship changes are not as complex as those involving multiple subjects in real-world scenarios (Huang and Chang, 2023), making them less challenging for the models' capabilities.

4 Making Language Models Better CoTempQA Responders

Previous research has demonstrated that the Chain-of-Thought (CoT) enables models to process complex reasoning tasks, such as mathematical and logical reasoning, in a step-by-step manner (Wei et al., 2022b). Motivated by this, we explore the application of CoT prompting to improve the capabilities of co-temporal reasoning in this section.

4.1 Math-reasoning CoT (Mr-CoT)

As indicated in Table 4, our research uncovers a mathematically oriented reasoning framework that can enrich the LLMs' understanding and handling of co-temporal reasoning tasks. Specifically, the WizardMath-70B model scores 30.1 overall, notably higher than the foundational LLaMA-70B model's score of 22.2. In light of this finding, we propose a Math-reasoning (Mr-CoT) instruction-based prompting, which can be used

Question	While Klaus Fuchs was working for Atomic Energy Research Establishment, which employer did Alexander Oppenheim work for during the identical time period?
Context	Alexander Oppenheim works for National University of Singapore from 1949 to 1965. Alexander Oppenheim works for University of Malaya from 1949 to 1965. Klaus Fuchs works for Atomic Energy Research Establishment from 1946 to 1950. Klaus Fuchs works for University of Birmingham from 1941 to 1944. ......
Gold Answer	**National University of Singapore and University of Malaya**
GPT-4	Answer: **National University of Singapore**
+ CoT	Reasoning: Klaus Fuchs worked for Atomic Energy Research Establishment from 1946 to 1950. Currently, Alexander Oppenheim worked for University of Ghana from 1968 to 1973. However, this does not overlap with the time Klaus Fuchs was at the Atomic Energy Research Establishment. Answer: **no information**
+ Fs	Answer: **University of Malaya**
+ Fs&CoT	Reasoning: Klaus Fuchs worked for Atomic Energy Research Establishment from 1946 to 1950. During this same period, Alexander Oppenheim worked for National University of Singapore, as his tenure with the university spanned from 1949 to 1965. Answer: **National University of Singapore**
+ Fs&Mr-CoT	The datetime can be formed as ((1946, None, None), (1950, None, None)). The content provided and related to the question can be structured as: (University of Malaya, (1949, None, None), (1965, None, None)) (National University of Singapore, (1949, None, None), (1965, None, None)) Given the ((1946, None, None), (1950, None, None)), compared with all contents related, we find that $[(1949, \text{None}, \text{None}) - (1965, \text{None}, \text{None})] \cap [(1946, \text{None}, \text{None}) - (1950, \text{None}, \text{None})] \neq \emptyset$ Reasoning: Klaus Fuchs worked for Atomic Energy Research Establishment from 1946 to 1950. Answer: **National University of Singapore and University of Malaya**

Table 6: Example inputs and outputs of GPT-4 with Zero-shot CoT (CoT) prompting, Few-shot (Fs) prompting, Few-shot CoT (Fs&CoT) prompting and Few-Shot Mr-CoT (Fs&Mr-CoT). Answers highlighted in blue are correct, whereas thoses marked in red are incorrect.

together with in-context learning and chain-of-thought prompting. As demonstrated in the bottom of Table 6, our framework consists of three steps: (1) establish the key datetime, (2) structure the relevant timeline, and (3) mathematically identify the overlap. This prompt aims to guide the LLMs towards approaching temporal reasoning problems through a mathematical perspective, aligning their problem-solving processes more closely with mathematical logic and principles.

4.2 Experimental Setup

We launch experiments under both zero-shot and few-shot settings. In the zero-shot CoT scenario, we use *Let's think step by step* (Kojima et al., 2022) after questions as the reasoning trigger. In contrast, the few-shot setting provides the model with several question-answer pairs as initial demonstrations. Specifically, for the few-shot CoT scenario, we manually create rationales for each task, which are used as demonstrations to guide the model in step-by-step reasoning. Further details on the instructions and demonstrations are available from Figure 3 to Figure 12 in Appendix C.

4.3 Results and Analysis

The results are presented in Table 5, and the output of GPT-4 to a range of prompts under different settings are shown in Table 6. From these tables, we can discover the following insights:

Inconsistency in the impact of existing CoT prompts on GPT-4 In the zero-shot scenario, improvements were inconsistent, with a notable 5.5 performance increase in the **Mix** task and a 3.8 decrease in the **Equal** task under the **OBQA** setting. This suggests that the impact of CoT prompts varies significantly based on the task type. Moreover, GPT-4 demonstrates an overall decline in performance on both **CBQA** and **OBQA** when complemented with CoT. In the few-shot scenario, while overall improvements exist due to CoT prompts, these are relatively modest, amounting to an average performance enhancement of 0.8 in **OBQA**. All results indicate that while existing CoT prompts can be beneficial, their effectiveness is nuanced and task-dependent.

Superiority of our proposed Mr-CoT Our method demonstrates significant superiority over existing reasoning enhancement strategies. Notably, MR-COT significantly enhances performance on the more challenging tasks, yielding improvements of 14.6, 11.4, and 13.5 on the tasks **Overlap**, **During**, and **Mix**, respectively in the **OBQA** setting. In the closed-book scenario, which is typically more challenging to improve, our method still achieves a 1.3 enhancement. However, it is observed that our method has a moderate effect on the **Equal** setting. We hypothesize that this is because this task is simple enough and does not require the additional complexity of mathematical reasoning. In such cases, this added complexity could be counterproductive. Despite these advancements, there is still a considerable gap compared to human-level reasoning, indicating the need for more effective methods to improve the model's co-temporal reasoning abilities.

5 Related Work

5.1 Temporal Reasoning Benchmarks

Temporal reasoning in natural language processing has seen significant advancements over the years. Early benchmarks, such as TimeBank (PUSTEJOVSKY, 2003), and TempEval-3 (UzZaman et al., 2012), lay the foundational work in this domain. They primarily focused on understanding temporal relationships between events in text, offering a preliminary framework for analyzing time in language models. However, recent years have witnessed a significant surge in developing time-sensitive question-answering datasets. These newer datasets, including MC-TACO (Zhou et al., 2019), SituatedQA (Zhang and Choi, 2021), TimeQA (Chen et al., 2021), TempLAMA (Dhingra et al., 2022), StreamingQA (Liska et al., 2022), RealtimeQA (Kasai et al., 2022), TempREASON (Tan et al., 2023) and Menatqa (Wei et al., 2023), represent a more nuanced approach to temporal reasoning. These datasets challenge models to answer questions grounded in specific times or events, thereby testing the models' ability to comprehend and reason with temporal information more dynamically. The introduction of benchmarks such as TRAM (Wang and Zhao, 2023) and TimeBench (Chu et al., 2023) marks a significant advancement, providing crucial platforms for temporal reasoning research. Despite these advancements, there has been a noticeable gap in exploring the concurrent nature of temporal events. Previous research has primarily focused on individual events or sequences of events in isolation, overlooking the complexity of scenarios where multiple events co-occur or interact over the same period. Our work aims to fill this gap by being the first to explore the concurrent nature of temporal events.

5.2 Temporal Reasoning over LLMs

To enhance the temporal reasoning capabilities of language models, previous methods either rely heavily on knowledge graphs to rank entities that satisfy the time-related queries (Han et al., 2021; Mavromatis et al., 2022; Liang et al., 2022; Chen et al., 2023) or are strictly dependent on the continual pre-training to strengthen models' abilities in certain temporal aspects (Tan et al., 2023; Yuan et al., 2023). The evolution of LLMs has demonstrated impressive ability in complex reasoning tasks (Chen, 2023), such as mathematical reasoning (Mishra et al., 2022) and logic reasoning (liu et al., 2023). In light of these advancements, recent methods shift towards a program-aided approach (Gao et al., 2023) to improve the performance of time-sensitive tasks, employing Python code as an intermediate logical step instead of natural language (Li et al., 2023). This method, while effective, relies heavily on external tools (Zhu et al., 2023b) and does not fully leverage the inherent capabilities of LLMs (Brown et al., 2020). The results from our COTEMPQA datasets reveal that existing LLMs, even with advanced strategies like Chain of Thought (Wei et al., 2022b), demonstrate limited efficacy in addressing the complexities inherent in co-temporal reasoning tasks. Meanwhile, our research highlights the significant role of mathematical abilities in co-temporal reasoning, offering a direction for future methodologies.

6 Conclusion

In this paper, we propose the COTEMPQA datasets to facilitate the investigation of under-explored co-temporal reasoning problem for large language models. Extensive experiments have shown a significant gap between existing advanced LLMs and human-level performance, even with the enhancement of reasoning approaches. We also discover that mathematical reasoning is crucial for understanding co-temporal events and propose a math-based strategy to improve LLMs' co-temporal reasoning. Reasoning on concurrent and intricate temporal relations remains an open research question, and we hope more enhancement to develop upon our COTEMPQA datasets.

7 Limitations

There are still some limitations in our work, which are listed below:

- For our open-book QA setting, we directly provide the subject's relevant facts in a structured format in the prompt. Recent work shows that LLM's performance in context-based reasoning was significantly weaker than in the former (Chu et al., 2023). In the future, we will employ some retrieval tools to construct prompts with more contextually rich information sources.
- We evaluate the co-temporal reasoning capabilities from the perspective of task performance. However, a more direct approach could involve analyzing how the model's neurons and hidden states are triggered (Zhang et al., 2023). This limitation is not unique to our study and is common in most evaluations of Large Language Models.

References

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. 2023. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. arXiv preprint arXiv:2305.13245.

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

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. 2023. A survey on evaluation of large language models.

Wenhu Chen. 2023. Large language models are few(1)-shot table reasoners. In Findings of the Association for Computational Linguistics: EACL 2023, pages 1120–1130, Dubrovnik, Croatia. Association for Computational Linguistics.

Wenhu Chen, Xinyi Wang, and William Yang Wang. 2021. A dataset for answering time-sensitive questions. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2).

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.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Haotian Wang, Ming Liu, and Bing Qin. 2023. Timebench: A comprehensive evaluation of temporal reasoning abilities in large language models. arXiv preprint arXiv:2311.17667.

Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. 2022. Time-aware language models as temporal knowledge bases. Transactions of the Association for Computational Linguistics, 10:257–273.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Pal: Program-aided language models.

Rujun Han, Xiang Ren, and Nanyun Peng. 2021. ECONET: Effective continual pretraining of language models for event temporal reasoning. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 5367–5380, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey.

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A Smith, Yejin Choi, and Kentaro Inui. 2022. Realtime qa: What's the answer right now? arXiv preprint arXiv:2207.13332.

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, volume 35, pages 22199–22213. Curran Associates, Inc.

Xingxuan Li, Liying Cheng, Qingyu Tan, Hwee Tou Ng, Shafiq Joty, and Lidong Bing. 2023. Unlocking temporal question answering for large language models using code execution.

Ke Liang, Lingyuan Meng, Meng Liu, Yue Liu, Wenxuan Tu, Siwei Wang, Sihang Zhou, Xinwang Liu, and Fuchun Sun. 2022. Reasoning over different types of knowledge graphs: Static, temporal and multi-modal. arXiv preprint arXiv:2212.05767.

Adam Liska, Tomas Kocisky, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, D'Autume Cyprien De Masson, Tim Scholtes, Manzil Zaheer, Susannah Young, et al. 2022. Streamingqa: A benchmark for adaptation to new knowledge over time in question answering models. In International Conference on Machine Learning, pages 13604–13622. PMLR.

Hanmeng liu, Zhiyang Teng, Ruoxi Ning, Jian Liu, Qiji Zhou, and Yue Zhang. 2023. Glore: Evaluating logical reasoning of large language models.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. 2023a. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. arXiv preprint arXiv:2308.09583.

Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. 2023b. Wizardcoder: Empowering code large language models with evol-instruct.

Costas Mavromatis, Prasanna Lakkur Subramanyam, Vassilis N. Ioannidis, Adesoji Adeshina, Phillip R Howard, Tetiana Grinberg, Nagib Hakim, and George Karypis. 2022. Tempoqr: Temporal question reasoning over knowledge graphs. Proceedings of the AAAI Conference on Artificial Intelligence, 36(5):5825–5833.

Swaroop Mishra, Matthew Finlayson, Pan Lu, Leonard Tang, Sean Welleck, Chitta Baral, Tanmay Rajpurohit, Oyvind Tafjord, Ashish Sabharwal, Peter Clark, and Ashwin Kalyan. 2022. LILA: A unified benchmark for mathematical reasoning. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 5807–5832, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

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.

J PUSTEJOVSKY. 2003. The timebank corpus. In Proceedings of Corpus Linguistics 2003, pages 647–656.

James Pustejovsky, Robert Ingria, Andrea Setzer, and Graham Katz. 2003. Timeml: Robust specification of event and temporal expressions in text.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, et al. 2023. Code llama: Open foundation models for code. arXiv preprint arXiv:2308.12950.

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.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

Naushad UzZaman, Hector Llorens, James F. Allen, Leon Derczynski, Marc Verhagen, and James Pustejovsky. 2012. Tempeval-3: Evaluating events, time expressions, and temporal relations. CoRR, abs/1206.5333.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledge base. Communications of the ACM, 57:78–85.

Yuqing Wang and Yun Zhao. 2023. Tram: Benchmarking temporal reasoning for large language models. arXiv preprint arXiv:2310.00835.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022a. Finetuned language models are zero-shot learners. In International Conference on Learning Representations.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In Advances in Neural Information Processing Systems, volume 35, pages 24824–24837. Curran Associates, Inc.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed H Chi, Quoc V Le, Denny Zhou, et al. 2022c. Chain-of-thought prompting elicits reasoning in large language models. In Advances in Neural Information Processing Systems.

Yifan Wei, Yisong Su, Huanhuan Ma, Xiaoyan Yu, Fangyu Lei, Yuanzhe Zhang, Jun Zhao, and Kang Liu. 2023. Menatqa: A new dataset for testing the temporal comprehension and reasoning abilities of large language models. arXiv preprint arXiv:2310.05157.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online. Association for Computational Linguistics.

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

Chenhan Yuan, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2023. Back to the future: Towards explainable temporal reasoning with large language models.

Michael Zhang and Eunsol Choi. 2021. SituatedQA: Incorporating extra-linguistic contexts into QA. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 7371–7387, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models.

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.

Victor Zhong, Weijia Shi, Wen tau Yih, and Luke Zettlemoyer. 2022. Romqa: A benchmark for robust, multi-evidence, multi-answer question answering.

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.

Xinyu Zhu, Junjie Wang, Lin Zhang, Yuxiang Zhang, Yongfeng Huang, Ruyi Gan, Jiaxing Zhang, and Yujiu Yang. 2023a. Solving math word problems via cooperative reasoning induced language models. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4471–4485, Toronto, Canada. Association for Computational Linguistics.

Xinyu Zhu, Cheng Yang, Bei Chen, Siheng Li, Jian-Guang Lou, and Yujiu Yang. 2023b. Question answering as programming for solving time-sensitive questions. arXiv preprint arXiv:2305.14221.

A Implementation Details

We utilize the OpenAI API[1] to evaluate all closed-source models, and for open-source models, we employ the transformers library (Wolf et al., 2020). In all our experiments, we set the temperature to 0 and the maximum length to 256. These experiments were conducted across a full range of scales for each evaluated model.

B The Details of Co-temporal Extraction

Building on our approach to structuring time-dependent facts from Wikidata, we delve into extracting co-temporal facts by identifying overlaps in the timestamps of different facts. Specifically, we compare a given fact $f_i = (s_i, r_i, o_i, t_{s_i}, t_{e_i})$ with another fact f_j in five distinct scenarios:

B.1 Scenario 1: $(\mathcal{S}, \mathcal{R}, \overline{\mathcal{O}})$

Original Fact: f_i as defined above.
Compared Fact: $f_j = (s_i, r_i, o_j, t_{s_j}, t_{e_j})$.
Explanation: In this scenario, the subject s_i and relation r_i remain constant, indicating the same subject in the same type of relationship. However, the object changes, where the subject is related to different objects during co-temporal periods.
Template: While <$subject1$> was holding the position of <$object1$>, which position did <$subject1$> hold during the same time span?

B.2 Scenario 2: $(\mathcal{S}, \overline{\mathcal{R}}, \overline{\mathcal{O}})$

Original Fact: f_i as above.
Compared Fact: $f_j = (s_i, r_j, o_j, t_{s_j}, t_{e_j})$.
Explanation: Here, the subject s_i stays constant while the relation and the object change. This scenario is crucial for identifying instances where a single subject is involved in different relationships with different objects concurrently.
Template: While <$subject1$> was holding the position of <$object1$>, which political party did <$subject1$> belong to simultaneously?

B.3 Scenario 3: $(\overline{\mathcal{S}}, \mathcal{R}, \mathcal{O})$

Original Fact: f_i as above.
Compared Fact: $f_j = (s_j, r_i, o_i, t_{s_j}, t_{e_j})$.
Explanation: This scenario reflects cases where the relationship and object remain constant, but the subject changes. It suggests different subjects simultaneously having the same type of relationship with the same object.
Template: While <$subject1$> was holding the position of <$object1$>, who also held the position of <$object1$> concurrently?

B.4 Scenario 4: $(\overline{\mathcal{S}}, \mathcal{R}, \overline{\mathcal{O}})$

Original Fact: f_i as above.
Compared Fact: $f_j = (s_j, r_i, o_j, t_{s_j}, t_{e_j})$.
Explanation: Only the relationship remains constant in this case, while both the subject and the object change. This scenario signifies instances where different subjects share a common relationship with different objects concurrently.
Template: While <$subject1$> was playing for <$object1$>, which team did <$subject2$> play for within the same time interval?

B.5 Scenario 5: $(\overline{\mathcal{S}}, \overline{\mathcal{R}}, \overline{\mathcal{O}})$

Original Fact: f_i as above.
Compared Fact: $f_j = (s_j, r_j, o_j, t_{s_j}, t_{e_j})$.
Explanation: This scenario represents completely distinct facts that overlap in time, with all quintuplet elements changing.
Template: While <$subject1$> was holding the position of <$object1$>, which employer did <$subject2$> work for during the same time?

C Prompts

The prompts and demonstrations can be found from Figure 3 to Figure 12.

---

[1] https://platform.openai.com/

Model	$(\mathcal{S},\mathcal{R},\overline{\mathcal{O}})$			$(\mathcal{S},\overline{\mathcal{R}},\overline{\mathcal{O}})$			$(\overline{\mathcal{S}},\mathcal{R},\mathcal{O})$			$(\overline{\mathcal{S}},\mathcal{R},\overline{\mathcal{O}})$			$(\overline{\mathcal{S}},\overline{\mathcal{R}},\overline{\mathcal{O}})$			Overall
	Acc.	**F_1**	**Avg.**	**Acc.**	**F_1**	**Avg.**	**Acc.**	**F_1**	**Avg.**	**Acc.**	**F_1**	**Avg.**	**Acc.**	**F_1**	**Avg.**	
The Closed Book Question Answer (CBQA) setting																
GPT-3.5-TURBO-0613	**7.3**	**11.9**	**9.6**	29.7	31.1	30.4	**3.0**	**4.5**	**3.7**	13.3	**30.6**	**22.0**	8.5	**21.8**	**15.1**	**16.3**
GPT-4-0613	3.5	8.0	5.7	**30.0**	**31.3**	**30.7**	2.3	4.4	3.3	**15.0**	24.5	19.7	**9.5**	15.1	12.3	14.5
The Open Book Question Answer (OBQA) setting																
GPT-3.5-TURBO	34.2	55.8	45.0	54.7	57.1	55.9	34.4	55.6	45.0	17.8	33.5	25.6	15.1	28.0	21.5	38.9
GPT-4	**57.0**	**74.7**	**65.9**	**68.0**	**70.8**	**69.4**	**58.0**	**73.3**	**65.6**	**29.9**	**48.8**	**39.4**	**23.5**	**39.8**	**31.7**	**54.7**
CODELLAMA-7B	4.3	**24.3**	**14.3**	6.5	23.6	15.0	1.3	15.4	8.4	1.7	9.9	5.8	2.2	14.3	8.3	10.5
WIZARDCODER-7B	3.3	17.9	10.6	16.6	**27.3**	22.0	3.6	16.7	10.1	1.7	8.2	4.9	2.3	12.5	7.4	11.2
LLAMA-7B	2.5	20.5	11.5	9.5	23.0	16.3	1.9	**20.5**	11.2	2.2	16.8	9.5	3.4	**18.7**	11.1	12.0
WIZARDMATH-7B	**7.2**	16.8	12.0	**22.8**	26.6	**24.7**	**4.5**	19.1	**11.8**	**7.5**	**17.9**	**12.7**	**7.1**	17.1	**12.1**	**14.8**
CODELLAMA-13B	6.0	26.0	16.0	7.7	26.9	17.3	2.0	17.9	10.0	1.0	16.2	8.6	1.7	16.7	9.2	12.4
WIZARDCODER-13B	5.0	16.6	10.8	19.6	30.2	24.9	**3.9**	**23.3**	**13.6**	**6.3**	15.1	**10.7**	**4.6**	12.6	8.6	13.9
LLAMA-13B	6.2	**28.6**	**17.4**	10.5	27.0	18.7	3.6	20.9	12.2	2.3	**17.2**	9.8	3.0	**17.6**	**10.3**	13.8
WIZARDMATH-13B	**11.2**	19.7	15.5	**30.9**	**34.0**	**32.5**	3.1	9.1	6.1	5.6	13.8	9.7	4.0	9.1	6.5	**14.4**
CODELLAMA-34B	9.9	42.4	26.2	19.6	38.3	29.0	4.7	27.1	15.9	4.3	25.6	14.9	3.6	21.6	12.6	20.0
WIZARDCODER-34B	11.0	22.6	16.8	35.2	38.0	36.6	**9.0**	27.4	18.2	7.5	15.9	11.7	7.4	16.0	11.7	19.2
LLAMA-70B	10.3	**43.6**	26.9	14.7	37.6	26.1	7.1	**31.7**	**19.4**	7.0	32.4	19.7	6.2	29.9	18.1	22.2
WIZARDMATH-70B	**18.3**	37.8	**28.1**	**49.8**	**53.4**	**51.6**	8.6	24.8	16.7	**20.1**	**37.3**	**28.7**	**17.4**	**30.1**	**23.8**	**30.1**

Table 7: Experimental results of different triple element types in COTEMPQA. The best performance is **bold**.

WikiData ID	**KB Relation Pairs**	**# Queries**	**Template**
P102-P102	political party & political party	475	While <subject> was a member of <object>, which political party did <subject> belong to within the same time interval?
P39-P39	position held & position held	1,017	While <subject> was holding the position of <object>, which position did <subject> hold during the same time span?
P108-P108	employer & employer	768	While <subject> was working for <object>, which employer did <subject> work for during the same time period?
P54-P54	member of sports team & member of sports team	204	While <subject> was playing for <object>, which team did <subject> play for at the same time?
P69-P69	educated at & educated at	258	While <subject> attended <object>, which school was <subject> attending during the identical time period?
P127-P127	owned by & owned by	75	While <subject> was owned by <object>, who was the owner of <subject> concurrently?
P102-P39	political party & position held	117	While <subject> was a member of <object>, which position did <subject> hold simultaneously?
P102-P108	political party & employer	101	While <subject> was a member of <object>, which employer did <subject> work for during the same time span?
P102-P69	political party & educated at	74	While <subject> was a member of <object>, which school was <subject> attending within the same time interval?
P39-P102	position held & political party	420	While <subject> was holding the position of <object>, which political party did <subject> belong to during the same time period?
P39-P108	position held & employer	380	While <subject> was holding the position of <object>, which employer did <subject> work for at the same time?
P108-P39	employer & position held	125	While <subject> was working for <object>, which position did <subject> hold during the identical time period?
P108-P69	employer & educated at	241	While <subject> was working for <object>, which school was <subject> attending concurrently?
P54-P69	member of sports team & educated at	77	While <subject> was playing for <object>, which school was <subject> attending simultaneously?
P69-P102	educated at & political party	187	While <subject> attended <object>, which political party did <subject> belong to during the same time span?
P69-P39	educated at & position held	95	While <subject> attended <object>, which position did <subject> hold within the same time interval?
P69-P108	educated at & employer	134	While <subject> attended <object>, which employer did <subject> work for during the same time period?

Table 8: Templates used for converting Wikidata facts into natural questions.

Question: While Valdis Dombrovskis was holding the position of European Commissioner for Trade, which position did Valdis Dombrovskis during the identical time period?
Only return the answer:

Figure 3: Default prompt for **C**losed-**B**ook **QA** (**CBQA**) in our proposed CoTempQA

Question: While Valdis Dombrovskis was holding the position of European Commissioner for Trade, which position did Valdis Dombrovskis during the identical time period?
Answer: Let's think step by step,

Figure 4: Zero-cot prompt for **C**losed-**B**ook **QA** (**CBQA**) in our proposed CoTempQA

Question: While Valdis Dombrovskis was holding the position of European Commissioner for Trade, which position did Valdis Dombrovskis during the identical time period?
Only return the answer: European Commissioner for Internal Market and Services
......
Question: While Eduard Jan Dijksterhuis was working for Leiden University, which employer did Eduard Jan Dijksterhuis work for during the same time span?
Only return the answer:

Figure 5: Few-shot prompt for **C**losed-**B**ook **QA** (**CBQA**) in our proposed CoTempQA (5-shot)

Question: While Valdis Dombrovskis was holding the position of European Commissioner for Trade, which position did Valdis Dombrovskis during the identical time period?
Answer: According to the fact, Valdis Dombrovskis became the European Commissioner for Trade on August 26, 2020. He also held the position of European Commissioner for Internal Market and Services from July 16, 2016, to October 12, 2020. This period overlaps with his tenure as Commissioner for Trade. Therefore, the answer is European Commissioner for Internal Market and Services.
......
Question: While Eduard Jan Dijksterhuis was working for Leiden University, which employer did Eduard Jan Dijksterhuis work for during the same time span?
Answer: According to the fact,

Figure 6: Few-shot&CoT prompt for **C**losed-**B**ook **QA** (**CBQA**) in our proposed CoTempQA (5-shot)

Question: While Valdis Dombrovskis was holding the position of European Commissioner for Trade, which position did Valdis Dombrovskis during the identical time period?
Answer: According to the context, Valdis Dombrovskis became the European Commissioner for Trade on August 26, 2020. The datetime can be formed (2020, 8, 26).
The content provided and related to the question can be structured as:
(Vice-President of the European Commission, (2019, 12, 1)).
(European Commissioner for Internal Market and Services, (2016, 6, 16), (2020, 10, 12)).
(European Commissioner for An Economy, (2019, 10, 1)).
(Prime Minister of Latvia, (2009, 3, 12), (2014, 1, 22)).
(Minister of Finance, (2002, 11, 7), (2004, 3, 9)).
Given the (2020, 8, 26), compared with all contents related, we find that $[(2016, 6, 16) - (2020, 10, 12)] \cap (2020, 8, 26) \neq \emptyset$. Therefore the answer is European Commissioner for Internal Market and Services.
......
Question: While Eduard Jan Dijksterhuis was working for Leiden University, which employer did Eduard Jan Dijksterhuis work for during the same time span?
Answer: According to the fact,

Figure 7: Few-shot&Mr-CoT prompt for **C**losed-**B**ook **QA** (**CBQA**) in our proposed CoTempQA (5-shot)

Answer the question based on the context:
Context: Valdis Dombrovskis holds the position of Vice-President of the European Commission in December 1, 2019.
Valdis Dombrovskis holds the position of European Commissioner for Internal Market and Services from July 16, 2016 to October 12, 2020.
Valdis Dombrovskis holds the position of European Commissioner for Trade in August 26, 2020.
Valdis Dombrovskis holds the position of European Commissioner for An Economy that Works for People in December 1, 2019.
Valdis Dombrovskis holds the position of Prime Minister of Latvia from March 12, 2009 to January 22, 2014.
Valdis Dombrovskis holds the position of Minister of Finance from November 7, 2002 to March 9, 2004.
Question: While Valdis Dombrovskis was holding the position of European Commissioner for Trade, which position did Valdis Dombrovskis during the identical time period?
Only return the answer:

Figure 8: Default prompt for **O**pen-**B**ook **QA** (**OBQA**) in our proposed CoTempQA

Answer the question based on the context:
Context: Valdis Dombrovskis holds the position of Vice-President of the European Commission in December 1, 2019.
Valdis Dombrovskis holds the position of European Commissioner for Internal Market and Services from July 16, 2016 to October 12, 2020.
Valdis Dombrovskis holds the position of European Commissioner for Trade in August 26, 2020.
Valdis Dombrovskis holds the position of European Commissioner for An Economy that Works for People in December 1, 2019.
Valdis Dombrovskis holds the position of Prime Minister of Latvia from March 12, 2009 to January 22, 2014.
Valdis Dombrovskis holds the position of Minister of Finance from November 7, 2002 to March 9, 2004.
Question: While Valdis Dombrovskis was holding the position of European Commissioner for Trade, which position did Valdis Dombrovskis during the identical time period?
Answer: Let's think step by step,

Figure 9: Zero-cot prompt for **O**pen-**B**ook **QA** (**OBQA**) in our proposed CoTempQA

Answer the question based on the context:
Context: Valdis Dombrovskis holds the position of Vice-President of the European Commission in December 1, 2019.
Valdis Dombrovskis holds the position of European Commissioner for Internal Market and Services from July 16, 2016 to October 12, 2020.
Valdis Dombrovskis holds the position of European Commissioner for Trade in August 26, 2020.
Valdis Dombrovskis holds the position of European Commissioner for An Economy that Works for People in December 1, 2019.
Valdis Dombrovskis holds the position of Prime Minister of Latvia from March 12, 2009 to January 22, 2014.
Valdis Dombrovskis holds the position of Minister of Finance from November 7, 2002 to March 9, 2004.
Question: While Valdis Dombrovskis was holding the position of European Commissioner for Trade, which position did Valdis Dombrovskis during the identical time period?
Only return the answer: European Commissioner for Internal Market and Services

......

Answer the question based on the context:
Context: Eduard Jan Dijksterhuis attended University of Groningen from 1911 to 1918.
Eduard Jan Dijksterhuis worked for Duke University School of Medicine from August 28, 1912 to April 28, 1923.
Eduard Jan Dijksterhuis works for Leiden University from July 5, 1954 to September 5, 1960.
Eduard Jan Dijksterhuis worked for American University of Armenia in August, 1911.
Eduard Jan Dijksterhuis worked for Austin College from July, 1936 to April, 1947.
Eduard Jan Dijksterhuis worked for Sonoma State University in July, 1932.
Eduard Jan Dijksterhuis worked for Fairfax Media in December 16, 1942.
Eduard Jan Dijksterhuis worked for Canadian Broadcasting Corporation in 1941.
Eduard Jan Dijksterhuis works for Utrecht University from May 1, 1953 to September 1, 1960.
Eduard Jan Dijksterhuis worked for Jean-Marie Le Pen in January, 1931.
Question: While Eduard Jan Dijksterhuis was working for Leiden University, which employer did Eduard Jan Dijksterhuis work for during the same time span?
Only return the answer:

Figure 10: Few-shot prompt for **O**pen-**B**ook **QA** (**OBQA**) in our proposed CoTempQA (5-shot)

Answer the question based on the context:
Context: Valdis Dombrovskis holds the position of Vice-President of the European Commission in December 1, 2019.
Valdis Dombrovskis holds the position of European Commissioner for Internal Market and Services from July 16, 2016 to October 12, 2020.
Valdis Dombrovskis holds the position of European Commissioner for Trade in August 26, 2020.
Valdis Dombrovskis holds the position of European Commissioner for An Economy that Works for People in December 1, 2019.
Valdis Dombrovskis holds the position of Prime Minister of Latvia from March 12, 2009 to January 22, 2014.
Valdis Dombrovskis holds the position of Minister of Finance from November 7, 2002 to March 9, 2004.
Question: While Valdis Dombrovskis was holding the position of European Commissioner for Trade, which position did Valdis Dombrovskis during the identical time period?
Answer: According to the context, Valdis Dombrovskis became the European Commissioner for Trade on August 26, 2020. He also held the position of European Commissioner for Internal Market and Services from July 16, 2016, to October 12, 2020. This period overlaps with his tenure as Commissioner for Trade. Therefore, the answer is European Commissioner for Internal Market and Services.

......

Answer the question based on the context:
Context: Eduard Jan Dijksterhuis attended University of Groningen from 1911 to 1918.
Eduard Jan Dijksterhuis worked for Duke University School of Medicine from August 28, 1912 to April 28, 1923.
Eduard Jan Dijksterhuis works for Leiden University from July 5, 1954 to September 5, 1960.
Eduard Jan Dijksterhuis worked for American University of Armenia in August, 1911.
Eduard Jan Dijksterhuis worked for Austin College from July, 1936 to April, 1947.
Eduard Jan Dijksterhuis worked for Sonoma State University in July, 1932.
Eduard Jan Dijksterhuis worked for Fairfax Media in December 16, 1942.
Eduard Jan Dijksterhuis worked for Canadian Broadcasting Corporation in 1941.
Eduard Jan Dijksterhuis works for Utrecht University from May 1, 1953 to September 1, 1960.
Eduard Jan Dijksterhuis worked for Jean-Marie Le Pen in January, 1931.
Question: While Eduard Jan Dijksterhuis was working for Leiden University, which employer did Eduard Jan Dijksterhuis work for during the same time span?
Answer: According to the context,

Figure 11: Few-shot&CoT prompt for **O**pen-**B**ook **QA** (**OBQA**) in our proposed COTEMPQA (5-shot)

Answer the question based on the context:
Context: Valdis Dombrovskis holds the position of Vice-President of the European Commission in December 1, 2019.
Valdis Dombrovskis holds the position of European Commissioner for Internal Market and Services from July 16, 2016 to October 12, 2020.
Valdis Dombrovskis holds the position of European Commissioner for Trade in August 26, 2020.
Valdis Dombrovskis holds the position of European Commissioner for An Economy that Works for People in December 1, 2019.
Valdis Dombrovskis holds the position of Prime Minister of Latvia from March 12, 2009 to January 22, 2014.
Valdis Dombrovskis holds the position of Minister of Finance from November 7, 2002 to March 9, 2004.
Question: While Valdis Dombrovskis was holding the position of European Commissioner for Trade, which position did Valdis Dombrovskis during the identical time period?
Answer: According to the context, Valdis Dombrovskis became the European Commissioner for Trade on August 26, 2020. The datetime can be formed (2020, 8, 26).
The content provided and related to the question can be structured as:
(Vice-President of the European Commission, (2019, 12, 1)).
(European Commissioner for Internal Market and Services, (2016, 6, 16), (2020, 10, 12)).
(European Commissioner for An Economy, (2019, 10, 1)).
(Prime Minister of Latvia, (2009, 3, 12), (2014, 1, 22)).
(Minister of Finance, (2002, 11, 7), (2004, 3, 9)).
Given the (2020, 8, 26), compared with all contents related, we find that $[(2016, 6, 16) - (2020, 10, 12)] \cap (2020, 8, 26) \neq \emptyset$.
Therefore the answer is European Commissioner for Internal Market and Services.

......

Answer the question based on the context:
Context: Eduard Jan Dijksterhuis attended University of Groningen from 1911 to 1918.
Eduard Jan Dijksterhuis worked for Duke University School of Medicine from August 28, 1912 to April 28, 1923.
Eduard Jan Dijksterhuis works for Leiden University from July 5, 1954 to September 5, 1960.
Eduard Jan Dijksterhuis worked for American University of Armenia in August, 1911.
Eduard Jan Dijksterhuis worked for Austin College from July, 1936 to April, 1947.
Eduard Jan Dijksterhuis worked for Sonoma State University in July, 1932.
Eduard Jan Dijksterhuis worked for Fairfax Media in December 16, 1942.
Eduard Jan Dijksterhuis worked for Canadian Broadcasting Corporation in 1941.
Eduard Jan Dijksterhuis works for Utrecht University from May 1, 1953 to September 1, 1960.
Eduard Jan Dijksterhuis worked for Jean-Marie Le Pen in January, 1931.
Question: While Eduard Jan Dijksterhuis was working for Leiden University, which employer did Eduard Jan Dijksterhuis work for during the same time span?
Answer: According to the context,

Figure 12: Few-shot&Mr-CoT prompt for **O**pen-**B**ook **QA** (**OBQA**) in our proposed CoTempQA (5-shot)