EvolveBench: A Comprehensive Benchmark for Assessing Temporal Awareness in LLMs on Evolving Knowledge

Anonymous ACL submission

Abstract

Large language models (LLMs) are trained on extensive historical corpora, but their ability to understand time and maintain temporal awareness of time-evolving factual knowledge remains limited. Previous studies often neglect the critical aspect of utilizing knowledge from various sources. To address this gap, we introduce EvolveBench, a comprehensive benchmark that evaluates temporal competence along five dimensions: Cognition, which examines the ability to recall and contextualize historical facts. Awareness, which tests the alignment between external inputs and the temporal context of a query. Trustworthiness, which assesses whether models can identify and appropriately refuse queries based on invalid or nonexistent timestamps. Understanding, which focuses on interpreting both explicit dates and implicit historical markers. Finally, reasoning evaluates the capacity to analyze temporal relationships and draw accurate inferences. Evaluating 15 widely used LLMs on EvolveBench shows that GPT-4 achieves the highest average EM score of 79.36, while the open-source Llama3.1-70B demonstrates notable strength in handling temporally misaligned contexts with an average score of 72.47. Despite these advances, all models still struggle with temporal misaligned context. Our code and dataset are available at https://anonymous.4open.science/r/ACL_2025.

1 Introduction

Large language models (LLMs) are trained on vast corpora spanning multiple historical periods. Yet, their ability to maintain temporal awareness—the capacity to track, interpret, and reason about time-evolving factual knowledge—remains a challenge (Xu et al., 2023; Hu et al., 2024). A fundamental question arises: Can LLMs accurately grasp the concept of time and effectively utilize knowledge across different historical eras? While previous studies have explored LLMs' temporal understanding (Jin et al., 2024; Jiang et al., 2024), they often overlook the critical aspect of knowledge utilization, which is essential for real-world applications.

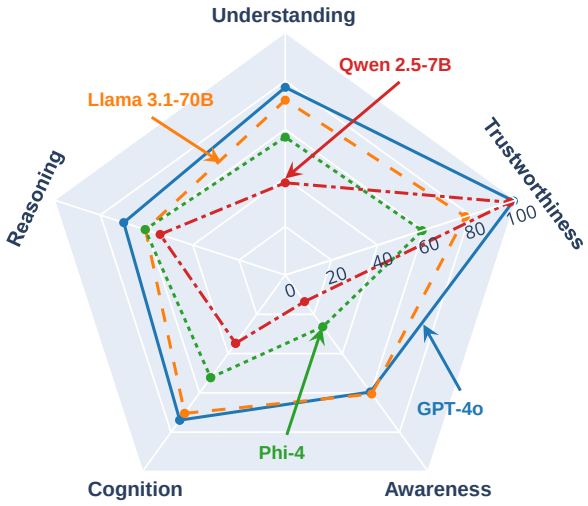

Figure 1: We comprehensively consider the capacity of cognition, awareness, trustworthiness, understanding, and reasoning when evaluating temporal awareness of LLMs on time-evolving knowledge.

Existing evaluations have primarily focused on assessing how LLMs perceive time (Fatemi et al., 2024; Wang and Zhao, 2024), but a more profound challenge lies in whether models can correctly apply time-sensitive knowledge (Dhingra et al., 2022; Mousavi et al., 2024; Sun et al., 2024) in dynamic contexts. Some studies have attempted to address this by evaluating LLMs' ability to integrate real-time information from external sources. For instance, Kasai et al. (2023) introduced a continuously updated knowledge base to test LLMs on time-sensitive queries, while Zhang et al. (2024) assessed their handling of rapidly changing news. Similarly, Tang et al. (2024) examined LLMs using evolving Wikipedia data. However, these approaches often assume external knowledge is accurate and aligned with queries, neglecting real-world

challenges such as temporal inconsistencies, conflicting information, and outdated knowledge (Su et al., 2024b; Xu et al., 2024). Without addressing these factors, current evaluations fail to capture the full complexity of temporal reasoning in LLMs.

To address this gap, we introduce EvolveBench, a novel benchmark designed to evaluate LLMs' temporal awareness of time-evolving factual knowledge across five key dimensions: cognition, which measures a model's ability to recall and contextualize historical facts; awareness, which tests whether external information aligns with a given query's temporal context; trustworthiness, which assesses the model's ability to recognize when a query references invalid or non-existent timestamps; understanding, which examines both explicit and implicit temporal expressions; and reasoning, which evaluates how well models analyze temporal relationships and infer changes over time. These dimensions collectively provide a holistic framework for assessing the temporal competence of LLMs.

We conduct extensive experiments on 15 widely used LLMs on EvolveBench to assess their performance. Results show that GPT-4 achieves the highest average EM score of 79.36, while Llama3.1-70B emerges as the strongest open-source model, scoring 72.47. Notably, Llama3.1-70B demonstrates superior performance in handling temporally misaligned contexts. Despite these advancements, all models still struggle with evolving factual knowledge, highlighting the need for further improvements in LLMs' temporal awareness.

- We establish a new paradigm for evaluating LLMs' temporal awareness, introducing a multi-dimensional framework that systematically assesses how models recall, verify, and reason about time-evolving knowledge.
- We present EvolveBench, the first benchmark designed to rigorously test LLMs across diverse historical contexts and real-world temporal inconsistencies, providing a more comprehensive assessment than existing methods.
- We extensively evaluate 15 state-of-the-art LLMs, uncovering fundamental limitations in their ability to handle temporally misaligned information and setting a new foundation for advancing temporal reasoning in LLMs.

By systematically assessing LLMs' ability to process and interpret time-evolving knowledge, our work lays a foundation for future advancements.

2 Benchmark Construction

This section describes constructing our benchmark (Figure 2) using Wikidata [1] (Vrandečić and Krötzsch, 2014). Table 1 shows the comparison of the previous benchmark. Each knowledge sample is represented as a triple tuple (S, P, A), where S is the subject (e.g., a person or entity name like Johns Hopkins University), P is the property, and $A = [a_1, a_2, \cdots, a_N]$ is a list of attribute values for that property, which change over time.

We collect time-evolving knowledge from four domains: countries, companies, athletes, and organizations. However, time data in the athlete domain is often inaccurate. For example, when a football player is on loan, the attribute values in Wikidata can become chaotic. We update the data with career information from Sofascore[2] to address this. In this benchmark, we set the knowledge cutoff date $T_{current}$ to December 31, 2024, and manually update attribute values from corresponding Wikipedia pages [3] for samples lacking updated knowledge.

2.1 Cognition of Temporal Knowledge

We propose two cognitive levels—timestamp and temporal interval—to evaluate the LLMs' ability to probe factual knowledge from its parameters.

For a given property P of a subject S, we require the model to probe the correct knowledge based on a specific timestamp T or temporal interval $[T_{start}, T_{end}]$. In our experiments, the temporal interval is randomly selected from the attribute list A, and the timestamp T is a random date between T_{start} and T_{end}. We consider the model to correctly recall factual knowledge only when the generated output y_{pred} matches the ground truth y_{turth}.

2.2 Awareness of Temporal Misalignment

In the second dimension, we evaluate how language models handle internal parameter knowledge when external knowledge is temporally misaligned with a timestamp in a user query. This evaluation focuses on "future" and "past" misaligned contexts.

For "future misalignment," we randomly select a past timestamp T_{past} from the attribute list A for the property P to construct the query. Then, we provide the up-to-date attribute $a_{current}$ with S and P to GPT-4[4], asking it to generate a paragraph $C_{current}$ that describes the knowledge tuple

[1] www.wikidata.org
[2] www.sofascore.com
[3] www.wikipedia.org
[4] https://platform.openai.com/

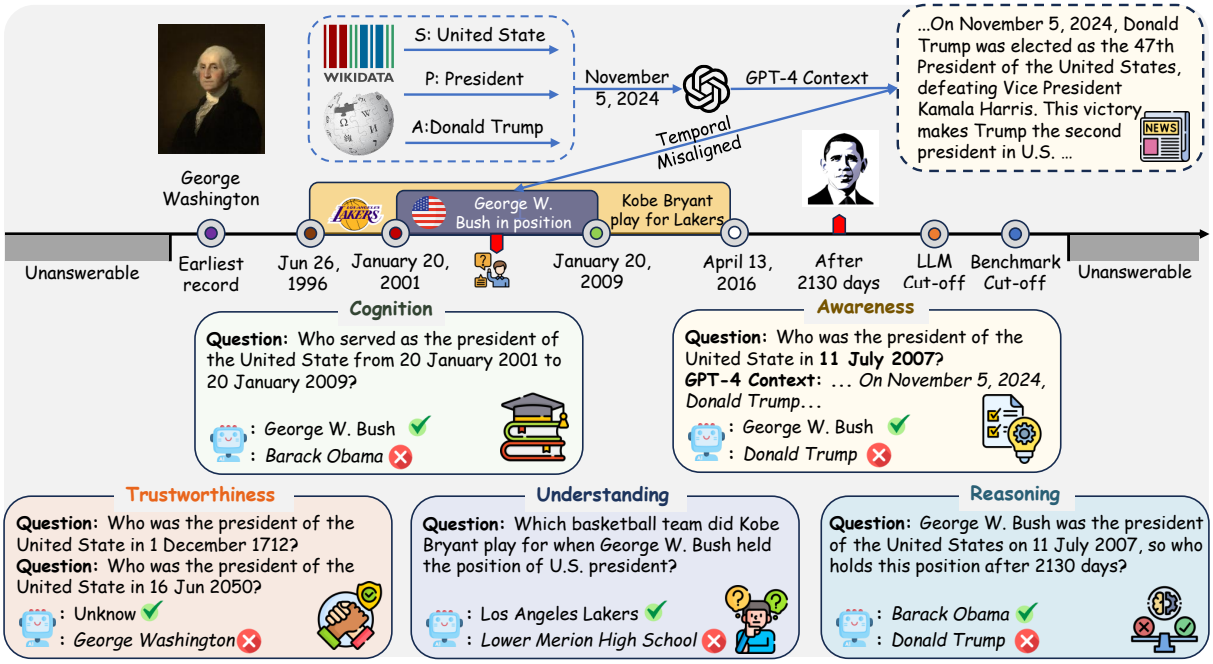

Figure 2: Overview of the construction of EvolveBench. We define five key dimensions: Cognition (Section 2.1), Awareness (Section 2.2), Trustworthiness (Section 2.3), Understanding (Section 2.4), and Reasoning (Section 2.5). These five aspects comprehensively evaluate the temporal awareness of the LLMs on time-evolving knowledge.

	Cogn.	Awar.	Trust.	Unde.	Reas.
TimeQA	✓	✗	✓	✓	✗
TEMPLAMA	✓	✗	✗	✗	✓
TRAM	✗	✗	✗	✓	✓
DyKnow	✓	✗	✗	✗	✗
EvolveBench	✓	✓	✓	✓	✓

Table 1: Our EvolveBench offers a more comprehensive evaluation of large language models' temporal awareness in handling time-evolving knowledge.

$(S, P, a_{current})$ in detail. Therefore, the timestamp in $C_{current}$ is temporally misaligned with T_{past} in the query, meaning the information is accurate but futuristic compared to the query timestamp. For "past misalignment," the timestamp in the user query is $T_{current}$. We provide a randomly selected past attribute a_{past} from A with S and P to GPT-4 and ask it to generate a paragraph C_{past} that describes the past knowledge tuple (S, P, a_{past}). This setup tests the model's ability to handle outdated information when responding to user queries.

This method simulates situations where a language model answers a query using the Retrieval-Augmented Generation (RAG) paradigm, primarily when misinformation exists in the retrieved content. We consider the model to correctly distinguish misinformation only when the model output y_{pred} matches the ground truth y_{truth}.

2.3 Trustworthiness of Unanswerable Date

We introduce trustworthiness as a third dimension to assess whether an LLM's answers hallucinate when the requested date is unanswerable. Specifically, if the timestamp T in a user query is earlier than the earliest record in an attribute list A for a given subject S and property P, or if it refers to a future date, the query is considered unanswerable.

We manually collect past unanswerable dates from the corresponding Wikipedia page to clarify a subject's S earliest historical time and address incomplete records in Wikidata. For example, various political entities have preceded the present Federal Republic of Germany. We select the day before Germany becomes a nation. For athletes, the unanswerable date is the day before they begin their careers. For companies or organizations, it is before their establishment date. For future unanswerable dates, we set the timestamp to December 31, 2050. The language model is considered correct in refusing to answer if it outputs "Unknown."

2.4 Understanding of Temporal Concept

This dimension evaluates how effectively LLMs interpret temporal concepts presented in different formats. In previous evaluations, we used an explicit time format (e.g., "DD Month YYYY") to represent time. For implicit time formats, we define

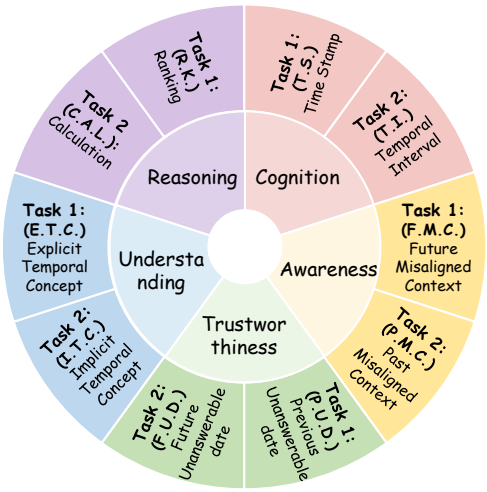

Figure 3: The specific subtasks evaluated within each capacity dimension. Their detailed construction is described in the subsections of Section 2.

temporal intervals $[T_{start}, T_{end}]$ based on historical events. For instance, the phrase "When Barack Obama was the president of the United States" represents the period from January 20, 2009, to January 20, 2017. We denote this implicit time representation as $T_{implict}$.

To avoid ambiguity, we exclude events that occur more than once. For example, "When Donald Trump was the president" refers to two different periods. In our dataset, we randomly select one $T_{implict}$ for each (S, P, A) tuple and prompt the language model to answer the factual questions.

2.5 Temporal Reasoning

Temporal reasoning involves analyzing the relationships between past events. We designed two subtasks to evaluate the LLMs' reasoning ability: ranking and calculation.

Given two past events, a_1, a_2, randomly selected from the attribute list A of a tuple (S, P, A). The ranking subtask requires the model to determine its correct chronological order. The model must first extract their timestamps from the input and then compare them to provide the final answer.

For two past events a_1, a_2, we randomly select two dates, T_1 and T_2, from their respective temporal periods $[T_{start}, T_{end}]$ and calculate the number of days, D_{elapse}, between them. Given T_1 and D_{elapse}, the calculation task requires the model to perform the necessary calculations and retrieve the correct answer, a_2, from its parameters. The language model is considered correct only if the output, y_{pred}, matches the ground truth, y_{turth}.

3 Experiments

3.1 Language Model for Evaluation

This paper evaluates several widely used large language models, including different sizes Llama 2 (Touvron et al., 2023), Llama 3 series (Grattafiori et al., 2024), Qwen2 (Yang et al., 2024), Qwen2.5 (and: et al., 2025), Phi-4 (Abdin et al., 2024), GPT-4 (OpenAI et al., 2024) and GPT-3.5 (Ye et al., 2023) on our benchmark. All models use greedy search in auto-regressive generation to eliminate the randomness introduced by sampling.

3.2 Evaluation Metrics

We evaluate the model's outputs using the Exact Match (EM) score for each subtask within a given capacity dimension. The model's capacity in this dimension is defined as the average EM score across all subtasks.

$$C_d = \frac{1}{N} \sum_{i=1}^{N} EM_i, \quad (1)$$

N is the number of subtasks in capacity dimension d, and EM_i is the EM score of the $i-th$ subtask.

3.3 Prompt Agreement

We designed three prompts for each subtask to reduce uncertainty from prompt variations, known as prompt agreement (Portillo Wightman et al., 2023) or the knowledge boundary (Wang et al., 2024; Yin et al., 2024) effect. These prompts convey the same meaning but differ in phrasing. The final score is the average of the EM scores from these prompts.

4 Experimental Results

4.1 Analysis of Main Results

Table 2 shows the main evaluation results. Figure 3 illustrates the subtasks for each capacity dimension. The **red** values in the bracket mean a negative effect, while **green** means a positive. We draw the following conclusions based on the data in Table 2.

LLMs perform better in cognition when queries are presented as temporal intervals. When evaluating the cognitive capacity of LLMs, we express the same historical event in user queries using timestamps and temporal intervals. For example, for "Steve Jobs served as the CEO of Apple.", the timestamp-based query would be "Who served as the CEO of Apple on 1 January 1998?" and the temporal interval-based query would be "Who served

Models	Cognition		Awareness		Trustworthiness		Understanding		Reasoning		Avg.
	T.S.	T.I.	F.M.C.	P.M.C.	P.U.D.	F.U.D.	E.T.C.	I.T.C.	R.K.	C.A.L.	
Model size under 10B											
Llama2-7B	39.63	24.39	0.00	15.04	56.30	16.46	41.06	27.64	79.47	19.51	33.33
Llama3-8B	44.51	51.22	16.87	38.82	62.40	2.03	58.13	47.76	84.96	25.00	45.19
Qwen2-7B	25.61	37.60	1.63	27.64	82.32	72.76	44.11	33.54	93.90	18.29	46.57
Llama3.1-8B	48.37	53.25	10.77	42.07	76.63	22.36	63.21	47.97	87.20	23.98	50.20
Qwen2.5-7B	32.93	36.59	3.05	23.98	95.93	98.17	47.76	28.25	92.68	15.65	51.04
Model size under 65B											
Llama2-13B	49.39	42.89	0.00	15.85	70.33	21.54	54.47	39.84	88.21	18.90	42.50
Phi-4	47.76	56.71	9.96	42.68	22.15	96.34	64.02	49.59	93.70	27.64	53.66
Model size under 100B											
Llama2-70B	55.28	58.54	0.61	21.95	60.98	21.14	64.43	51.63	95.53	26.42	47.79
Llama3-70B	64.63	71.14	57.32	40.45	75.00	28.86	72.36	67.89	84.96	36.99	62.51
Qwen2-72B	53.66	54.88	14.63	32.32	95.73	92.07	48.98	48.58	97.15	27.64	59.78
Llama3.1-70B	67.28	73.58	**61.18**	59.96	77.24	78.66	73.37	70.73	90.24	30.69	72.47
Llama3.3-70B	65.85	68.09	57.11	51.02	79.47	68.29	71.14	70.33	94.72	30.89	69.56
Qwen2.5-72B	55.49	61.18	26.63	38.82	97.15	83.54	68.50	55.49	**97.15**	30.08	64.88
Proprietary language models											
GPT-4o	**71.95**	**75.81**	57.32	**61.79**	**98.17**	**99.39**	**79.67**	**75.00**	95.12	**44.72**	**79.36**
GPT-3.5-turbo	64.02	70.53	1.42	22.36	68.29	61.79	75.41	59.76	94.31	33.13	57.54

Table 2: Evaluation results of the recent widely used LLMs. We highlight the best with **boldface** and underline the second-best. Figure 3 shows the task of each capacity dimension. Among the evaluated LLMs, the GPT4-o performs best in our benchmark, while Llama3.1 70B is the best-performing open-source language model.

as the CEO of Apple from 1 September 1997 to 23 August 2011?". As shown in Table 2, most language models perform better with temporal interval-based queries. This phenomenon is likely because such queries provide more temporal context, helping the model identify the most relevant knowledge. However, even the best-performing GPT-4o still fails to recall about 25% of factual knowledge. This result highlights the importance of up-to-date knowledge in model generation.

Figure 4 shows that all language models follow a similar trend in recalling world knowledge. They perform well on information about heads of state (average EM score of 80) but struggle with knowledge about companies (average EM score of 56), organizations (average EM score of 26), and athletes (average EM score of 43). This result is likely because detailed time-related data, like company personnel changes or sports club transfers, is scarce in public sources like Wikipedia. In contrast, information about heads of state is more widely available across various knowledge bases.

LLMs are prone to be misled by temporal misaligned context. LLMs perform worse when queries are accompanied by temporally misaligned context, compared to the T.S. and T.I. columns of the Cognition section (Table 2). This misalignment hampers the model's recall of correct knowledge from its parameters, which decreases EM scores.

To control for variations in accuracy when recalling knowledge at different timestamps, we used the same timestamp in the queries for the experiment in Figure 5. The only difference was whether the input query included the relevant but temporal misaligned text. In the left part of Figure 5, the EM scores of the five language models show a significant decline, with an average drop of 47.66%. Llama3.1-70B performs the best in distinguishing a context from a future date that doesn't align with the query's timestamp. In contrast, the right part of Figure 5 shows a smaller decline, with an average EM drop of only 18.17%.

We conclude that LLMs are prone to being misled by temporally misaligned contexts. They are more sensitive to detecting outdated texts from the 'past' than to identifying texts from the 'future.' Among the evaluated models, Llama3.1-70B shows the best ability to handle temporal misalignment, with the smallest average EM score drop of 10%.

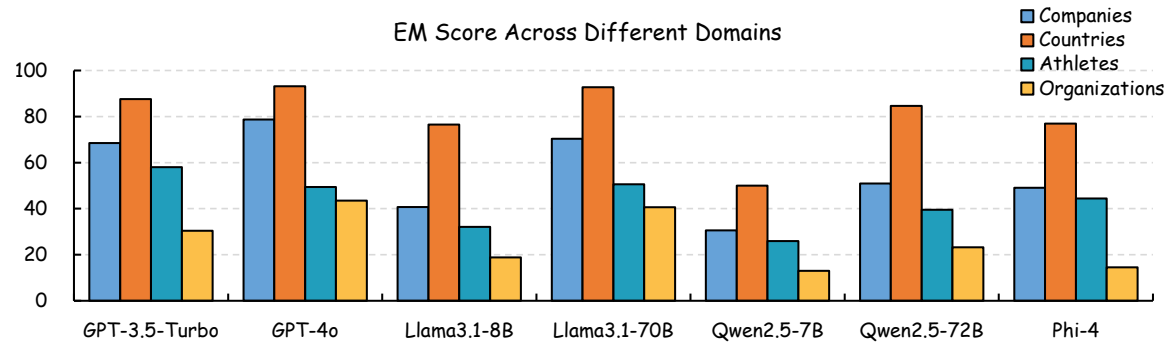

Figure 4: The EM scores reflect the cognitive capacity of various language models across four factual knowledge domains when queried with temporal intervals. All models exhibit higher accuracy in recalling knowledge about heads of state, while their recall of information about athletes and organizations is comparatively weaker.

	Cont. ↓	**Oth.** ↓	**Irrel.** ↓
	w/ Misaligned Context		
GPT-4o	14.6	13.6	14.4
Llama3.3-70B	15.0	14.2	13.6
Llama3.1-70B	3.1	16.1	19.7
Qwen2.5-72B	64.2	6.5	2.6
Phi-4	77.9	9.55	2.6
	w/o Misaligned Context		
GPT-4o	0.8 (-13.8)	13.6 (+0)	13.6 (-0.8)
Llama3.3-70B	1.6 (-13.4)	18.5 (+4.3)	14.0 (+0.4)
Llama3.1-70B	2.2 (-0.8)	16.1 (+0)	14.4 (-5.3)
Qwen2.5-72B	2.6 (-61.6)	20.5 (+14.0)	21.3 (+18.7)
Phi-4	7.1 (-70.7)	20.5 (+11.0)	24.6 (+22.0)

Table 3: Error analysis when provide with future context (Left of Figure 5). Despite providing relevant context, a significant portion of questions are still answered with unexpected responses (**Oth.** and **Irrel.**).

	Corr. ↑	**Cont.** ↓	**Oth.** ↓
	w/ Time Information		
GPT-4o	57.3	14.6	28.1
Llama3.3-70B	57.1	15.0	27.9
Llama3.1-70B	61.1	3.0	35.8
Qwen2.5-72B	26.6	64.2	9.1
Phi-4	9.9	77.8	12.2
	w/o Time Information		
GPT-4o	45.9 (-11.4)	39.6 (+25.0)	14.4 (-13.6)
Llama3.3-70B	55.4 (-1.6)	26.4 (+11.4)	18.1 (-9.8)
Llama3.1-70B	59.1 (-2.0)	21.3 (+18.3)	19.5 (-16.3)
Qwen2.5-72B	10.1 (-16.5)	88.2 (+24.0)	1.6 (-7.5)
Phi-4	8.1 (-1.8)	86.2 (+8.3)	5.7 (-6.5)

Table 4: LLMs show performance degradation when temporal information is removed from the context. This indicates that while temporal information helps distinguish misaligned context, it is still not effective enough.

Table 3 provides a detailed error analysis of the experiment where queries reference the past but are given future context. **Corr.** refers to correct answers, **Cont.** to context-based answers, **Oth.** to other answers, and **Irrel** to irrelevant ones. Despite the significant performance drop, many responses remain unanticipated, even with relevant context, either unrelated to the question or containing incorrect values from the attribute list A but not derived from the provided context. This result highlights the need to explore how models integrate external information with their own knowledge.

LLMs are better at rejecting questions with unanswerable past dates than those with future dates. In this experiment, we set future unanswerable dates in a query to 1 October 2050, while past unanswerable dates are based on the earliest historical record in a specific factual knowledge tuple (S, P, A). According to the Trustworthiness column in Table 2, most language models find it easier to refuse questions with past unanswerable dates despite Phi-4. This result may be because these dates that do not exist in the history of specific tuples were never present in the model's pre-training data, making LLMs more confident in refusing such questions. In contrast, for future dates, LLMs are uncertain whether their knowledge is up-to-date enough to handle those dates.

Among the fifteen models evaluated, GPT-4 and Qwen2.5-7B perform best at refusing unanswerable questions. This is likely due to their instruction tuning, which enhances their safety features.

The model's ability to interpret implicit time expressions depends directly on its accuracy in recalling the historical time of an event. To evaluate the model's understanding of implicit time

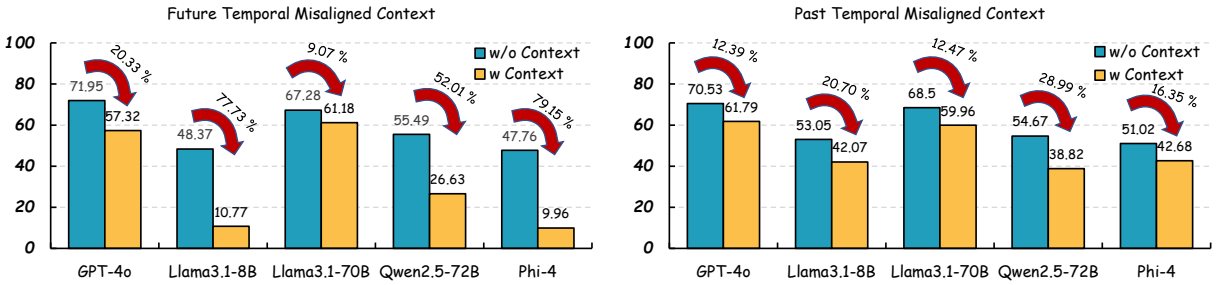

Figure 5: The EM scores of various language models when queried with temporally misaligned context. The left figure shows cases where the context date and information are later than the query date, while the right figure shows cases where the context date and information are earlier. Compared to perceiving the futurity of texts from the 'future,' language models are more sensitive to detecting the obsolescence of texts from the 'past.'

	Corr. ↑	Oth. ↓	Irrel. ↓
w/ Time Information			
GPT-4o	95.1	3.2	1.6
Llama3.3-70B	94.7	3.8	1.4
Llama3.1-70B	90.2	7.9	1.8
Qwen2.5-72B	97.2	2.2	0.6
Phi-4	93.7	5.7	0.6
w/o Time Information			
GPT-4o	87.0 (-8.1)	8.3 (+5.1)	4.7 (+3.1)
Llama3.3-70B	79.3 (-15.4)	19.9 (+16.1)	0.8 (-0.6)
Llama3.1-70B	82.1 (-8.1)	15.0 (+7.1)	2.9 (+1.1)
Qwen2.5-72B	92.1 (-5.1)	6.9 (+4.7)	1.0 (+0.4)
Phi-4	85.0 (-8.7)	14.4 (+8.7)	0.6 (+0.0)

Table 5: Removing temporal information from the ranking task of reasoning leads to performance degradation in all LLMs, highlighting the challenge of mapping entity names to timestamps and comparing their order.

	Corr. ↑	Oth. ↓	Irrel. ↓
Describe in Days			
GPT-4o	44.7	40.0	15.2
Llama3.3-70B	30.9	56.3	12.8
Llama3.1-70B	30.7	53.7	15.7
Qwen2.5-72B	30.1	51.8	18.1
Phi-4	27.6	49.2	23.2
Describe in Years			
GPT-4o	59.6 (+14.8)	27.2 (-12.8)	13.2 (-2.0)
Llama3.3-70B	47.4 (+16.5)	40.9 (-15.5)	11.8 (-1.0)
Llama3.1-70B	45.9(+15.2)	40.5 (-13.2)	13.6 (-2.0)
Qwen2.5-72B	43.9 (+13.8)	39.2 (-12.6)	16.9 (-1.2)
Phi-4	36.6 (+9.0)	41.9 (-7.3)	21.5 (-1.6)

Table 6: All LLMs show significant performance gains when the calculation task is described in 'Year,' highlighting that multiplication and division are more challenging for the model than addition and subtraction.

concepts, we use "country" as the subject S to represent historical events. As shown in Figure 4, the model performs well at recalling factual knowledge about heads of state across various domains.

The understanding column in Table 2 shows that when the model accurately remembers the temporal intervals of historical events (e.g., GPT-4 and Llama 3.1-70B, achieving an EM score of 90% in the country domain), it effectively uses this information to recall facts from parameters. Its performance is comparable to cases with explicit time expressions. However, when the model fails to remember these time intervals accurately (e.g., Qwen2.5-7B and Phi-4), it struggles to use implicit cues, resulting in a significant performance drop compared to explicit time expressions.

Compared to ranking tasks, calculations are more challenging for LLMs in temporal reasoning. When the input prompt includes temporal information, all models perform well on the ranking task, with the Qwen series 7B model achieving an EM score above 90. However, the calculation task is more challenging for all models. Unlike the ranking task, which directly compares chronological order from the input, the calculation task requires the model first to compute the correct year and date and then retrieve the relevant knowledge. The Reasoning column in Table 2 shows that even the best-performing GPT-4 achieved only an EM score of 44.72, a 53% drop compared to the ranking task. While recent open-source models match GPT-4 in ranking tasks, a gap remains in the calculation task. This evaluation highlights the requirements for further improvement in temporal reasoning.

4.2 Importance of Temporal Information

In this section, we evaluate how temporal information affects the awareness and reasoning capacity of

	Corr. ↑	Oth. ↓	Irrel. ↓
w/o Context			
GPT-4o	72.0	14.4	13.6
Llama3.1-70B	67.3	18.3	14.4
Qwen2.5-72B	55.5	23.2	21.3
Phi-4	47.8	27.6	24.6
w Retrieved Context			
GPT-4o	63.6 (-8.3)	16.7 (+2.2)	19.7 (+6.1)
Llama3.1-70B	59.6 (-7.7)	19.9 (+1.6)	20.5 (+6.1)
Qwen2.5-72B	42.9 (-12.6)	31.3 (+8.1)	25.8 (+4.5)
Phi-4	39.4 (-8.3)	31.9 (+4.3)	28.7 (+4.1)
w Generated Context			
GPT-4o	57.3 (-14.6)	28.3 (+13.8)	14.4 (+0.8)
Llama3.1-70B	61.2 (-6.1)	19.1 (+0.8)	19.7 (+5.3)
Qwen2.5-72B	26.6 (-28.9)	70.7 (+47.6)	2.6 (-18.7)
Phi-4	10.0 (-37.8)	87.4 (+59.8)	2.6 (-22.0)

Table 7: EM score of LLMs with retrieved or generated context as input: Most models experience a more significant performance drop in generated than in retrieved. Demonstrate that the more relevant the input document is to the query, the more likely the model will be misled.

the language model. In the experiments in Table 4, we removed the temporal information from the temporal misaligned context while keeping other settings the same. We found that temporal information is crucial for LLMs to identify misaligned contexts. Without it, all models showed a decline in performance (20% performance degradation at the correct rate). The Llama3 series again demonstrated the most substantial ability to detect temporal misalignment, with only a 3% EM score drop.

We also removed the temporal information from the input prompt for the ranking task, requiring LLMs to rank based only on the attribute's name. The results in Table 5 show that temporal information is essential for ranking historical events. Without it, the model struggles to recall the temporal details and compare the events' chronological order in a single reasoning step.

4.3 Difficulty of Mathematical Operations

This section evaluates the difficulty of different mathematical operations for language models. We changed the problem description to simplify temporal reasoning from days to years. Instead of asking the model what the value will be after a certain number of days, we directly ask how many years later the attribute of the tuple (S, P, A) will change from a_1 to another value. By providing year information directly, we reduce the model's calculation difficulty, as language models no longer need to convert days into years and then retrieve the knowledge from its parameters.

Table 6 shows that all language models benefit from the LLM-friendly problem description, with a 43% average improvement in EM scores and a decline in error rates for "Other" and "Irrelevant" categories across the five best-performing models. These results suggest that multiplication and division, especially when converting days to years, are more challenging for recent large language models than addition and subtraction.

4.4 Comparison with RAG Method

We also explore the temporal awareness of LLMs within the retrieval-augmented generation (RAG) paradigm. This section uses processed English Wikipedia data from HuggingFace[5] as our knowledge base. The text is divided into chunks of up to 1000 characters for retrieval and encoded into embedding vectors. We use dense vector (Lewis et al., 2020) search as the retriever.

Table 7 shows that the performance drop in using retrieved-context is less severe than in using GTP-4o generated context. This result is because the retriever often retrieves irrelevant paragraphs due to randomly generated past dates within the temporal interval $[T_{start}, T_{end}]$, which is rarely found in the corpus. This makes it easier for LLMs to detect irrelevant content than highly related but temporally misaligned contexts. The performance drop in C_{RAG} and C_{Gen} underscores the need to enhance the temporal awareness of LLMs.

5 Conclusion

This paper presents a benchmark EvolveBench for evaluating large language models' temporal awareness of time-evolving factual knowledge. We believe that to stay aligned with the dynamic nature of the world, LLMs must excel in cognition, awareness, trustworthiness, understanding, and reasoning. Our experiments show that GPT-4 leads across all five dimensions, while Llama3.1-70B demonstrates the most robustness in handling temporally misaligned information. Among five aspects, the awareness of time-evolving knowledge remains the most challenging for LLMs. Our benchmark and findings offer valuable insights for future research.

[5] https://huggingface.co/datasets/wikimedia/wikipedia

Limitations

This study presents valuable results in applying LLMs to time-evolving factual knowledge and temporal reasoning. However, the model's temporal reasoning and time-sensitive handling of multiple external documents remain areas for further investigation. The current implementation focuses on one specific temporal misaligned context, and future research will aim to refine the model's adaptability and reasoning capabilities when confronted with multiple external contexts.

Ethical Considerations

All the pre-trained language models used in our paper are downloaded from the Huggingface publicly released model card, and we strictly follow the user license. The data contained in our benchmark are collected from publicly available knowledge bases like Wikidata or Wikipedia, and we use this information only for academic research. We also tried to minimize bias in the evaluation queries when constructing evaluation in each capacity dimension.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, et al. 2024. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

Qwen and:, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, et al. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Himanshu Beniwal, Dishant Patel, Kowsik Nandagopan D, Hritik Ladia, Ankit Yadav, and Mayank Singh. 2024. Remember this event that year? assessing temporal information and understanding in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16239–16348, Miami, Florida, USA. 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)*.

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.

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, et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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

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.

Yuxin Jiang, Yufei Wang, Chuhan Wu, Wanjun Zhong, Xingshan Zeng, Jiahui Gao, Liangyou Li, Xin Jiang, Lifeng Shang, Ruiming Tang, Qun Liu, and Wei Wang. 2024. Learning to edit: Aligning LLMs with knowledge editing. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4689–4705, Bangkok, Thailand. Association for Computational Linguistics.

Zhuoran Jin, Pengfei Cao, Chenhao Wang, Zhitao He, Hongbang Yuan, Jiachun Li, Yubo Chen, Kang Liu, and Jun Zhao. 2024. RWKU: Benchmarking real-world knowledge unlearning for large language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Jungo Kasai, Keisuke Sakaguchi, yoichi takahashi, Ronan Le Bras, Akari Asai, Xinyan Velocity Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. 2023. Realtime QA: What's the answer right now? In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Seyed Mahed Mousavi, Simone Alghisi, and Giuseppe Riccardi. 2024. DyKnow: Dynamically verifying

time-sensitive factual knowledge in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8014–8029, Miami, Florida, USA. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Gwenyth Portillo Wightman, Alexandra Delucia, and Mark Dredze. 2023. Strength in numbers: Estimating confidence of large language models by prompt agreement. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 326–362, Toronto, Canada. Association for Computational Linguistics.

Zhaochen Su, Juntao Li, Jun Zhang, Tong Zhu, Xiaoye Qu, Pan Zhou, Yan Bowen, Yu Cheng, and Min Zhang. 2024a. 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.

Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yanshu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. 2024b. $\texttt{ConflictBank}$: A benchmark for evaluating the influence of knowledge conflicts in LLMs. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Kai Sun, Yifan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. 2024. Head-to-tail: How knowledgeable are large language models (LLMs)? A.K.A. will LLMs replace knowledge graphs? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 311–325, Mexico City, Mexico. Association for Computational Linguistics.

Hexiang Tan, Fei Sun, Wanli Yang, Yuanzhuo Wang, Qi Cao, and Xueqi Cheng. 2024. Blinded by generated contexts: How language models merge generated and retrieved contexts when knowledge conflicts? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6207–6227, 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.

Wei Tang, Yixin Cao, Yang Deng, Jiahao Ying, Bo Wang, Yizhe Yang, Yuyue Zhao, Qi Zhang, Xuanjing Huang, Yugang Jiang, and Yong Liao. 2024. Evowiki: Evaluating llms on evolving knowledge. *Preprint*, arXiv:2412.13582.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

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

Yuhao Wang, Yusheng Liao, Heyang Liu, Hongcheng Liu, Yanfeng Wang, and Yu Wang. 2024. MM-SAP: A comprehensive benchmark for assessing self-awareness of multimodal large language models in perception. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9192–9205, Bangkok, Thailand. Association for Computational Linguistics.

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.

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. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1434–1447, Singapore. Association for Computational Linguistics.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565, Miami, Florida, USA. Association for Computational Linguistics.

Yan Xu, Mahdi Namazifar, Devamanyu Hazarika, Aishwarya Padmakumar, Yang Liu, and Dilek Hakkani-Tur. 2023. KILM: Knowledge injection into encoder-decoder language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5013–5035, Toronto, Canada. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, , et al. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, Jie Zhou, Siming Chen, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. *Preprint*, arXiv:2303.10420.

Xunjian Yin, Xu Zhang, Jie Ruan, and Xiaojun Wan. 2024. Benchmarking knowledge boundary for large language models: A different perspective on model evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2270–2286, Bangkok, Thailand. Association for Computational Linguistics.

Zhihan Zhang, Yixin Cao, Chenchen Ye, Yunshan Ma, Lizi Liao, and Tat-Seng Chua. 2024. Analyzing temporal complex events with large language models? a benchmark towards temporal, long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1588–1606, Bangkok, Thailand. Association for Computational Linguistics.

Zihan Zhang, Meng Fang, Ling Chen, Mohammad-Reza Namazi-Rad, and Jun Wang. 2023. How do large language models capture the ever-changing world knowledge? a review of recent advances. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8289–8311, Singapore. Association for Computational Linguistics.

A Appendix

A.1 Related Work

The reasoning capacity (Huang and Chang, 2023) of the recent large language model and its expertise in using the knowledge (Zhang et al., 2023) that is internally stored in parameters or from external retrieval has received attention recently.

Temporal Reasoning Benchmarks The recent research community has yielded multiple evaluation benchmarks to assess LLMs' temporal reasoning abilities. Benchmarks such as TimeQA (Chen et al., 2021), MenatQA (Wei et al., 2023), and TEMPREASON (Tan et al., 2023) mainly focus on temporal reasoning in the context provided. Other benchmarks like ToT (Fatemi et al., 2024), TRAM (Wang and Zhao, 2024) and COTEMPQA (Su et al., 2024a) demonstrate that mathematical capacity is essential in handling temporal relationships. The commonality of these works is that they emphasize reasoning while overlooking the critical role knowledge plays in LLMs' temporal awareness.

Knowledge Utilization of LLMs The large language model uses knowledge to answer the user's query in two ways. Benchmarks like TEMPLAMA (Dhingra et al., 2022), DyKnow (Mousavi et al., 2024), and TempUN (Beniwal et al., 2024) treat LLMs as knowledge repositories and use knowledge stored in the language model's parameters to answer the user's query. For frequently changing knowledge, datasets like TCELongBench (Zhang et al., 2024) and REALTIMEQA (Kasai et al., 2023) build an external knowledge base to support language models in acquiring updated information. However, in realistic application scenarios, the retrieved data from an external knowledge base may not be consistent with the temporal context of a particular query.

Knowledge Conflict The vast pre-trained corpus and fixed parameters cause language models to encounter internal and external knowledge conflicts (Xu et al., 2024; Xie et al., 2024) when processing time-evolving knowledge. Although previous works have investigated the behavior of the language model when encountering counterfactual (Longpre et al., 2021; Tan et al., 2024) knowledge conflict, the situation in which the external context temporally misaligns with a user's query is far from well-studied. In this situation, the internal and external knowledge conflicts exist simultaneously.

	Subject	Property
Countries	Countries	President
Organizations	International Org.	Chairperson
	University	President
Athletes	Basketball	Club name
	Football	
	Formula 1	
Companies	Top 500	CEO

Table 8: Detailed subject and property example for each knowledge domain.

	Nums of Subjects	Nums of Queries
Countries	47	780
Organizations	22	230
Athletes	27	270
Companies	36	360
Total	132	1640

Table 9: We collected 132 subjects with time-varying properties and manually constructed 1,640 queries for the five capabilities in our benchmark.

This paper introduces a benchmark that comprehensively evaluates LLMs' temporal awareness of language models on time-evolving knowledge of five novel key capacity dimensions: cognition, awareness, trustworthiness, understanding, and reasoning. Our benchmark simultaneously considers the fundamental capacity of reason over the temporal relationship and the complicated scenario in handling internal and external knowledge conflicts. Table 1 shows the comparison between other related benchmarks.

A.2 Benchmark detail

Our benchmark, EvolveBench, collects factual knowledge from four domains: countries, organizations, athletes, and companies. Building on the Dyknow benchmark (Mousavi et al., 2024), we added over 40 new entities, including renowned universities and global companies. As detailed in Section 2, we manually update the data using Wikipedia and Sofascore to ensure the accuracy of the attribute lists for each knowledge tuple (S, P, A). Table 8 and Table 9 shows the detail of our EvolveBench.

A.3 Prompt List

The following table describes the detailed prompt used in our evaluation. The bottom of Figure 2 shows five example cases of our benchmark.

Prompt A.1: Cognition (Time Stamp)

System Prompt: You are a knowledgeable assistant who can answer factual questions.

User Prompt: Today Date: {date}.

Given a question, you should answer it using your own knowledge based on today's date ({date}). Remember, your answer must contain only the name, with no other words.

QUESTION: The {property} of {subject} is current held by?

Your answer:

Prompt A.2: Cognition (Temporal Interval)

System Prompt: You are a knowledgeable assistant who can answer factual questions.

User Prompt: Given a question, you should answer it using your own knowledge based on the temporal interval. Remember, your answer must contain only the name, with no other words.

QUESTION: Who served as {property} of {subject} from $\{T_{start}\}$ to $\{T_{end}\}$?

Your answer:

Prompt A.3: Awareness (Future Misaligned Context)

System Prompt: You are a knowledgeable assistant who can answer factual questions.

User Prompt: Today Date: {date}.

Given a question and its relevant context, you should answer it using your own knowledge or the knowledge provided by the context. Remember, the provided context may not necessarily be up-to-date to answer the question, and your answer must contain only the name, with no other words.

CONTEXT: {Future temporal misaligned context}

QUESTION: The {property} of {subject} is current held by??

Your answer:

Prompt A.4: Awareness (Past Misaligned Context)

System Prompt: You are a knowledgeable assistant who can answer factual questions.

User Prompt: Today Date: 1 January 2025.

Given a question and its relevant context, you should answer it using your own knowledge or the knowledge provided by the context. Remember, the provided context may not necessarily be up-to-date to answer the question, and your answer must contain only the name, with no other words.

CONTEXT: {Past temporal misaligned context}

QUESTION: The {property} of {subject} is current held by??

Your answer:

Prompt A.5: Trustworthiness (Previous Unanswerable date)

System Prompt: You are a knowledgeable assistant who can answer factual questions.

User Prompt: Today Date: {date}.

Given a question, you should answer it using your own knowledge. Remember, please output 'Unknown' only if the answer does not exist. Otherwise, output the name only.

QUESTION: The {property} of {subject} is current held by??

Your answer:

Prompt A.6: Trustworthiness (Future Unanswerable date)

System Prompt: You are a knowledgeable assistant who can answer factual questions.

User Prompt: Today Date: 1 October 2050.

Given a question, you should answer it using your own knowledge. Remember, please output 'Unknown' only if the answer does not exist. Otherwise, output the name only.

QUESTION: The {property} of {subject} is current held by??

Your answer:

Prompt A.7: Understanding (Explicit Temporal Concept)

System Prompt: You are a knowledgeable assistant who can answer factual questions.

User Prompt: You should answer the question using your knowledge and reasoning capacity. Remember, your answer must contain only the name, with no other words.

QUESTION: Who served as {property} of {subject} from $\{T_{start}\}$ to $\{T_{end}\}$?

Your answer:

Prompt A.8: Understanding (Implicit Temporal Concept)

System Prompt: You are a knowledgeable assistant who can answer factual questions.

User Prompt: You should answer the question using your knowledge and reasoning capacity. Remember, your answer must contain only the name, with no other words.

QUESTION: Who served as the {property-1} of {subject-1} when {attribute-2} served as the {property-2} of {subject-2}?

Your answer:

Prompt A.9: Reasoning (Ranking)

System Prompt: You are a knowledgeable assistant who can answer factual questions.

User Prompt: You should answer the question using your knowledge and reasoning capacity. Remember, your answer must contain only the name, with no other words.

QUESTION: {attribute-1} and {attribute-2} served as the {property} of {subject}, respectively. Can you identify which one the former {property} was?

Your answer:

Prompt A.10: Reasoning (Calculation)

System Prompt: You are a knowledgeable assistant who can answer factual questions.

User Prompt: You should answer the question using your knowledge and reasoning capacity. Remember, your answer must contain only the name, with no other words.

QUESTION: {attribute-1} served as the {property} of {subject}. Can you identify who occupied this position before {num-of-days} days?

Your answer: