Temporal Referential Consistency: Do LLMs Favor Sequences Over Absolute Time References?

Ashutosh Bajpai[1,2], Tanmoy Chakraborty[1]

[1]Indian Institute of Technology Delhi, India
[2]MongoDB, Inc., Delhi, India
{eez228482,tanchak}@ee.iitd.ac.in

Abstract

The increasing acceptance of large language models (LLMs) as an alternative to knowledge sources marks a significant paradigm shift across various domains, including time-sensitive fields such as law, healthcare, and finance. To fulfill this expanded role, LLMs must not only be factually accurate but also demonstrate consistency across temporal dimensions, necessitating robust temporal reasoning capabilities. Despite this critical requirement, efforts to ensure temporal consistency in LLMs remain scarce including noticeable absence of endeavors aimed at evaluating or augmenting LLMs across temporal references in time-sensitive inquiries. In this paper, we seek to address this gap by introducing a novel benchmark entitled temporal referential consistency, accompanied by a resource `TEMP-ReCon` designed to benchmark a wide range of both open-source and closed-source LLMs with various linguistic contexts characterized by differing resource richness (including English, French, and Romanian). The findings emphasis that LLMs do exhibit insufficient temporal referent consistency. To address this, we propose `UnTRaP`, a reasoning path alignment-based model that aims to enhance the temporal referential consistency of LLMs. Our empirical experiments substantiate the efficacy of `UnTRaP` compared to several baseline models.

1 Introduction

Temporal reasoning encompasses the understanding of time-related relationships, including time-time, time-event, and event-event correlations (Tan et al., 2023). This capability allows Large Language Models (LLMs) (OpenAI et al., 2024; Team et al., 2024) to process and interpret the evolving nature of information, which is essential in fields sensitive to temporal changes, such as finance, healthcare, and legal studies. By enhancing LLMs with robust temporal reasoning abilities, their analytical competencies are significantly improved, facilitating complex tasks such as timelines summarization (Sojitra et al., 2024), disease progression (Lu and Naseem, 2024), managing contracts negotiations (Narendra et al., 2024), and conducting analysis on historical text (Zeng, 2024).

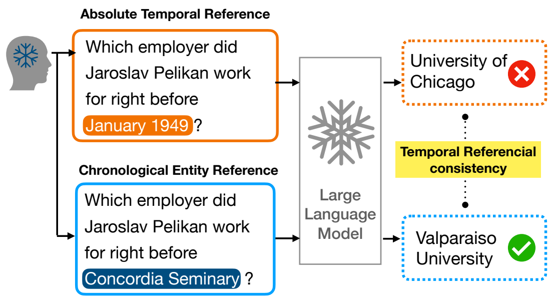

Figure 1: An illustration of temporal referential consistency.

Consistency. Consistency in LLMs ensures reliable and coherent responses across diverse inputs, making them essential for applications such as decision support systems. Knowledge Bases (KBs), built on manually engineered schemas, have historically been used for information retrieval to ensure factual and consistent answers (Petroni et al., 2019). Conversely, inconsistent factuality is reported as a prevalent issue in LLMs, especially in autoregressive settings (Tam et al., 2023a). Ravichander et al. (2020) evaluated consistency using paired probes, while Elazar et al. (2021) expanded this analysis by examining and improving LLM consistency across different types of factual knowledge, thereby defining consistent factuality in LLMs.

Temporal Consistency. Recently, Bajpai et al. (2024) expanded the evaluation of consistent factuality across the temporal axis, marking it as a preliminary attempt in assessing consistent temporal reasoning in LLMs. However, their definition of temporal consistency is limited to surface-form alterations (paraphrasing) of time-sensitive queries,

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 17618–17636
November 4-9, 2025

neglecting consistency across various temporal references such as an event or absolute time. To address this gap in the literature, we aim to formalize two research questions in this study.

- **RQ1:** To what extent do LLMs possess the capability for temporal referential consistency?
- **RQ2:** How can LLMs be improved to address inadequate temporal referential consistency?

As illustrated in Figure 1, temporal referential consistency expects identical responses across absolute (e.g., January 1994) and chronological temporal references, which rely on implicit temporal information (e.g., an event like Concordia Seminary). To this end, we introduce `TEMP-ReCon`, a novel multilingual dataset, and expand existing metrics (defined in Section 4.1) to quantitatively benchmark LLMs for temporal referential consistency across languages. Our study focuses on English, French, and Romanian, chosen for their varied resource richness, with Romanian identified as a low-resource language. Findings indicate that LLMs exhibit inadequate performance regarding temporal referential consistency across languages.

To address this issue, we propose a novel method – `UnTRaP (Unifying Temporal Reasoning Pathways)`. First, we define two reasoning pathways a model can utilize to arrive at the anticipated answer despite divergent temporal references: event-oriented and time-oriented reasoning pathways (defined in Section 3). We hypothesize that alignment between reasoning pathways within the LLM's hyperspace will enable the model to consistently respond to variations in temporal references. Therefore, we adopt a supervised instruction-tuning approach for this alignment. While several attempts have been made to identify, evaluate, or optimize reasoning paths (Chung et al., 2022; Yuan et al., 2023), existing methods lack a unified approach to enhance consistency. Our novel method, `UnTRaP`, addresses this gap by unifying reasoning paths to achieve improved consistency. We compare `UnTRaP` with six baselines and demonstrate that it outperforms the best baseline (Tan et al., 2023) by 9.06, and 5.47 percentage points, respectively, for temporal referential consistency and temporally referential consistent factuality.

To summarize, our contributions are as follows[1]:

- We propose `TEMP-ReCon`, a multilingual resource to evaluate a new benchmark: temporal referential consistency in LLMs. To achieve this, we build upon the metrics proposed by Bajpai et al. (2024) for performance benchmarking.
- Our findings indicate that LLMs demonstrate superior performance in handling chronological temporal references as opposed to absolute temporal references, thereby resulting in greater temporal referential inconsistency across languages.
- We propose `UnTRaP`, which improves the temporal referential consistency of LLMs and empirically outperforms robust baselines across multiple languages.

[1]Source code and dataset are available at `https://github.com/ab-iitd/untrap`.

	Train	**Dev**	**Test**
Time Range	1015-2023	634-2023	1001-2022
#Total Queries	13,014	4,437	4,426

Absolute Temporal Reference Query
Which employer did Jaroslav Pelikan work for right before January 1949?
Chronological Temporal Reference Query
Which employer did Jaroslav Pelikan work for right before Concordia Seminary?
Correct Answer
Valparaiso University
Time-Oriented Reasoning Pathway
Because jaroslav pelikan worked for concordia seminary from january 1949 to january 1953, and right before january 1949, jaroslav pelikan worked for valparaiso university.
Event-Oriented Reasoning Pathway
Because jaroslav pelikan worked for concordia seminary from january 1949 to january 1953, and right before concordia seminary, jaroslav pelikan worked for valparaiso university.

Table 1: Dataset statistics for `TEMP-ReCon` along with a representative instance of the dataset.

2 Dataset

Despite extensive research in temporal reasoning, no dataset currently exists to specifically support evaluation of LLMs for temporal referential consistency. Among notable datasets, TEMPREASON (Tan et al., 2023) emerges as a comprehensive benchmark, offering a multifaceted approach to temporality across an extended temporal period. TEMPREASON is a Q/A style dataset that categorizes temporal reasoning into three distinct types: time-time (L1), time-event (L2), and event-event (L3). It expects time, event, and event as a response from the model when provided with time, time, and event information in a query. Consequently, we sourced this dataset to develop a novel resource for evaluating temporal referential consistency in LLMs, designated as `TEMP-ReCon`[2] (Temporal Referential Consistency). We select the instances of event-event category (L3) relations from TEMPREASON due to its provision of relations among events alongside the temporal periods associated with them. An event-event temporal query involves identifying an answer-object chronologically related to a query-object, based on a subject-relation-object triplet and directional cues like "before" or "after."

`TEMP-ReCon`. An instance of `TEMP-ReCon` comprises five fields: absolute temporal reference query, chronological temporal reference query, correct answer, time-oriented reasoning pathway, and event-oriented reasoning pathway. To construct the chronological temporal reference-based query, we utilize the default L3 question from the dataset, subsequently replacing the query event with the appropriate absolute temporal reference corresponding to the time associated with the event to obtain chronological temporal reference query. Additionally, we devise the event and time-oriented reasoning paths based on information provided under the field *fact_context*, present in source data. *fact_context* states all the temporal facts for a query. Algorithm 1 presents details on data construction.

Multilingual Expansion. Further, we extend the dataset to non-english languages to support the evaluation of multilingual temporal referential consistency. Inspired by Bajpai and Chakraborty (2025), we select French and Romanian due to their varying levels of resource availability in comparison to English. Romanian and French possess 98.3% and 78% fewer speakers than English, respectively. We utilize the T5 model (Raffel et al., 2023) to facilitate the automatic translation from English to Romanian and French. We rely on the source dataset for the data quality standard. Additionally, we report the mean Translation Success Rate (TSR) of 99.4% and the mean Back-Translation Accuracy (BTA) of 71.49% and 45.49%, based on BERT-Score and ChrF++ (Popović, 2015), respectively, to evaluate translation quality. To report inter-translator agreement, we employed NLLB-200 (Team et al., 2022) alongside T5, achieving mean translation agreement scores of 93.75 and 62.48 as measured using BERTScore and BLEU-3 based metrics. Dataset statistics, along with a representative instance, is detailed in Table 1 (Reasers can refer to Appendix B for more details on proposed data).

[2]The resource was curated by a male NLP expert, within the 30-40 age bracket.

Algorithm 1 Data Construction Approach

1: **Step 1: Selection of Data Points**
2: Select data points from the event-event category of TEMPREASON, leveraging its provision of event relations and temporal periods.
3: **Step 2: Adding ’right’ Keyword**
4: Add the keyword "right" before the reference terms "before/after" in the query to restrict probing to a one-to-one setting and access the immediate entity before or after the provided temporal reference.
5: **Step 3: Generating Chronological Temporal Reference Query**
6: Transform the query using the added keyword "right" to produce chronological temporal reference queries.
7: **Step 4: Generating Absolute Temporal Reference Query**
- Extract the reference entity from the query using reference terms ’before/after’.
- Search the ‘fact_context‘ field in TEMPREASON for the extracted entity to retrieve its "from" and "to" time periods.
- Based on the keyword ("before/after"), select the "from" or "to" time for the entity.
- Replace the reference entity in the query with the absolute time to create the absolute temporal reference query.

8: **Step 5: Constructing Reasoning Pathways**
9: Build two reasoning pathways:
10: **Step 5.1: Time-Oriented Reasoning Pathway**
- Extract the absolute time reference from the absolute temporal reference query.
- Search the ‘fact_context‘ for facts corresponding to $t+1$ or $t-1$ based on "after" or "before" type of query.
- Extract the fact containing the correct answer, remove "from" and "to" time information from it, and concatenate using the phrase "right before/after [absolute time reference]."
- Prefix with the keyword "because" to construct the final time-oriented reasoning pathway.

11: **Step 5.2: Event-Oriented Reasoning Pathway**
- Extract the event reference from the chronological temporal reference query.
- Search the ‘fact_context‘ for facts corresponding to $t+1$ or $t-1$ based on "after" or "before" type of query.
- Extract the fact containing the correct answer, remove "from" and "to" time information, and concatenate it using the phrase "right before/after [chronological temporal reference]."
- Prefix with the keyword "because" to construct the final event-oriented reasoning pathway.

3 Proposed Methodology

3.1 Preliminaries

Let us consider a dataset D that comprises M pairs of referential paraphrases of temporal queries, denoted as (q_x^a, q_x^c), along with an associated answer a_x. In this context, q_x^a represents the temporal query with absolute temporal reference, while q_x^c represents the temporal query with chronological

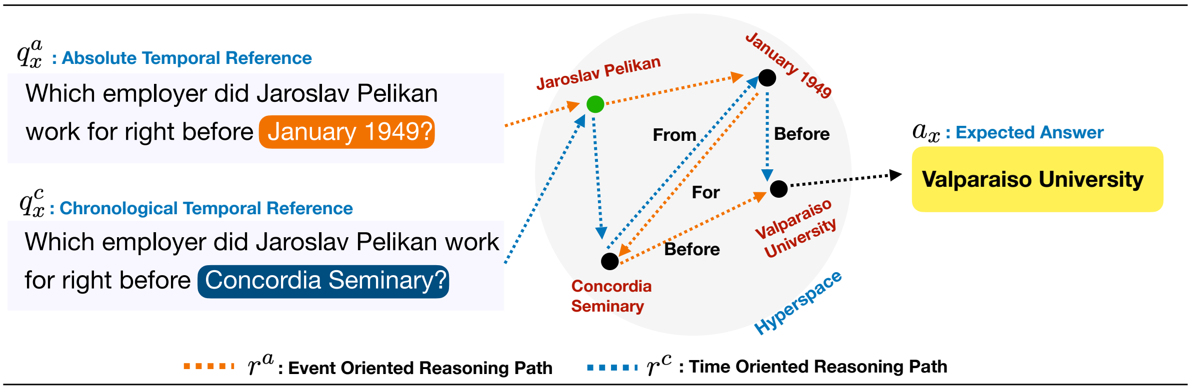

Figure 2: The architectural formulation of `Unifying Temporal Reasoning Pathways` (UnTRaP). r^a: an event-oriented reasoning path that sequences time-event:event-event to address absolute temporal reference queries. r^c: a time-oriented reasoning path that follows event-time:time-event for chronological temporal reference queries.

temporal reference, where $x \in M$. Within a classical supervised fine-tuning framework, a model Θ is subjected to the loss objective L_{sft}, as articulated in Equation 1. The primary objective of this framework is to jointly maximize the probability of obtaining a_x given either q_x^a or q_x^c.

$$\mathcal{L}_{sft} = -\sum_{x=1}^{M} logP(a_x | \, q_x^a) + logP(a_x | \, q_x^c) \quad (1)$$

Bajpai et al. (2024) extended the earlier formulation to a multi-task (MT) fine-tuning framework. This approach demonstrates that the inclusion of an explicit consistency task significantly influences task transfer, thereby enhancing the temporally consistent factuality within the model. The model, denoted as Θ, is subject to joint fine-tuning through an MT objective that encompasses both factuality and consistency. The loss objective pertaining to the consistency task, denoted as L_{con}, is presented in Equation 2. Consistency is framed as a binary generative task, wherein the objective is to maximize the probability of the ground truth class labels $True$ and $False$, corresponding to two positive and antagonistic time-sensitive referential paraphrases, respectively. Here, the class label c is defined within the set $\{True, False\}$.

$$\mathcal{L}_{con} = -\sum_{x=1}^{N} logP(c_x \, | \, (q_x^a, q_x^c)) \quad (2)$$

Therefore, the final loss objective of the MT approach, L_{total}, is delineated in Equation 3. In this formulation, L_{fact} represents the loss objective for the primary task of temporal factuality, consistent with Equation 1. Furthermore, they have introduced consistent time-sensitive reinforcement learning (CTSRL) to enhance the model Θ by implementing an independent reward mechanism for both tasks within the MT approach.

$$\mathcal{L}_{total} = \mathcal{L}_{fact} + \mathcal{L}_{con} \quad (3)$$

3.2 `Unifying Temporal Reasoning Pathways` (UnTRaP)

Thus far, the objective presents simplistic assumptions that establish a direct association between the answer a_x and the query q_x. However, it neglects to assess or enhance the logical deductions– termed time-sensitive reasoning pathways– that link an answer to a query. Such deductions may also be multi-step in nature. To this end, as illustrated in Figure 2, we define two time-sensitive reasoning pathways, r^a and r^c, for accessing the answer a_x given queries q_x^a and q_x^c. The first time-sensitive reasoning pathway, r^a, given a query with absolute temporal reference, adheres to the event-oriented order of *time-event:event-event*. For instance, consider the query: *Which employer did Jaroslav Pelikan work for right before January 1949?* This query provides an absolute temporal reference (January 1949) to ascertain the correct answer, an event (Valparaiso University). Consequently, a reasoning pathway may first lead to another event, denoted as event (Concordia Seminary), utilizing the specified time (January 1949) and then proceed from event (Concordia Seminary) to the anticipated event (Valparaiso University). Thus, r^a for the given example can be articulated as: *Jaroslav Pelikan worked for Concordia Seminary from January 1949 to January 1953, and prior to Concordia Seminary, Jaroslav Pelikan was employed by Valparaiso University.*

$$P^a = \arg\max_{a_x \in A} p\left(a_x,\, r_x^a \mid q_x^a\right) \quad (4)$$

$$P^c = \arg\max_{a_x \in A} p\left(a_x,\, r_x^c \mid q_x^c\right) \quad (5)$$

Whereas another time-sensitive reasoning pathway r^c, given a query with chronological temporal reference, adheres to the time-oriented sequence of *event-time:time-event* order. Consider the same case within a chronological temporal reference setting. The query is - *Which employer did Jaroslav Pelikan work for immediately prior to Concordia Seminary?*, where the chronological temporal reference, an event (Concordia Seminary), is utilized to ascertain the correct answer, which is another event (Valparaiso University). In this scenario, we first identify an absolute temporal reference, time (January 1949), corresponding to the given event, and subsequently follows the pathway from this time to the expected event (Valparaiso University). Thus, the reasoning pathway r^c for the given example can be articulated as *Jaroslav Pelikan worked for Concordia Seminary from January 1949 to January 1953, and immediately prior to January 1949, he was employed by Valparaiso University.*

$$\mathcal{L}_{\texttt{UnTRaP}} = -\sum_{x=1}^{M} logP^a(a_x,\, r_x^a \mid q_x^a) + logP^c(a_x,\, r_x^c \mid q_x^c) \tag{6}$$

The objective is to establish unified time-sensitive reasoning pathways across temporal references. It is anticipated that the absolute temporal reference-based query adheres to the event-oriented path, while the chronological temporal reference-based query aligns with the time-oriented path. This cross alignment between the two pathways aims to enhance referential consistency and maximize temporal factuality. The likelihood estimation for absolute temporal reference and chronological temporal reference is denoted as P^a and P^c, as presented in Equations 4 and 5, respectively. Meanwhile, A represents the entire vocabulary space encompassing the anticipated answer a_x. Consequently, this leads to the final loss objective $L_{\texttt{UnTRaP}}$ detailed in Equation 6.

4 Experiments

4.1 Experimental Setting

We primarily utilize LLaMA3.1-8B (Grattafiori et al., 2024) for all experimental procedures. We employ LoRA-based supervised instruction-tuning (IT) (Ouyang et al., 2022; Hu et al., 2021) for implementing UnTRaP on the base language, English. Furthermore, we finetune $UnTRaP^*$ using training instances across multiple languages. We utilize a test set to benchmark all the results presented in this study. Details including prompts presented in appendices A and D.

Baselines. The experimental baselines include several established approaches. In-Context Learning (ICL) (Brown et al., 2020), Semantic ICL (Liu et al., 2022), Semantic Chain-of-Thoughts Reasoning (Semantic CoT) builds on standard CoT (Wei et al., 2022), SFT Model (Ouyang et al., 2022), SFT + TSRL (Tan et al., 2023), and CoTSeLF (Bajpai et al., 2024) (see Appendix A for details.)

Evaluation Metrics. We employ and refine the definition of evaluation metrics established by Elazar et al. (2021); Bajpai et al. (2024) to evaluate the model's temporal referential consistency. The first metric, referred to as *Temporal Referential Factual Deviation*, quantifies the disparity in temporal factuality between queries based on absolute and chronological temporal references as depicted in Equation 8. This evaluation is grounded in two established metrics: Exact Match (EM), presented in Equation 7, and similarly for F1-Score (F1) to measure partial correctness, which assess temporal factuality across temporal references. Here, (*) is a placeholder for absolute or chronological temporal reference and a' is the generated answer. The subsequent metric, termed *Temporal Referential Consistency*, employs a binary estimation – assigning a value of one or zero – to determine whether the model's responses to a given pair of queries remain identical or differ across temporal references, presented in Equation 9.

$$EM_x^* = \begin{cases} 1, & \text{if } a_x = a'^*_x \\ 0, & \text{if } a_x \neq a'^*_x \end{cases} \tag{7}$$

Temporal Referential Factual Deviation

$$= \frac{1}{M}\left(\sum_{x=1}^{M} em_x^a - \sum_{x=1}^{M} em_x^c\right) \tag{8}$$

Temporal Referential Consistency

$$= \frac{1}{M}\sum_{x=1}^{M} \begin{cases} 1, & \text{if } a'^a_x = a'^c_x \\ 0, & \text{if } a'^a_x \neq a'^c_x \end{cases} \tag{9}$$

The third and final metric, designated as *Temporally Referential Consistent Factuality*, serves as a composite measure of the preceding two metrics, establishing a more stringent criterion for model responses to be both temporally consistent and factually accurate across absolute and chronological temporal references, presented in Equation 10.

Model	Method	Temp-Ref-Fact						Temp-Ref-Cons	Temp-Ref-Cons-Fact
		EM			F1.				
		CTR	ATR	Dev.	CTR	ATR	Dev.		
Open-Source LLMs									
LLaMA3.1-8B	ICL	7.95	5.01	-2.94	18.82	13.40	-5.42	26.77	2.76
	Semantic ICL	12.94	10.23	-2.71	20.96	19.76	-1.20	20.70	6.82
	Semantic CoT	17.84	13.51	-4.34	24.58	23.10	-1.48	23.32	10.42
Mistral-v1	ICL	4.27	3.29	-0.98	16.61	10.15	-6.46	18.53	1.81
	Semantic ICL	11.95	7.79	-4.16	22.50	18.25	-4.25	18.12	5.08
	Semantic CoT	17.21	12.29	-4.92	24.84	21.76	-3.08	22.10	9.67
Vicuna-7b-v1.5	ICL	2.59	0.88	-1.71	12.70	6.79	-5.91	14.21	0.41
	Semantic ICL	7.36	5.28	-2.08	15.95	13.61	-2.34	11.32	3.37
	Semantic CoT	11.72	5.17	-6.55	17.47	12.52	-4.95	14.71	4.20
Bloomz-7b1	ICL	1.35	0.65	-0.70	8.08	5.84	-2.24	38.84	0.23
	Semantic ICL	3.14	2.80	-0.34	7.49	8.51	1.02	23.81	0.93
	Semantic CoT	7.72	4.08	-3.64	14.85	12.09	-2.76	12.36	2.42
Closed-Source LLMs									
Gemini 2.5 Pro	ICL	36.4	25.4	-11.0	47.5	39.0	-8.5	33.8	19.2
	Semantic ICL	38.4	29.4	-9.0	49.2	38.2	-11.0	35.2	23.4
	Semantic CoT	40.2	36.4	-3.8	49.37	48.18	-1.19	42.2	29.0

Table 2: Default performance of various LLMs on temporal referential consistency across methods – ICL, Semantic ICL, and Semantic CoT, for English test set. We choose a 3-shot setting across methods. Temp-Ref-Fact (Dev.): temporal factuality deviation across references (in %), Temp-Ref-Cons: temporal referential consistency (in %), Temp-Ref-Cons-Fact: temporally referential consistent factuality (in %). EM is an exact match-based metric. CTR and ATR represents chronological and absolute temporal references.

Model	Method	Temp-Ref-Fact						Temp-Ref-Cons	Temp-Ref-Cons-Fact
		EM			F1.				
		CTR	ATR	Dev.	CTR	ATR	Dev.		
English	ICL	7.95	5.01	-2.94	18.82	13.40	-5.42	26.77	2.76
	Semantic ICL	12.94	10.23	-2.71	20.96	19.76	-1.20	20.70	6.82
	Semantic CoT	17.84	13.51	-4.34	24.58	23.10	-1.48	23.32	10.42
French	ICL	2.14	3.02	0.88	10.95	9.47	-1.48	10.94	0.95
	Semantic ICL	8.58	8.33	-0.25	17.77	16.53	-1.24	10.21	4.13
	Semantic CoT	7.68	6.91	-0.77	19.95	17.90	-2.05	19.30	3.23
Romanian	ICL	2.32	2.28	-0.04	7.07	6.79	-0.28	8.65	0.77
	Semantic ICL	11.56	8.35	-3.21	16.85	14.54	-2.31	12.18	6.57
	Semantic CoT	12.78	9.55	-3.23	18.68	16.42	-2.26	20.40	7.37

Table 3: Default performance of LLaMA3.1-8B on temporal referential consistency across languages- English, French, and Romanian. We choose a 3-shot setting across various methods.

Temporal Referential Consistent Factuality

$$= \frac{1}{M}\sum_{x=1}^{M}\begin{cases}1, & \text{if } a'^{a}_{x} = a'^{c}_{x} \\ & \text{and if } a'^{a}_{x} = a^{c}_{x} \\ 0, & \text{if } a'^{a}_{x} \neq a'^{c}_{x}\end{cases} \tag{10}$$

4.2 Experimental Results

Temporal Referential Consistency Across LLMs. We start with benchmarking temporal referential consistency across distinct contemporary LLMs, including the multilingual model – LLaMA3.1-8B, the English-dominant instruction-tuned model Vicuna (Zheng et al., 2023), the French and German fluent Mistral (Jiang et al., 2023), and the cross-lingual specialized Bloomz (Muennighoff et al., 2023). We utilize three prompting strategies: ICL, Semantic ICL, and Semantic CoT. Table 2 shows that all models exhibit poor performance, with metrics ranging from -0.34% to -4.34% for EM-based temporal referential factual deviation, 11.32% to 38.84% for temporal referential consistency, and 0.23% to 10.42% for temporally referential consistent factuality. Notably, models perform better on chronological reference queries than on absolute temporal references. Additionally, prompting strategies significantly influence performance, with Semantic CoT enhancing temporally referential consistent factuality while exacerbating the performance gap for EM-based temporal factuality. We use an English test set for this experiment. Additionally, we employ, Gemini-2.5-Pro, one of the most contemporary benchmarks in closed-source LLMs. We observe

Method	Temp-Ref-Fact						Temp-Ref-Cons	Temp-Ref-Cons-Fact
	EM			F1.				
	CTR	ATR	Dev.	CTR	ATR	Dev.		
ICL	7.95	5.01	-2.94	18.82	13.40	-5.42	26.77	2.76
Semantic ICL	12.94	10.23	-2.71	20.96	19.76	-1.20	20.70	6.82
Semantic ICL	17.84	13.51	-4.34	24.58	23.10	-1.48	23.32	10.42
SFT (IT)	13.30	11.99	-1.31	21.26	20.89	-0.37	32.99	8.40
IT + TSRL ↑	14.03	11.54	-2.49	22.22	21.82	-0.39	33.57	8.20
MT-IT	8.40	7.77	-0.63	22.04	21.46	-0.57	24.22	3.84
CoTSeLF (MT-IT + CTSRL) ⇑	9.69	8.54	-1.15	21.68	21.14	-0.53	25.94	4.7
UnTRaP	17.30	16.85	**-0.44**	25.01	25.00	**-0.01**	**42.63**	**13.67**
$\Delta_{\texttt{UnTRaP}}$ −↑	3.27 ↑	5.31 ↑	2.04 ↑	2.79 ↑	3.18 ↑	0.39 ↑	9.06 ↑	5.47 ↑
$\Delta_{\texttt{UnTRaP}}$ −⇑	7.61 ↑	8.31 ↑	0.70 ↑	3.33 ↑	3.86 ↑	0.53 ↑	16.69 ↑	8.97 ↑

Table 4: Experimental results for UnTRaP across metrics – Temp-Ref-Fact-Dev: temporal factuality deviation across references (in %), Temp-Ref-Cons: temporal referential consistency (in %), Temp-Ref-Cons-Fact: temporally referential consistent factuality (in %), in comparison to multiple baselines on test data with LLaMA3.1-8B. We opt for a 3-shot setting across contextual methods. $\Delta_{\texttt{UnTRaP}}$ −↑: improvement by UnTRaP over TSRL, and $\Delta_{\texttt{UnTRaP}}$ −⇑: improvement by UnTRaP over CoTSeLF.

Language	Methods	M1	M2	M3	M4
French	Semantic CoT	-0.77	-2.05	19.30	3.23
	$UnTRaP^{*}$	0.27	0.55	31.99	10.96
Romanian	Semantic CoT	-3.23	-2.26	20.40	7.37
	$UnTRaP^{*}$	-0.39	0.5	36.42	11.5
English	Semantic CoT	-4.34	-1.48	23.32	10.42
	$UnTRaP^{*}$	0.54	0.74	39.36	12.83
$\overline{\Delta}$		2.38 ↑	1.33 ↑	14.91 ↑	4.75 ↑

Table 5: $UnTRaP^{*}$ over Semantic CoT baseline across languages and metrics – M1:Temp-Ref-Fact-Dev (EM), M2:Temp-Ref-Fact-Dev (F1), M3:Temp-Ref-Cons, and M4: Temp-Ref-Cons-Fact with LLaMA3.1-8B.

that it also demonstrates suboptimal performance regarding temporal references despite significant improvements over counterparts in open-source category.

Suboptimal Performance Across Languages. Next, we benchmark temporal referential consistency across distinct languages. We employ LLaMA3.1-8B and consider English, French, and Romanian for this experiment across three prompting strategies: ICL, Semantic ICL, and Semantic CoT. Table 3 indicates that the model demonstrates suboptimal performance, with metrics ranging from -0.04% to -4.34% for EM-based temporal referential factual deviation, 8.65% to 26.77% for temporal referential consistency, and 0.77% to 10.42% for temporally referential consistent factuality across languages. The models perform better on chronological reference queries than on absolute temporal references across languages. Additionally, there appears to be no significant correlation between the model's temporal referential consistency traits and the resource richness of a language, potentially attributable to the model's extremely poor performance across languages.

Model	Methods	M1	M2	M3	M4
LLaMA3.1-8B	Semantic CoT	-4.34	-1.48	23.32	10.42
	UnTRaP	**-0.45**	**-0.01**	**42.63**	**13.67**
LLaMA2-13B-Chat	Semantic CoT	-2.03	-0.91	13.1	2.55
	UnTRaP	**-1.00**	**0.26**	**43.81**	**12.92**

Table 6: UnTRaP over Semantic CoT baseline across base-models and metrics – M1:Temp-Ref-Fact-Dev (EM), M2:Temp-Ref-Fact-Dev (F1), M3:Temp-Ref-Cons, and M4: Temp-Ref-Cons-Fact for English.

4.2.1 Improvements with UnTRaP

The exhaustive comparison of UnTRaP with baseline methods demonstrates the efficacy of aligning event- and time-oriented reasoning pathways alongside corresponding temporal references, particularly in contrast to the multi-task reinforcement learning-based method, CoTSeLF, as illustrated in Table 4. Notably, UnTRaP achieves enhancements of 16.69 and 8.97 percentage points over CoTSeLF for temporal referential consistency and temporal referential consistent factuality, respectively. Furthermore, UnTRaP exhibits improvements of 9.06 and 5.47 percentage points over the temporal factuality-based baseline, TSRL for temporal referential consistency and temporal reference consistent factuality, respectively. Consistent improvements in temporal factuality across temporal references demonstrate that UnTRaP enhances consistency without compromising factual accuracy. English variant of TEMP-ReCon is used for this experiment.

Generalization Across Languages. Results presented in Table 5, demonstrate that $UnTRaP^{*}$ consistently outperforms the Semantic CoT baseline across all metrics and languages – French, Romanian, and English. Notably, $UnTRaP^{*}$ ex-

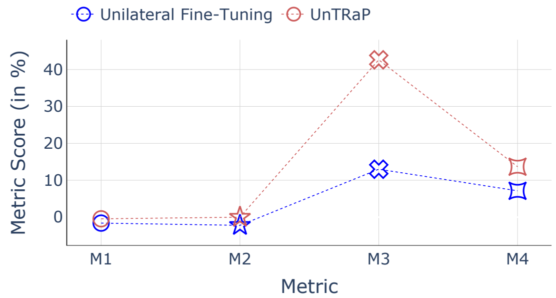

Figure 3: A comparison between UnTRaP with unilateral (event-oriented) reasoning path fine-tuning utilizing only absolute temporal reference training instances. M1: Temp-Ref-Fact-Dev[EM], M2:Temp-Ref-Fact-Dev[F1], M3: Temp-Ref-Cons, M4: Temp-Ref-Cons-Fact.

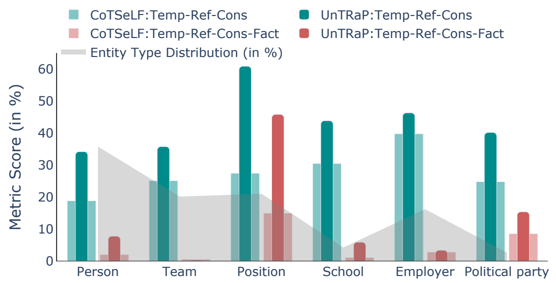

Figure 4: Temporally referential consistency across various entity types in the test data for UnTRaP compared to the baseline CoTSeLF.

hibits substantial improvements in low-resource languages such as Romanian, achieving temporal referential consistency score of 36.42 compared to 20.40 from the Semantic CoT baseline – a significant gain of 14.91. Similarly, across high-resource languages like English, $UnTRaP^{*}$ showcases robust generalization, achieving temporal referential consistent accuracy improvement of 4.75.

Scalability Across Base Models. Table 6 highlights the strong generalization capability of UnTRaP across diverse LLMs, including both base and chat models with parameter sizes – 8B and 13B. For LLaMA2-13B-Chat, temporal referential consistency scores improve to 43.81 (+30.71), demonstrating its ability to scale across different architectures and parameter configurations.

Unilateral Reasoning Path Fine-Tuning. Here, we aim to assess the impact of unilateral reasoning path fine-tuning in comparison to UnTRaP. We fine-tune the base model exclusively with instances derived from absolute temporal references, alongside the event-oriented reasoning path r^a. In Figure 3, a notable decline of 29.64 percentage points in temporal referential consistency is observed when contrasted with UnTRaP, while no significant alteration in temporal referential factual deviation is detected. This finding is crucial and substantiates the argument for UnTRaP, as widely employed unilateral reasoning path fine-tuning methods may substantially undermine the temporal referential consistency of the model.

5 Error Analysis

Qualitative Analysis

Entity Type: Position

- **Query**
  - ABT: Which position did St John Brodrick, 1st Earl of Midleton hold right before November 1885?
  - CBT: Which position did St John Brodrick, 1st Earl of Midleton hold right before Member of the 23rd Parliament of the United Kingdom?
- **Ground Truth Answer**: member of the 22nd parliament of the united kingdom
- **CoTSeLF Response (Baseline)**
  - ABT: member of the 23rd parliament of the united kingdom X
  - CBT: member of the 22nd parliament of the united kingdom ✓
- **UnTRaP Response (Proposed Model)**
  - ABT: member of the 22nd parliament of the united kingdom ✓
  - CBT: member of the 22nd parliament of the united kingdom ✓

Entity Type: Team

- **Query**
  - ABT: Which team did Manuel Hartmann play for right after January 2010?
  - CBT: Which team did Manuel Hartmann play for right after TuS Koblenz?
- **Ground Truth Answer**: FC Ingolstadt 04
- **CoTSeLF Response (Baseline)**
  - ABT: fc köln X
  - CBT: fc bayern munich X
- **UnTRaP Response (Proposed Model)**
  - ABT: fc ingolstadt 04 ✓
  - CBT: fc köln X

Here we evaluate the performance of the new model across various entity types to identify key error-contributing factors. In Figure 4, we compare the temporal referential consistency and temporally referential consistent factuality among six entity types in the test data for UnTRaP and the baseline, CoTSeLF. The correlation score between entity-type distribution and the new model's temporal referential consistency is -0.15, suggesting that entity sample size does not influence performance. In contrast, the correlation score between CoTSeLF and the proposed model's temporally referential consistent factuality is 0.95, indicating consistent factual performance distribution across entity types. These findings suggest that `UnTRaP` **does not specifically target performance disparities among entity types within the foundation model**. Notably, the model shows disproportionate improvements of 33.44 and 30.86 percentage points for temporal referential consistency and consistent factuality for the 'position' entity type. However, it fails to produce significant enhancements for the 'team' name entity type, **highlighting the need for entity-specific adjustments within** `UnTRaP`. An explicit example of such a case is presented in the qualitative analysis box, where `UnTRaP` improves upon the baseline for the Position entity type but fails in the case of the Team entity type.

We also provide extended results in Appendix A, hyperparameters in Appendix C, and commentaries on DeepSeek-R1, and domain adaptation and the applicability in closed-source models in Appendix A and Appendix E, respectively.

6 Related Work

Research on temporal reasoning finds its roots in pioneering efforts such as TimeBank (Pustejovsky et al., 2003) and TempEval (Verhagen et al., 2010), laying the groundwork for understanding temporal expressions and their relationships in text, enabling more nuanced comprehension of temporal data. The evolution of Temporal Knowledge Graph Completion (TKGC) has led to the emergence of significant QA datasets, such as TEQUILA (Jia et al., 2018), TimeQuestions (Jia et al., 2021), and CronQuestions (Saxena et al., 2021). Simultaneously, the consistency of LLMs (Liu et al., 2024) has been rigorously examined, yielding methods to assess inconsistencies across various tasks, including question answering (Kassner et al., 2020; Lee and Kim, 2024), dialogue comprehension (She et al., 2024; Tang et al., 2022; Wang et al., 2022), summarization (Tam et al., 2023b; Luo et al., 2024; Laban et al., 2023), and natural language inference (NLI) (Li et al., 2019). Moreover, Jain et al. (2023) conducted benchmarking of several LLMs using existing time-sensitive datasets.

The rapid deployment of language models in the public domain accentuates the imperative for both temporal accuracy and consistency within generated responses. Given this necessity, several time-sensitive QA datasets, such as TEMPLAMA (Dhingra et al., 2022) and TEMPREASON (Tan et al., 2023), have been established to evaluate and benchmark the temporal reasoning capabilities inherent in LLMs. Another pivotal contribution, TGQA (Temporal Graph QA) (Xiong et al., 2024), was proposed as a framework for the text-to-Temporal Graph translation task. Notably, TEMPREASON serves as a comprehensive benchmark for temporal reasoning, covering diverse temporal periods and three levels of temporal relations. Bajpai et al. (2024) introduced the TEMP-COFAC dataset, revealing significant temporal consistency deficits in medium-sized LLMs. While Kougia et al. (2024) assessed temporal consistency for relation extraction in biomedical text, their approach primarily focuses on surface changes in time-sensitive queries, neglecting consistency across temporal references. In contrast, our dataset, `TEMP-ReCon`, addresses this critical measure of temporal consistency.

7 Conclusion

This paper presented a novel benchmark – temporal referential consistency, along with a dataset `TEMP-ReCon` designed to assess the sound temporal reasoning capabilities of LLMs. Our analysis revealed that all examined LLMs, across various linguistic contexts, demonstrate inadequate temporal referential consistency. Furthermore, we proposed `UnTRaP` which unifies time- and event-oriented reasoning pathways across temporal references to enhance the model's temporal referential consistency. Empirical findings indicated that `UnTRaP` outperforms several baseline models. The contributions and findings aim to augment public trust in LLMs, particularly in critical domains such as law and healthcare, thereby facilitating the adoption of LLMs for time-sensitive information processing.

8 Limitations

The proposed dataset, referred to as `TEMP-ReCon` is designed for the evaluation of temporal referential consistency in large language models (LLMs). It is important to note that `TEMP-ReCon` is derived from an existing dataset, TEMPREASON, and as such, we are reliant on the quality measures established by the original authors. Temporal referential consistency constitutes a specific aspect of temporal consistency (across temporal references); consequently, it necessitates integration with other prominent methodologies to effectively measure temporal consistency and to derive comprehensive metrics for LLM evaluation.

Furthermore, recent literature has highlighted certain limitations inherent in the source dataset, such as its anchoring on a singular entity type, namely, person names, and a constrained range of entity types, which TEMP-ReCon does not rectify. The proposed model, UnTRaP employs a fine-tuning-based approach that necessitates access to the parameters of the foundational model. Consequently, the use of closed-source LLMs may be infeasible, as any benefits realized by UnTRaP could inadvertently extend to these models. Additionally, the persistent challenges associated with computational resources for stacking LLMs further constrain our ability to experiment with models of larger parameter sizes.

9 Ethical Considerations

The dataset sourced in this study, referred to as TEMPREASON, is derived from the Wikidata Knowledge Base and Wikipedia articles, thereby adhering to the respective licensing agreements of these platforms. Specifically, Wikidata operates under a public domain[3] dedication, while Wikipedia articles are licensed under the Creative Commons Attribution-ShareAlike 3.0 Unported License[4]. Consequently, our dataset will be disseminated under the same license as Wikidata, in accordance with the intentions of the original authors of TEMPREASON.

It is important to delineate that the scope of the TEMP-ReCon is strictly confined to the scientific investigation of temporal referential consistency within language models. Although TEMPREASON may encompass information pertaining to entities and their interrelations that are publicly accessible, any biases inherent in the selection of entities or the authenticity of content of the source data do not necessarily represent the perspectives or positions of the authors of this paper. In this study, all data annotators were compensated fairly for their contributions. The lack of women's representation in TEMP-ReCon annotations reflects global gender disparity in research and innovation, though the authors do not share this perspective. We genuinely support gender and racial equality in these fields. Additionally, no generative AI tools were used to create this content, aside from those for spell checking and grammar correction.

[3]https://www.wikidata.org/wiki/Wikidata:Licensing

[4]https://en.wikipedia.org/wiki/Wikipedia:Copyrights

Acknowledgments

Tanmoy Chakraborty acknowledges the support of Google GCP Grant, IBM-IITD AI Horizons network and Rajiv Khemani Young Faculty Chair Professorship in Artificial Intelligence.

References

Ashutosh Bajpai and Tanmoy Chakraborty. 2025. Multilingual llms inherently reward in-language time-sensitive semantic alignment for low-resource languages. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(22):23469–23477.

Ashutosh Bajpai, Aaryan Goyal, Atif Anwer, and Tanmoy Chakraborty. 2024. Temporally consistent factuality probing for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15864–15881, Miami, Florida, USA. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2022. Scaling instruction-finetuned language models. *Preprint*, arXiv:2210.11416.

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.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Aaron Grattafiori, Abhimanyu Dubey, and Others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Raghav Jain, Daivik Sojitra, Arkadeep Acharya, Sriparna Saha, Adam Jatowt, and Sandipan Dandapat. 2023. Do language models have a common sense regarding time? revisiting temporal commonsense reasoning in the era of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6750–6774, Singapore. Association for Computational Linguistics.

Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. 2018. Tequila: Temporal question answering over knowledge bases. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 1807–1810, New York, NY, USA. Association for Computing Machinery.

Zhen Jia, Soumajit Pramanik, Rishiraj Saha Roy, and Gerhard Weikum. 2021. Complex temporal question answering on knowledge graphs. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 792–802, New York, NY, USA. Association for Computing Machinery.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Nora Kassner, Benno Krojer, and Hinrich Schütze. 2020. Are pretrained language models symbolic reasoners over knowledge? In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 552–564, Online. Association for Computational Linguistics.

Vasiliki Kougia, Anastasiia Sedova, Andreas Stephan, Klim Zaporojets, and Benjamin Roth. 2024. Analysing zero-shot temporal relation extraction on clinical notes using temporal consistency. *Preprint*, arXiv:2406.11486.

Philippe Laban, Wojciech Kryscinski, Divyansh Agarwal, Alexander Fabbri, Caiming Xiong, Shafiq Joty, and Chien-Sheng Wu. 2023. SummEdits: Measuring LLM ability at factual reasoning through the lens of summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9662–9676, Singapore. Association for Computational Linguistics.

Yanggyu Lee and Jihie Kim. 2024. Evaluating consistencies in llm responses through a semantic clustering of question answering. *Preprint*, arXiv:2410.15440.

Tao Li, Vivek Gupta, Maitrey Mehta, and Vivek Srikumar. 2019. A logic-driven framework for consistency of neural models. *Preprint*, arXiv:1909.00126.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2024. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *Preprint*, arXiv:2308.05374.

Haohui Lu and Usman Naseem. 2024. Can large language models enhance predictions of disease progression? investigating through disease network link prediction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17703–17715, Miami, Florida, USA. Association for Computational Linguistics.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2024. Factual consistency evaluation of summarization in the era of large language models. *Expert Systems with Applications*, 254:124456.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Savinay Narendra, Kaushal Shetty, and Adwait Ratnaparkhi. 2024. Enhancing contract negotiations with LLM-based legal document comparison. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 143–153, Miami, FL, USA. Association for Computational Linguistics.

OpenAI, Josh Achiam, and Others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *Preprint*, arXiv:1909.01066.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

James Pustejovsky, Patrick Hanks, Roser Saurí, Andrew See, Rob Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, and Marcia Lazo. 2003. The timebank corpus. *Proceedings of Corpus Linguistics*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer. *Preprint*, arXiv:1910.10683.

Abhilasha Ravichander, Eduard Hovy, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. 2020. On the systematicity of probing contextualized word representations: The case of hypernymy in BERT. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 88–102, Barcelona, Spain (Online). Association for Computational Linguistics.

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

Shuaijie She, Shujian Huang, Xingyun Wang, Yanke Zhou, and Jiajun Chen. 2024. Exploring the factual consistency in dialogue comprehension of large language models. 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 6087–6100, Mexico City, Mexico. Association for Computational Linguistics.

Daivik Sojitra, Raghav Jain, Sriparna Saha, Adam Jatowt, and Manish Gupta. 2024. Timeline summarization in the era of llms. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2657–2661, New York, NY, USA. Association for Computing Machinery.

Derek Tam, Anisha Mascarenhas, Shiyue Zhang, Sarah Kwan, Mohit Bansal, and Colin Raffel. 2023a. Evaluating the factual consistency of large language models through news summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5220–5255, Toronto, Canada. Association for Computational Linguistics.

Derek Tam, Anisha Mascarenhas, Shiyue Zhang, Sarah Kwan, Mohit Bansal, and Colin Raffel. 2023b. Evaluating the factual consistency of large language models through news summarization. *Preprint*, arXiv:2211.08412.

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.

Xiangru Tang, Arjun Nair, Borui Wang, Bingyao Wang, Jai Desai, Aaron Wade, Haoran Li, Asli Celikyilmaz, Yashar Mehdad, and Dragomir Radev. 2022. CONFIT: Toward faithful dialogue summarization with linguistically-informed contrastive fine-tuning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5657–5668, Seattle, United States. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Gemini Team, Petko Georgiev, and Others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Marc Verhagen, Roser Saurí, Tommaso Caselli, and James Pustejovsky. 2010. SemEval-2010 task 13: TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 57–62, Uppsala, Sweden. Association for Computational Linguistics.

Bin Wang, Chen Zhang, Yan Zhang, Yiming Chen, and Haizhou Li. 2022. Analyzing and evaluating faithfulness in dialogue summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4897–4908, Abu

Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Siheng Xiong, Ali Payani, Ramana Kompella, and Faramarz Fekri. 2024. Large language models can learn temporal reasoning. *Preprint*, arXiv:2401.06853.

Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. 2023. Scaling relationship on learning mathematical reasoning with large language models. *Preprint*, arXiv:2308.01825.

Yifan Zeng. 2024. Histolens: An llm-powered framework for multi-layered analysis of historical texts – a case application of yantie lun. *Preprint*, arXiv:2411.09978.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

A Extended Results

Experimental Settings. We primarily utilize LLaMA3.1-8B (Grattafiori et al., 2024) for all experimental procedures, including implementing our UnTRaP model. We opt for supervised instruction-tuning (IT) (Ouyang et al., 2022) to enhance the temporally referential consistency of the pre-trained model. This approach aligns with parameter-efficient fine-tuning (PEFT) techniques, such as low-rank adaptation (LoRA) (Hu et al., 2021), which provide substantial advantages for instruction tuning within cost-effective infrastructures. By leveraging advancements in model fine-tuning alongside the train set of TEMP-ReCon, we aim to fine-tune a task-specific adapter. We utilize a test set to benchmark all the results presented in this study. Our experimental framework uses a three-shot methodology for non-gradient update baselines such as in-context learning (ICL) and chain-of-thought (CoT) reasoning. Further, we use EM-based temporal factuality measure to calculate temporal referential consistency and temporally referential consistent factuality.

Baselines. Here is the details on all the baselines used in this work -

- **In-Context Learning** (ICL) (Brown et al., 2020): The model is primed with a set of instances from training data, which serve as demonstrations, accompanied by a query.
- **Semantic In-Context Learning** (Semantic ICL) (Liu et al., 2022): The Semantic-ICL approach enhances this process by incorporating the retrieval of semantically similar examples for queries from training data.
- **Semantic Chain-of-Thoughts Reasoning** (Semantic CoT): We augment the standard CoT (Wei et al., 2022) approach by drawing semantically aligned examples and respective reasoning steps for this baseline.
- **SFT Model** (Instruction Tuning (Ouyang et al., 2022): The model is fine-tuned on a supervised dataset of instruction, input and corresponding responses, helping it better understand and follow specific instructions.
- **SFT + Time-Sensitive Reinforcement Learning** (TSRL) (Tan et al., 2023): TSRL positively reinforces the model for identifying the correct temporal response among various temporal possibilities. In this approach, we utilize a closed-vocabulary framework, restricting the model to its factual knowledge only.
- **CoTSeLF: MT-IT + Consistent Time-Sensitive Reinforcement Learning** (Bajpai et al., 2024): An explicit binary task to enhance temporal consistency is fine-tuned in conjunction with the primary objective of ensuring temporal factuality followed by a reward-based framework to reinforce the model.

Unilateral Reasoning Path-based Fine-Tuning. The detailed results for the temporal referential consistency for the UnTRaP compared to unilateral reasoning path based fine-tuning is presented in Table 7.

Extended Analysis of Commercial LLMs. Additionally, in our extended evaluation of closed-source LLMs such as GPT-4 (OpenAI et al., 2024) and Claude-3-Opus[5], we observe that they also demonstrate suboptimal performance regarding temporal references, presented in Table 8. The temporal referential factuality deviation for GPT is quantified at -6.40 and -4.69 for EM and F1-

[5]https://www.anthropic.com/news/claude-3-family

Method	Temp-Ref-Fact-Dev		Temp-Ref-Cons	Temp-Ref-Cons-Fact
	EM	F1.		
UnTRaP	-0.45	-0.01	42.63	13.67
Unilateral Fine-tuning	-1.61	-2.22	12.99	7.14

Table 7: Detailed Results: a comparison between UnTRaP with unilateral (event-oriented) reasoning path fine-tuning utilizing only absolute temporal reference-based training instances.

based metrics, respectively, representing the lowest performance among all LLMs assessed in this study, thereby indicating an urgent need for improvement. For benchmarking commercial LLMs, we randomly subsample the test data to a thousand instances due to limited access attributed to financial constraints.

Error Analysis. The detailed results for the temporal referential consistency across entity types is presented in Table 9.

DeepSeek-R1 Qualitative Analysis. DeepSeek-R1 is a flagship model, showcasing strong reasoning capabilities across domain and tasks. We put the DeepSeek-R1 on test for our scenario. We test the model qualitatively across temporal directions before/after and temporal references- absolute and chronological. The reasoning capabilities of DeepSeek-R1, as demonstrated in its responses, presented in Table 10, to questions about the head coach history of Legia Warsaw, showcase several significant challenges with accuracy, consistency, and contextual understanding:

- **Inconsistent Temporal Reasoning:** DeepSeek-R1 fails to align its answers with the factual timeline provided. For example, in response to the first question, it incorrectly states that Edmund Zientara left Legia Warsaw in 1980, when the actual timeline indicates he was head coach from July 1969 to July 1971. Similarly, Józef Zwierzyna, mentioned as succeeding Zientara in 1980, does not appear in the factual data provided, showing a lack of alignment with the given information.
- **Circular Errors in Temporal Placement:** The second question's response claims that Zientara became head coach in July 1971, which contradicts the provided timeline, where Tadeusz Chruściński was listed as the coach after Zientara post-July 1971.
- **Fabrication of Unverified Data:** In the third question, the model completely fabricates a timeline, stating that Lucjan Brychczy was the coach before Zientara (1983–1985), which is inconsistent with any factual evidence supplied and temporally misaligned. This suggests the model introduces information outside the scope of the given context, undermining its reliability.
- **Misinterpretation of Predecessor-Successor Relationships:** The response to the fourth question incorrectly states that Zientara managed the team right before July 1969 (1968–1969), despite the clear evidence that Jaroslav Vejvoda was coach until June 1969 and Zientara only started in July 1969. This demonstrates poor reasoning when handling temporal predecessor-successor relationships, failing to factor transitional periods.

The findings necessitate a more rigorous and focused approach, even within flagship reasoning models such as DeepSeek-R1, to effectively address nuanced challenges in temporal reasoning, with specific emphasis on ensuring temporal referential consistency.

B Dataset and Quality

Dataset. The formulation of queries using the event-event (L3) approach adheres to ten distinct templates, each pertaining to unique entities across six overarching categories of named entities, specifically: person, position, employer, political party, team, and school. The TEMP-ReCon aligns with the original constituents of the source data (TEMPREASON), accompanied by comprehensive data statistics, as presented in Table 11.

The TEMPREASON dataset, widely recognized as a standard benchmark in the field of temporal reasoning. Developed using multiple established predecessors, this dataset serves as the foundation for the novel TEMP-ReCon framework, consists of 13,014 unique queries across various temporal references, with a test set comprising 4,426 instances. The event-event category within the dataset spans six distinct entity types and encompasses ten knowledge base (KB) relations. Additionally, the dataset provides extensive time-period coverage, ranging from the year 634 to 2023. An event-event type of temporal query follows the

Model	Method	Temp-Ref-Fact						Temp-Ref-Cons	Temp-Ref-Cons-Fact
		EM			F1.				
		CTR	ATR	Dev.	CTR	ATR	Dev.		
Closed-Source LLMs									
GPT-4	Zero-Shot	22.5	16.1	-6.40	31.89	27.20	-4.69	27.20	9.60
Claude-3-Opus	Zero-Shot	13.9	12.8	-1.10	24.93	23.26	-1.67	23.20	8.00

Table 8: Zero-shot performance of Commercial LLMs on temporal referential consistency.

Model	Metric	Person	Team	Position	School	Employer	Political party
	#Count of Queries	1583	892	930	187	717	117
CoTSeLF	Temp-Ref-Cons	18.83	25.11	27.42	30.48	39.75	24.79
	Temp-Ref-Cons-Fact	2.02	0.56	14.95	1.07	2.79	8.55
UnTRaP	Temp-Ref-Cons	34.18	35.76	60.86	43.85	46.3	40.17
	Temp-Ref-Cons-Fact	7.77	0.34	45.81	5.88	3.35	15.38

Table 9: Temporally referential consistency across various entity types in the test data for UnTRaP compared to the baseline CoTSeLF along with the count of queries in test set for respective entity types.

structure in which, given a triplet of subject, relation, and query-object, the model is expected to identify answer-object, which is chronologically related to query-object in either forward or backward directions. Key phrases, such as 'after' or 'before,' indicate the anticipated direction of order.

TEMP-ReCon's Quality. To ensure the proposed dataset's high quality, we first evaluated the translation success rate (TSR), which measures successful translations from source to target languages. An automated language detection method, langdetect[6] applied to the entire test dataset revealed an remarkable TSR of 99.40%. Additionally, we assessed translation quality using a back-translation[7] evaluation method (Bajpai and Chakraborty, 2025). A sample of 100 queries from TEMP-ReCon was randomly selected for each translated language, considering both absolute and chronological temporal references. These queries were back-translated into English and compared with their original texts. For quantitative assessment, we employed BERTScore-based F1-score and ChrF++ based metrics, resulting in a mean BERTScore (F1) of 71.49% and mean ChrF++ score of 45.49% between the original and back-translated instances, focusing solely on queries across temporal references.

To evaluate inter-translator agreement, we utilized NLLB-200-1.3B alongside T5 on these 100 test samples. Translations were generated using NLLB for French and Romanian languages and then subsequently compared against T5 outputs to measure agreement at semantic and syntactical levels. This evaluation employed two metrics, BERTScore and BLEU-3, yielding mean translation agreement scores of 93.75 and 62.48, respectively across languages. The individual agreement scores for French were 93.38 (BERTScore) and 57.61 (BLEU-3), while Romanian achieved 94.13 (BERTScore) and 67.36 (BLEU-3). These results demonstrate high semantic alignment across both languages, with Romanian showcasing slightly stronger syntactic agreement compared to French. This highlights that T5 translation capability to achieve robust translation when compared to the recent benchmarks cush as NLLB.

C Hyperparameters

In this study, we utilize the Python-based PyTorch[8] library to implement all fine-tuning approaches. Additionally, to reproduce reinforcement learning-based baselines, we employed the Python-based library trlx[9]. The Hugging Face platform was utilized to stack open-source models, including the base model LLaMA3.1-8B, which served as the foundation for experimentation. For the execution of these experiments, an NVIDIA A100 GPU[10] with 80GB of memory was employed. We set the values of lora_r and lora_alpha at 16 and 32, re-

[6] https://pypi.org/project/langdetect
[7] 4 https://translate.google.com/

[8] https://pytorch.org
[9] https://github.com/CarperAI/trlx
[10] https://www.nvidia.com/en-in/data-center/a100/

Wikkipedia Facts		
Coaches of Legia Warsaw: Czech Republic Jaroslav Vejvoda (July 1966 – June 1969) Poland Edmund Zientara (July 1969 – July 1971) Poland Tadeusz Chruściński (July 1971 – May 1972) Source: `https://en.wikipedia.org/wiki/Legia_Warsaw`		
DeepSeek-R1 Responses		
Direction	**Reference**	**Q/A**
After	Chronological	**Question:** Who was the head coach of the team Legia Warsaw right after Edmund Zientara? **Answer:** After Edmund Zientara left Legia Warsaw in 1980, the next head coach was Józef Zwierzyna, who took over the team in 1980 and managed Legia until 1981. Here's a quick timeline: Edmund Zientara (1978–1980) Józef Zwierzyna (1980–1981) Would you like more details about Zwierzyna's tenure or other Legia coaches?
After	Absolute	**Question:** Who was the head coach of the team Legia Warsaw right after July 1971? **Answer:** After July 1971, the head coach of Legia Warsaw was Edmund Zientara. He took over the team in July 1971 and managed Legia until the end of the 1972 season. Zientara had previously played for Legia and later became a notable coach, leading the team to a Polish Cup victory in 1973 (after the 1972–73 season). His tenure was part of a transitional period for the club in the early 1970s. Would you like more details about his coaching career or Legia's performance during that time?
Before	Chronological	**Question:** Who was the head coach of the team Legia Warsaw right before Edmund Zientara? **Answer:** The head coach of Legia Warsaw right before Edmund Zientara was Lucjan Brychczy. Here's the context: Lucjan Brychczy served as Legia's head coach from 1983 to 1985. Edmund Zientara took over in 1985 and coached the team until 1987. Brychczy is a club legend, having also been a prolific striker for Legia in the 1950s–1970s before transitioning into coaching. Would you like details on their tenures or achievements? Who was the head coach of the team Legia Warsaw right before July 1969?
Before	Absolute	**Question:** Who was the head coach of the team Legia Warsaw right before July 1969? **Answer:** The head coach of Legia Warsaw right before July 1969 was Edmund Zientara. He managed the team from 1968 to 1969 and was succeeded by Józef Walczak in July 1969. Zientara was a notable figure in Polish football, both as a player and a coach, and he later had a successful managerial career, including leading the Poland national team in the 1980s. Would you like more details about his tenure at Legia or his career in general?

Table 10: Qualitative analysis of DeepSeek-R1 on temporal placements across direction and references.

KB Relation	#Queries	Template	Anticipated General Entity Type
member of sports team	2524	Which team did <subject> play for before/after <object> ?	team
position held	2538	Which position did <subject> hold before/after <object> ?	position
employer	1991	Which employer did <subject> work for before/after <object> ?.	employer/company
political party	433	Which political party did <subject> belong to before/after <object> ?	political party
head coach	1051	Who was the head coach of <subject> before/after <object> ?	person
educated at	594	Which school was <subject> attending before/after <object> ?	school
chairperson	1881	Who was the chair of <subject> before/after <object> ?	person
head of government	1535	Who was the head of the government of <subject> before/after <object> ?	person
head of state	268	Who was the head of the state of <subject> before/after <object> ?	person
owned by	199	Who was the owner of <subject> before/after <object> ?	person

Table 11: The structure of the event-event (L3) category from TEMPREASON is outlined in the following table. The first three columns are sourced from TEMPREASON.

spectively, for the LoRA adapter and fine-tuned the `UnTRaP` for 20 epochs across all supervised fine-tuning methods. We convert the TEMP-ReCon's training data into the instruction format proposed by (Taori et al., 2023) (an instance is presented in Figure 6). Furthermore, we applied 8-bit quantization to optimize computational efficiency and accelerate training. An alpha value of 0.66, as recommended by the authors of CoTSeLF, was utilized to run the baseline. Benchmarking temporal referential consistency across models is performed using the default parameter settings of the base models for temperature, top_p, and top_k, along with num_beams set to 2 and max_new_tokens set to 30. During fine-tuning, we set the max token cut_off limit to 256.

We selected 2000 random parallel instances ($\tilde{1}5\%$) from the training data to construct the paired sentences dataset for the consistency task for CoTSeLF baseline. A positive pair across temporal references and an antagonist pair each occur within the temporal framework.

D Prompts

An example of a few-shot prompt utilized in in-context settings is illustrated in Figure 5, encompassing various languages. Furthermore, a representative instance of the training data employed during the fine-tuning process is depicted in Figure 6.

E Commentaries

Domain Adaptation. Temporal referential consistency can be adapted across various domains, particularly in time-sensitive fields such as finance and law. For instance, queries such as *"How did the US economy perform during 2008-2016?"* and *"How did the US economy perform under Barack Obama?"* are semantically identical yet employ differing temporal references: a specific time period (2008-2016) versus a notable individual (Barack Obama). It is imperative that responses generated by models exhibit consistency in addressing such queries across diverse temporal references.

In the context of finance, the accurate and consistent retrieval of economic indicators and financial data is fundamental for informed decision-making. Inconsistencies in responses provided by language models (LLMs) to diverse temporal reference inquiries regarding the same subject can result in suboptimal investment strategies, misinformed clientele, or inaccuracies in regulatory filings. Similarly, in the legal domain, variations in responses to queries that reference the same legal principle or case but through different temporal references may evoke confusion and increase the risk of legal malpractice. Consequently, establishing temporal referential consistency in time-sensitive domains is critical for enhancing accuracy, reliability, and risk mitigation. It is anticipated that future research will extend or adapt the findings of the current study across various domains.

Applicability For Closed-Source Models. The commercial models, such as GPT-4 and Claude-3, exhibit suboptimal performance in terms of temporal referential consistency. This limitation necessitates the development of strategies, including the proposed UnTRaP or analogous enhancements, aimed at improving this aspect of temporal consistency. The UnTRaP introduces an alignment-based methodology that seeks to unify reasoning paths across various temporal references through the refinement of model parameters. Due to the restrictions placed on direct architectural access in models like GPT-4 and Claude-3, it is imperative that such methodologies demonstrate their effectiveness solely on open-source models.

Considering that the predominant architecture for contemporary commercial large language models is based on decoder-based transformers, the UnTRaP approach can be effectively applied during either the fine-tuning stage or through reinforcement learning from human feedback, serving as an extended objective alongside existing methodologies. Furthermore, closed-source models can benefit from a similar methodological framework implemented via non-gradient update approaches that do not necessitate access to the model's parameters. However, we have chosen to adopt a fine-tuning-based methodology, as context-based approaches have been shown to yield suboptimal performance compared to fine-tuning methodologies.

```
Consider the following examples-

Question: Which employer did Olivier Pironneau work for right before Inria?
Answer: The correct answer is University of Cambridge

Question: Which employer did Jacob Tamarkin work for right after Perm State
University?
Answer: The correct answer is Saint Petersburg State Polytechnical University

Question: Which employer did Alan Perlis work for right before Yale University?
Answer: The correct answer is Carnegie Mellon University

Now, please respond accurately to the below question.
Question: Which employer did Jaroslav Pelikan work for right before Concordia
Seminary ?
Answer: The correct answer is
```

(a) English

```
Considérez les exemples suivants-
Question: Quel poste a occupé Sir Gabriel Goldney, 1er baronet, après avoir été
député du 19e Parlement du Royaume-Uni?
Réponse: La bonne réponse est Membre du 20e Parlement du Royaume-Uni

Question: Quelle position a occupé John Ramsbottom devant le député du 12e Parlement
du Royaume-Uni?
Réponse: La bonne réponse est Membre du 11e Parlement du Royaume-Uni

Question: Qui était président de l'Université d'État de Pennsylvanie avant Milton
Stover Eisenhower?
Réponse: La bonne réponse est James Milholland

Maintenant, veuillez répondre avec précision à la question ci-dessous.
Question: Pour quel employeur Jaroslav Pelikan a-t-il travaillé avant le Concordia
Seminary?
Réponse: La bonne réponse est
```

(b) French

```
Luați în considerare următoarele exemple-
Întrebare: Ce poziţie a avut Sir Gabriel Goldney, primul baronet, după ce a fost
deputat în cel de-al 19-lea Parlament al Regatului Unit?
Răspuns: răspunsul corect este Membru al celui de-al 20-lea Parlament al Regatului
Unit

Întrebare: Ce poziţie a avut John Ramsbottom înainte de a fi deputat în cel de-al
12-lea Parlament al Regatului Unit?
Răspuns: răspunsul corect este Membru al celui de-al 11-lea Parlament al Regatului
Unit

Întrebare: Cine a fost preşedintele Universităţii de Stat a Pennsylvania chiar
înainte de Milton Stover Eisenhower?
Răspuns: răspunsul corect este James Milholland

Acum, vă rugăm să răspundeți cu acuratețe la întrebarea de mai jos.
Întrebare: Pentru ce angajator a lucrat Jaroslav Pelikan înainte de Seminarul
Concordia?
Răspuns: Răspunsul corect este
```

(c) Romanian

Figure 5: A three-shot prompt that is used to assess temporal referential consistency of LLMs in experiments involving in-context learning (ICL) methods.

```
Below is an instruction that describes a task, paired with an input that provides
further context. Write a response that appropriately completes the request.

### Instruction:
Answer the following question accurately.

### Input:
Which employer did Jaroslav Pelikan work for right before Concordia Seminary?

### Response:
The correct answer is Valparaiso University because Jaroslav Pelikan worked for
Concordia Seminary from January 1949 to January 1953, and immediately prior to
January 1949, he was employed by Valparaiso University.
```

Figure 6: An example from the training data for chronological temporal reference-based queries, along with the time-oriented reasoning path used in UnTRaP's instruction tuning.