Temporally Consistent Factuality Probing for Large Language Models

Ashutosh Bajpai[1,2], **Aaryan Goyal**[1,*], **Atif Anwer**[1,*], **Tanmoy Chakraborty**[1]

[1]Indian Institute of Technology Delhi, India
[2] Wipro Research, India
{eez228482, aaryan.goyal.ee120, atif.anwer.ee120,tanchak}@ee.iitd.ac.in

Abstract

The prolific use of Large Language Models (LLMs) as an alternate knowledge base requires them to be factually consistent, necessitating both correctness and consistency traits for paraphrased queries. Recently, significant attempts have been made to benchmark datasets and metrics to evaluate LLMs for these traits. However, structural simplicity (subject-relation-object) and contemporary association in their query formulation limit the broader definition of factuality and consistency. In this study, we introduce `TeCFaP`, a novel **Te**mporally **C**onsistent **Fa**ctuality **P**robe task to expand the consistent factuality probe in the temporal dimension. To this end, we propose `TEMP-COFAC`, a high-quality dataset of prefix-style English query paraphrases. Subsequently, we extend the definitions of existing metrics to represent consistent factuality across temporal dimension. We experiment with a diverse set of LLMs and find most of them performing poorly on `TeCFaP`. Next, we propose a novel solution `CoTSeLF` (**Co**nsistent-**T**ime-**Se**nsitive **L**earning **F**ramework) combining multi-task instruction tuning (MT-IT) with consistent-time-sensitive reinforcement learning (`CTSRL`) to improve temporally consistent factuality in LLMs. Our experiments demonstrate the efficacy of `CoTSeLF` over several baselines.

1 Introduction

Large Language Models (LLMs) are pivotal in propelling the advancement of Artificial General Intelligence (AGI) by acquiring self-learning capabilities for complex tasks (Ge et al., 2023). A key development within LLMs is the ability for *temporal reasoning* - comprehending, processing, and reasoning about time-related concepts, temporal dependencies, chronological sequences, and the nuanced, consistent temporal relationship of events. This ability is vital for a myriad of domain-specific tasks, including but not limited to summarizing timelines, tracking disease progression (medical), scheduling events (planning), managing contracts (legal), historical analysis (archaeology), and identifying tasks dependencies (project management).

Why is consistent factuality important? Knowledge bases (KBs) were the foremost choice in factual knowledge retrieval tasks before the appearance of LLMs. A KB is a structured database containing a collection of facts (subject, relation, object) (Lan et al., 2021). There has been a surge of interest in using pre-trained LLMs as KBs. AlKhamissi et al. (2022) presented an extensive review on significant developments (Petroni et al., 2019) (Dhingra et al., 2022) (Heinzerling and Inui, 2021) in this direction. One of the biggest appeals of using LLMs as KBs is that a query can be written in natural language instead of relying on a specific KB schema (Elazar et al., 2021). Due to the complex nature of language, the semantic meaning can be expressed in multiple surface forms. Accurate and consistent retrievals, despite a change in surface form, are two fundamental traits required from LLMs to replace KBs. Built on a manually engineered schema that dictates the possible set of entities and relations, KBs ensure factual and consistent answers (AlKhamissi et al., 2022). On the contrary, inconsistent factuality is reported as a widespread problem in LLMs, especially in autoregressive setting (Tam et al., 2022).

Probing consistent factuality. Usually, factuality (accuracy) is used as a widespread metric in probing LLMs to check linguistic capabilities such as commonsense (Zhang et al., 2020; Forbes et al., 2019) and reasoning (Talmor et al., 2020; Kassner et al., 2020). On the other hand, an increase in public access of LLMs requires them to consistently respond to user-specific diversities in surface forms of a query. Ravichander et al. (2020) measured consistency through paired probes. Elazar et al. (2021) extended the work by investigating and improving

[*]Equal contribution

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 15864–15881
November 12-16, 2024

the consistency of LLMs behavior across different factual knowledge types.

Our novel task and dataset. So far, the consistent factuality probing in LLMs has been predominantly contemporary – the existence of a subject and an object associated via a relation is coeval, i.e., a typical query from PARAREL dataset (Elazar et al., 2021) is – *X was born in Y* in which subject *X* (*person*) and object *Y* (*location*) are connected via relation *born-in*. Note that the subject and object are contemporaneous; both exist simultaneously. As information keeps on getting generated, maintained and lost over time, the above formulation is insufficient in capturing the temporal association between the queried entity (subject) and expected value entity (object), therefore providing a poor representation of the overall consistent factuality of LLMs.

To address this, we present **Te**mporally **C**onsistent **Fa**ctuality **P**robe (TeCFaP), a novel task accompanied by a new dataset, TEMP-COFAC. TeCFaP seeks to exploit the temporal association among entities via a *subject-relation* pair to represent temporally consistent factuality of LLMs. It defines the query structure in the form of (*key_object*, *subject-relation*, *value_object*). The proposed formulation expands the probe in the temporal dimension. As we observe in Figure 1, the space is three-dimensional – *subject*, *relation*, and *time*. Since time has a directional attribute, we expect temporal association to be either in a forward or backward direction. For example, the query *Hybrid Theory was released by linkin park just before [Meteora]* is a *forward* direction probe where a *key_object* (Hybrid Theory) placed at time t is temporally associated with a *value_object* (Meteora) in time-space at $(t + 1)$ via a subject (*Linkin Park*) and a relation (*release-by*). At present, our probe is limited to only strict associations where the expected *value_object* is either located at $(t+1)$ or $(t - 1)$ step in temporal space w.r.t *key_object* at time t. Inspired by Pustejovsky et al. (2005), strict association is achieved via trigger words such as *immediately before*, *right after*, *soon after*, etc. We observe the destitute performance of LLMs on TeCFaP metrics (defined in Section 3) – temporal factuality, temporal consistency, and temporally consistent factuality are in [0.95% - 3.63%], [13.28% - 64.87%] and [0% to 2%] range, respectively in zero-shot setting.

Our proposed model. We present CoTSeLF, a

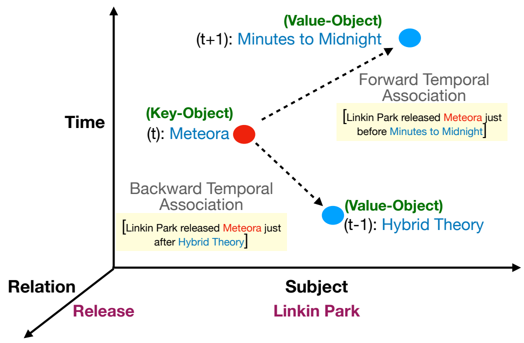


Figure 1: Symbolic representation of the TeCFaP objective. An entity *key_object* holds a temporal relationship with another entity *value_object* via a *subject-relation* pair in either direction – forward or backward.

Consistent-**T**ime-**Se**nsitive **L**earning **F**ramework, built on a multi-task instruction-tuned (MT-IT) framework followed by consistent-time-sensitive reinforcement learning (CTSRL) to improve temporally consistent factuality in LLMs. We compare CoTSeLF with several baselines and show it outperforming the best baseline (Tan et al., 2023) by 12.7%, 10.9% and 90.4%, respectively, for temporal factuality, temporal consistency, and temporally consistent factuality.

Contributions. In short, we make the following contributions through this study[*]:

- We establish the need for temporally consistent factuality in LLMs and propose TeCFaP, a novel task (Section 3).
- We create TEMP-COFAC, a novel dataset consisting of 66 diverse *subject-relation* pairs and 8 paraphrase samples each for forward and backward temporal association for a given *subject-relation* pair (Section 2).
- We experiment with a diverse set of LLMs. Our experiments and analyses highlight how LLMs poorly perform on TeCFaP (Section 5).
- We propose CoTSeLF, a framework to improve temporally consistent factuality in LLMs (Section 4). Our experiments highlight that CoTSeLF surpasses the recent baseline models (Section 5). We further analyze how the probabilistic space evolves under CoTSeLF (Appendix A.2.4).

2 The TEMP-COFAC Dataset

Here, we present a novel English prefix-style TEMP-COFAC dataset with a temporal range of 1526-2022. Inspired by Elsahar et al. (2018), we semi-

[*]Source code and dataset are available at https://github.com/ab-iitd/tecfap

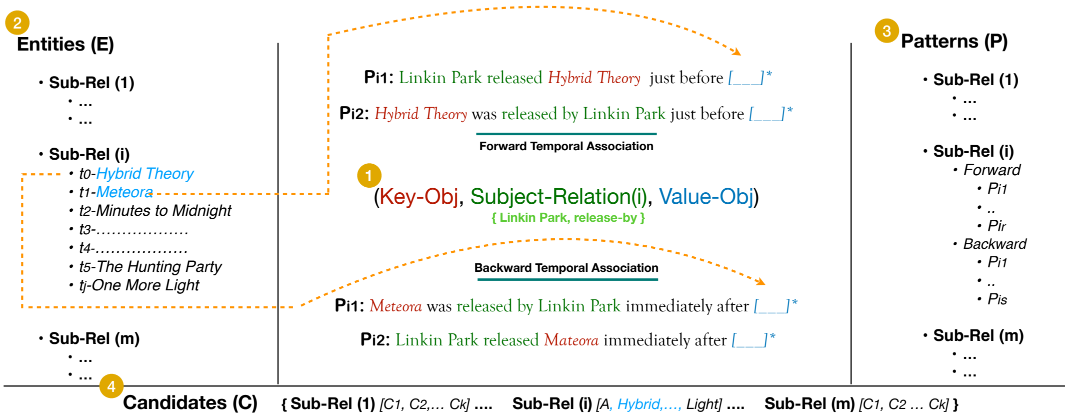


Figure 2: The architectural framework of `TEMP-COFAC` – (1) a set of diverse *subject-relation* pairs, (2) a sequence of entities which are temporally connected via a given *subject-relation* pair, (3) a set of paraphrase templates with a placeholder for *key_object* and *value_object* developed from *subject-relation* pairs, and (4) a closed vocabulary candidate set developed from possible entity space for a given *subject-relation* pair.

automatically[*] curate a diverse set of base subject and relations pairs *subject-relation*. Next, we define E_i, a strict temporally ordered set of entities associated with i^{th} *subject-relation* pair. An entity can act as a *key_object* or a *value_object* relative to the role of another entity positioned right next/before it. We then create *base_patterns* for both forward and backward directions following Petroni et al. (2019). Afterwards, a set of patterns P_i is constructed by employing paraphrasing techniques (Bhagat and Hovy, 2013) on *base_patterns* in forward and backward directions followed by a candidate set C_i.

Construction approach. The `TEMP-COFAC` resource is constructed by three NLP experts[*] with a mean cross-annotator agreement of 4.84 ± 0.39 and 4.86 ± 0.49 out of 5 maximum on Likert scale (Likert, 1932) for factuality and consistency, respectively (refer Appendix A.1 for more details). Following Elazar et al. (2021), our construction process broadly follows a four-step procedure described in Figure 2.

(i) First, we define a set of m diverse *subject-relation* pairs randomly collected from varied domains – entertainment, technology, politics, automobiles and corporate. Using annotators' linguistic expertise, we then define *base_patterns* for forward and backward directions, i.e., *Linkin Park released [X] just before [Y]* and, *Linkin Park released [X] just after [Y]* are two *base_patterns* examples of the forward and backward temporal associations, respectively, where [X] and [Y] are the placeholders for a *key_object* and a *value_object*, respectively.

(ii) Next, we manually curate a set of entities E_i ($\forall i = 0, \ldots, m$) for the i^{th} *subject-relation* pair such that they have temporal association with it in ascending temporal sequence $t = 0, \ldots, j$, where, entity e_i^t precedes e_i^{t-1} and is followed by e_i^{t+1} in the temporal space. The cardinality of E_i, represented by j, varies for each *subject-relation* pair.

(iii) A set of paraphrased patterns P_i ($\forall i = 0, 1, \ldots, m$) is constructed through an application of an online paraphraser tool, Quillbot[*] on *base_patterns*. Set P_i defines r and s number of paraphrases in the forward and backward directions, respectively. For simplicity, here we consider uniform values of r and s to be 8.

(iv) Finally, a constrained candidate set C_i is developed as an unordered set of words from E_i.

Table 1 summarizes the statistics of `TEMP-COFAC`. Readers can refer to Appendix A.1 for more detail about `TEMP-COFAC`, including annotation quality, temporal and entity type distributions.

3 TeCFaP Task Structure

We pose `TeCFaP` as a sentence completion task that aligns with the general objective of an LM. With prefix style, paraphrase templates for a *subject_relation* pair are filled with only *key_object*, and we expect the model to generate *value_object*. Next, we define `TeCFaP` metrics to evaluate LLMs.

[*]We leverage `GPT-4` to extend manually constructed initial set of base subject-relation pairs.

[*]All of them are male with their ages ranging between 25-35 years.

[*]`https://quillbot.com/`

# Subject-Relation pairs	66
# Paraphrase patterns	1056
# Forward patterns	528
# Backward patterns	528
Avg # pattern per relation-subject	18
# Entities	700
#Unique entity types	11
Min # entities per relation-subject	2
Max # Entities per relation-subject	16
Avg # entities per relation-subject	10.6
# Unique samples in dataset	10144

Table 1: High-level statistics of the TEMP-COFAC dataset.

Temporal factuality and temporal consistency. We extend the metrics defined by Elazar et al. (2021) to TeCFaP. The first metric, *temporal-factuality*, captures the accuracy of a model across the temporal direction. In contrast with their definition of exact-match accuracy, we define factuality as soft accuracy, a ratio of the number of continuous matches of words in actual and generated *value_object* for a sample. It helps capture the partially correct generation as well. Next, we define its sub-classification across temporal directions. The forward *temporal-factuality* measures the accuracy in the forward temporal direction where *value_object* is located at $t+1$ time step for a given *key_object* at t. In backward *temporal-factuality*, the *value_object* is located at $t-1$ time step for a given *key_object* at t.

The second metric is *temporal-consistency*. Given a pair of prefix-style paraphrases for a *subject_relation* filled with an identical *key_object*, an identical *value_object* should be generated. The metric estimate is binary (one or zero) for a given pair of such paraphrases if the model's responses are identical or different. Forward and backward directions paraphrases contribute to forward and backward *temporal-consistency*, respectively.

Temporally consistent factuality. The third composite metric, *temporally-consistent-factuality*, is a stricter version of *temporal-factuality*, requiring a model to be consistent and factual across the temporal direction. It reports the *temporal-factuality* only if the responses are identical from all prefix-style paraphrases for a given *subject_relation* and *key_object* in a particular temporal direction. Paraphrases in forward and backward directions contribute to forward and backward *temporally-consistent-factuality*, respectively.

Other metrics. The quality of patterns is measured using *temporal-succ_patt*, indicating the percentage of patterns yielding a correct *value_object* at least once during the probe. Furthermore, *temporal-succ_objs* is introduced to measure the model's temporal knowledge by reporting the percentage of *value_objects* accurately generated at least once. Further, we define *temporal-know_cons* and *temporal-unk_cons* as metrics to measure *temporal-consistency* of the fraction of patterns which generated correct *value_object* at least once and the fraction of patterns which never responded with a correct *value_object* for a *subject_relation*, respectively. All four metrics are then classified into forward and backward temporal directions.

4 Consistent-Time-Sensitive Learning Framework (CoTSeLF)

Equipped with the advancements in model fine-tuning combined with the recent baseline in temporal reasoning, We begin with a base pre-trained LLM and apply supervised multi-task instruction-tuning to develop a multi-task instruction-tuned (MT-IT) model. Subsequently, we apply time and consistency-sensitive reinforcement learning to the MT-IT model to enhance its temporally consistent factuality capabilities.

Motivation. Instruction tuning (IT) (Zhang et al., 2023) is a cost-effective, efficient technique for developing specialized models, as evidenced by the parameter-efficient fine-tuning (PEFT) (Mangrulkar et al., 2022) approach such as low-rank adaptation (LoRA) (Hu et al., 2021), which offers significant benefits for instruction tuning in LLMs within low-cost infrastructures. Additionally, the combination of RL and supervised fine-tuning enhances model performance in several domain-specific tasks, notably temporal reasoning (Tan et al., 2023). Furthermore, advances in multi-task learning led to several breakthroughs in modeling multiple objectives simultaneously (Zhang and Yang, 2022).

Multi-Task Instruction-Tuning (MT-IT). We consider a multi-objective optimization problem that simultaneously improves model's factuality and consistency, denoted by tasks k1 and k2, respectively. In k1, we apply a standard sentence completion task where an incomplete sentence filled with *key_object*, augmented with a context and task instruction, is passed as input to allow the model complete the sentence with expected *value_object*. Whereas k2 is a binary task predicting true or false if two sentences are paraphrased. We start with transforming the TEMP-COFAC dataset into an instruction-based dataset in line with the

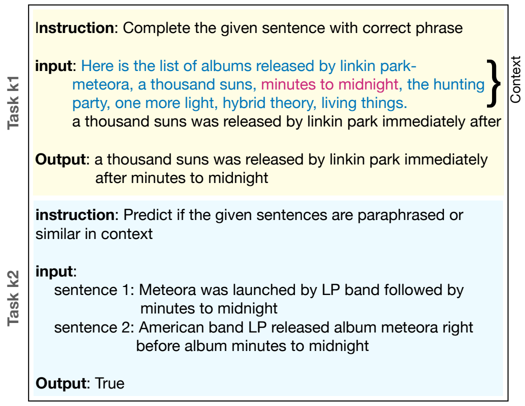


Figure 3: An instruction-based sample from training data for MT-IT model. Task k1: Generative sentence completion; Task k2: Binary paraphrase prediction.

formats proposed by Wang et al. (2023); Taori et al. (2023) (Figure 3). Next, we consider a base pre-trained LLM with parameter θ and then apply LoRA instruction-based supervised fine-tuning where the objective is to maximize $p(o^{k1}|s^{k1}, i^{k1}, c)$ for improving model's factual capability, where s^{k1}, i^{k1}, c and, o^{k1} are instruction, input, context and output, respectively for task k1. We further add another objective of maximizing $p(o^{k2}|s^{k2}, i^{k2})$ to improve the model's consistency in extending the framework to multi-task learning setup (Figure 3), where i^{k2} is a pair of sentences randomly sampled positively and negatively to predict the output o^{k2} as true or false, respectively.

Consistent Time-Sensitive Reinforcement Learning (CTSRL). Multiple *value_object*s for a given *key_object* are possible by excluding the temporal direction. We introduce CTSRL, aimed at modeling consistent sensitivity towards time and overcoming the limitations of the binary temporal characterization inherent in TSRL (Tan et al., 2023). CTSRL is employed to further fine-tune θ through the joint modeling of both k1 and k2. The reward mechanism is devised as a linear amalgamation of temporal and consistence sensitivities aspects. Initially, $\texttt{CTSRL}_{Discrete}$ is conceptualized in alignment with TSRL's binary reward framework, awarding a positive discrete reward of one for accurate predictions and zero for incorrect responses. Additionally, α is defined to act as the weighting parameter for the consistence sensitivity reward component.

$$R_d(x) = (1-\alpha)R_d^t(x) + \alpha R_d^c(x) \tag{1}$$

$$R_d^\lambda(x) = \begin{cases} P_d^\lambda(x), & \textit{if } O_g(\theta(x)) = O_l(x), \\ N_d^\lambda(x), & \textit{otherwise}. \end{cases} \tag{2}$$

For a given input x, the overall reward function $R_d(x)$ in the $\texttt{CTSRL}_{Discrete}$ setting is presented in Equation 1, where $R_d^t(x)$ and $R_d^c(x)$ are reward contribution for tasks k1 and k2, respectively. In Equation 2, λ is used as a task indicator where $P_d^\lambda(x)$ is a positive reward score for consistence sensitivity component if λ equals to c. Similarly, $N_d^\lambda(x)$ is a negative reward score. $O_g(\theta(x))$ and $O_l(x)$ are the generated *value_object* and ground-truth label *value_object*, respectively for given input x. In the case of $\texttt{CTSRL}_{Discrete}$, a positive reward score equal to one is assigned in case of correctly generated output and zero otherwise for both tasks k1 and k2.

We further define another variant, $\texttt{CTSRL}_{Smooth}$, to model continuous and relative properties of time. The temporal sensitivity reward is a continuous function where a positive reward has a maximum function value equal to one. However, the negative reward is proportional to the relative distance of the predicted answer from the correct answer in the temporal axis. The goal is to penalize incorrect answers that are distant from the correct answer more severely than the incorrect predictions that are nearby on the temporal axis. There is no change in the reward component for consistence sensitivity except for releasing the constraint of the parameter $alpha$, indicating that $\texttt{CTSRL}_{Smooth}$ is an unweighted linear combination of continuous temporal and discrete consistence sensitivity reward components.

$$R_s(x) = R_s^t(x) + R_s^c(x) \tag{3}$$

$$N_s^t(x) = \begin{cases} \dfrac{|t_{O_l} - t_{O_g}|}{t_n - t_{O_l}}, & \textit{if } t_{O_g} > t_{O_l}, \\ \dfrac{|t_{O_l} - t_{O_g}|}{t_{O_l}}, & \text{otherwise} \end{cases} \tag{4}$$

Equation 3 presents reward function $R_s(x)$ in $\texttt{CTSRL}_{Smooth}$ for a given input x, where $R_s^t(x)$ and $R_s^c(x)$ are the reward contribution for tasks k1 and k2, respectively. For the temporal sensitivity task (k1), the positive reward score is assigned as a value equal to one in case of correctly generated output, and the negative reward score is the relative distance of the wrong prediction from the correct prediction in the temporal axis as presented in Equation 4. The symbol t_{O_l} is the time step of the ground-truth label, t_{O_g} represents the time step for generated output, and t_n is the end time step of entity sequence for that *subject-relation* pair.

Models	Temp-fact			Temp-cons			Temp-cons-fact		
	Avg	**Bwd**	**Fwd**	**Avg**	**Bwd**	**Fwd**	**Avg**	**Bwd**	**Fwd**
GPT-J [6B]	3.63	6.88	0.37	64.87	66.29	63.46	1.48	2.73	0.22
Falcon [7B]	2.99	5.74	0.24	41.52	40.42	42.63	1.08	2.03	0.13
LLaMA [7B]	1.48	0.89	2.07	13.28	12.52	14.03	0	0	0
LLaMA [13B]	1.11	1.01	1.21	15.65	16.33	14.96	0.2	0.17	0.24
LLaMA2 [7B]	0.95	0.73	1.17	22.34	21.95	22.73	0.08	0	0.17
LLaMA2 [13B]	1.13	0.97	1.27	13.34	14.41	12.26	0.09	0	0.19

Table 2: Zero-shot performance in open vocabulary setting on `TeCFaP` across various LLMs. *Temp-fact*: temporal factuality (in %), *Temp-cons*: temporal consistency (in %), *Temp-cons-fact*: temporally consistent factuality (in %). *Fwd* and *Bwd* are the forward and backward direction probe, respectively (*Avg*, an average of both directions).

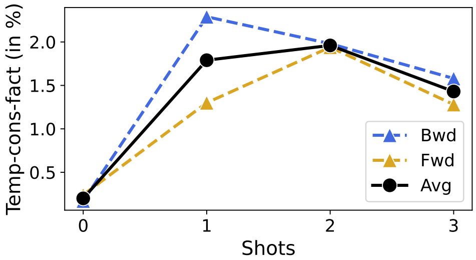


Figure 4: Results for temporally consistent factuality (*Temp-cons-fact*) in k-shot (k=1,2,3) ICL setup with `LLaMA`[13B] in an open vocabulary setting.

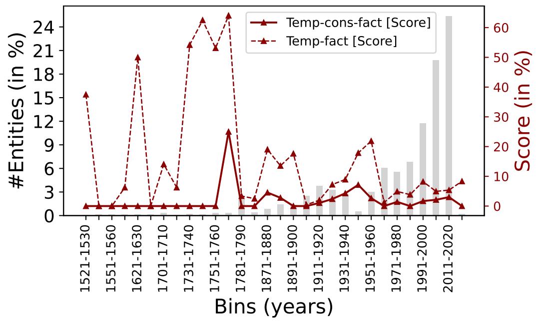


Figure 5: Average temporally consistent factuality and temporal factuality (second y-axis) in an open vocabulary and two-shot setting across temporal bins of Entities (bin size: 10 years) with `LLaMA`[13B].

5 Experimental Results

Experimental setup. The families of `GPT-J`*, `Falcon` (Almazrouei et al., 2023), and `LLaMA` (Touvron et al., 2023) are considered for evaluation on `TeCFaP`. Primarily, we evaluate LLMs in an open vocabulary setting where the next token is sampled from the entire vocabulary. Next, we conduct experiments in an in-context learning setup, followed by a closed vocabulary setting. Finally, the `CoTSeLF` efficacy evaluation is conducted.

Temporally consistent factuality on `TeCFaP`. We start with a comparison of various LLMs in zero-shot setting. The task is to correctly complete a sentence with the expected *value_object* given an instruction followed by an input, i.e., *"complete the given sentence with the correct phrase: Meteora was released by Linkin Park immediately after"*. We observe the destitute performance of all the experimented LLMs over both *temporal-factuality* and *temporally-consitent-factuality* in the range of [0.95% – 3.63%] and [0% – 1.48%], respectively (Table 2). We notice that `GPT-J` and `Falcon` tend to be highly consistent in the range of [41.52% – 64.87%] while being miserably factually incorrect compared to the `LLaMA` model in the same range of parameters size. Additionally, various families of LLMs behave differently regarding their sensitivity towards temporal direction, but it doesn't significantly correlate with temporal consistency. Further, a preliminary evaluation of `TeCFaP` in zero-shot setting for commercial LLMs such as `GPT-4` (OpenAI et al., 2023) and `Claude-3` is (Anthropic, 2024) presented in Appendix A.4.

In-context setup. In-context Learning (ICL) helps off-the-shelf LLMs solve unseen tasks without the requirement of fine-tuning (Dong et al., 2023). We provide k randomly-drawn examples from the same *subject_relation* pair as supplementary context in a k-shot ICL setting, i.e., a one-shot example is as follows: *"complete the given sentence with the correct phrase: Meteora was released by Linkin Park immediately after => Hybrid Theory. American band LP released Minutes to Midnight immediately after =>"*. This test evaluates `LLaMA[13B]` and varies k in the range [1-3]. In a two-shot setup, we observe absolute percentage points improvement of 1.76 in *temporally-consistent-factuality* (Figure 4). At the same time, the improvements of 7.01 and 18.22 percentage points are noted in contributory metrics *temporal-factuality* and *temporal-consistency*, respectively (refer to Appendix A.2.1 for more details).

Figure 5 presents *temporal-factuality* across the temporal distribution of entities. Findings reveal that *temporal-factuality* for entities belonging to the historical period (1500-1800) is significantly

*https://huggingface.co/EleutherAI/gpt-j-6b

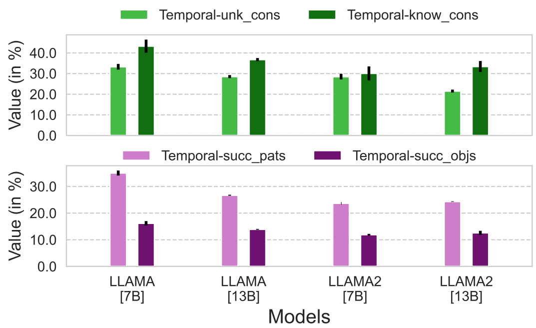


Figure 6: Qualitative results in an open vocabulary with two-shot ICL setup across LLaMA variants. Error bars represent divergence across temporal directions – forward and backward.

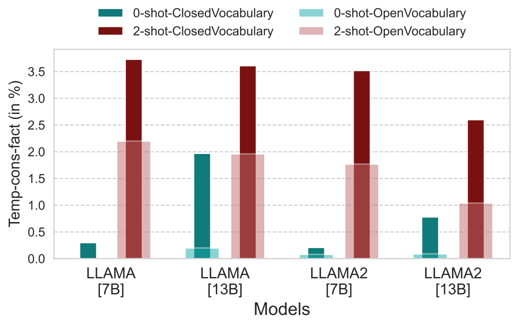


Figure 7: A comparison of average *temporally-conistent-factuality* between open and closed vocabulary settings across LLaMA variants in zero-shot and 2-shot.

higher, with an average of 25.27% compared to 9.08% for the entities in the contemporary period (after 1800). At the same time, the presence of multiple sources of the same information in LLMs pre-trained dataset for the contemporary period leads to better *temporally-consistent-factuality*.

Additionally, in Figure 6, qualitative analysis reveals that LLaMA[7B] attains best *temporal-know_cons* at 43.29% with a divergence of 10.00 percentage points between known and unknown temporal consistencies across LLaMA variants. On the other hand, 35.04% patterns yield a correct *value_object* at least once in contrast to only 16.18% *value_objects*, which were predicted correctly once in the entire probe for a model.

Closed vocabulary setup. The next word generated by an LM can still be the right placement given its general objective to maximize the semantic expectation irrespective of the expected *value_object*. Therefore, we also conduct experiments in a closed vocabulary setting where the sample space is reduced to a candidate set[*] (defined in Section 2) during generation. This approximation helps set up the probe as KB fact extraction from a given possible facts space, thus maximizing the behavioral expectation of LM as KBs. With an improvement in the range [1.77% - 2.08%], LLaMA[13B] has best scores of 1.97% and 3.51% for *temporally-consistent-factuality* in one and two shots setting, respectively, across variants of LLaMA under closed vocabulary setting (Figure 7). We observe a similar trend for *temporal-factuality* and *temporal-consistency*, presented in Appendix A.2.2.

Improvements with `CoTSeLF`. We conduct all experiments in an open vocabulary setting (assuming no access to candidate sets during inferences) with LLaMA[13B]. First, TEMP-COFAC is vertically split (test ratio: 0.3) to produce a train set and a test set containing 46 and 26 subject-relation pairs, respectively. The random vertical split ensures the stricter evaluation of CoTSeLF as the test set contains only unseen *subject-relation* pairs. We broadly categorize this evaluation into four categories: (i) the default performance of the model in zero-shot and two-shot (ICL) setup, (ii) variants of instruction-tuned (IT) models based on the presence of a context along with the novel multi-task IT model, (iii) IT (with context) followed by TSRL; a strong baseline model, and (iv) CoTSeLF, a combined strategy of MT-IT followed by variants of novel CTSRL method. Empirically, we set α as 0.66 in formulating a discrete variant of CTSRL (significance of α is presented in Appendix A.2.5).

In Table 3, the additional context provided with an input improves *temporal-factuality* by 5.22 percentage points for an IT model. The MT-IT model improves *temporal-factuality*, *temporal-consistency* and *temporally-consitent-factuality* by 7.7%, 6.4% and 90.2%, respectively, compared to the IT model. We further observe improvements of 12.7%, 10.9% and 90.4% in *temporal-factuality*, *temporal-consistency* and *temporally-consitent-factuality*, respectively, with CoTSeLF over the baseline, indicating that improving a model's temporal consistency also positively impacts its temporal factuality. However, the inaccessibility of GPT-4 architecture limits us to assess the efficacy of CoTSeLF's on this model. (the probabilistic space evolution under CoTSeLF including the ablations for scalability are presented in Appendix A.2.4, & A.3).

Significance of `CTSRL` over TSRL. We perform this ablation in two different scenarios by immo-

[*]A set of restricted tokens are generated by employing byte pair encoding (Sennrich et al., 2016) on candidate set.

Models	Setting (open vocab)	Temp-fact			Temp-cons			Temp-cons-fact		
		Avg	Bwd	Fwd	Avg	Bwd	Fwd	Avg	Bwd	Fwd
Default	Zero-shot	$0.40_{\pm 0.00}$	0.28	0.52	$16.31_{\pm 0.00}$	15.22	17.40	$0.00_{\pm 0.00}$	0.00	0.00
	Two-shot [ICL]	$4.95_{\pm 0.40}$	4.26	5.63	$31.57_{\pm 1.50}$	29.23	33.91	$0.53_{\pm 0.18}$	0.50	0.57
IT	Without context	$11.53_{\pm 0.08}$	10.36	12.69	$25.08_{\pm 0.06}$	26.76	23.41	$2.59_{\pm 0.30}$	2.67	2.50
	+ Context	$16.87_{\pm 0.17}$	16.79	16.96	$34.39_{\pm 0.92}$	34.45	34.31	$1.75_{\pm 0.15}$	2.11	1.38
MT-IT	+ Context	$18.17_{\pm 0.15}$	17.83	18.52	$36.60_{\pm 0.36}$	36.49	36.70	$3.33_{\pm 0.41}$	2.82	3.85
Δa		1.3 ↑			2.21 ↑			1.58 ↑		
Baseline	IT (+context) + TSRL	$16.60_{\pm 0.51}$	17.04	16.15	$33.23_{\pm 0.38}$	33.28	33.18	$2.28_{\pm 0.44}$	2.88	1.67
CoTSeLF	MT-IT+CTSRL$_{Discrete}$	$\mathbf{18.72}^{P_1}_{\pm 0.25}$	19.16	18.29	$\mathbf{36.88}^{P_2}_{\pm 0.49}$	37.58	36.17	$\mathbf{4.34}^{P_3}_{\pm 0.27}$	4.45	4.22
	MT-IT+CTSRL$_{Smooth}$	$18.16_{\pm 0.05}$	18.11	18.21	$36.07_{\pm 0.48}$	37.09	35.04	$3.89_{\pm 0.92}$	3.74	4.05
Δb		2.12 ↑			3.65 ↑			2.06 ↑		

$P_1 = 0.003, P_2 = 0.003, P_3 = 0.003$

Table 3: Experimental results of CoTSeLF across – *temporal-factuality*, *temporal-consistency*, and *temporally-consistent-factuality* (in %), in comparison to multiple baselines on test data with LLaMA[13B] (average scores over three runs). Δa: improvements of MT-IT over an IT model, Δb: improvements of CoTSeLF over a baseline model. (P_1, P_2, P_3): p-values at CoTSeLF's best scores compared to baseline model with $n = 3$ and one-tailed test.

Model	Temp-fact	Temp-cons	Temp-cons-fact
Base SFT Model: IT			
TSRL	16.60	33.23	2.28
CTSRL	**17.6**	**34.27**	**2.74**
Base SFT Model: MT-IT			
TSRL	17.55	34.5	3.68
CTSRL	**18.72**	**36.88**	**4.34**

Table 4: Comparison of CTSRL$_{Discrete}$ and TSRL by immobilizing the SFT model across two settings: IT and MT-IT model.

bilizing the SFT model. In the first scenario, an IT model serves as the foundation [comparing (IT + TSRL) with (IT + CTSRL$_{Discrete}$)], while in the second scenario, the basis is the MT-IT model [comparing (MT-IT + TSRL) with (MT-IT + CTSRL$_{Discrete}$)]. These experiments are meticulously executed under identical conditions, as meticulously outlined in Table 3.

The findings are delineated in Table 4. CTSRL$_{Discrete}$ exhibits superior performance over TSRL on all metrics: *temporal-factuality*, *temporal-consistency*, and *temporally-consistent-factuality* in both experimental settings. Results from this ablation study further underscore the distinct advantage of a preference of CTSRL over TSRL.

6 Error Analysis

A causal analysis is conducted to determine any correlation between data characteristics and failure cases. Figure 8 shows a strong correlation between entity-type and *temporally-consistent-factuality*. Entity types not exclusively attached to a *subject-relation* pair, such as movie names and geographical locations, perform poorly compared to other entities, such as satellite, person, and software

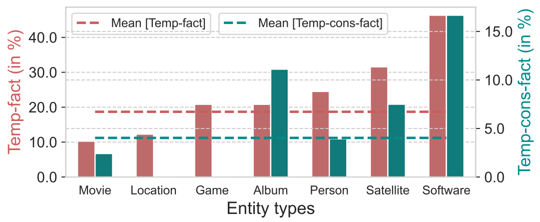


Figure 8: Temporally consistent factuality across various entity-types present in test data for CTSRL$_{Discrete}$.

names. It would be interesting to enhance both TEMP-COFAC and CTSRL formulation to include such data characteristics explicitly. For more detail on this, readers can refer to Appendix A.2.3.

7 Related Work

A noteworthy advancement in temporal factuality involves analyzing and updating LLMs to address the obsolescence of their factual knowledge over time (Hu et al., 2024; Vu et al., 2023). These approaches concentrate on amending the factuality of evolving temporal relationships without rectifying the current inaccuracies and inconsistencies. On the other hand, consistency in KBs was extensively studied, with developments around data and methods to benchmark the degree of inconsistencies (Hansen and Jaumard, 2000; Andersen and Pretolani, 2001; Thimm, 2013) across several tasks; QA (Kassner et al., 2020), reading comprehension (Antol et al., 2015; Rajpurkar et al., 2016; Ribeiro et al., 2019), summarizing (Xie et al., 2021; Roit et al., 2023) and NLI (Li et al., 2019).

Notable works in temporal information extraction include TimeBank (Pustejovsky et al., 2003) and TimeEval (Verhagen et al., 2010). The annotations defined in these datasets are primarily for time-event and events relationships such as before/after. The remarkable advancement in studying temporal reasoning over KGs led to the development of datasets such as TEQUILA (Jia et al., 2018a), TimeQuestions (Jia et al., 2021), and CronQuesions (Saxena et al., 2021), probing KG's response in ranking entities for a given temporal query. Subsequently, datasets like TEMPLAMA (Dhingra et al., 2022) and StreamingQA (Liška et al., 2022) were curated for temporal reasoning in LMs. The TSQA (Chen et al., 2021) dataset has addressed the drawback of the limited time span of queries, but it defines only time-event relationships. TEMP-REASON (Tan et al., 2023) captures multiple facets of temporal reasoning – time-time, time-event, and event-event relationships over a longer time span. Our dataset, TEMP-COFAC, addresses the lack of consistent temporal reasoning evaluation and improves upon TEMP-REASON by offering a diverse set of anchor queries through sixty-six unique sub-rel-obj triplets across eleven entity types, overcoming TEMP-REASON's limitation of single-entity-type anchors and coverage of only six entity types. (Appendix A.1).

8 Conclusion

This paper presented a new task TeCFaP with a novel resource, TEMP-COFAC, to evaluate temporally consistent factuality in large language models. The contribution continued to present a novel solution CoTSeLF based on multi-task instruction tuning (MT-IT) combined with consistent-time-sensitive reinforcement learning (CTSRL) to improve LLMs temporally consistent factuality. We observed that CoTSeLF outperforms the baselines to improve *temporally-consistent-factuality* in LLMs. The contribution and findings in this paper would help better understand and enhance LLMs' underlying capabilities around the tasks which require consistent temporal reasoning and deductions.

9 Limitations

The scope of TEMP-COFAC is to evaluate LLMs for their temporally consistent factuality capabilities. TEMP-COFAC comprehensively covers entity-entity temporal relations but falls short of stating entity-time and time-time aspects of temporal relations. Therefore, it should be applied along with other prominent datasets to estimate the overall temporal reasoning capabilities of LLMs.

Moreover, the underlying rationale for the smooth variant of CTSRL posits that a fundamental characteristic of time is its relativity. The objective is to impose greater penalties on incorrect responses that significantly deviate from the correct answer compared to those incorrect predictions that are proximal on the temporal scale. This premise suggests that such a formulation will compel the model to assimilate the relative aspect of time, thereby enhancing its efficiency in addressing queries necessitating relative temporal responses (before/after). While the smooth variant of CoTSeLF appears promising at the conceptual level, it has not surpassed the performance of its discrete counterparts. A comprehensive ablation study on smooth variants is earmarked for future investigation.

While temporal relation settings such as t+/-1 scenarios necessitate precision, as only a single correct response is viable for any given query contrary to the open temporal relation settings (one-to-many), it is our fervent hope that the insights derived from this study will inspire further research to test TeCFaP toward addressing multi-hop temporal queries (t+/-n) in forthcoming endeavors. On the other hand, the challenges related to computational assets to stack LLMs continue. Due to asset limitations, we may not utilize more extensive or commercially accessible LLMs to comprehensively evaluate on TeCFaP and approve on the off chance that the preferences of CoTSeLF are also advantageous to those models.

Furthermore, despite efforts to apply higher quality standards, TEMP-COFAC relies on human annotation and is therefore prone to annotation errors. Moreover, the base language of TEMP-COFAC is English; therefore, it falls short in measuring consistent temporal factuality for other languages, particularly low-resource ones. Extensions to multilingual setting or resource-poor languages are left to future research.

10 Ethics Statement

The TEMP-COFAC is based on the Wikipedia and open world wide web knowledge sources. Wikipedia articles are licensed under a Creative Commons Attribution-ShareAlike 4.0 International License* (CC BY-SA 4.0) and its knowledge base is in the public domain. We will release TEMP-COFAC under same licence too. The experiments are conducted with all open source LLMs. Authors do not intend to introduce biases in any form to LLMs

*https://en.wikipedia.org/wiki/Wikipedia:FAQ/Copyright while applying fine-tuning methods.

The subjects and entities selected to be part of TEMP-COFAC are prone to unintended human biases during construction. The authors do not propagate any views/opinions, products, or representations of these subjects or entities in any form. The fact that women do not find representation during TEMP-COFAC annotations should be seen as a symptom of the gender disparity in research and innovation worldwide, but this is not the authors' view. We support gender and racial equality in research and innovation with the utmost sincerity. Furthermore, no generative AI-based content creation tools or applications were used to create this artifact, except for specialized support for spell checking, grammar correction, and paraphrasing.

Acknowledgement

Tanmoy Chakraborty acknowledges the support of the IBM-IITD AI Horizons network.

References

Badr AlKhamissi, Millicent Li, Asli Celikyilmaz, Mona Diab, and Marjan Ghazvininejad. 2022. A review on language models as knowledge bases.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of open language models.

KimAllan Andersen and Daniele Pretolani. 2001. Easy cases of probabilistic satisfiability. *Annals of Mathematics and Artificial Intelligence*, 33:69–91.

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

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433.

Rahul Bhagat and Eduard Hovy. 2013. Squibs: What is a paraphrase? *Computational Linguistics*, 39(3):463–472.

Louis Castricato, Alex Havrilla, Shahbuland Matiana, Duy V. Phung, Aman Tiwari, Jonathan Tow, and Maksym Zhuravinsky. 2023. trlX: A scalable framework for RLHF.

Wenhu Chen, Xinyi Wang, and William Yang Wang. 2021. A dataset for answering time-sensitive questions.

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.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A survey on in-context learning.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models.

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. 2018. T-REx: A large scale alignment of natural language with knowledge base triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Maxwell Forbes, Ari Holtzman, and Yejin Choi. 2019. Do neural language representations learn physical commonsense?

Yingqiang Ge, Wenyue Hua, Jianchao Ji, Juntao Tan, Shuyuan Xu, and Yongfeng Zhang. 2023. Openagi: When llm meets domain experts.

Pierre Hansen and Brigitte Jaumard. 2000. *Probabilistic Satisfiability*, pages 321–367. Springer Netherlands, Dordrecht.

Benjamin Heinzerling and Kentaro Inui. 2021. Language models as knowledge bases: On entity representations, storage capacity, and paraphrased queries.

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.

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*.

Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. 2018a. 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, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. 2018b. Tequila: Temporal question answering over knowledge bases. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18. ACM.

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 and Knowledge Management*, CIKM '21. ACM.

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.

Yunshi Lan, Gaole He, Jinhao Jiang, Jing Jiang, Wayne Xin Zhao, and Ji-Rong Wen. 2021. A survey on complex knowledge base question answering: Methods, challenges and solutions.

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

R. Likert. 1932. A technique for the measurement of attitudes. *Archives of Psychology*, 22 140:55–55.

Adam Liška, Tomáš Kočiský, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, Cyprien de Masson d'Autume, Tim Scholtes, Manzil Zaheer, Susannah Young, Ellen Gilsenan-McMahon, Sophia Austin, Phil Blunsom, and Angeliki Lazaridou. 2022. Streamingqa: A benchmark for adaptation to new knowledge over time in question answering models.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and

Sebastian Riedel. 2019. Language models as knowledge bases?

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*.

James Pustejovsky, Robert Knippen, Jessica Moszkowicz, and Roser Saurí. 2005. Temporal and event information in natural language text. *Language Resources and Evaluation*, 39:123–164.

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

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.

Marco Tulio Ribeiro, Carlos Guestrin, and Sameer Singh. 2019. Are red roses red? evaluating consistency of question-answering models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6174–6184, Florence, Italy. Association for Computational Linguistics.

Paul Roit, Johan Ferret, Lior Shani, Roee Aharoni, Geoffrey Cideron, Robert Dadashi, Matthieu Geist, Sertan Girgin, Léonard Hussenot, Orgad Keller, Nikola Momchev, Sabela Ramos, Piotr Stanczyk, Nino Vieillard, Olivier Bachem, Gal Elidan, Avinatan Hassidim, Olivier Pietquin, and Idan Szpektor. 2023. Factually consistent summarization via reinforcement learning with textual entailment feedback.

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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2020. oLMpics-on what language model pre-training captures. *Transactions of the Association for Computational Linguistics*, 8:743–758.

Derek Tam, Anisha Mascarenhas, Shiyue Zhang, Sarah Kwan, Mohit Bansal, and Colin Raffel. 2022. Evaluating the factual consistency of large language models through summarization.

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.

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`.

Matthias Thimm. 2013. Inconsistency measures for probabilistic logics. *Artificial Intelligence*, 197:1–24.

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

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.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2023. Freshllms: Refreshing large language models with search engine augmentation.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions.

Yuexiang Xie, Fei Sun, Yang Deng, Yaliang Li, and Bolin Ding. 2021. Factual consistency evaluation for text summarization via counterfactual estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 100–110, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2023. Instruction tuning for large language models: A survey.

Xikun Zhang, Deepak Ramachandran, Ian Tenney, Yanai Elazar, and Dan Roth. 2020. Do language embeddings capture scales? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4889–4896, Online. Association for Computational Linguistics.

Yu Zhang and Qiang Yang. 2022. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(12):5586–5609.

A Appendix

A.1 Extended Description for TEMP-COFAC

In this section we continue to present details on TEMP-COFAC dataset. TEMP-COFAC is the first dataset of its kind in the temporal consistency domain, providing a strictly homogeneous sequence of entities for diverse subject-relation pairs. The conditioning on events that occurred without recurrence for strict temporal relationships along with the prefix formulation of key-object, sub-rel, value-object posed a severe challenge and required significant human evaluations at the level of *subject-relation* pairs during construction.

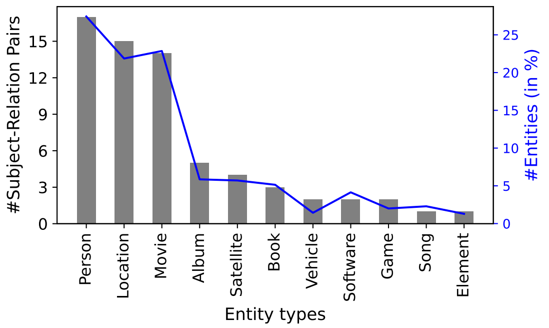


Figure 9: Entity types distribution in TEMP-COFAC resource as % of overall entities in data and representation across subject-relation (*subject-relation*) pairs.

Entity Types Distribution. We present the overall entity types distribution across different *subject-relation* pairs in Figure 9. TEMP-COFAC has broader representations for various entity types such as person names, countries, books, movie and album names, satellites, vehicles, and software.

Temporal Distribution. The temporal distribution of entities is presented in Table 5. We observe that the entities have a more comprehensive range of temporal representations in TEMP-COFAC dataset spanning in the range of 1500 to 2022. Most of the pre-training datasets that LLMs are trained on have a cutoff year of 2020, with few exceptions. Therefore, we consider that 99% of all entities in TEMP-COFAC dataset must belong to a year equal to or less than 2020. It can be observed that there is a skewed temporal distribution of entities in favor of entities in the contemporary period (the year 1800-) compared to the historical period (the year 1500-1800). It is noted that 57% of all entities are from 1991 to 2020.

Bins (year)	#Entities	Bins (year)	#Entities
1521-1530	0.08	1881-1890	1.42
1531-1540	0.16	1891-1900	1.34
1551-1560	0.16	1901-1910	2.52
1601-1610	0.16	1911-1920	3.79
1621-1630	0.16	1921-1930	3.31
1651-1660	0.16	1931-1940	2.76
1701-1710	0.32	1941-1950	0.55
1711-1720	0.08	1951-1960	3.00
1731-1740	0.24	1961-1970	6.07
1741-1750	0.16	1971-1980	5.60
1751-1760	0.32	1981-1990	6.86
1761-1770	0.32	1991-2000	11.75
1781-1790	2.05	2001-2010	19.79
1791-1800	0.39	2011-2020	25.39
1871-1880	0.87	2021-2030	0.24

Table 5: Temporal distribution of entities (in %) in TEMP-COFAC with a bin size of 10 years.

Evidences from Pre-training Data. TeCFaP is a novel task of evaluating the consistent temporal relationship between entities in LLMs. Here, we present a few manually extracted evidence from pre-training data to support that the objective of TeCFaP is fairly expected from LLMs. We consider Wikipedia for this test as it is a part of the pre-training dataset for most of the recent LLMs, including LLaMA. Evidences are manually extracted from Wikipedia for an entity pair *hybrid-theory and meteora* for a *subject-relation* pair *Linkin Park* and *Release*. Figure 10 presents the evidence with their sources in the pre-training dataset. Given these sentences, a human can easily find the temporal relation between *hybrid-theory* and *meteora*. Therefore, it is a fair ask from LLMs to learn the temporal relation between entities given such sentences.

Annotations Quality. We ensure the high quality of TEMP-COFAC resource by applying it to execute cross-annotator agreement experiment. To assess the qualitative measure of factuality, We randomly select a hundred samples of filled patterns across different *subject-relation* pairs. The two reviewers split the data in half and reviewed the samples that the other annotator produced during construction. They score the agreements on a scale of 5-point (1-lowest agreement and 5-complete agreement on factuality of filled pattern) Likert scale (Likert, 1932). Similarly, to give the qualitative measure for consistency of patterns created, we again randomly sample hundreds of filled patterns and pair these positively with another paraphrased filled pattern sampled randomly from respective *subject-relation*. The reviewers repeat the similar scoring methodology in factuality assessment (1-

- Hybrid Theory (2000) was certified diamond by the RIAA in 2005. The band's second album, Meteora (2003), reached number one on the Billboard 200 album chart, as did its third album, Minutes to Midnight (2007). **Ref - https://en.wikipedia.org/wiki/Chester_Bennington**

- The sound of later Linkin Park albums would involve experimentation with classical instruments such as strings and piano, both of which, along with the same elements of electronica from Hybrid Theory, are prominently included in the band's second studio album, Meteora. **Ref- https://en.wikipedia.org/wiki/Hybrid_Theory**

- Formed in 1996, Linkin Park rose to international fame with their debut studio album, Hybrid Theory (2000) ... Their second album, Meteora (2003), continued the band's success. The band explored experimental sounds on their third album, Minutes to Midnight (2007). **Ref- https://en.wikipedia.org/wiki/Linkin_Park**

Figure 10: A few extracts from Wikipedia for a *subject-relation* pair (linkin park - release-by). We find the presence of sentences in pre-training data (Wikipedia) of LLMs, which defines the temporal relationship among entities associated with the given *subject-relation* pair.

lowest agreement and 5-complete agreement if the two patterns are paraphrased). We observe a mean agreement of 4.84 ± 0.39 out of 5 maximum for factuality and a mean agreement of 4.86 ± 0.49 on similar lines for consistency.

TEMP-COFAC Coverage. We present the TEMP-COFAC comparison with prior datasets in Table 6. Some of the columns data reused from Tan et al. (2023) comparison of datasets. TEMPREASON is one of the most comprehensive temporal reasoning datasets, with 21K queries of event-event probe type in the QA setting. Here are the six types of entities it has considered during automatic construction from Wikipedia: Person, School, Political party, Company, Position, and Sports team. A significant drawback of TEMP-REASON is that all the queries are anchored around just person names, either as a subject or as an object. I.e., *Which team did <subject> play for before/after oj?* or *Who was the head of the government of <subject> before/after oj?*.

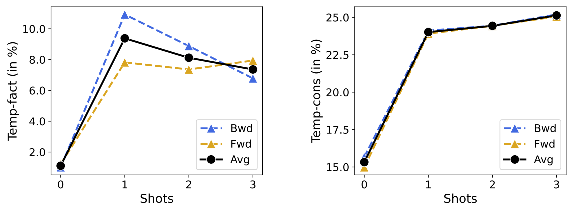
In comparison, the proposed TEMP-COFAC has 10.1K prefix-style queries and covers eleven diverse entity types from several domains: Movies, Geographical location, Games, Albums, Persons, Satellites, Software, Books, Vehicles, Songs, and Elements. Here, the queries are anchored around eleven entity types via sixty-six diverse subject-relation pairs, such as *Linkin Park released Hybrid Theory just before* ___ or *The FIFA U-17 World Cup was hosted by Canada immediately after* ___.

The dataset will facilitate the further development around various aspects of consistent temporal reasoning such as consistent event sequencing, multi-hop sequence-based QA, consistent multi-reasoning QA, and architectural study of LLMs through temporal explainability of inconsistent behaviour via paraphrased queries.

A.2 Extended Results

A.2.1 ICL Setting Cont.

Here, we continue from the result section and present results for *temporal-factuality* and *temporal-consistency* in ICT setting (Figure 11).



(a) Temporal factuality (b) Temporal consistency

Figure 11: Results for k-shot (k=1,2,3) ICL setup with LLaMA[13B] in an open vocabulary setup across – (a) *Temp-fact*: temporal factuality, (b) *Temp-cons*: temporal consistency.

A.2.2 Closed Vocabulary Setting Cont.

We continue from the Section 5 to present the results for metrics *temporal-factuality* and *temporal-consistency* in closed vocabulary setting in Figure 12. Three-shot with closed vocab setting significantly outperforms by attaining maximum scores of 12.79% and 55.06% for *temporal-factuality* and *temporal-consistency* respectively across various variants and settings.

A.2.3 Error Analysis Cont.

The positive case is presented in Table 7 where CoTSeLF improves the temporal consistent factuality over a baseline model. It can be observed that The temporal consistency-driven factual improvement makes substantial changes both in the model temporal consistency and factuality.

A.2.4 Probabilistic Space Analysis

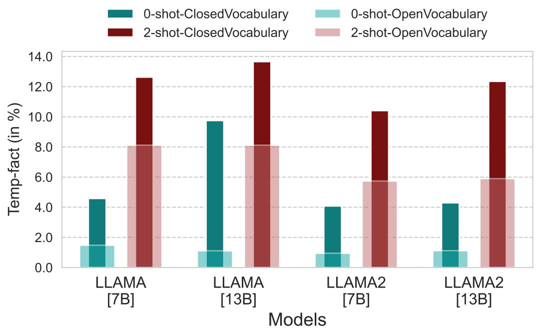
We conduct an exploration for the evidences of improvement in model consistency through its probabilistic space analysis. The divergence in probability distribution for the next word shall be minimal for identical intent paraphrases to generate consistent *value_object* for given

Dataset	Format	Knowledge Source	Paraphrases	Temporal Sequence of Homogeneously Grouped Entities	Time Coverage	Size
TEMPCOFAC	Prefix-Style:Language	Human Annotated/Open Web	✓	✓	1526-2022	10.1K
TEMPREASON(Tan et al., 2023)	QA:Language	Wikipedia/Wikipedia	✗	✗	634-2023	52.8K
TEMPLAMA(Dhingra et al., 2022)	QA:Language	Wikipedia	✗	✗	2010-2020	50k
Time-SensitiveQA(Chen et al., 2021)	QA:Language	Wikipedia/Wikipedia	✗	✗	1367-2018	41.2k
StreamingQA(Liška et al., 2022)	QA:Language	WMT	✗	✗	2007-2020	147k
TempQuestions(Jia et al., 2018b)	QA:KG	Freebase	✗	✗	NA	1.2k
TimeQuestions(Jia et al., 2021)	QA:KG	Wikipedia	✗	✗	NA	16.1k
CronQuestions(Saxena et al., 2021)	QA:KG	Wikipedia	✗	✗	34-2021	410k

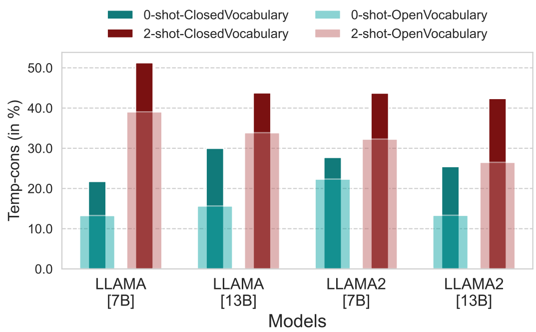
Table 6: Comparison of TEMPCOFAC with prior datasets.

S.No.	Paraphrases	Baseline Model	$\texttt{CTSRL}_{\texttt{Discrete}}$
1	the academy award for best picture was won by the departed immediately before	no country for old men	no country for old men
2	the departed won the best picture academy award just before	the reader	no country for old men
3	the departed won the oscar for best picture right before	the curious case of benjamin	no country for old men
4	the departed won the best picture academy award succeeded by	the curious case of benjamin	no country for old men
5	movie the departed won the academy award for best picture right before	the curious case of benjamin	the artist
6	the departed won the oscar for best picture immediately prior to	the curious case of benjamin	no country for old men
7	the academy award for best picture was won by the departed soon before	the curious case of benjamin	no country for old men
8	the departed was awarded the academy award for best picture immediately prior to	the curious case of benjamin	the departed

Table 7: A comparison of output generated by a baseline model and the discrete variant of CTSRL model for the sentence completion task. Where, *subject-relation* pair is "academy award for best picture - win by", *key_object* is "the departed," and an expected *value_object* is "no country for old men." The rest of the settings are similar as in the case of Table 3.



(a) Temporal Factuality



(b) Temporal Consistency

Figure 12: A comparison between open and closed (using restricted candidate set) vocabulary settings across variants of LLaMA model in zero-shot and in-context learning setup with k-shot (k=3).

key_object. Similarly, the divergence shall be wider for different intent paraphrases. We use positive and agnostic paraphrases as notations to denote paraphrases with identical intent and paraphrases with different intent. We apply random hard negative sampling while considering the agnostic paraphrases. Here is an example of a positive and agnostic paraphrases.

Positive paraphrases
P1: *hybrid theory was released by linkin park just before*
P2: *linkin park released hybrid theory immediately before*
Agnostic paraphrases
P1: *hybrid theory was released by linkin park just before*
P2: *linkin park released hybrid theory immediately after*

The KL divergence metric is widely used to compare probabilistic distributions. We calculate the KL divergence of subsequent word's probability distribution between positive and agnostic paraphrases, respectively. The objective is to maximize the difference between these two scores. We perform this experiment on all the entities for randomly selected ten *subject-relation* from test data. We randomly sample five pairs of positive and respective agnostic sentences. The scores are then averaged over a *subject-relation* pair, presented in Table 8. We compare the scores between the default LLaMA[13B] model and $\texttt{CTSRL}_{Discrete}$ model.

It is evident from Table 8 that the CoTSeLF improves the average difference between the KL divergence scores of positive paraphrase and agnostic paraphrase by the value of 0.18 nats. We also observe a positive change in eight *subject-relation* pairs out of a total ten in the experiment. This analysis helps explain why the CoTSeLF achieves better *temporally-consistent-factuality*.

Subject-Relation	Default			CTSRL$_{\text{Discrete}}$			Δ
(ID)	PP	AP	Diff (A)	PP	AP	Diff (B)	(A - B)
34	1.42	1.52	0.09	1.67	2.10	0.43	0.34
1	5.01	4.95	-0.06	5.51	5.19	-0.32	-0.26
19	1.99	2.01	0.03	2.77	2.91	0.14	0.11
42	1.79	1.72	-0.07	2.89	3.36	0.47	0.54
28	3.98	4.07	0.09	4.07	4.6	0.54	0.45
49	1.62	1.64	0.02	2.07	2.13	0.05	0.03
36	1.14	1.29	0.15	2.96	2.91	-0.05	-0.20
53	3.09	2.92	-0.17	2.26	2.17	-0.08	0.08
9	2.21	2.22	0.02	3.22	3.42	0.20	0.18
8	2.43	2.41	-0.02	2.15	2.65	0.50	0.51
Average	2.47	2.48	0.01	2.96	3.15	0.19	**0.18**

Table 8: The results present a comparison between the default LLaMA[13B] model and CTSRL$_{Discrete}$ variant (a fine-tuned LLaMA[13B] model) of CoTSeLF in probabilistic space. Where PP and AP are the KL divergence score between the next word's probability distribution of positive paraphrases and agnostic paraphrases, respectively, Diff is the score difference between agnostic paraphrases (AP) and positive paraphrases (PP), and Δ is the difference between the two models differentials outcome. A positive value of Δ represents that the CoTSeLF has a broader separation of divergence in agnostic paraphrases compared to positive paraphrases, reflecting a more consistent model.

A.2.5 Significance Test for Alpha

We carry out the experiment on significance of parameter α used in CTSRL$_{Discrete}$ formulation. The experimental and data setting is the same as in section 5 for MT-IT + CTSRL$_{Discrete}$ formulation. The results are presented in Figure 13 for three different values of α (= 0.5, 0.66 and 0.75) across all three metrics *temporal-factuality*, *temporal-consistency* and *temporally-consitent-factuality* respectively. We observe the optimal performance at α=0.66 for *temporal-factuality* and *temporally-consitent-factuality*. Whereas *temporal-consistency* further improves as we increase the value of α. We have selected α as 0.66 for the main results presented in Table 3 based on the outcome of this ablation.

A.2.6 Significance Test for Sample Size

in Section 5, we analyzed sampling efficiency across temporal variations in entity size and their corresponding performance, as depicted in Figure 5. Furthermore, an additional experiment was undertaken to examine the relationship between the distribution of entity types and their respective performance within the test data set. Our findings indicate a notable negative correlation, quantified as -0.61, between the distribution of entity types and *temporal-factuality*. Likewise, a negative correlation, valued at -0.42, was observed in relation to the distribution of entity types and *temporally-consistent-factuality*. It was discerned that performance pertaining to domain-specific entity types, such as Software, Satellites, and Games, surpassed that of more general entity types, like Location and Person, with the data distribution being inversely skewed towards the latter.

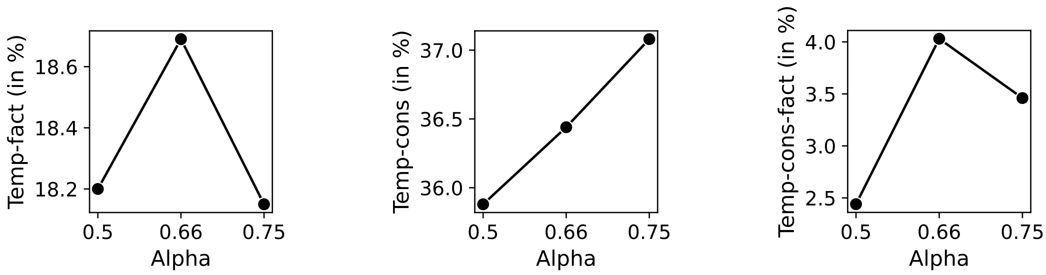
A.3 Discussion on CoTSeLF's Scalability

In the majority of instances, the solutions that rely on parameters fine-tuning, regardless of the specific large language model (LLM) architecture involved, demonstrate scalability and adaptability to diverse models without being constrained by the number of parameters. This ablation study is performed to investigate the scalability of the proposed solution, CoTSeLF, specifically its capacity to scale according to the model parameter size. For this purpose, the CoTSeLF is implemented in two different LLMs configurations, one with a reduced parameter size, LLaMA[7B] and another with an increased parameter size, LLaMA[30B]. However, it was necessary to adjust the quantization from 8-bit to 4-bit for LLaMA[30B] and MT-IT's LoRA adapter both due to computational resource constraints. The experiment is conducted in same conditions as outlined in Table 3

The findings from Table 9 reaffirm CoTSeLF's efficacy for models across a spectrum of parameter sizes, inclusive of those with large dimensions. A pronounced correlation between CoTSeLF's performance and the sizes of the parameters was noted. Moreover, it was observed that enhancements attributed to CoTSeLF are significant across all evaluated metrics with LLaMA[30B] despite an increment in quantization level.

A.4 Discussion on Commercial LLMs

This section presents the evaluation results for two prominent commercial LLMs, GPT-4 and Claude-3. We conducted this experiment with a limited randomly selected 100 samples from the TEMP-COFAC test set due to financial limitations. We set up the probe in zero-shot, where the task is to complete a sentence with the correct phrase.



(a) Temporal Factuality (b) Temporal Consistency (c) Temporally Consistent Factuality

Figure 13: The results present the significance of a parameter alpha (α) in the formulation of $\texttt{CTSRL}_{Discrete}$ across three metrics: (a) temporal factuality, (b) temporal consistency, and (c) temporally consistent factuality respectively in percentages. The rest of the settings for this experiment are the same as in the case of Table 3.

Model [Parameter Size] [Quantization]	Temp-fact	Temp-cons	Temp-cons-fact
LLaMA[7B][8-bit]	8.10	24.07	0.00
LLaMA[13B][8-bit]	18.72	36.88	4.34
LLaMA[30B][4-bit]	23.05	46.55	6.04

Table 9: CoTSeLF's ($\texttt{CTSRL}_{Discrete}$) results across various parameter sizes of LLaMA.

To this purpose, an instruction (Instruction: "complete the given sentence with the correct phrase") is provided along with the input to the model. Subsequently, we provide a concise examination of the potential for employing CoTSeLF within the realm of commercial LLMs applications.

A.4.1 GPT-4 Evaluation Results

We use 'GPT-4-0125-preview' version of GPT-4 for this experiment. In a zero-shot setting, the *temporal-factuality* stands at 36.00% (Table 10). Trained with a trillion of parameters, in absolute terms, the *temporal-factuality* in GPT-4 is not very impressive, and requires interventions such as CoTSeLF to improve it. An example of each of the positive and negative responses is mentioned in Table 11.

A.4.2 Claude-3 Evaluation Results

In conducting a parallel investigation with another leading commercial LLM, Claude-3, under identical conditions, we operationalized the 'CLAUDE-3-OPUS-20240229' version for this assessment. The *temporal-factuality* of Claude-3 is noted at 20.00% (Table 10), marking a notable decline in performance relative to GPT-4.

Due to a minimal sample set, we are short of reporting either the *temporally-consistent-factuality* or *temporal-consistency* for this test. Metrics like temporal consistency and temporally consistent factuality require a sufficient number of pair of samples for a *subject-relation* to calculate those. The other factors are temporal direction and entity pairs, which are to be considered while calculating. The 100 individual samples might be enough to gauge the preliminary assessment of temporal factuality but statistically insufficient for reporting temporally consistent factuality.

Model	Settings	Temp-fact
GPT-4	GPT-4-0125-preview	36.00
Claude-3	Claude-3-Opus-20240229	20.00

Table 10: Temporal factuality (in %) of two commercial LLMs, GPT-4 and Claude-3 through the minimal sample set.

A.4.3 CoTSeLF's Applicability

Applicability across commercial LLMs. Commercial model's such as GPT-4 and Claude-3's *temporal-factuality* remains suboptimal, thus necessitating strategies like CoTSeLF or similar enhancements to augment its *temporal-factuality*. CoTSeLF advances an RL-based fine-tuning methodology that refines the model's parameters via a custom-made fine-tuning procedure. Almost universally, advancements in parameter fine-tuning methodologies are contingent upon access to the model's architecture. GPT-4 and Claude-3, by restricting direct architectural access, necessitate that such methodologies could demonstrate their effectiveness only on open-source models. Thus, we could not assess CoTSeLF's efficacy on GPT-4 or Claude-3 models due to the inaccessibility of its architecture for open fine-tuning.

S.No.	Prompt	Correct Answer	GPT-4 Response
1	Android Gingerbread was released by google immediately after android	Froyo	Froyo
2	Indian band Euphoria released album Sharnagat right after album	Item	MeHFuz

Table 11: An example of each of the positive and negative response from GPT-4 in zero-shot setting.

Considering that GPT-4 adhered to a decoder-based transformer architecture (up to GPT-2) before the later versions transitioned to a commercial model, there is an intense anticipation that CoTSeLF, particularly CTSRL, will serve as an efficacious approach to augmenting the closed-source model's capabilities in improving *temporally-consistent-factuality* as in case of various versions of LLaMA model under experimentation.

An alternative approach involves examining methods that do not necessitate access to the model's architecture for parameter updates, acknowledging that typically, strategies not involving parameter modifications exhibit suboptimal performance compared to those that do, like CoTSeLF. Future research should focus on enhancing the *temporally-consistent factuality* of commercial LLMs through techniques that might bypass the need for gradient updates.

Applicability across domains. The establishment of consistent temporal reasoning capabilities is poised to profoundly influence their applicability across a spectrum of complex fields, including healthcare and the legal sector. Specifically, within the legal domain, the temporally reliant nature of case precedents and statutory amendments demands the reliable and precise retrieval of information. LLMs must avoid offering divergent interpretations of precedents for specific case types to prevent severe repercussions and undermine trust in the judicial system. The precedence of rules, articles, and cases, being temporally bound, demands their consistent retrieval when employing LLMs in legal settings. Similarly, in healthcare, the prognostication of diseases is time-sensitive, mandating consistent and precise forecasts for diagnosing and formulating treatment plans, leading to enhanced trust in the usage of AI-based medical solutions. Techniques like CoTSeLF could be tailored to meet such requirements.

A.5 Hyperparameters

All codes were composed utilizing PyTorch. We utilized the Huggingface[*] repository for stacking the open-source LLMs. The implementation of RL is carried out through trlX, a python-based library (Castricato et al., 2023). A PEFT-based method LoRA with 8-bit quantization is used for both instruction-tuned and PPO-based RL models. We apply minimal pre-processing on generated outputs from LLMs such as filtering of articles or special characters (at max up to one word) before retrieving the *value_object*. The maximum sequence length of 256 is used across all experiments. An NVIDIA A100 GPU (80 GB)[*] was used to train the model.

[*] https://huggingface.co/

[*] https://www.nvidia.com/en-in/data-center/a100/