Counterfactual-Consistency Prompting for Relative Temporal Understanding in Large Language Models

Anonymous ACL submission

Abstract

Despite the advanced capabilities of large language models (LLMs), their temporal reasoning ability remains underdeveloped. Prior works have highlighted this limitation, particularly in maintaining temporal consistency when understanding event relations. For example, models often confuse mutually exclusive temporal relations like "before" and "after" between events and make inconsistent predictions. In this work, we tackle the issue of temporal inconsistency in LLMs by proposing a novel counterfactual prompting approach. Our method generates counterfactual questions and enforces collective constraints, enhancing the model's consistency. We evaluate our method on multiple datasets, demonstrating significant improvements in event ordering for both explicit and implicit events, as well as in temporal commonsense understanding, by effectively addressing temporal inconsistencies.

1 Introduction

Despite the impressive capabilities of LLMs, a line of research (Jain et al., 2023; Chu et al., 2023) has highlighted that these models often lack temporal reasoning abilities. This is especially true for *relative* event understanding, where the goal is to infer temporal relationships between events or within an event in the passage, without depending on *absolute* time indicators (e.g., specific dates).

The primary challenge is that LLMs lack *temporal consistency* in their responses (Qiu et al., 2024; Chen et al., 2024). Temporal consistency is defined as the model's ability to ensure that conflicting timelines do not co-exist. For instance in Figure 1-(a), if the model is temporally inconsistent, mutually exclusive temporal relations like "before" and "after" are sometimes confused when ordering events, leading to contradictory predictions—such as stating that Event A happens both before and after Event B in the same context. While events with time indicators are often addressed

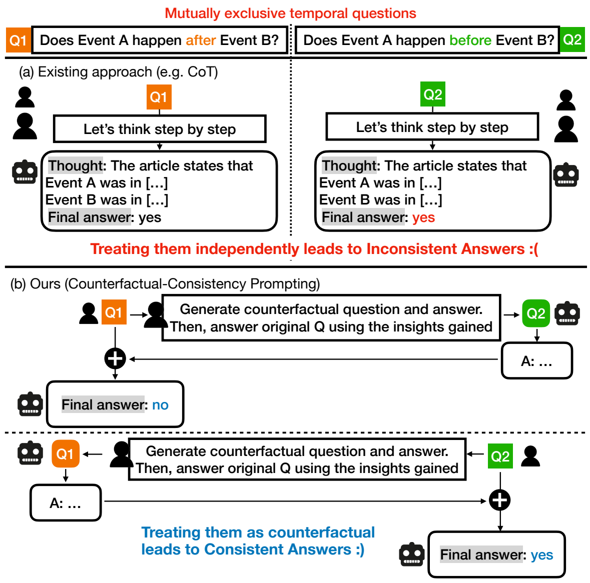

Figure 1: Our approach generates counterfactual questions to address the temporal inconsistency in LLMs.

with mathematical reasoning (Zhu et al., 2023; Su et al., 2024), no existing work has successfully tackled the challenge of temporal inconsistency in the events' relative relationship without requiring explicit time markers. Chain-of-thought (CoT) reasoning (Wei et al., 2022), which primarily aids mathematical and symbolic reasoning (Sprague et al., 2024), is also reported to fail to solve such inconsistency (Qiu et al., 2024). Considering temporal consistency is fundamental in temporal reasoning, its absence in LLM can undermine key tasks like planning (Sakaguchi et al., 2021; Zhang et al., 2024). These observations highlight the need for alternative reasoning skills to achieve temporal consistency.

This study answers the following research question: **Can we prompt LLMs to elicit the ability to mitigate temporal inconsistency?** Inspired by counterfactual augmentation, where models are exposed with lexically similar, but typically label-flipping pairs in training (Kaushik et al., 2020), we extend it to LLMs to generate temporally *coun-*

terfactual questions: We introduce lexically small interventions to the original input (e.g. before to after) that drastically affect its temporal semantics. By providing these questions and self-generated answers alongside the original input, the model would rely less on lexical similarities and better understand the semantics.

To this end, we propose a novel counterfactual-consistency prompting (CCP), designed to enhance the temporal consistency of LLMs, as described in Figure 1-(b). CCP first generates temporal counterfactual exemplars and then applies the insights gained to address the original temporal question. This method is particularly effective in relative event understanding because the counterfactual exemplars not only encourage the model to understand different temporal semantics but also directly impose temporal constraints. For instance, if the model states that "Event A happens after Event B" and also recognizes that "Event A happens before Event B", the conflict forces the model to collectively re-weight the validity of these two statements.

We show performance gain of CCP across multiple relative event understanding tasks. Our effectiveness in mitigating temporal inconsistencies is further demonstrated by our inconsistency metric.

2 Related Work

2.1 Self-correction with Ensemble in LLMs

Studies have demonstrated the advantages of producing multiple predictions and aggregating them to self-correct the initial answer, such as self-consistency (Wang et al., 2023) and multiagent-debate (Du et al., 2024). However, they can lead to errors as they solely rely on feedback from a single question. Our distinction is that we aggregate the answer from multiple questions, which provides more diverse reasoning on the original question.

The most relevant work to ours is Chen et al. (2024) in logical reasoning. While they aggregate the fixed set of questions and predefined rules to correct the answer consistently, such fixed settings limit flexibility across contexts. In contrast, our correction is done by dynamically generating related temporal questions, providing a more adaptable evaluation.

2.2 Counterfactual Data Augmentation

The goal of the augmentation is generating a counterfactual instance by making minimal edits like lexical changes while keeping others unchanged (Huang et al., 2019; Kaushik et al., 2020; Wang and Culotta, 2020). This discourages models from relying too much on lexical similarity or dissimilarity. As retraining with augmented data may not be practical for large models, we apply this approach during inference instead.

3 Method

Given a context C and corresponding question Q, our goal is to provide an accurate answer while maintaining temporal consistency.

Our idea is to apply the insights from the counterfactual exemplars to answer the original question. We want the model to approximate the *temporal constraints* from the counterfactuals. Examples of temporal counterfactuals are in Table 1. For example, if the model establishes from a counterfactual exemplar [1-2] that "Event e_1 happens [r_2: before] Event e_2", it is causally constrained to predict the original question [1-1] that "Event e_1 cannot happen [r_1: after] Event e_2":

$$r_2(e_1, e_2) \in \mathcal{V} \implies r_1(e_1, e_2) \notin \mathcal{V} \quad (1)$$

where $r(e_a, e_b)$ represents the temporal relation r between events e_a and e_b, and $\mathcal{V}$ represents the set of coherent temporal relations with the context.

We start by generating counterfactual questions, $Q^{c_1}...Q^{c_i}$, by making small but impactful temporal interventions to the original question. Given the broad range of counterfactual questions beyond what a rule-based system can fully capture, we design the model to autonomously generate these counterfactual questions. We guide the model to follow the counterfactual aspects provided through in-context learning (ICL) to control the relevance of the generated questions, where the full prompts are in Appendix G.1. After the counterfactual questions are made, the model also self-generates auxiliary answers $Y^{c_1}...Y^{c_n}$. We note that we use exemplars irrespective of whether they are label-flipping or label-preserving, as both contribute to improving the model's robustness (Zhou et al., 2022).

However, there is a risk when LLMs may fail to answer the counterfactual questions correctly. In this case, their direct use propagates errors to the original question. As a proxy for determining whether the generated prediction can be trusted, existing works aggregate multiple predictions of the same question (Wang et al., 2023; Madaan

Index	Relation	Counterfactual	Examples
1-1	$r_1(e_1, e_2)$	Original	Do they got thanked after they help someone maintain their home?
1-2	$r_2(e_1, e_2)$	$r_1 \longrightarrow r_2$	Do they got thanked *before* they help someone maintain their home?
1-3	$r_1(e_3, e_2)$	$e_1 \longrightarrow e_3$	Do they *uncovered the dark side of life* after they help someone maintain their home?
2-1	$r_1(e_1)$	Original	Are they still helping people?
2-2	$r_2(e_1)$	$r_1 \longrightarrow r_2$	Are they *no longer* helping people?
3-1	$r_1(e_1)$	Original	Do they help someone maintain their home every month?
3-2	$r_2(e_1)$	$r_1 \longrightarrow r_2$	Do they help someone maintain their home *twice a month*?

Table 1: Examples of original and counterfactual question types across different relations. The examples illustrate how counterfactual questions modify the semantics regarding temporal relations (r_1,r_2) between events (e_1,e_2,e_3).

et al., 2024; Du et al., 2024). Formally, the refined prediction Y is derived by re-weighting the probability distribution P of previous predictions $Y_1, ..., Y_n$ from the same question as: $P(Y) = f(P(Y_1), ..., P(Y_n))$ where f is an aggregation function such as majority voting or LLM itself.

Our distinction is to aggregate predictions from both the original and counterfactual questions. We design the model to re-weight the counterfactual answer distributions across the questions.

$$P(Y) = f(P(Q, Y), P(Q^{c_1}, Y^{c_1}), ..., P(Q^{c_n}, Y^{c_n})) \tag{2}$$

For instance, even if the model wrongly predicts the relation as 'after' in a counterfactual, collectively considering the possibility of the relation 'before' can re-weight the effect of the constraint. Full prompts are provided in Appendix G.

This re-evaluation approach improves robustness against potential errors in generated answers. The analysis in Subsection 4.6 shows such self-correction outperforms a baseline directly leveraging counterfactuals without aggregation.

4 Experiments

4.1 Datasets

Among publicly available datasets, we selected three based on two criteria: (1) the task focuses on understanding relative event relationships without absolute time indicators, and (2) the temporal inconsistency on the dataset can be evaluated.

TempEvalQA-Bi (Qiu et al., 2024) involves ordering two explicit events in time, assessing temporal consistency in mutually exclusive question pairs. **TRACIE** (Zhou et al., 2021) expands the event-event ordering to implicit events, testing whether the hypothesis logically follows the story. We finally added **MCTACO** (Zhou et al., 2019) regarding the diverse event-related temporal properties. The dataset covers broader aspects like event duration or frequency, as illustrated in Appendix A. We modified the multiple-choice setting of MCTACO into a binary question-answering task for consistency evaluation, presenting each answer candidate separately to determine if it fits the context. Dataset statistics and examples are in Appendix A.

4.2 Metrics

Along with accuracy (ACC) and F1 scores to assess overall performance, we introduce the inconsistency metric (INC) as a main evaluation measure for temporal inconsistency. We define the INC as the percentage of inconsistent predictions. An inconsistency is counted when at least one incorrect answer is found within a group of minimally dissimilar questions with slight modifications in their temporal semantics, while all other aspects remain unchanged.

TempEvalQA-Bi directly provides this metric. For TRACIE, we manually group questions that are counterfactual to each other. We adapt INC in MCTACO by grouping original multiple-choice candidates by question. An inconsistency is counted if at least one incorrect answer exists among the candidates for a given question.

4.3 Evaluation Settings and Baselines

For models, we used open-source LLM **Llama-3 8B** and **70B** (AI@Meta, 2024), and API-based LLM **GPT-4o-mini** and **GPT-4o** (OpenAI et al., 2024).

For baselines, we employ a 3-shot setting across all configurations. First, we compare CCP with standard prompting (**SP**) that directly answers the question without intermediate steps, and **CoT**, which incorporates step-by-step reasoning to derive the answer. Next, we consider methods that aggregate multiple predictions of the same question. Self-**Consistency** (Wang et al., 2023) predicts one question multiple times and performs majority voting. Self-**Reflect** methods (Madaan et al., 2024; Shinn et al., 2024) iteratively refine own

		TempevalQA-Bi			TRACIE			MCTACO		
		ACC	F1	INC (↓)	ACC	F1	INC (↓)	ACC	F1	INC (↓)
Llama -3-8B	SP	65.4	63	57.6	57.4	66.9	75.2	77.7	69.4	59.8
	CoT	69.6	70.6	50	63	64.9	56	77.6	69.8	63.4
	Consistency	70.8	71.2	49.6	64.9	67.3	57.8	77.5	69.0	61.1
	Reflection	63.6	63.9	44.6	62.5	55.7	55.5	77.4	69.7	76.4
	Debate	67.6	65.2	52.2	63.6	66	53.2	37.4	31.6	88.1
	CCP	**75.9**	**75.2**	**32.7**	**68.8**	**70.4**	**39.8**	**82.9**	**80.4**	**56.0**

Table 2: Performance comparison on the test set of relative event understanding tasks. Other models are in Table 5.

predictions. Multi-agent **Debate** (Du et al., 2024) leverages both majority vote and reflection. More details on evaluation settings are in Appendix B.

4.4 Main results

Table 2 highlights the performance of our method compared to baseline methods on relative event understanding tasks. Compared to SP, the CoT baseline is not usually effective and often worsens performance. Advanced baselines, Consistency, Reflect, and Debate, also fail to consistently reduce inconsistencies or achieve competitive accuracy. In contrast, CCP steadily outperforms these baselines, significantly reducing temporal inconsistencies across all datasets and achieving notable improvements in ACC and F1 scores. The full results on other models are available in Table 5 in Appendix C.

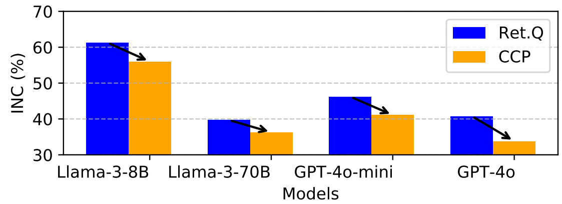

Figure 2: Comparison between counterfactual example collection methods on MCTACO.

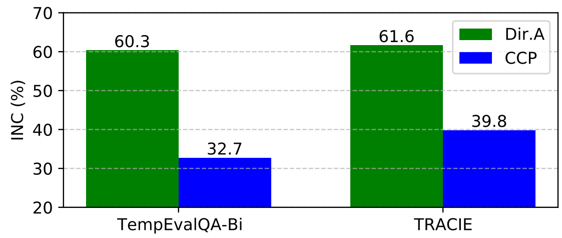

Figure 3: Comparison between different counterfactual leveraging methods with the Llama-3-8B model.

4.5 Generated vs Retrieved Questions

We designed our method to generate counterfactual questions to handle diverse temporal relations. To support this claim, we compared our generative setting with ‘Ret.Q’, where counterfactual questions were retrieved from the other questions in the same question group. We evaluated the methods on MCTACO which covers various aspects of events.

As shown in Figure 2, generating counterfactual questions proved more effective for all models. These results suggest that our method performs better in event understanding with diverse relations, where the dataset cannot often provide high-quality counterfactual questions. Notably, CCP outperforms the Ret.Q baseline even though our method occasionally produces incorrect answers. We also note that CCP is more practical since Ret.Q assumes the questions in the test set are visible.

4.6 Robustness on Counterfactual Exemplars

To validate our robustness against wrong counterfactual exemplars, we conducted a comparative analysis of two methods: answering directly from counterfactual exemplars Dir.A) versus leveraging the aggregation step to re-evaluate them (CCP). We conducted experiments on the TempEvalQA-Bi and TRAIE datasets, which provide question pairs involving mutually exclusive (counterfactual) temporal scenarios. In the Dir.A implementation, the answer to the counterfactual question is flipped and directly used as the response to the original question. The results in Figure 3 demonstrate that CCP consistently outperforms Dir.A, supporting our robustness by the collective evaluation.

5 Conclusion

We targeted the temporal inconsistency in relative event understanding with LLMs by proposing a prompting approach using counterfactual questions. This encourages the model to focus more on the temporal relations and collectively evaluate its answer with imposed constraints. Experiments with the INC metric show that our approach mitigates inconsistency and improves overall performance.

6 Limitation

Our method showed limited performance improvement when time indicators, such as specific years (e.g., 1980), are involved in temporal understanding. This is implied from our evaluations on event-time ordering and time-time ordering tasks, as shown in Appendix E. The findings suggest that arithmetic reasoning is essential for grounding timelines with absolute time indicators, as emphasized in prior studies (Su et al., 2024; Zhu et al., 2023).

We focused on pointwise and pairwise event reasoning to highlight the model's struggles with basic temporal reasoning due to consistency issues. We anticipate future work expanding our approach to more complex listwise ordering like event schema prediction (Zhang et al., 2024).

References

AI@Meta. 2024. Llama 3 model card.

Meiqi Chen, Yubo Ma, Kaitao Song, Yixin Cao, Yan Zhang, and Dongsheng Li. 2024. Improving large language models in event relation logical prediction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9451–9478.

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

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

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*.

Yi Fang, Moxin Li, Wenjie Wang, Hui Lin, and Fuli Feng. 2024. Counterfactual debating with preset stances for hallucination elimination of llms.

Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. 2019. Reducing sentiment bias in language models via counterfactual evaluation. *arXiv preprint arXiv:1911.03064*.

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.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2024. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.

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, Mohammad 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, Jan 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 H. 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 B. 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. 2024. Gpt-4 technical report.

Yifu Qiu, Zheng Zhao, Yftah Ziser, Anna Korhonen, Edoardo Ponti, and Shay Cohen. 2024. Are large language model temporally grounded? 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 7064–7083, Mexico City, Mexico. Association for Computational Linguistics.

Keisuke Sakaguchi, Chandra Bhagavatula, Ronan Le Bras, Niket Tandon, Peter Clark, and Yejin Choi. 2021. proScript: Partially ordered scripts generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2138–2149, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2024. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36.

Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2024. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. *arXiv preprint arXiv:2409.12183*.

Romain Storaï and Seung-won Hwang. 2024. Harp: Hesitation-aware reframing in transformer inference pass.

Zhaochen Su, Jun Zhang, Tong Zhu, Xiaoye Qu, Juntao Li, Min Zhang, and Yu Cheng. 2024. Timo: Towards better temporal reasoning for language models. *arXiv preprint arXiv:2406.14192*.

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

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Zhao Wang and Aron Culotta. 2020. Identifying spurious correlations for robust text classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3431–3440.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

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

Liang Yao. 2024. Large language models are contrastive reasoners.

Xinliang Frederick Zhang, Nicholas Beauchamp, and Lu Wang. 2024. Narrative-of-thought: Improving temporal reasoning of large language models via recounted narratives. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16507–16530, Miami, Florida, USA. Association for Computational Linguistics.

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

Ben Zhou, Kyle Richardson, Qiang Ning, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2021. Temporal reasoning on implicit events from distant supervision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1361–1371.

Jing Zhou, Yanan Zheng, Jie Tang, Li Jian, and Zhilin Yang. 2022. FlipDA: Effective and robust data augmentation for few-shot learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8646–8665, Dublin, Ireland. Association for Computational Linguistics.

Xinyu Zhu, Cheng Yang, Bei Chen, Siheng Li, Jian-Guang Lou, and Yujiu Yang. 2023. Question answering as programming for solving time-sensitive questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12775–12790.

Appendices

A Data Summary

Table 3 summarizes the dataset statistics used in this study. The numbers of official test samples are reported. Due to the budget, we evaluated Llama-3-8B on the full test set, GPT-4o-mini and Llama-3-70B on a random sample of up to 2,000 test set instances and GPT-4o on 1,000 test set instances.

Additionally, the number of temporal relations considered in each dataset is included in Table 3. TempEvalQA-Bi and TRACIE focus mainly on the before-after relation. MCTACO includes diverse temporal relations, and the number of annotated candidates is reported. The questions in MCTACO are categorized into 5 question types, and examples for each type are provided in Figure 4.

	#Test	#Temporal relations
TempEvalQA-Bi	448	2
TRACIE	4248	2
MCTACO	9442	1-19

Table 3: Dataset Statistics. For TempEvalQA-Bi, the numbers represent the total number of questions. For TRACIE, the numbers refer to the number of story-hypothesis pairs. For MCTACO, the numbers reflect question-and-answer candidate pairs.

B Details of Evaluation Settings

This section outlines the detailed evaluation settings, including hyperparameters, resources, efficiency, and parsing methods. We use greedy decoding for **SP**, **CoT**, and **CCP**. For **Consistency**, **Reflect**, and **Debate**, we adopt the approach from Wang et al. (2023), employing top-k sampling with $k = 40$ and a temperature of 0.5 for the LLaMA model. For GPT-based models, we set the temperature to 0.7. **Consistency** samples 40 outputs from the decoder. **Reflect** refines the output iteratively for two iterations, including the initial output. In **Debate**, three agents engage in a debate over two rounds(Du et al., 2024). The implementations of the latter two baselines (**Reflect**, **Debate**) are based on the GitHub repository [1] from Du et al. (2024). Single-run performances are reported.

We note that our method prompts 3 times: for counterfactual question generation, counterfactual answer generation, and original question's answer

[1] https://github.com/composable-models/llm_multiagent_debate

Event Duration
P. However, more recently, it has been suggested that it may date from earlier than Abdalonymus' death. Q. How long has it existed? (A) 2,000 hours (B) 2,000 years (C) 1 year (D) thousands of years (E) centuries (F) months

Event Frequency
P. Most of us have seen steam rising off a wet road after a summer rainstorm. Q. How often does it rain in the summer? (A) 333.33 times (B) every other minute (C) a couple times (D) every month (E) once a hour (F) once a year

Stationarity
P. She renews in Ranchipur an acquaintance with a former lover , Tom Ransome , now a dissolute alcoholic. Q. Is she still in Ranchipur? (A) yes (B) no

Event Ordering
P. Some of the people who took advantage of her through a questionable loan program were sent to jail. Q. What happened after they were put in jail? (A) they went to the store (B) they repented (C) even some people took these steps

Typical Time
P. Durer's father died in 1502, and his mother died in 1513. Q. When did Durer die? (A) 40 years later (B) 360 years later (C) 4545 (D) 40 seconds later (E) April 6, 1528

Figure 4: Examples of MCTACO Question Types. MCTACO covers various temporal relation categories including event duration, frequency, stationarity, ordering, and typical time.

generation, whose efficiency is compatible with or even more efficient than the three baselines. We also note that the Consistency baseline of Llama-3-70B cannot be reported due to its computation inefficiency.

For resources, we used the Transformers library (Wolf et al., 2020) and vLLM (Kwon et al., 2023) with 4 RTX A6000 GPUs for Llama-3 models. We used Openai API [2] for GPT models. For output parsing, the models generate the final answer after the phrase "`Final answer:`". Counterfactual exemplars are generated by modifying the questions, hypotheses, and candidate answers for each dataset.

[2] `platform.openai.com`

Models	Methods	MCTACO	
		EM	F1
GPT-4o-mini	CoT	41.1	56.7
	MCQA-CoT	39.4	60.2
	CCP	**58.9**	**78.6**
GPT-4o	CoT	50.3	67.2
	MCQA-CoT	51.0	63.1
	CCP	**66.2**	**80.2**

Table 4: Performance comparison on MCTACO with multiple-choice question answering setting.

C Details of Main Results

Table 5 shows the performance of our method compared with baseline methods on relative event understanding tasks. The results show that our method outperforms the baselines across the board.

To demonstrate that our solution extends beyond binary question answering to multiple-choice question answering (MCQA), we evaluated the performance of GPT models using the original MCTACO evaluation setting (Zhou et al., 2019). While our primary evaluation decomposed the multiple-choice format into binary questions to measure inconsistency, it can be reconstructed for multiple-choice evaluation. We additionally introduced a baseline for MCQA (MCQA-CoT) that provides the context, question, and all candidate answers, generating one or more correct answers step-by-step. The results in Table 4 indicate that 1) question formulation has a negligible impact on the model's performance (CoT vs MCQA-CoT), and 2) our method (CCP) outperforms the MCQA-CoT baseline on multiple-choice tasks, demonstrating its effectiveness in handling this question type.

D Further Analysis

D.1 Number of In-context Learning Examples

Our approach inevitably introduces additional counterfactual examples during in-context learning (ICL), leading to a higher total number of shots compared to the baseline. To ensure a more competitive baseline, we increased the total number of shots in the baseline. In the MCTACO dataset and with the Llama model, we additionally experimented with the 12-shot CoT, which includes 12 passage (P)-question (Q)-candidate (C) pairs, and compared them with our 3-shot. We note that our 3-shot examples include 3 passage-question pairs

		TempevalQA-Bi			TRACIE			MCTACO		
		ACC	F1	INC (↓)	ACC	F1	INC (↓)	ACC	F1	INC (↓)
Llama -3-8B	SP	65.4	63	57.6	57.4	66.9	75.2	77.7	69.4	59.8
	CoT	69.6	70.6	50	63	64.9	56	77.6	69.8	63.4
	Consistency	70.8	71.2	49.6	64.9	67.3	57.8	77.5	69.0	61.1
	Reflection	63.6	63.9	44.6	62.5	55.7	55.5	77.4	69.7	76.4
	Debate	67.6	65.2	52.2	63.6	66	53.2	37.4	31.6	88.1
	CCP	**75.9**	**75.2**	**32.7**	**68.8**	**70.4**	**39.8**	**82.9**	**80.4**	**56.0**
Llama -3-70B	SP	76.6	78.6	39.7	79.9	79.7	29.6	85.2	81.8	43.5
	CoT	80.4	82	31.3	80.1	80	31.8	85.9	82.2	46.9
	Consistency	-	-	-	-	-	-	-	-	-
	Reflection	77	77.9	35.3	80	78.3	30.3	80.6	73	56.5
	Debate	81	82.8	32.6	81.6	80.7	25.9	85.3	81.4	45.9
	CCP	**87.3**	**87.9**	**19.2**	**86.5**	**86.1**	**12.0**	**89.1**	**87.3**	**36.3**
GPT-4o -mini	SP	78.8	76.4	36.6	74.6	71.3	38.2	76.0	63.1	65.8
	CoT	81.3	79.9	29	73.2	68.5	42.7	80.9	73.7	58.9
	Consistency	85.5	85.5	21.9	73.6	68.8	42.8	78.9	69.4	60.6
	Reflection	86.8	86.9	22.8	74.4	70.9	39.1	74.8	60.2	68.5
	Debate	86.4	86.4	24.6	73	67.1	44.5	78.3	68.2	61.0
	CCP	**88.8**	**88.7**	**19.6**	**82.5**	**81.2**	**20.2**	**88.7**	**86.8**	**41.1**
GPT-4o	SP	86.4	85.8	20.1	80.1	78.6	27.0	79.7	70.9	60.5
	CoT	90.4	90	17.4	80.2	78.1	32.4	84.4	80	49.7
	Consistency	91.7	91.5	14.7	80.1	77.7	31.4	82.9	77.3	49.7
	Reflection	93.1	93	11.2	82.7	80.9	26.6	80.0	72.2	55.4
	Debate	90.8	90.6	11.2	80.6	77.9	32.8	81.4	74.6	52.2
	CCP	**93.8**	**93.8**	**8.0**	**85.8**	**84.7**	**17.6**	**91.0**	**89.6**	**33.8**

Table 5: The full performance comparison results on the relative event understanding tasks. Our prompting methods, which leverage self-generated exemplars as the temporal constraint, outperform baselines across the board.

		TimeQA		TimexNLI	
		ACC	F1	ACC	F1
Llama 3-8B	3 shot	**34.3**	40.8	68.0	65.3
	CoT 3 shot	32.3	38.4	**74.0**	**73.3**
	CCP 3 shot	34	**41.5**	67.3	62.2
GPT-4o -mini	3 shot	40	52.36	86.4	85.3
	CoT 3 shot	**43.3**	**56.75**	**90.4**	**90.3**
	CCP 3 shot	41	53.59	90.3	90.0

Table 6: Performance comparison on TimeQA and TimexNLI.

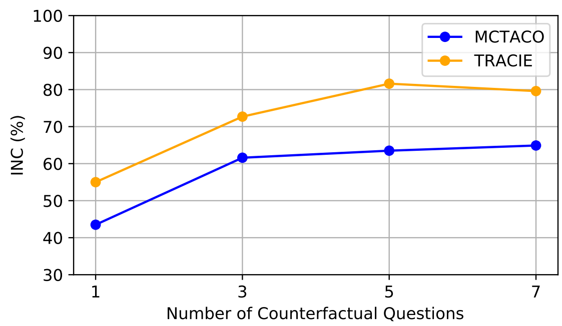

Figure 5: Inconsistency changes with the different number of counterfactual questions. The Llama-3-8B model is used.

and 11 candidates.

The results in Table 7 demonstrate that our method significantly outperforms the CoT, even with the increased number of examples in the baseline (INC score: 60.0 for CoT vs. 56.0 for Ours). This indicates that the performance gains are not simply due to the inclusion of more examples but are primarily driven by leveraging temporal constraints through counterfactual questions to enhance reasoning.

Additionally, we tested whether our approach benefits from additional ICL examples. The results in the last row of Table 7 confirm this, showing an improvement in INC score from 56.0 to 49.8, further validating the potential performance gain

	MCTACO				
	#P-Q	#C	ACC	F1	INC
CoT	3	3	77.6	69.8	63.4
CoT	12	12	78.9	72.2	60.0
CCP	3	11	82.9	80.4	56.0
CCP	12	26	**85.0**	**82.2**	**49.8**

Table 7: Performance comparison of Llama-3.1-8B on MCTACO with the different number of ICL examples.

of our method.

D.2 Number of Counterfactual Questions

We analyze the impact of varying the number of counterfactual questions on performance by testing with 1, 3, 5, and 7 questions. As shown in Figure 5, performance is highest with a single counterfactual question, with degradation observed as the number increases. The result aligns with findings from prior studies, where incorporating excessive counterfactual or contrastive information in prompts often results in diminished performance (Yao, 2024; Fang et al., 2024; Storaï and Hwang, 2024). For instance, multiple information degrade the performance in arithmetic and symbolic reasoning (Yao, 2024), multi-hop and commonsense question answering (Fang et al., 2024), and speculative decoding (Storaï and Hwang, 2024). These results indicate that such a phenomenon is not specific to our method but highlights a broader limitation in LLMs' ability to process multiple counterfactual information effectively.

E Details of Limitations

Table 6 shows the experimental results for temporal reasoning tasks where the understanding of temporal indicators are involved. We evaluated our method on TimeQA (Chen et al., 2021), the event-time ordering task, and TimexNLI-T1 (Thukral et al., 2021), the time-time ordering task.

F Usage of AI Assistants

ChatGPT was employed to generate answers in the prompt examples.

G Prompt Templates

We list the prompts that we used.

G.1 Prompt Templates for Generating Counterfactual Questions

To generate the counterfactual questions, we use the prompts provided below. We control question aspects by tailoring ICL examples and prompting the model to follow few-shot examples with the prompt: (“`following previous examples`”).

TempEvalQA-Bi

[System]
You are an expert in temporal knowledge. Generate one related question to help determine the correctness of the original question, following the previous examples.
[User]
Question: Is Farkas sent into space on board the Soyuz before McBride on board the Orbiter Challenger?
[Assistant]
Is Farkas sent into space on board the Soyuz after McBride on board the Orbiter Challenger?
[User]
Question: Is McBride on board the Orbiter Challenger after Farkas was made a brigadier general?
[Assistant]
Is McBride on board the Orbiter Challenger before Farkas was made a brigadier general?
[User]
Question: Is Farkas was appointed military attache at the Hungarian embassy before he was made a brigadier?
[Assistant]
Is Farkas was appointed military attache at the Hungarian embassy after he was made a brigadier?

TRACIE

[System]
You are an expert in temporal knowledge. Generate one related hypothesis to help determine the plausibility of the original hypothesis, following previous examples.
[User]

Hypothesis: Chad looked for his baseball cap starts after he got off the ride.
[Assistant]
Chad looked for his baseball cap starts before he got off the ride.
[User]
Hypothesis: Tom's friend says he is not gay starts before One day he noticed someone else asking
[Assistant]
Tom's friend says he is not gay starts after One day he noticed someone else asking
[User]
Hypothesis: Liza read a book on economics in the library starts before she got to the library
[Assistant]
Liza read a book on economics in the library starts after she got to the library

MCTACO

[System]
You are an expert in temporal commonsense knowledge. Generate related candidate answers that are either plausible or implausible that help determine the correctness of the original candidate answer following the previous examples.
[User]
Passage: The majority religion during the centuries of Ottoman rule, though a significant Christian minority remained. Question: What happened before Islam was the majority religion? Candidate answer: christianity was the majority religion.
[Assistant]
The Ottoman Empire had just begun / The spread of Byzantine influence in the region / The emergence of Islam as a minor religion
[User]
Passage: It's hail crackled across the comm, and Tara spun to retake her seat at the helm. Question: How long was the storm? Candidate answer: 6 years.
[Assistant]
an hour / a week / a month
[User]
Passage: His counter-attack with Dayak warriors drove the Chinese out of Bau and across the Sarawak border. Question: What time did the battle end? Candidate answer: 7:00 PM.
[Assistant]
a minute before it started / midnight / 5:00 AM

G.2 Prompt Templates for CCP and Baselines

We only show the SP, CoT, and CCP since Consistency is the method that leverages CoT multiple times, and the Reflect and Debate follow the implementation of Du et al. (2024) after CoT. Though we use 3-shot examples, we provide 1 example for simplicity.

Prompt templates used to generate answers for the TempevalQA-Bi dataset are provided below.

TempevalQA-Bi: SP

[System]
You are an expert in temporal knowledge. Answer the question according to the article. Answer the question yes or no (Format: final answer: yes/no)
[User]
Article: Farkas, an air force captain, was sent into space on board the Soyuz 36 on May 26, 1980. He spent six days aboard the Salyut 6 spacecraft with three Soviet astronauts, Valery Kubasov, Leonid Popov and Valery Riumin. McBride, 54, of Lewisburg, West Virginia, was part of a seven-member crew aboard the Orbiter Challenger [...]
Question: Is Farkas sent into space on board the Soyuz before McBride on board the Orbiter Challenger?
[Assistant]
Final answer: yes

TempevalQA-Bi: CoT

[System]
You are an expert in temporal knowledge. Answer the question according to the article. Let's think step by step to answer the question. The answer should be "yes" or "no" followed by "Final answer: ".
[User]
Article: [...]
Question: Is Farkas sent into space on board the Soyuz before McBride on board the Or-

biter Challenger?
[Assistant]
Let's think step by step. Farkas was sent into space on May 26, 1980 and McBride was on board the Orbiter Challenger in October 1984. May 26, 1980 is before October 1984. So the answer is: yes.
Final answer: yes

TempevalQA-Bi: CCP

[System]
You are an expert in temporal knowledge. Answer the question according to the article. To improve understanding, follow these steps: Answer the related question to better understand the timeline (Format: answer: [yes/no]) Using the answers to the related questions as the additional information, answer the original question yes or no (Format: final answer: [yes/no]).
[User]
Article: [...]
Original question: Is Farkas sent into space on board the Soyuz before McBride on board the Orbiter Challenger?
Related question: Is Farkas sent into space on board the Soyuz after McBride on board the Orbiter Challenger?
Answer the related question
[Assistant]
Farkas was sent into space on board the Soyuz 36 on May 26, 1980. McBride was part of the crew on board the Orbiter Challenger in October 1984. Since 1980 is before 1984, no, Farkas was not sent into space after McBride.
answer: no
[User]
Now, answer the original question. Question: Is Farkas sent into space on board the Soyuz before McBride on board the Orbiter Challenger?
[Assistant]
Farkas was sent into space on May 26, 1980 and McBride was on board the Orbiter Challenger in October 1984. May 26, 1980 is before October 1984. So the answer is: yes.
Final answer: yes

Prompt templates used to generate answers for the TRACIE dataset are provided below.

TRACIE: SP

[System]
You are an expert in temporal knowledge. You can understand the implicit temporal relationships between events to make judgments. Your task is to determine the relationship between the story and hypothesis. A 'positive' label means the hypothesis can be inferred from the story or logically sound. A 'negative' label means the hypothesis cannot be inferred from the story. You must answer only with the labels 'positive' or 'negative'. (Format: final answer: positive/negative)
[User]
Story: Chad looked for his baseball cap starts after he got off the ride story: Chad had gone to an amusement park. He was riding on the roller coaster.
Hypothesis: Chad looked for his baseball cap starts after he got off the ride.
[Assistant]
Final answer: positive

TRACIE: CoT

[System]
You are an expert in temporal knowledge. You can understand the implicit temporal relationships between events to make judgments. Your task is to determine the relationship between the story and hypothesis. A 'positive' label means the hypothesis can be inferred from the story or logically sound. A 'negative' label means the hypothesis cannot be inferred from the story. You must give final answer only with the labels 'positive' or 'negative' after the "Final answer:" prompt. Let's think step by step to evaluate the relationship.
[User]
Story: [...]
Hypothesis: Chad looked for his baseball cap starts after he got off the ride.
[Assistant]
Let's think step by step:
The story states that Chad's baseball cap fell off during the roller coaster ride.
It also mentions that Chad found the cap

after he got off the ride. To find something, one typically needs to look for it first. Since Chad found the cap after getting off the ride, he must have started looking for it after getting off the ride. The hypothesis directly aligns with this sequence of events.
Final answer: positive

TRACIE: CCP

[System]
You are an expert in temporal knowledge. You can understand the implicit temporal relationships between events to make judgments. Your task is to determine the relationship between the story and hypothesis. A 'positive' label means the hypothesis can be inferred from the story or logically sound. A 'negative' label means the hypothesis cannot be inferred from the story. You must give final answer only with the labels 'positive' or 'negative'. To improve understanding, follow these steps: Evaluate the related hypothesis to help clarify the story's timeline (Format: answer: [positive/negative]). Using answers to the related hypothesis as the additional information, evaluate the original hypothesis (Format: final answer: [positive/negative]).
[User]
Story: [...]
Hypothesis: Chad looked for his baseball cap starts after he got off the ride.
Related hypothesis: Chad looked for his baseball cap starts before he got off the ride.
Answer the related hypothesis
[Assistant]
The hypothesis conflicts with the story, as Chad finds the cap after the ride.
Answer: negative
[User]
Now, evaluate the original hypothesis: Chad looked for his baseball cap starts after he got off the ride.
[Assistant]
Chad rides the roller coaster while wearing a baseball cap, which falls off during the ride. After getting off the ride, Chad finds the cap. The hypothesis suggests that Chad started looking for the cap after getting off the ride, which aligns with the story's sequence of events.
Final answer: positive

Prompt templates used to generate answers for the MCTACO dataset are provided below.

MCTACO: SP

[System]
You are expert in temporal commonsense knowledge. Your goal is to determine if the candidate answer to a given question is plausible based on the information in the passage. Answer the related candidate answers then answer the candidate answer. Provide a 'yes' or 'no' response to indicate whether the original candidate answer is correct. (Format: final answer: yes/no) If the passage does not provide explicit details, provide an educated guess based on what is common or plausible in similar situations
[User]
Passage: the majority religion during the centuries of Ottoman rule, though a significant Christian minority remained.
Question: What happened before Islam was the majority religion?
Candidate answer: the end of white-minority rule.
[Assistant]
Final answer: no

MCTACO: CoT

[System]
You are an expert in temporal commonsense knowledge. Your goal is to determine if the candidate answer to a given question is plausible based on the information in the passage. Let's think step by step and provide a final 'yes' or 'no' response to indicate whether the original candidate answer is correct. If the passage does not provide explicit details, provide an educated guess based on what is common or plausible in similar situations. The final answer should be in the format: "Final answer: yes" or "Final answer: no".
[User]
Passage: the majority religion during the centuries of Ottoman rule, though a significi-

cant Christian minority remained.
Question: What happened before Islam was the majority religion?
Candidate answer: christianity was the majority religion
[Assistant]
Let's think step by step.
Step 1: Analyze the passage. The passage states that during the centuries of Ottoman rule, Islam was the majority religion, while a significant Christian minority remained. This implies that before the Ottoman Empire (which was Islamic), there could have been a different majority religion.
Step 2: Historical context. Before the rise of the Ottoman Empire, large parts of the Middle East, North Africa, and parts of Europe were under the rule of the Byzantine Empire, which was a Christian empire. Additionally, many regions that later came under Ottoman rule were predominantly Christian.
Step 3: Plausibility of the candidate answer. Given that Christianity was the dominant religion in many regions before the Ottoman Empire and its Islamic rule, it is plausible that Christianity was the majority religion before Islam in those areas.
Final answer: yes

MCTACO: CCP

[System]
You are an expert in temporal commonsense knowledge. Your goal is to determine if the candidate answer to a given question is plausible based on the information in the passage. To improve understanding, follow these steps: generate related candidate answers that are either plausible or implausible (Format: answer: yes/no) to help in determining the correctness of the original candidate answer. Provide a 'yes' or 'no' response to indicate whether the original candidate answer is correct (Format: final answer: yes/no). If the passage does not provide explicit details, provide an educated guess based on what is common or plausible in similar situations
[User]
Passage: the majority religion during the centuries of Ottoman rule, though a significant Christian minority remained.
Question: What happened before Islam was the majority religion?
original candidate answer: christianity was the majority religion
Related candidate: the end of white-minority rule
Can the candidate answer the given question? yes or no
[Assistant]
This is incorrect because the passage discusses events during Ottoman rule, not before it.
answer: no
[User]
Now, can the candidate answer the given question? candidate answer: christianity was the majority religion
[Assistant]
This is plausible because Islam would have started to emerge before becoming the majority.
Final answer: yes

H Scientific Artifacts

We used existing scientific artifacts for research purposes, and the use of existing artifacts was consistent with their intended applications.

TempEvalQA-Bi: MIT license
TRACIE: Apache-2.0 license
Llama-3: custom commercial license
OpenAI API: Apache-2.0 license