Are Large Language Models Temporally Grounded?

[1]Yifu Qiu [1]Zheng Zhao [1]Yftah Ziser
[2]Anna Korhonen [1]Edoardo M. Ponti [1]Shay B. Cohen
[1]Institute for Language, Cognition and Computation, University of Edinburgh
[2]Language Technology Lab, University of Cambridge
{yifu.qiu,zheng.zhao,yftah.ziser,eponti,scohen}@ed.ac.uk

Abstract

Are Large Language Models (LLMs) temporally grounded? Since LLMs cannot perceive and interact with the environment, it is impossible to answer this question directly. Instead, we provide LLMs with textual narratives and probe them with respect to their common-sense knowledge of the structure and duration of events, their ability to order events along a timeline, and self-consistency within their temporal model (e.g., temporal relations such as *after* and *before* are mutually exclusive for any pair of events). We evaluate state-of-the-art LLMs (such as LLaMA 2 and GPT-4) on three tasks reflecting these abilities. Generally, we find that LLMs lag significantly behind both human performance as well as small-scale, specialised LMs. In-context learning, instruction tuning, and chain-of-thought prompting reduce this gap only to a limited degree. Crucially, LLMs struggle the most with self-consistency, displaying incoherent behaviour in at least 27.23% of their predictions. Contrary to expectations, we also find that scaling the model size does not guarantee positive gains in performance. To explain these results, we study the sources from which LLMs may gather temporal information: we find that sentence ordering in unlabelled texts, available during pre-training, is only weakly correlated with event ordering. Moreover, public instruction tuning mixtures contain few temporal tasks. Hence, we conclude that current LLMs lack a consistent temporal model of textual narratives.[1]

1 Introduction

Recent large language models (LLMs) like GPT-4 have shown impressive progress on various downstream tasks in natural language processing (Brown et al., 2020; Chung et al., 2022; Workshop et al., 2023; Chowdhery et al., 2022; OpenAI, 2023). However, a notorious deficiency of these models is the *lack of grounding* (Yun et al., 2021; Abdou et al., 2021; Carta et al., 2023; Mahowald et al., 2023). LLMs, which are solely trained with textual data, acquire knowledge from distributional patterns of language. Humans, on the other hand, perceive and interact with the world through the dimensions of time and space. Accurately grounding intelligent systems within the physical world is crucial for developing a versatile form of intelligence that is truly embodied (Dasgupta et al., 2022; Driess et al., 2023; Carta et al., 2023), as well as capable of sample-efficient learning and causal discovery through mental simulations (LeCun, 2022).

In this work, we focus on temporal reasoning as a case study for establishing to what extent LLMs are grounded in the absence of perception and action. We first present a framework (Section 2) for evaluating temporal grounding, where we decompose this property into three fundamental aspects: 1) commonsense knowledge about the typical duration, occurrence, frequency, and structure of events, 2) ability to order events along a timeline and 3) consistency based on temporal constraints. Then, we carefully choose and curate three benchmarks (Section 3.1) for capturing each of these abilities: McTACO (Zhou et al., 2019), CaTeRS (Mostafazadeh et al., 2016) and TempEvalQA (Llorens et al., 2015), respectively.

We conduct an extensive empirical evaluation of the temporal reasoning capabilities of state-of-the-art LLMs, including both open-source (e.g., LLaMA 2) and proprietary (e.g., GPT-4). In addition, we investigate the effect of recent techniques that facilitate generalisation to new tasks with ablations: we compare zero-shot inference, few-shot prompting using in-context learning, scaling in terms of model parameters and the number of few-shot examples, and chain-of-thought prompting.

Firstly, we find that all the tasks requiring temporal grounding remain extremely challenging for

[1]Code, datasets, and all LLMs' outputs are available at https://github.com/yfqiu-nlp/temporal-llms.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 7061–7080
June 16-21, 2024 ©2024 Association for Computational Linguistics

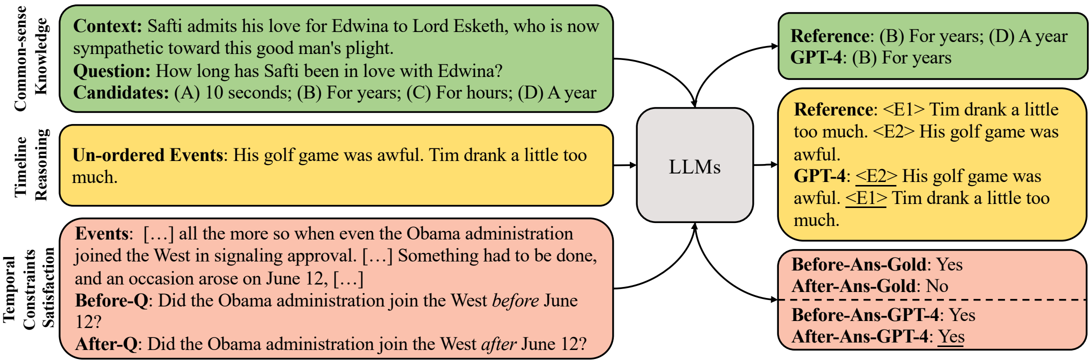


Figure 1: Examples from three datasets to evaluate the temporal grounding of LLMs based on common-sense knowledge about events, timeline reasoning, and temporal constraints satisfaction. We include the predictions from the best-performing model, GPT-4, with in-context learning. We highlight wrong predictions with underline.

LLMs (Section 4), even for the recently released state-of-the-art GPT-4. In particular, LLMs lag significantly behind human performance and even behind specialised, small-scale, fine-tuned language models in temporal commonsense knowledge and event ordering. Most strikingly, however, the results reported on our TempEvalQA-Bi benchmark (Section 4.3) reveal that the predictions of LLMs are not consistent with respect to mutually exclusive temporal relations. For example, LLaMA-65B and `text-davinci-003` display a percentage of inconsistent predictions at 78.13% and 69.64%, respectively. Secondly, we observe that recent techniques introduced for boosting LLMs' performance, including few-shot in-context learning (Section 4.5) and chain of thought (Section 4.6), have a positive but limited effect. Thirdly, scaling the examples for in-context learning and model size does not necessarily improve the model performance on temporal tasks (Section 4.5).

Finally, to understand the reason for the underwhelming performance of LLMs in temporal tasks, we attempt to determine whether enough temporal information is provided to them during their training pipeline (Section 5). We find that in the unlabelled texts available during pre-training, sentence ordering and event ordering are weakly correlated. Hence, LLMs assign only a slightly higher probability to sentences describing a ground-truth event series in temporal order, rather than an inverted order. This is especially the case whenever explicit temporal markers (e.g., *before* and *after*) are missing. We also find that most training mixtures for instruction tuning lack temporal tasks. Thus, temporal grounding cannot emerge from the current LLM training paradigm.

2 Motivation and Framework

While the question of whether language models contain a spatial map of the world has been thoroughly investigated (Louwerse and Zwaan, 2009; Mikolov et al., 2013; Rahimi et al., 2017; Faisal and Anastasopoulos, 2022, *inter alia*), *temporal* grounding of language models remains an under-explored area. Recently, Gurnee and Tegmark (2023) found that a linear probe can recover the year of existence of an entity from its representation. From this evidence, they conclude that LLMs can learn (temporal) world models; however, this seems unwarranted given that 'world model' defines the ability of grounded agents to internally simulate physical dynamics (Ha and Schmidhuber, 2018). On the other hand, since LLMs are incapable of action and perception, this ability cannot be assessed directly.

Instead, to study whether LLMs are temporally grounded, we assess their ability to reason about textual narratives, which contain accounts of series of events. Events are usually denoted by verbal (e.g. *to ride*) or nominal expressions (e.g. *a discussion*). The temporal structure of an event is partly inherent to such expressions, i.e. lexical aspect (or *Aktionsart*; Vendler, 1967), and partly determined by their textual context. The events mentioned in a text are connected by (asymmetrical) temporal relations (Pustejovsky et al., 2003), which can be expressed explicitly (e.g., through adverbial markers such as *after* or *before*) or implicitly.

In particular, we focus on 3 aspects of temporal reasoning illustrated with examples from Figure 1:

- **Commonsense Knowledge about Events**: A grounded model should exhibit commonsense knowledge with respect to the typical duration, occurrence, frequency, and structure of events.

For instance, when presented with the question "*How long has Safti been in love with Edwina?*", the model should recognise that "*for years*" is a more probable answer than "*10 seconds*", due to the typical duration of this emotional state.

- **Ordering Events along a Timeline**: Such a model should also be able to infer the chronology of events from narratives (e.g., news) where they are not presented in temporal order. For example, the model should deduce that "*Tim drank a little too much.*" occurs before that "*His golf game was awful.*", due to the causal relation between these two events.
- **Satisfying Temporal Constraints**: A grounded model should maintain *self-consistency* in its predictions. One important constraint is that contradicting timelines cannot co-exist. If, with respect to a certain context, the model classifies "*The Obama administration joined the West before June 12*" as true, then it must classify the opposite statement "*The Obama administration joined the West after June 12*" as false.

3 Temporal Grounding Evaluation

We probe the temporal grounding of LLMs from the GPT and LLaMA families, which collectively represent a wide range of LLM types. The models we tested can be grouped into two categories: 1) foundation models (e.g., LLaMA 1 and 2; Touvron et al. 2023a,b) pre-trained on unlabelled texts, and 2) LLMs fine-tuned on instruction-following (e.g., Alpaca and `text-davinci-002`) or conversational objectives (e.g., GPT-4; OpenAI 2023 and LLaMA-2-chat; Touvron et al. 2023b). We use the OpenAI official API and the `transformers` library (Wolf et al., 2020) to conduct all experiments.

3.1 Benchmarks for Temporal Grounding

For the evaluation, we choose two off-the-shelf benchmarks, McTACO (Zhou et al., 2019) and CaTeRS (Mostafazadeh et al., 2016), and curate a third benchmark starting from TempEvalQA (Llorens et al., 2015) to measure the three aspects of temporal grounding identified in Section 2. Table 1 showcases an example from each benchmark.

McTACO is a benchmark for evaluating temporal commonsense knowledge through multiple choice question answering. McTACO consists of 13K tuples in the form of (context, question, candidate answers). McTACO examples fall within one of five categories: *duration* (how long an event lasts), *temporal ordering* (typical order of events), *typical time* (when an event happens customarily), *frequency* (how often an event occurs), and *stationarity* (whether a state holds for a very long time or indefinitely).

CaTeRS is a benchmark for event ordering, consisting of 1684 instances. This task involves identifying events mentioned in a text and arranging them in chronological order. To solve this task, a model must rely on explicit clues as well as commonsense knowledge about relations between events, to reason about the underlying timeline.

TempEvalQA-Bi, derived from TempEvalQA (Llorens et al., 2015), is our curated dataset for assessing a model's self-consistency in temporal reasoning. We opt for TempEvalQA due to its evaluation simplicity, facilitated by annotations for temporal modifiers in questions enabling easy construction of negative question-answer triples, and high quality, with all triples being human-annotated. Secondly, TempEvalQA boasts high quality, with all triples being human-annotated. TempEvalQA comprises 294 question–answer triples in the format of (context, question, yes/no answer), focusing on temporal relations between pairs of events. To construct TempEvalQA-Bi, we selectively choose question–answer pairs that involve the *before*/*after* temporal relation, because these pairs allow us to easily construct the samples evaluating model's consistency: we can swap the temporal relation to its opposite to create negative question–answer pairs. For instance, if the original pair is ("*Is* E1 *after* E2?", "*yes*"), the corresponding negative pair will be ("*Is* E1 *before* E2?", "*no*"), and vice versa. After this post-processing, TempEvalQA-Bi consists of 224 test examples with 448 question–answer pairs. We assess the model's performance by evaluating its predictions on both the original and negative pairs, considering the model's answer as correct only if it accurately predicts both pairs.

3.2 Inference of LLMs

We now provide details of performing inference with LLMs on these three benchmarks.

Multiple-choice Question Answering is used for benchmarking LLMs' performance on McTACO. For GPT models, which do not always provide access to output probabilities, we simply provide the context, question, and *all* the candidate answers in a single prompt and generate one or more answers. For LLaMA models, a more robust setup consists in presenting the context, question, and *each* of

Models	McTACO Zero-shot Strict Acc.	McTACO Zero-shot F1	McTACO Few-shot Strict Acc.	McTACO Few-shot F1	CaTeRS Pair Acc.
RoBERTa	43.62	72.34			
TemporalBART					**77.06**
Human	**75.80**				
GPT-4	**28.45**	35.88	**50.15**	**65.27**	**60.51**
text-davinci-003	26.05	**48.30**	33.56	65.04	53.47
LLaMA-7B	14.39$_{2.82}$	35.30$_{15.18}$	20.17$_{2.46}$	22.39$_{5.07}$	3.76$_{4.58}$
Alpaca-7B	**21.75**$_{5.22}$	**52.17**$_{9.69}$	**30.03**$_{10.11}$	44.10$_{18.36}$	10.37$_{4.91}$
LLaMA-13B	15.67$_{3.42}$	36.59$_{14.69}$	24.37$_{6.08}$	34.99$_{19.01}$	5.27$_{5.51}$
LLaMA-33B	17.24$_{3.36}$	33.20$_{15.07}$	29.70$_{4.79}$	47.57$_{8.36}$	14.38$_{10.77}$
LLaMA-65B	18.14$_{5.63}$	46.83$_{6.51}$	26.13$_{12.15}$	**47.84**$_{2.65}$	**21.02**$_{10.27}$
LLaMA-2-7B	11.16$_{1.55}$	**42.55**$_{12.29}$	21.74$_{3.83}$	32.94$_{17.56}$	5.85$_{2.06}$
LLaMA-2-13B	15.69$_{3.49}$	39.35$_{15.55}$	**29.75**$_{0.69}$	**43.21**$_{2.51}$	16.26$_{5.75}$
LLaMA-2-70B	19.12$_{3.58}$	33.51$_{9.75}$	27.77$_{2.35}$	37.20$_{3.71}$	21.61$_{8.39}$
LLaMA-2-chat-7B	20.74$_{3.45}$	28.73$_{4.48}$	23.00$_{3.56}$	31.50$_{10.18}$	26.32$_{2.09}$
LLaMA-2-chat-13B	**22.22**$_{0.13}$	31.67$_{9.38}$	28.90$_{1.04}$	41.63$_{5.97}$	30.27$_{3.02}$
LLaMA-2-chat-70B	20.84$_{2.08}$	26.42$_{5.98}$	27.18$_{4.9}$	34.37$_{7.75}$	**30.55**$_{21.87}$

Table 1: Average model performance (standard deviations as subscripts). Left: McTACO for evaluating temporal commonsense reasoning in LLMs. **Strict Acc.** and **F1** follow the definitions in Zhou et al. (2019). Right: CaTeRS results for few-shot prompting. **Pair Acc.** stands for pairwise accuracy. The results for three prompt templates are shown in Table 5 for McTACO and in Table 6 for CaTeRS.

candidate answers separately in multiple prompts. We calculate the likelihood of decoding "*True*" and "*False*" and assign the label with the higher value to each candidate answer.

Sequence-to-Sequence Generation. We follow Lin et al. (2021) to formulate the evaluation for CaTeRS as a sequence-to-sequence problem. The model takes the event series as the input and temporally sorts it as the output. We only test prompting with in-context learning for CaTeRS and not zero-shot inference because sorting the unordered events requires the model to also generate and sort special event tokens, which are unseen during pretraining.

Yes/No Question Answering, in the TempEvalQA-Bi benchmark, requires the model to predict either Yes or No as the answer for each question. To probe the models' predictions, we use 1) greedy decoding for GPT models, where we generate the token with the maximum likelihood to complete a given prompt. We then map all possible verbal tokens (e.g., yes/no, true/false and 1/0) into binary form. 2) Likelihood-based evaluation for LLaMA models: similar to Ruis et al. (2022) an Liu et al. (2023a), we append "*yes*" and "*no*" as the answer in our prompt template. We then estimate the likelihood of both with LLMs. The label with the higher likelihood value is taken as the model's prediction.

4 Results and Discussion

In this section, we describe our main results on three datasets: McTACO, CaTeRS, TempEvalQA-Bi. We also analyse LLMs' sensitivity to prompting in temporal reasoning, scaling behaviour, and the impact of chain-of-thought prompting.

4.1 Temporal Commonsense: McTACO

Evaluation Settings. First, we probe temporal commonsense knowledge. Table 1 shows each model's strict accuracy and F1 score on McTACO. Following Zhou et al. (2019), strict accuracy is the percentage of questions where the model predicts *all* the correct candidate answers, whereas the F1 score is calculated from all question–answer pairs. Figure 2 shows the strict accuracy split by question category.

Discussion. In Table 1, GPT-4 outperforms all tested LLMs, confirming its state-of-the-art ability in temporal commonsense knowledge among LLMs. Within the LLaMA family, LLaMA-2-chat-13B and Alpaca-7B achieve the best performance in zero-shot and few-shot experiments, respectively. Alpaca-7B is comparable to LLaMA-2-chat-13B also in the zero-shot setting. We attribute its remarkable performance to its in-domain temporal tasks during instruction tuning (for more discus-

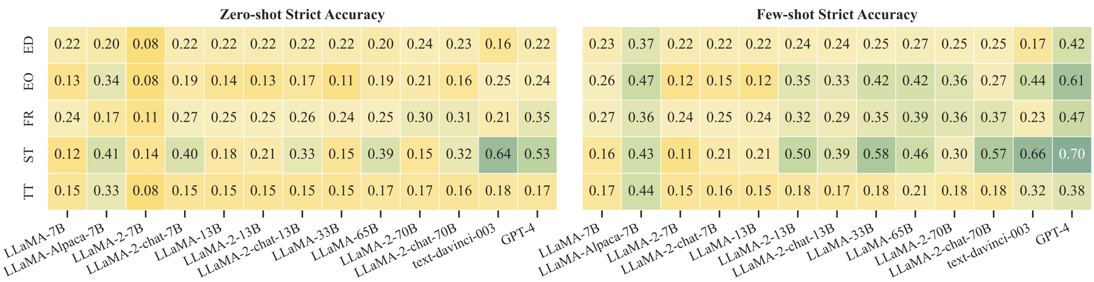


Figure 2: Strict accuracy of all tested LLMs for different reasoning categories in McTACO. The rows represent all reasoning categories in McTACO, which are Event Duration (ED), Event Ordering (EO), Frequency (FR), Stationary (ST) and Typical Time (TT). The columns represent several LLMs.

sion, see Section 5.2). We also notice that all LLaMA-2-chat models considerably improve their pre-trained counterparts, indicating the importance of advanced tuning (e.g., conversational-style tuning and reinforcement learning with human feedback, RLHF; Touvron et al. 2023b) in addition to pre-training. However, even when prompting with in-context learning, the top-performing GPT-4 lags behind the fine-tuned RoBERTa baseline in terms of F1 score and human performance in terms of strict accuracy. This finding highlights the shortcomings of state-of-the-art LLMs in learning temporal commonsense knowledge. Inspecting the performance by question category, plotted as a heatmap in Figure 2, reveals that questions requiring commonsense knowledge of Typical Time and Event Duration are the most challenging. Here, GPT-4 achieves a success rate of only around 40%. On the other hand, questions related to Stationarity are relatively easier, with most LLMs showing adequate performance using in-context learning.

4.2 Event Ordering: CaTeRS

We evaluate the ability of LLMs to order events on CaTeRS: results are displayed in Table 1.

Evaluation Settings. We follow Lin et al. (2021) in using pairwise accuracy as a metric. This measures how many pairs of events in the output are ordered correctly by a model. This metric validates each model in two respects: 1) the model is asked to extract the events from the given text correctly. And 2) the model must order the given set of events according to their chronology, producing a semi-structured output. We also include TemporalBART from Lin et al. (2021) as a baseline. TemporalBART is a BART-Large model fine-tuned on re-constructing manually corrupted event series.

Discussion. First, we confirm again the positive gains from conversational tuning and RLHF by observing the large advantage of LLaMA-2-chat models over LLaMA-2 models. However, we do not observe improvements by increasing the number of in-context learning examples (see Figure 3b) or model parameters (see Figure 3a). The proprietary models, `text-davinci-003` and GPT-4, achieve higher pairwise accuracy than all LLaMA models; however, they lag behind a small fine-tuned model, TemporalBART, by a large margin. Our comparative analysis in Appendix C also reveals the decline in model performance as the complexity of reasoning tasks increases, in line with recent findings in (Dziri et al., 2023). These outcomes indicate the challenging nature of CaTeRS for current LLMs.

4.3 Satisfaction of Temporal Constraints: TempEvalQA-Bi

In Table 2, we report the performance of LLMs on our curated bi-directional TempEvalQA-Bi dataset.

Evaluation Metrics. We evaluate the models in terms of their accuracy (**Acc.**), where the model is required to correctly predict a yes/no answer for a question with a *before*/*after* relation (see Table 1 for an example). We also include a metric, namely the percentage of inconsistent predictions (**Inc.**), to shed further light on the model behaviour. This measures the number of times a model predicts the *same* yes/no label for both the original question and a version where the temporal relation is flipped (everything else remaining the same). For example, if models predict *yes* for a question with a *before* relation, ideally they should predict *no* for the corresponding question with *after*. We can compare the two metrics, **Acc.** and **Inc.**, to probe the model's ability to answer correctly and consistently, respectively.

Discussion We first observe that most models perform poorly. The best model in the LLaMA family (LLaMA-2-chat-70B with in-context learning) can

Models	Zero-shot		Few-shot	
	Acc. (↑)	**Inc.** (↓)	**Acc.** (↑)	**Inc.** (↓)
GPT-4	**64.29**	**31.25**	**67.41**	**27.23**
text-davinci-003	27.68	69.64	33.93	62.05
text-davinci-002	16.52	77.83	36.16	60.71
davinci	14.73	79.02	13.39	79.91
LLaMA-7B	$3.42_{2.46}$	$94.79_{3.64}$	$3.27_{0.51}$	$94.94_{1.57}$
LLaMA-Alpaca-7B	$10.12_{2.29}$	$83.63_{5.08}$	$13.10_{6.00}$	$77.23_{7.37}$
LLaMA-13B	$0.60_{0.68}$	$97.77_{3.49}$	$0.60_{0.68}$	$99.25_{0.51}$
LLaMA-33B	$1.34_{1.34}$	$98.22_{1.18}$	$14.73_{7.74}$	$83.33_{9.43}$
LLaMA-65B	$\mathbf{14.14}_{5.17}$	$\mathbf{83.48}_{6.14}$	$\mathbf{31.99}_{1.57}$	$\mathbf{60.42}_{4.47}$
LLaMA-2-7B	$0.15_{0.26}$	$99.85_{0.26}$	$11.90_{0.52}$	$85.12_{2.62}$
LLaMA-2-13B	$5.65_{3.3}$	$92.86_{3.81}$	$13.69_{7.63}$	$83.63_{8.00}$
LLaMA-2-70B	$6.55_{2.01}$	$92.41_{3.13}$	$29.76_{2.73}$	$65.77_{2.02}$
LLaMA-2-chat-7B	$13.84_{7.63}$	$83.33_{7.82}$	$23.51_{2.20}$	$70.09_{0.77}$
LLaMA-2-chat-13B	$22.92_{4.03}$	$72.91_{5.58}$	$31.69_{3.22}$	$62.95_{3.57}$
LLaMA-2-chat-70B	$\mathbf{38.54}_{3.04}$	$\mathbf{58.03}_{2.36}$	$\mathbf{46.42}_{1.18}$	$\mathbf{48.96}_{2.01}$

Table 2: Average model performance (standard deviations as subscripts) evaluated on our curated bi-directional TempEvalQA benchmark. **Acc.** and **Inc.** stand for accuracy and the percentage of inconsistent predictions. (↑)/(↓) indicate that higher / lower values are better, respectively. The detailed results for three prompt templates are presented in Table 4.

only solve 46% of questions correctly. Moreover, if we flip the temporal relation of the question, most models fail to convert their output predictions accordingly, resulting in significant percentages of inconsistent predictions. This observation suggests that inconsistent predictions are the main cause of low performance on this task.

On the other hand, solid improvements result from advanced tuning, including instruction/conversation tuning and RLHF, enabling the model's predictions to be more consistent. The Alpaca-7B model dramatically reduces the rate of inconsistencies and improves accuracy over LLaMA-7B. The same holds true for LLaMA-2 and LLaMA-2-chat, `text-davinci` and `text-davinci-002/002`. In-context learning can also be beneficial, but it yields mixed results for some models, such as LLaMA-7B and 13B.

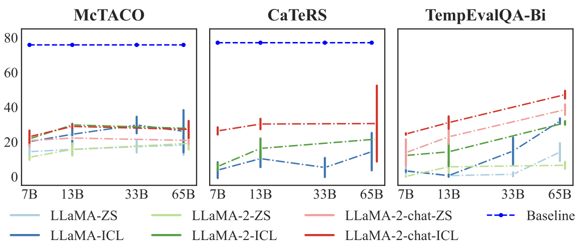
Finally, GPT-4 again achieves state-of-the-art accuracy by solving around 67% of our curated questions. Although it significantly reduces the percentage of inconsistent predictions compared to other models, it still predicts the same answer for around 27% of the questions and their flipped versions.

4.4 Sensitivity to Different Prompts

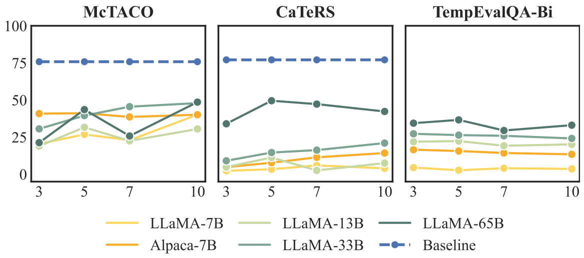
Large language models have been shown to be sensitive to the wording of their prompts (Webson and Pavlick, 2022; Ruis et al., 2022). Similar to Ruis et al. (2022), to test the randomness introduced by this sensitivity in our experiments, we manually curate three different prompt templates and measure the variation of performance across these different wordings for the LLaMA family.

We experiment with three different kinds of prompt templates: 1) *semi-structured prompt*: we indicate the instruction, context, and pairs of input and output (for in-context learning) in a sequence, separated by special symbols (e.g., newline). 2) *Natural prompt*: we interleave the context and input–output pairs in the instruction, to make the template closer to natural language. 3) *Text continuation prompt*: given that LLaMA is not trained with instruction-following tasks, we introduce a text-continuation style prompting where we only ask the model to complete the input. We provide all our prompt templates in Table 3.

We report the standard deviation of model performance in the result tables for each of the three datasets. Standard deviations range from 0.13 for the LLaMA-2-chat-13B model to 5.63 for LLaMA-



(a) Scaling the model parameters.



(b) Scaling the examples of in-context learning.

Figure 3: The performance curve for scaling experiments. We report the strict accuracy for McTACO, pairwise accuracy for CaTeRS and accuracy for TempEvalQA-Bi. (a): The error bars show the standard deviation over three prompt templates. (b): The baseline for McTACO is Human, and for CaTeRS is TemporalBART.

65B when considering all templates in McTACO in zero-shot prompting. The variation caused by prompt templates is slightly higher for zero-shot results in TempEvalQA-Bi, where the standard deviation ranges from 0.26 for LLaMA-2-7B to 7.63 for LLaMA-2-chat-7B. The few-shot experiments tend to be more varied, because of the increased length of each prompt due to adding additional exemplars. Overall, these findings show that the prompt wording does not affect the main conclusions of our experiments. However, they also confirm that LLMs—especially large-scale ones—are brittle with respect to prompt wording.

4.5 Scaling Behaviour

Scaling the Model Parameters. We perform scaling experiments (refer to Figure 3a) to explore the influence of parameter increments on the performance of LLMs. We exclusively execute this ablation within the LLaMA family due to the unavailability of information on model sizes for the proprietary GPT models. The largest 70B LLaMA-2-chat model can generally achieve the best performance on our three datasets. However, for many models, performance is only weakly correlated with their size. For example, the LLaMA-13B model, rather than the smaller LLaMA-7B one, performs the worst on both McTACO and TempEvalQA-Bi. We also observe a general trend such that LLM performance tends to saturate and cannot further increase after the parameter size exceeds 13B on McTACO and CaTeRS. Finally, comparing the LLMs with the much smaller baselines of TemporalBART and RoBERTa shows that model specialisation is more impactful than an increase in scale for temporal grounding.

Scaling the Few-shot Examples. We begin by conducting scaling experiments (please refer to Figure 3b) to investigate the impact of the number of few-shot examples (k) on model performance. We focus our experiments on the LLaMA family. The results on McTACO and CaTeRS indicate that increasing k generally leads to a slight performance improvement, albeit minimal. Conversely, we observe a slight decrease in performance on TempEvalQA-Bi when increasing k. This decrease can be attributed to our choice of increasing only the number of question–answer pairs for the same context, as using multiple contexts would surpass the model's maximum input limit. Notably, McTACO and CaTeRS exhibit true scaling with few-shot examples, supporting the idea that examples must be diverse to contribute incrementally to performance (Min et al., 2022).

4.6 Chain-of-thought Prompting

Finally, we investigate the impact of chain-of-thought prompting (Wei et al., 2022) on model performance in TempEvalQA-Bi, a challenging reasoning task. We apply chain-of-thought prompting to LLaMA and GPT-4 models in combination with in-context learning. The prompts we constructed are listed in Appendix D. Figure 4 shows that chain-of-thought prompting reduces inconsistent predictions for all models; however, the improvement of reasoning consistency does not always translate into increased accuracy. For example, we observe a drop in accuracy for LLaMA-7B, LLaMA-2-70B and LLaMA-2-70B-chat models. We also notice that the improvement in the state-of-the-art GPT-4 is marginal. These results indicate that current training and inference paradigms may struggle with complex temporal reasoning tasks requiring self-consistency.

5 Sources of Temporal Information

In this section, we examine two key aspects: firstly, the extent of temporal information provided by

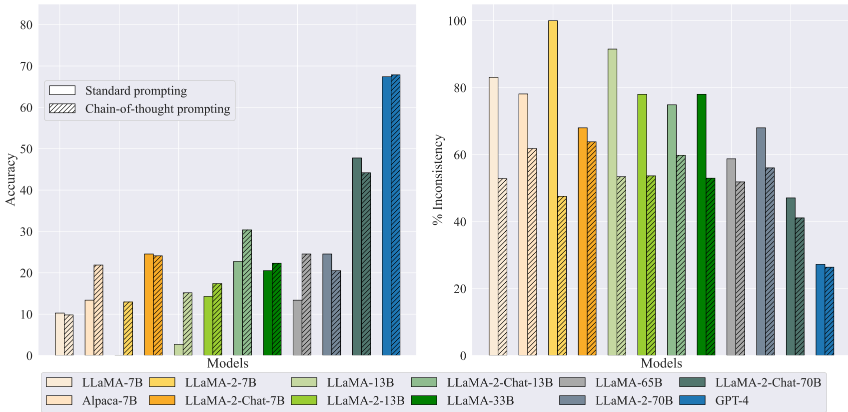


Figure 4: Comparison between chain-of-thought prompting (Wei et al., 2022) and the standard few-shot prompting on TempEvalQA-Bi for all tested LLMs.

pre-training, and secondly, the significance of supervised examples in addressing information gaps within the pre-trained model.

5.1 How Much Temporal Information Does Pre-training Provide?

Overall, the results from Section 4 indicate that LLMs are not sufficiently temporally grounded. We speculate that this is due to not being exposed to temporal information during their training pipeline. To verify this, we first focus on the pre-training stage. We investigate whether the order in which events are presented in human-written texts provides a clue about their actual chronology. We measure the correlation between the temporal relations (i.e., *before*/*after*) annotated in TempEvalQA's training articles and the textual order of the events they refer to. This reveals that only 55.98% of event pairs (1866 out of 3333) occur in the text in accordance with their temporal order. The Matthews correlation coefficient of 0.09 indicates a weak correlation (Matthews, 1975; Davenport Jr and El-Sanhurry, 1991).

We then study whether, as a result of this property of the data, LLMs prefer temporally ordered or unordered descriptions of the same event series, by comparing their log-likelihoods. We randomly sample 100 instances from the CaTeRS testing set and manually curate two paraphrases for each ground-truth sequence of events. For the first paraphrase, we keep the same sentence order as the event temporal order. For the second paraphrase, we permute the order of the sentences. All the resulting paraphrases are manually adjusted to be coherent and grammatically correct to exclude confounding factors. We present one example:

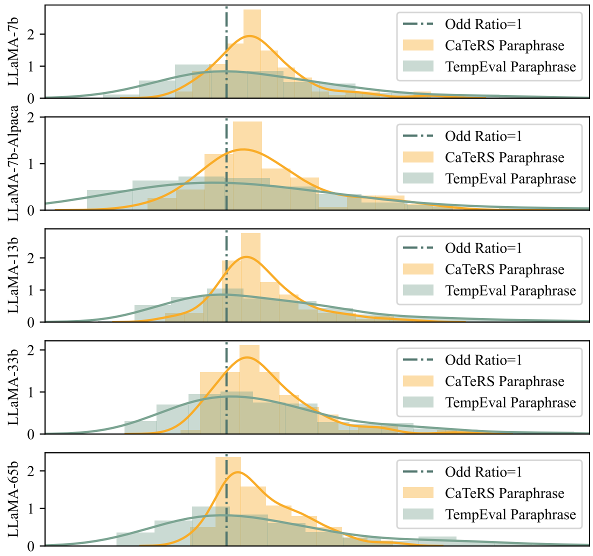


Figure 5: Density plot of the odds ratio under several LLMs (rows) for differently ordered paraphrases in CaTeRS (orange) and TempEvalQA-Bi (green).

Ground-truth Order: [E1] *Mary was very stressed about the job.* [E2] *Mary got the job.*
Ordered Paraphrase: *Mary was very stressed about the job, but she got it.*
Inverted Paraphrase: *Mary got the job. She was very stressed about it.*

We then create another subset of paraphrases from TempEvalQA-Bi, retaining explicit temporal relation markers such as *before* and *after*.

We plot the distribution of the length-normalised odds ratios between corresponding paraphrases in Figure 5. Firstly, we find that the odds ratio is higher than 1 on average for the CaTeRS paraphrases. This means that the probability of temporally ordered sequences is only slightly higher than their permuted counterparts. Thus, foundation models have a mild bias in preferring sequences of sentences that follow the temporal order of events. For the examples with explicit temporal relation markers (i.e., *before* and *after*) from TempEvalQA-Bi, the odds ratio falls around 1. Thus, the model's preference for temporally ordered and unordered sequences tends to be equalised. These results demonstrate that it may be difficult for foundation models to identify the ground-truth event ordering based on their pre-training information, as this provides a weak signal about the temporal dynamics of real-world events.

5.2 Importance of Supervised Examples

Next, we investigate whether subsequent fine-tuning might compensate for the lack of temporal information during pre-training. The current open-sourced instruction-tuning datasets commonly used to fine-tune LLMs indeed tend to include some temporal tasks. For example, Super-Natural Instruction (Mishra et al., 2022; Wang et al., 2022), one of the most comprehensive massively multi-task benchmarks, has only 2 temporal reasoning tasks in its training set out of 756 in its `default` split.[2] In Table 1 and Table 2, we observe Alpaca-7B consistently outperforming LLaMA-7B and being almost comparable with LLaMA-33B. In addition, we found that small-scale models fine-tuned on many in-task examples, such as RoBERTa and TemporalBART in Table 1, often surpass or are competitive with large-scale, state-of-the-art LLMs such as GPT-4. This lends hope that the gap with human performance can be partly filled with sufficient supervision; however, there remains an insurmountable barrier that even providing more examples cannot overcome: GPT-4 performance plateaus even increasing in-context examples (see Figure 3). Even chain-of-thought prompting does not increase its self-consistency (see Figure 4). We hypothesise that only equipping LLMs with perception and action in a simulated or physical environment might further ground them sufficiently to enhance their temporal reasoning.

6 Related Work

Grounding of Language Models. Recently, a budding interest has emerged in probing whether large language models are grounded (Yun et al., 2021; Abdou et al., 2021; Carta et al., 2023; Mahowald et al., 2023; Chandu et al., 2021) with respect various properties. Abdou et al. (2021) found a correspondence between physical colour space and the embeddings of colour terms. Ebert et al. discovered that object trajectories in an environment correlate with the embeddings of motion verbs. In contrast, Yun et al. (2021) showed that vision-and-language pre-training fails to further ground lexical representations. Improving the grounding of LLMs is also an active area of research. Integrating the modelling of subject–object interactions enables better grounding during training (Merullo et al., 2022; Carta et al., 2023). Liu et al. (2023c) uses a computational physics engine as a simulator for improving how language models reason about the physical world.

Temporal Reasoning. Temporal reasoning has always been widely concerned in the NLP community (Hu et al., 2023; Kynoch and Latapie, 2023; Liu et al., 2023b; Lin et al., 2021; Llorens et al., 2015). To improve models' ability in temporal reasoning, previous works mainly fine-tune the language models on the task-specific datasets (Lin et al., 2021; Cai et al., 2023; Zhou et al., 2022; Ma et al., 2023). More recent works attempt on improving large language model's temporal ability using various prompting methods and show the improvements (Wei et al., 2022; Xiong et al., 2024; Yao et al., 2024) or training with specific data mixtures (Tan et al., 2023b; Yuan et al., 2023). However, the extent to which temporal information can be learned through existing training paradigms to facilitate the successful temporal reasoning remains an unresolved problem (Wang and Zhao, 2023; Tan et al., 2023a; Nylund et al., 2023).

7 Conclusions

In this study, we analyse LLMs' temporal grounding ability. We probes LLMs on tasks in temporal knowledge, timeline ordering, and internal consistency through temporal constraints. Despite advancements, LLMs fall short compared to smaller models and human performance. Increasing model scale or in-context examples doesn't consistently improve performance. Instead, instruction-tuning and chain-of-thought prompts enhance accuracy and consistency, but with diminishing benefits at larger scales. To address the limitations, we suggest exploring novel paradigms involving action and perception in physical or simulated environments for more robust temporal grounding.

[2] `task1507_boolean_temporal_reasoning`, `task389_torque_generate_temporal_question`.

Limitations

Although we carefully select human-annotated benchmarks for evaluation, there are still some known issues. Questions in McTACO (Zhou et al., 2019) may contain duplicate candidates, and some of them have all true or false candidates. Models that have not been trained on in-domain data may not be robust to this noise in evaluation. CaTeRS (Mostafazadeh et al., 2016) contains a small proportion of highly similar or duplicated instances. A high-quality benchmark for systematically evaluating the recent model's temporal grounding ability is still essential but missing.

It is not clear whether the training of GPT models is contaminated with data leakage from these datasets. This fact may pose a potential bias in overestimating the advantage of GPT models. Additionally, we confirm that McTACO annotation is not included in the instruction tuning of LLaMA-Alpaca; however, other tasks have texts sourced from the same corpus as McTACO. Hence, the McTACO's evaluation could also be biased to favour the Alpaca model.

Ethical Considerations

We do not expect any ethical concerns to be raised with respect to this work.

Acknowledgements

We thank the reviewers for their useful feedback. We are grateful for an Apple AI/ML scholarship awarded to Yifu Qiu. Zheng Zhao is supported by the UKRI Centre for Doctoral Training in Natural Language Processing (EP/S022481/1). We appreciate the use of computing resources through the Baskerville cluster at the University of Birmingham.

References

Mostafa Abdou, Artur Kulmizev, Daniel Hershcovich, Stella Frank, Ellie Pavlick, and Anders Søgaard. 2021. Can language models encode perceptual structure without grounding? A case study in color. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 109–132, Online. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Bibo Cai, Xiao Ding, Zhouhao Sun, Bing Qin, Ting Liu, Baojun wang, and Lifeng Shang. 2023. Self-supervised logic induction for explainable fuzzy temporal commonsense reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):12580–12588.

Thomas Carta, Clément Romac, Thomas Wolf, Sylvain Lamprier, Olivier Sigaud, and Pierre-Yves Oudeyer. 2023. Grounding large language models in interactive environments with online reinforcement learning.

Khyathi Raghavi Chandu, Yonatan Bisk, and Alan W Black. 2021. Grounding 'grounding' in NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4283–4305, Online. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling language modeling with pathways.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Ishita Dasgupta, Christine Kaeser-Chen, Kenneth Marino, Arun Ahuja, Sheila Babayan, Felix Hill, and Rob Fergus. 2022. Collaborating with language models for embodied reasoning. In *NeurIPS 2022 Foundation Models for Decision Making Workshop*.

Ernest C Davenport Jr and Nader A El-Sanhurry. 1991. Phi/phimax: review and synthesis. *Educational and psychological measurement*, 51(4):821–828.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. 2023. PaLM-E: An embodied multimodal language model.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jian, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Sean Welleck, Xiang Ren, Allyson Ettinger, Zaïd Harchaoui, and Yejin Choi. 2023. Faith and fate: Limits of transformers on compositionality. *ArXiv*, abs/2305.18654.

Dylan Ebert, Chen Sun, and Ellie Pavlick. 2022. Do trajectories encode verb meaning? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2860–2871, Seattle, United States. Association for Computational Linguistics.

Fahim Faisal and Antonios Anastasopoulos. 2022. Geographic and geopolitical biases of language models.

Wes Gurnee and Max Tegmark. 2023. Language models represent space and time.

David Ha and Jürgen Schmidhuber. 2018. Recurrent world models facilitate policy evolution. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 2455–2467, Red Hook, NY, USA. Curran Associates Inc.

Xuming Hu, Junzhe Chen, Xiaochuan Li, Yufei Guo, Lijie Wen, Philip S Yu, and Zhijiang Guo. 2023. Do large language models know about facts? *arXiv preprint arXiv:2310.05177*.

Brandon Kynoch and Hugo Latapie. 2023. Recallm: An architecture for temporal context understanding and question answering. *arXiv preprint arXiv:2307.02738*.

Yann LeCun. 2022. A path towards autonomous machine intelligence. Version 0.9.2, 2022-06-27.

Shih-Ting Lin, Nathanael Chambers, and Greg Durrett. 2021. Conditional generation of temporally-ordered event sequences. 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 7142–7157, Online. Association for Computational Linguistics.

Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2023a. We're afraid language models aren't modeling ambiguity.

Jason Xinyu Liu, Ziyi Yang, Ifrah Idrees, Sam Liang, Benjamin Schornstein, Stefanie Tellex, and Ankit Shah. 2023b. Grounding complex natural language commands for temporal tasks in unseen environments. In *Conference on Robot Learning*, pages 1084–1110. PMLR.

Ruibo Liu, Jason Wei, Shixiang Shane Gu, Te-Yen Wu, Soroush Vosoughi, Claire Cui, Denny Zhou, and Andrew M. Dai. 2023c. Mind's eye: Grounded language model reasoning through simulation. In *The Eleventh International Conference on Learning Representations*.

Hector Llorens, Nathanael Chambers, Naushad UzZaman, Nasrin Mostafazadeh, James Allen, and James Pustejovsky. 2015. SemEval-2015 task 5: QA TempEval - evaluating temporal information understanding with question answering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 792–800, Denver, Colorado. Association for Computational Linguistics.

Max M Louwerse and Rolf A Zwaan. 2009. Language encodes geographical information. *Cognitive Science*, 33(1):51–73.

Congda Ma, Kotaro Funakoshi, Kiyoaki Shirai, and Manabu Okumura. 2023. Coherent story generation with structured knowledge. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 681–690, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2023. Dissociating language and thought in large language models: a cognitive perspective.

Brian W Matthews. 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451.

Jack Merullo, Dylan Ebert, Carsten Eickhoff, and Ellie Pavlick. 2022. Pretraining on interactions for learning grounded affordance representations. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 258–277, Seattle, Washington. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations:

What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, Dublin, Ireland. Association for Computational Linguistics.

Nasrin Mostafazadeh, Alyson Grealish, Nathanael Chambers, James Allen, and Lucy Vanderwende. 2016. CaTeRS: Causal and temporal relation scheme for semantic annotation of event structures. In *Proceedings of the Fourth Workshop on Events*, pages 51–61, San Diego, California. Association for Computational Linguistics.

Kai Nylund, Suchin Gururangan, and Noah A Smith. 2023. Time is encoded in the weights of finetuned language models. *arXiv preprint arXiv:2312.13401*.

OpenAI. 2023. GPT-4 technical report.

James Pustejovsky, José M Castano, Robert Ingria, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R Radev. 2003. TimeML: Robust specification of event and temporal expressions in text. In *Proceedings of the 5th International Workshop on Computational Semantics (IWCS-5)*, pages 337–353.

Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. 2017. A neural model for user geolocation and lexical dialectology. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 209–216, Vancouver, Canada. Association for Computational Linguistics.

Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2022. Large language models are not zero-shot communicators.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2023a. Towards benchmarking and improving the temporal reasoning capability of large language models. *arXiv preprint arXiv:2306.08952*.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2023b. Towards robust temporal reasoning of large language models via a multi-hop qa dataset and pseudo-instruction tuning. *arXiv preprint arXiv:2311.09821*.

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. 2023a. LLaMA: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Zeno Vendler. 1967. *Linguistics in philosophy*. Cornell University Press.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, Seattle, United States. Association for Computational Linguistics.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. 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.

BigScience Workshop et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.

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

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Chenhan Yuan, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2023. Back to the future: Towards explainable temporal reasoning with large language models. *arXiv preprint arXiv:2310.01074*.

Tian Yun, Chen Sun, and Ellie Pavlick. 2021. Does vision-and-language pretraining improve lexical grounding? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4357–4366, Punta Cana, Dominican Republic. 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, Hong Kong, China. Association for Computational Linguistics.

Bo Zhou, Yubo Chen, Kang Liu, Jun Zhao, Jiexin Xu, Xiaojian Jiang, and Qiuxia Li. 2022. Generating temporally-ordered event sequences via event optimal transport. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1875–1884, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

A All Prompt Templates

We provide the three prompt templates we design for our experiments in Table 3.

B Detailed Results for All Prompt Templates

In Table 4, we present the complete results based on all three prompts on the TempEvalQA-bi dataset. Additionally, comprehensive results utilising all three distinct prompts for McTACO are provided in Table 5. We only list the likelihood-based evaluation results for LLaMA models in Table 1 for simplicity, but we observe the same trend for both decoding-based and likelihood-based evaluation. Fialy, detailed results using all three different prompts for CaTeRS can be found in Table 6.

C Detailed Evaluation for CaTeRS

In line with (Lin et al., 2021), we extend our evaluation to a subset of the CaTeRS testing set, specifically focusing on instances with more than 3 input events (referred to as the Long group; Figure 6). We compare the performance of top-performing LLMs in the CaTeRS evaluation and the specialized baseline, TemporalBART, between the complete testing set and this subset.

We note a consistent trend wherein most models exhibit decreased performance in the more demanding subset of CaTeRS, except for LLaMA-2-70B. However, it's evident that LLMs struggle with the more intricate timeline reasoning tasks. For instance, TemporalBART experiences a relative drop of 3.06%, while GPT-4 and LLaMA-2-chat-70B show larger declines of 5.3% and 4.58%, respectively.

D Chain-of-thought Reasoning Examples

We provide an example of a model's input using chain-of-thought prompting on TempEvalQA-Bi in Table 7. Table 8, Table 9, and Table 10 showcase a selection of generations obtained from LLaMA-65B and GPT-4 models using chain-of-thought prompting on the TempEvalQA-bi dataset.

McTACO	Structured	Answer the following multiple-choice question with candidate answers according to the given passage. There can be multiple correct answers. Passage: {passage} Question: {question} Candidate answers: {candidates} The answer is
	Natural	Based on the information presented in the passage “{passage}”, answer the multiple-choice question “{question}” with following candidate answers “{candidate}”. There can be multiple correct answers. The correct answer(s) is/are:
	Text continuation	Finish the following text: According to the passage “{passage}”, the correct answer(s) to the multiple-choice question “{question}” with following candidate answers “{candidates}” is/are:
CaTeRS	Structured	Following the given template to order the events according to temporality: *[Input–Output Templates]* Input: {source} Output:
	Natural	Based on temporality in the given events “{source}”, arrange the events in temporal order. The order is:
	Text continuation	Finish the following text: According to the temporality in the given events “{source}”, the temporal order of the events is:
TempEvalQA-Bi	Structured	Answer the question according to the article. Only answer yes or no. Article: {events}. Question: {question}. The answer is:
	Natural	Based on the information presented in the article “{events}”, answer the question “{question}” with yes or no. The answer is:
	Text continuation	Finish the following text: Article: {events}" The answer to the yes or no question “{question}” according to the the article is:

Table 3: The three templates, i.e., Structured prompt, Natural prompt, Text continuation prompt, used for LLMs' inference in each dataset.

Models		Prompt 1				Prompt 2				Prompt 3			
		Likelihood		Decoding		Likelihood		Decoding		Likelihood		Decoding	
		Acc. (↑)	**Inc.** (↓)	**Acc.** (↑)	**Inc.** (↓)	**Acc.** (↑)	**Inc.** (↓)	**Acc.** (↑)	**Inc.** (↓)	**Acc.** (↑)	**Inc.** (↓)	**Acc.** (↑)	**Inc.** (↓)
LLaMA-7B	ZS	6.25	90.63	10.71	82.35	1.79	96.43	0.45	92.16	2.23	97.32	0	60
	ICL	2.68	96.43	3.57	91.96	3.57	93.30	8.93	82.03	3.57	95.09	2.23	86.67
Alpaca-7B	ZS	12.05	82.14	12.95	80.36	7.59	89.29	1.79	92.04	10.71	79.46	0	80.49
	ICL	17.41	73.66	11.61	80.80	6.25	85.71	10.71	77.31	15.63	72.32	9.38	76.67
LLaMA-13B	ZS	0.45	99.55	4.02	90.41	1.34	93.75	0	94.23	0	100	0	96.15
	ICL	1.34	98.66	2.68	95.98	0.45	99.55	1.34	85.11	0	99.55	0	100
LLaMA-33B	ZS	2.68	96.88	11.16	86.16	0.89	99.11	5.36	84.83	0.45	98.66	0	100
	ICL	19.20	78.57	17.41	79.46	5.80	94.20	17.86	72.93	19.20	77.23	24.55	69.72
LLaMA-65B	ZS	19.64	78.13	9.38	73.39	13.39	82.14	4.91	84.34	9.38	90.18	0	100
	ICL	32.14	63.84	33.93	62.50	30.36	62.05	5.36	87.60	33.48	55.36	27.68	58.20
LLaMA-2-7B	ZS	0	100	0	100	0.45	99.55	0	1	0	100	0	100
	ICL	11.61	87.05	0	100	11.6	82.14	4.91	77.7	12.5	86.16	11.16	80.35
LLaMA-2-13B	ZS	9.38	88.84	10.26	78.05	3.12	96.43	0	97.42	4.46	93.3	0.45	76.47
	ICL	18.75	78.57	12.5	76.76	4.91	92.85	4.02	92.27	17.41	79.46	13.83	80.6
LLaMA-2-70B	ZS	6.7	92.41	0	100	8.48	89.28	11.16	78.92	4.46	95.54	0.9	44.44
	ICL	26.78	67.86	14.29	80.38	32.14	63.83	25	62.63	30.36	65.63	28.57	66.21
LLaMA-2-7B-chat	ZS	20.98	75.44	21.88	73.66	5.8	91.07	5.8	93.12	14.73	83.48	8.04	78.34
	ICL	20.98	70.98	19.64	72.94	24.55	69.64	20.98	72.65	25	69.64	20.98	68.49
LLaMA-2-13B-chat	ZS	23.21	72.76	8.92	80.77	18.75	78.57	15.17	77.42	26.79	67.41	0.89	81.82
	ICL	30.8	62.94	21.88	72.64	29.01	66.52	23.21	71.95	35.26	59.38	30.8	62.5
LLaMA-2-70B-chat	ZS	41.96	55.35	13.39	66.67	37.5	58.92	30.8	59.38	36.16	59.82	0.45	60
	ICL	47.32	46.88	37.05	47.64	46.86	49.1	47.32	45.7	45.09	50.89	47.77	48.66

Table 4: Detailed Model performance evaluated on our curated bi-directional TempEvalQA-bi benchmark with different prompt templates. **Acc.** and **Inc.** stand for accuracy and in-consistent prediction rate. (↑)/(↓) indicate that higher / lower values are better, respectively.

Models		Prompt 1		Prompt 2		Prompt 3	
		Strict Acc. (↑)	**F1** (↑)	**Strict Acc.** (↑)	**F1** (↑)	**Strict Acc.** (↑)	**F1** (↑)
LLaMA-7B	ZS	17.64	19.06	12.69	37.72	12.84	49.12
	ICL	22.15	27.56	20.95	22.18	17.42	17.42
Alpaca-7B	ZS	27.63	61.19	17.64	41.92	19.97	53.41
	ICL	40.92	62.59	20.95	25.86	28.23	43.84
LLaMA-13B	ZS	19.07	20.77	15.69	39.20	12.24	49.80
	ICL	19.07	19.40	23.05	29.41	31.01	56.16
LLaMA-33B	ZS	17.94	18.24	20.20	32.98	13.59	48.39
	ICL	33.71	45.04	31.01	56.90	24.40	40.78
LLaMA-65B	ZS	23.35	51.30	18.92	39.36	12.16	49.84
	ICL	34.31	44.83	31.91	48.91	12.16	49.80
LLaMA-2-7B	ZS	9.38	28.36	12.23	49.94	11.86	49.35
	ICL	17.56	18.22	25.07	52.38	22.59	28.23
LLaMA-2-13B	ZS	19.14	21.48	15.76	46.72	12.16	49.84
	ICL	30.55	40.32	29.35	44.66	29.35	44.66
LLaMA-2-70B	ZS	21.92	26.45	20.34	29.45	15.09	44.63
	ICL	28.9	38.38	29.35	40.18	25.07	33.05
LLaMA-2-7B-chat	ZS	23.72	29.47	16.96	32.8	21.54	23.93
	ICL	20.04	23.77	26.95	43.04	22	27.69
LLaMA-2-13B-chat	ZS	22.07	25.32	22.29	27.25	22.29	42.44
	ICL	27.7	34.74	29.5	45.08	29.5	45.07
LLaMA-2-70B-chat	ZS	23.04	26.98	18.91	20.18	20.57	32.11
	ICL	31.15	40.15	28.68	37.4	21.7	25.56

Table 5: Detailed model performance on McTACO using all prompt templates. **Strict Acc.** and **F1** stand for strict accuracy and the F-1 score as in (Zhou et al., 2019). (↑) indicates higher values are better.

Category	Models	T1 PAcc.	T2 PAcc.	T3 PAcc.
LLaMA	LLaMA-7B	8.9	2.3	0.09
	LLaMA-Alpaca-7B	13.86	4.75	12.5
	LLaMA-13B	0	4.82	10.99
	LLaMA-33B	26.78	9.05	7.32
	LLaMA-65B	15.54	33.98	13.54
	LLaMA-2-7B	8.15	4.19	5.21
	LLaMA-2-13B	15.67	10.83	22.28
	LLaMA-2-70B	20.97	30.3	13.56
	LLaMA-2-7B-chat	23.99	28.04	26.93
	LLaMA-2-13B-chat	27.1	33.12	30.6
	LLaMA-2-70B-chat	17.46	55.79	18.39

Table 6: Detailed Model performance evaluated on our curated CaTeRS benchmark with different prompt templates. **T1/2/3** and **PAcc.** stand for Template 1/2/3 and pair-wise accuracy.

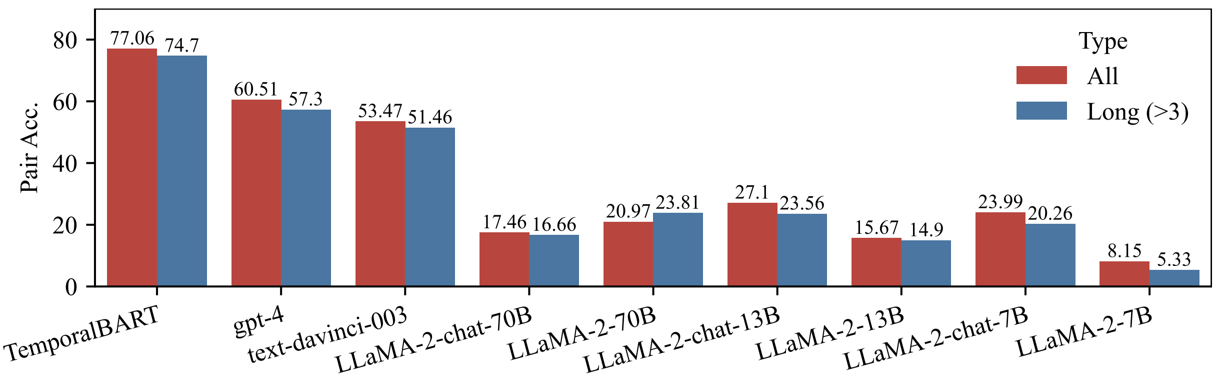


Figure 6: Performance of 8 top-performing LLMs and TemporalBART in all CaTeRS testing instances (All) and instances with more than 3 input events (Long).

Answer the question according to the article. Give step-by-step explanations and then answer yes or no.
Article: Tired of being sidelined, Hungarian astronaut Bertalan Farkas is leaving for the United States to start a new career, he said Saturday. B̈eing 48 is too early to be retired,ä fit-looking Farkas said on state TV's morning talk show. With American astronaut Jon McBride, Farkas set up an American-Hungarian joint venture called Orion 1980, manufacturing space-travel related technology. Farkas will move to the company's U.S. headquarters. 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 in October 1984 and later served as assistant administrator for congressional relations for NASA. Farkas expressed the hope he one day follow in the footsteps of fellow astronaut John Glenn, who at 77 is about to go into space again. On May 22, 1995, Farkas was made a brigadier general, and the following year he was appointed military attache at the Hungarian embassy in Washington. However, cited by District of Columbia traffic police in December for driving under the influence of alcohol, Farkas was ordered home and retired.
Question: Is Farkas sent into space on board the Soyuz before McBride on board the Orbiter Challenger?
The step-by-step explanation and answer is: 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.
Question: Is McBride on board the Orbiter Challenger after Farkas was made a brigadier general?
The step-by-step explanation and answer is: McBride was on board the Orbiter Challenger in October 1984. Farkas was made a brigadier general on May 22, 1995. October 1984 is before May 22, 1995. So the answer is: no.
Question: Is Farkas was appointed military attache at the Hungarian embassy before he was made a brigadier?
The step-by-step explanation and answer is: Farkas was made a brigadier general on May 22, 1995. He was appointed military attache at the Hungarian embassy in the following year. So the answer is: no.
Answer the question according to the article. Give step-by-step explanations and then answer yes or no.
Article: {events}.
Question: {question}.
The step-by-step explanation and answer is:

Table 7: Example model input using Chain-of-thought prompts for TempEvalQA-Bi. The input consists of *demonstrations* and the actual prompt.

Event: *Well. I started this blog years ago and never kept up with it. In fact, I sort of forgot about it. Now, I am done with school and married and I actually have time on my hands! (Gasp! Imagine that!) And, I really want to get back into the swing of writing, so I thought blogging might be a good way to start. I feel like an introduction is in order, so here's a bit about me. My name is Kathleen. I am 22 years old, and I live in Portland, Maine with my husband Brett. We are originally from Portland, Oregon but lived in Seattle for 4 years during college. I work as an office manager in a small software company, so I tend to have a bit of time on my hands. Hence, the blogging. I'll probably talk a lot about all of the changes that have happened in my life in the past year, since that's what is on my mind right now: getting married in July, moving to Maine in November, changing jobs several times, etc. So I will get right down to it. I am pretty freaking grumpy today. Well, to be honest, I've been really grumpy for a few weeks and I can't seem to get un -grumpy, which I hate. Feeling out of control, especially of myself, is not something I like. I guess I'm coming to realize that life, and especially marriage, is NOT easy. Maybe I assumed that since we dated for over 4 years before we got married, that things would be easy. Everyone says that the first year of marriage is the hardest, and for the first six months I laughed that off. ""Hard?"" I would scoff. ""No way! I love being married. It's wonderful! It's amazing! blah, blah, blah."" Now it's like the crap has hit the fan and I am transformed into a grumpy, nagging, mean, yelling wife. Ugh. I don't want to be like this but every time I resolve to be different I fall right back into it. Choosing your battles is a lot harder than it seems. I think that's a lot of what I'm learning; that I really do have to choose my battles, because if I pick on every little thing, he won't listen at all when the big things come along and I really need him to listen. Heh. That's my psychological epiphany for the day.*
Q: Did the getting married happen after they dated? **LLaMA-65B:** They dated for over 4 years before they got married. They dated for over 4 years happened before they got married. So the answer is: yes. **GPT-4:** Kathleen mentioned that they dated for over 4 years before they got married. So the answer is yes.
Q: Did the getting married happen before they dated? **LLaMA-65B:** Getting married happened in July. Dating happened before getting married. July is before getting married. So the answer is: no. **GPT-4:** They dated for over 4 years before getting married. So the answer is no.

Table 8: Example of chain-of-thought for LLaMA-65B. Chain-of-thought reasoning processes are highlighted. We also highlight the answers from the models to indicate whether they are correct or incorrect. Both LLaMA-65B and GPT-4 consistently provided correct answers for both questions.

Event: *P.F. Chang's China Bistro Inc. has provided new details about a security breach discovered in June, saying customer data may have been stolen from 33 restaurants in 16 states, including eight in California. The restaurant chain said Monday that credit card numbers, expiration dates and, in some cases, cardholder names were stolen over eight months. However, the chain has not yet determined if ""any specific cardholder's credit or debit card data was stolen by the intruder,"" according to Chief Executive Rick Federico. P.F. Chang's confirmed the data breach June 13, three days after the U.S. Secret Service alerted the chain that its credit card processing systems may have been hacked. The company said the breach occurred between Oct. 19, 2013, and June 11, 2014. The intrusion was first reported by security blogger Brian Krebs, who said on his website that banks reported data from thousands of customers had been pilfered from locations in Florida, Maryland, New Jersey, Pennsylvania, Nevada and North Carolina, and was being sold online. Federico said P.F. Chang's conducted an internal investigation with the help of forensic data experts, and determined the 33 locations affected and specific time frames the credit card processing system was compromised for each. He also said the company has been processing card data securely since June 11.*

Q: Did the determined happen before the processing?

LLaMA-65B: 1) The determined happened between Oct. 19, 2013 and June 11, 2014. 2) The processing happened after June 11, 2014. 3) Oct. 19, 2013 and June 11, 2014 is before June 11, 2014. So the answer is: yes.

GPT-4: The determination of the affected locations and time frames happened after the processing was compromised. So the answer is no.

Q: Did the determined happen after the processing?

LLaMA-65B: The company said the breach occurred between Oct. 19, 2013, and June 11, 2014. The determined was determined on June 13. So the answer is: yes.

GPT-4: The company determined the affected locations and specific time frames after processing the card data securely. So the answer is yes.

Table 9: Example of chain-of-thought for LLaMA-65B. Chain-of-thought reasoning processes are highlighted. We also highlight the answers from the models to indicate whether they are correct or incorrect. LLaMA-65B successfully answered the first question accurately but exhibited inconsistency in answering the second question. On the other hand, GPT-4 failed to provide correct answers for both questions.

Event: *The Great Depression was a severe worldwide economic depression in the decade preceding World War II. The timing of the Great Depression varied across nations, but in most countries it started in 1930 and lasted until the late 1930s or middle 1940s. It was the longest, deepest, and most widespread depression of the 20th century. In the 21st century, the Great Depression is commonly used as an example of how far the world's economy can decline. The depression originated in the U.S., after the fall in stock prices that began around September 4, 1929, and became worldwide news with the stock market crash of October 29, 1929 (known as Black Tuesday). The Great Depression had devastating effects in countries rich and poor. Personal income, tax revenue, profits and prices dropped, while international trade plunged by more than 50%. Unemployment in the U.S. rose to 25%, and in some countries rose as high as 33%. Cities all around the world were hit hard, especially those dependent on heavy industry. Construction was virtually halted in many countries. Farming and rural areas suffered as crop prices fell by approximately 60%. Facing plummeting demand with few alternate sources of jobs, areas dependent on primary sector industries such as cash cropping, mining and logging suffered the most. Some economies started to recover by the mid-1930s. In many countries, the negative effects of the Great Depression lasted until after the end of World War II. Start. Economic historians usually attribute the start of the Great Depression to the sudden devastating collapse of US stock market prices on October 29, 1929, known as Black Tuesday; some dispute this conclusion, and see the stock crash as a symptom, rather than a cause, of the Great Depression. . . .*
Q: Did the depression start before middle 1940s? **LLaMA-65B:** The depression started after the fall in stock prices that began around September 4, 1929. 2. Middle 1940s is after September 4, 1929. 3. So the answer is: no. **GPT-4:** The Great Depression started in the decade preceding World War II, and in most countries, it began in 1930. It lasted until the late 1930s or middle 1940s. So the answer is yes.
Q: Did the depression start after middle 1940s? **LLaMA-65B:** The depression originated in the U.S., after the fall in stock prices that began around September 4, 1929, and became worldwide news with the stock market crash of October 29, 1929 (known as Black Tuesday). So the answer is: no. **GPT-4:** The depression started in the decade preceding World War II, which means it began before the 1940s. In most countries, the Great Depression started in 1930 and lasted until the late 1930s or middle 1940s. So the answer is no.

Table 10: Example of chain-of-thought for LLaMA-65B. Chain-of-thought reasoning processes are highlighted. We also highlight the answers from the models to indicate whether they are correct or incorrect. LLaMA-65B demonstrated accuracy in answering the second question while displaying inconsistency in answering the first question. Conversely, GPT-4 provided correct responses for both questions.