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

THE UNIVERSITY of EDINBURGH

[1]Institute for Language, Cognition and Computation, University of Edinburgh

[2]Language Technology Lab, University of Cambridge

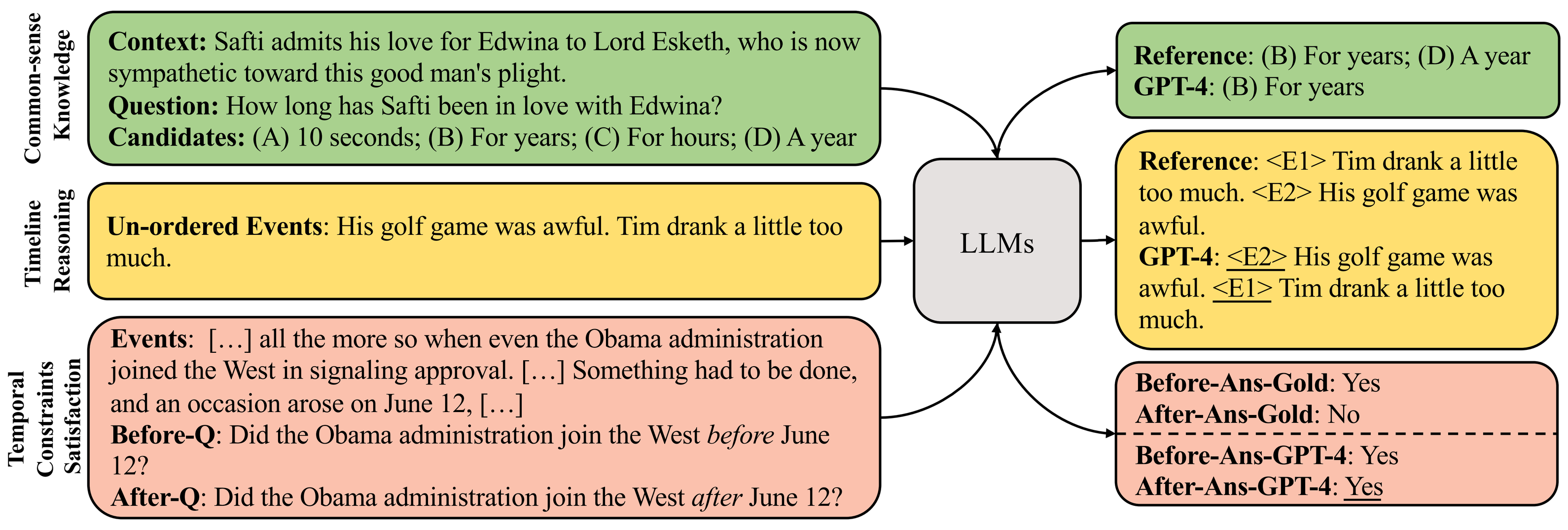
UNIVERSITY OF CAMBRIDGE

Overview

- This work evaluates the **temporal grounding** of large language models (LLMs) like GPT-4 and LLaMA by probing their ability to reason about textual narratives involving events.
- It tests three key aspects: models' **commonsense knowledge about events**, their **ability to order events on a timeline**, and their **ability to satisfy temporal constraints**.
- The study utilizes three benchmarks - McTACO, CaTeRS, and TempEvalQA-Bi - to evaluate each of the three aspects respectively.
- Results show that **LLMs struggle significantly on all three temporal reasoning abilities** compared to humans and specialized models, with recent techniques like few-shot prompting, scaling, and chain-of-thought prompting providing only limited improvements.

Tasks for Evaluation

This study proposes a framework to evaluate temporal grounding by decomposing it into three fundamental abilities. We provide an illustration using examples below. We highlight wrong predictions with underline.



The expectation is that a truly grounded model with temporal understanding should perform well across all three abilities.

Evaluation Setup

Three benchmarks are used to evaluate the temporal grounding abilities of LLMs:

1. **McTACO** - A multiple-choice question answering dataset to assess commonsense temporal knowledge across categories like event duration, ordering, time, frequency, and stationarity.
2. **CaTeRS** - An event ordering task where models must arrange events from a narrative into the correct chronological sequence by reasoning over causal and temporal cues.
3. **TempEvalQA-Bi** - A binary question-answering dataset derived from TempEvalQA to test self-consistency. Models must maintain mutual exclusivity between contradictory "before/after" relations for event pairs.

Multiple strategies are employed based on model type and benchmark format:

- **Multiple-choice QA**: LLMs generate answers by ranking provided candidates.
- **Sequence-to-Sequence**: For event ordering, models take events as input and temporally sort them as output.
- **Yes/No QA**: Models predict yes/no by ranking likelihoods or through greedy decoding.

Materials

(a) Paper

(b) Github Repo

Main Results

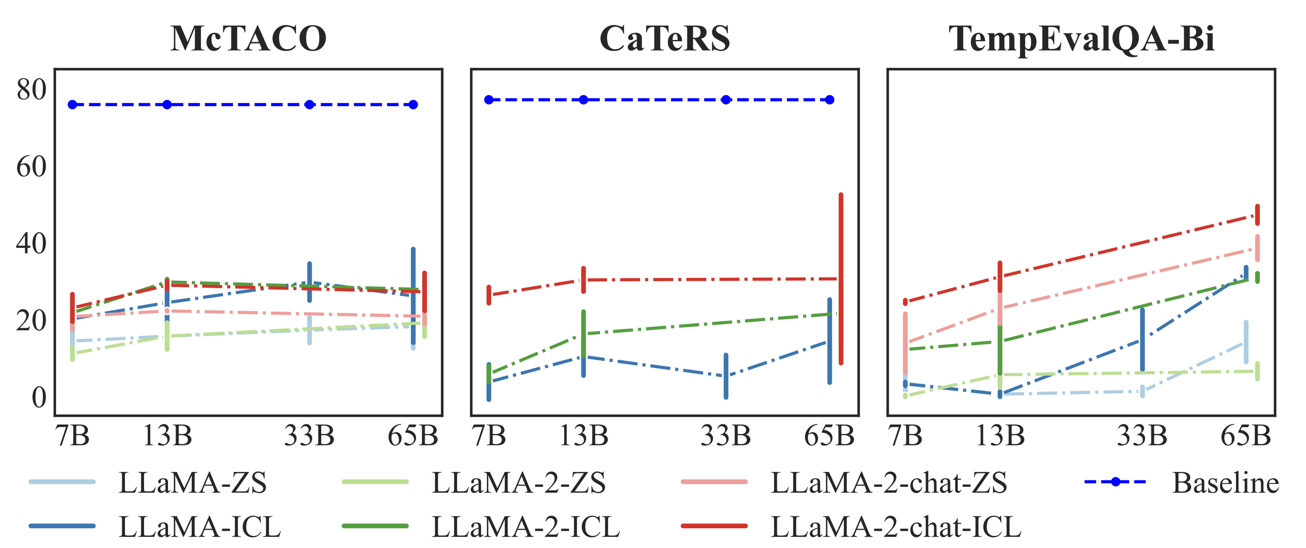
Models	McTACO Zero-shot		McTACO Few-shot		CaTeRS
	Strict Acc.	F1	Strict Acc.	F1	Pair Acc.
RoBERTa	43.62	72.34	-	-	-
TemporalBART	-	-	-	-	**77.06**
Human	**75.80**	-	-	-	-
GPT-4	**28.45**	35.88	**50.15**	**65.27**	**60.51**
text-davinci-003	26.05	**48.30**	33.56	65.04	53.47
LLaMA-7B	$14.39_{2.82}$	$35.30_{15.18}$	$20.17_{2.46}$	$22.39_{5.07}$	$3.76_{4.58}$
Alpaca-7B	$\mathbf{21.75}_{5.22}$	$\mathbf{52.17}_{9.69}$	$\mathbf{30.03}_{10.11}$	$44.10_{18.36}$	$10.37_{4.91}$
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	$18.14_{5.63}$	$46.83_{6.51}$	$26.13_{12.15}$	$\mathbf{47.84}_{2.65}$	$\mathbf{21.02}_{10.27}$
LLaMA-2-7B	$11.16_{1.55}$	$\mathbf{42.55}_{12.29}$	$21.74_{3.83}$	$32.94_{17.56}$	$5.85_{2.06}$
LLaMA-2-13B	$15.69_{3.49}$	$39.35_{15.55}$	$\mathbf{29.75}_{0.69}$	$\mathbf{43.21}_{2.51}$	$16.26_{5.75}$
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	$\mathbf{22.22}_{0.13}$	$31.67_{9.38}$	$28.90_{1.04}$	$41.63_{5.97}$	$30.27_{3.02}$
LLaMA-2-chat-70B	$20.84_{2.08}$	$26.42_{5.98}$	$27.18_{4.9}$	$34.37_{7.75}$	$\mathbf{30.55}_{21.87}$

Table 1. Average model performance (standard deviations as subscripts). Left: McTACO for evaluating temporal commonsense reasoning in LLMs. Right: CaTeRS results for few-shot prompting. **Pair Acc.** stands for pairwise accuracy.

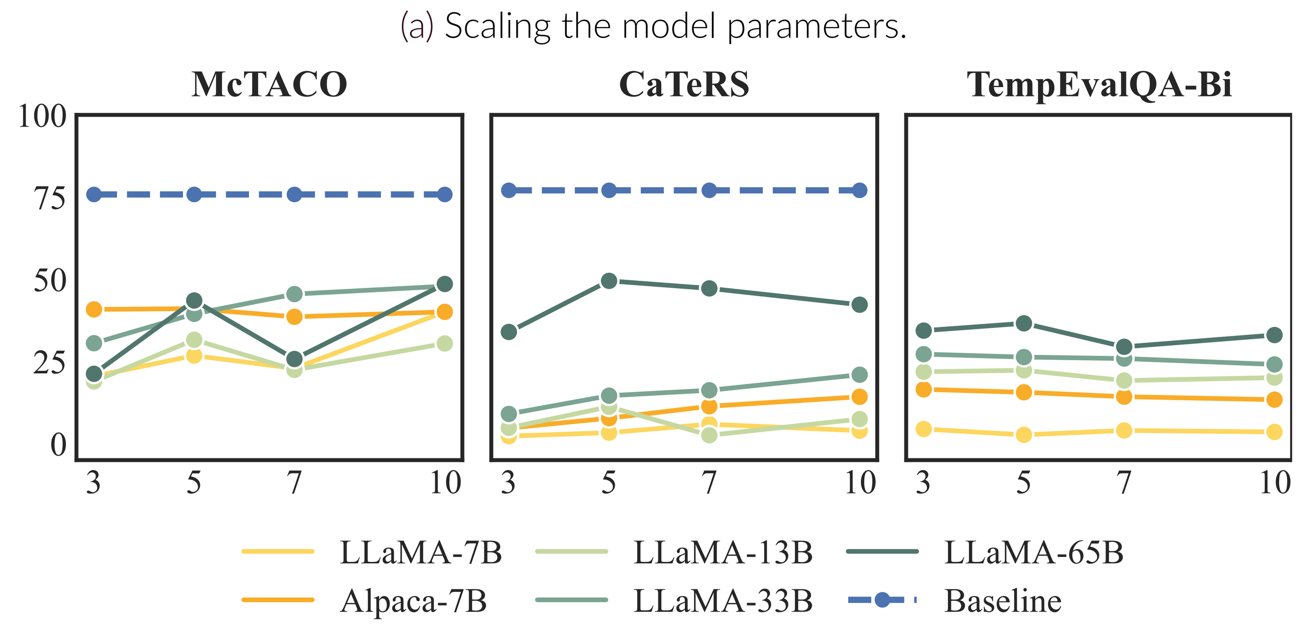
Models	Zero-shot		Few-shot	
	Acc. (↑)	Inc. (↓)	Acc. (↑)	Inc. (↓)
GPT-4	**64.29**	**31.25**	**67.41**	**27.23**
text-davinci-003	27.68	69.64	33.93	62.05
text-davinci-002	16.52	77.83	36.16	60.71
davinci	14.73	79.02	13.39	79.91
LLaMA-7B	$3.42_{2.46}$	$94.79_{3.64}$	$3.27_{0.51}$	$94.94_{1.57}$
LLaMA-Alpaca-7B	$10.12_{2.29}$	$83.63_{5.08}$	$13.10_{6.00}$	$77.23_{7.37}$
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.

Analysis



(a) Scaling the model parameters.



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

Figure 2. 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.

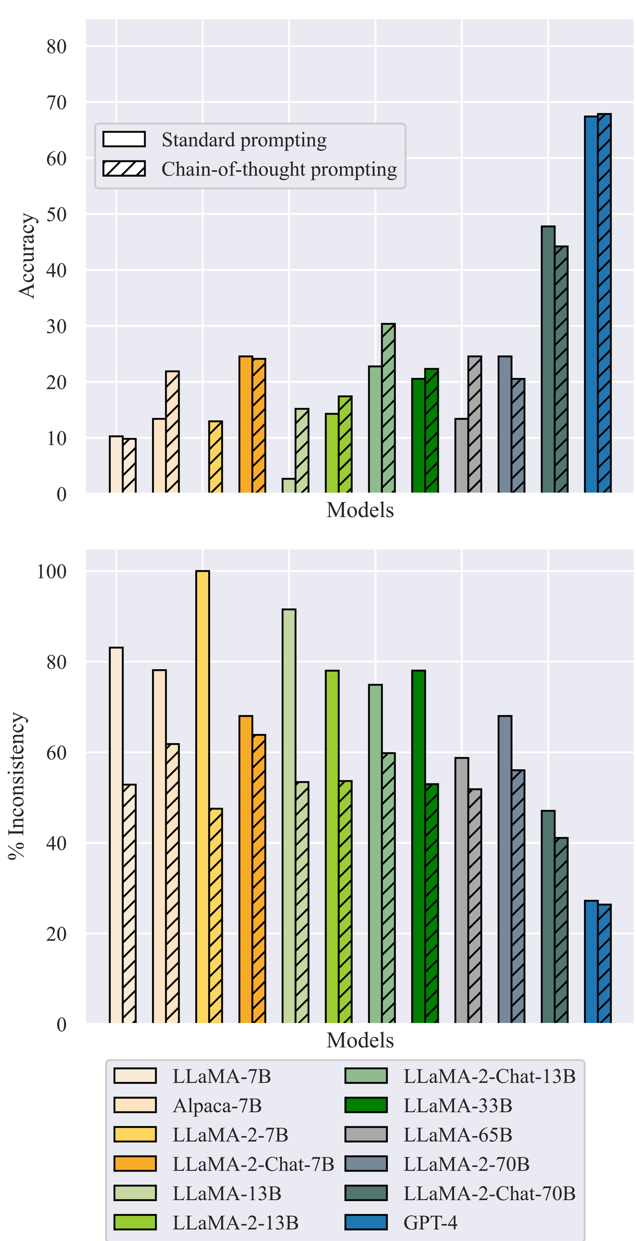


Figure 3. Comparison between chain-of-thought prompting and the standard few-shot prompting on TempEvalQA-Bi for all tested LLMs.

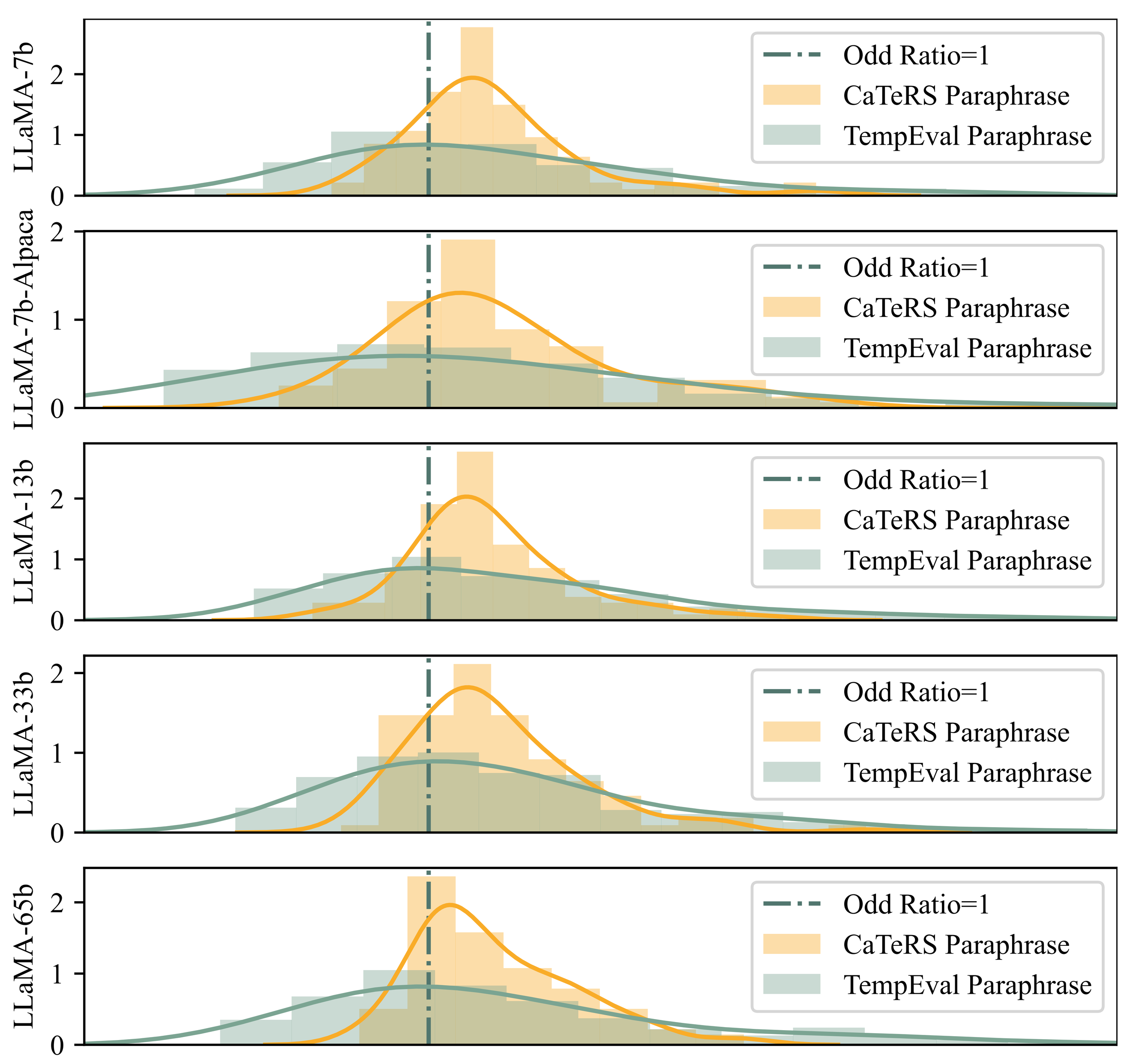


Figure 4. Density plot of the odds ratio under several LLMs (rows) for differently ordered paraphrases in CaTeRS (orange) and TempEvalQA-Bi (green). The odds ratio represents the likelihood of temporally ordered sequences compared to their permuted counterparts.