Are Large Language Models Temporally Grounded?

Presenter : Chenpei Xie

Contents

1 **Introduction**

What Does Time Mean? How Do Humans Learn About Time?

2 **Basic Information**

Framework, Benchmarks, LLMs, Techniques

3 **Result**

Performances

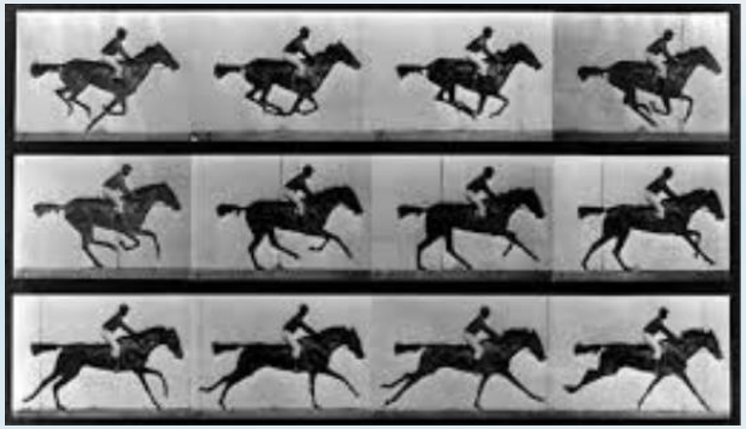
4 **conclusion, further information and discussion**

lack or not?

1. What Does Time Mean?

3 dimensions→ 4 dimensions

live like super observer:



Perception 👉 Analysis 👉Inference

1.1 How Do Humans Learn About Time? Time?

explicit & implicit time clues

Early Development

- Physical maps
- human senses

Cultural Teachings

- word and work
- feedback training

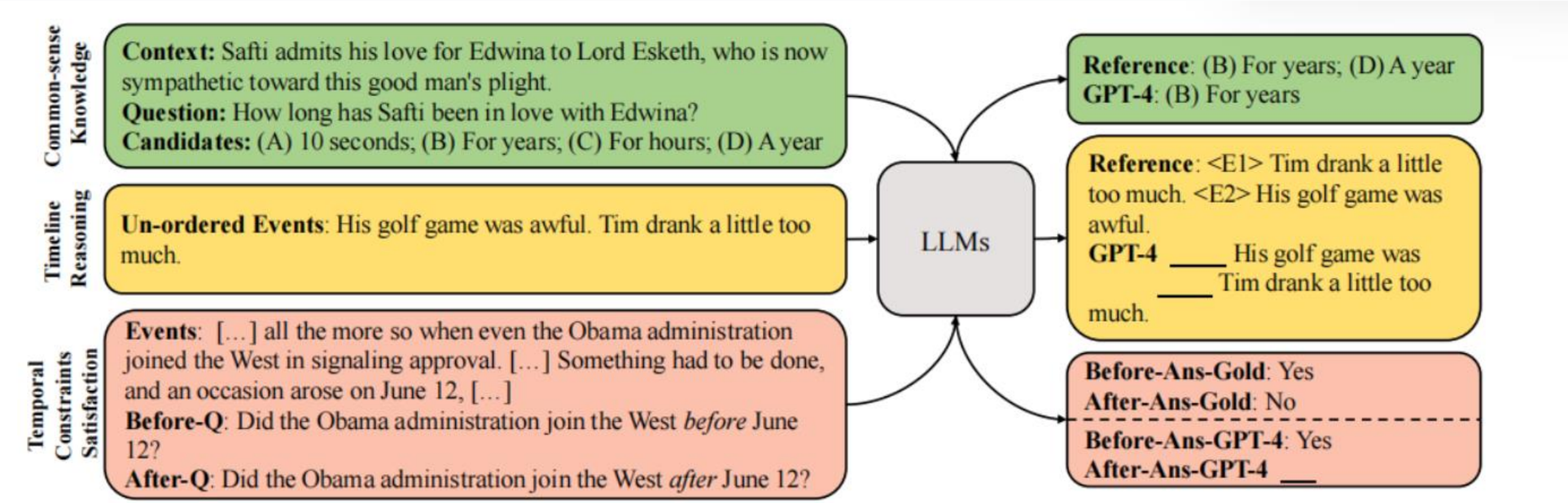
Personal Experiences

- emotional perception
- information exchange

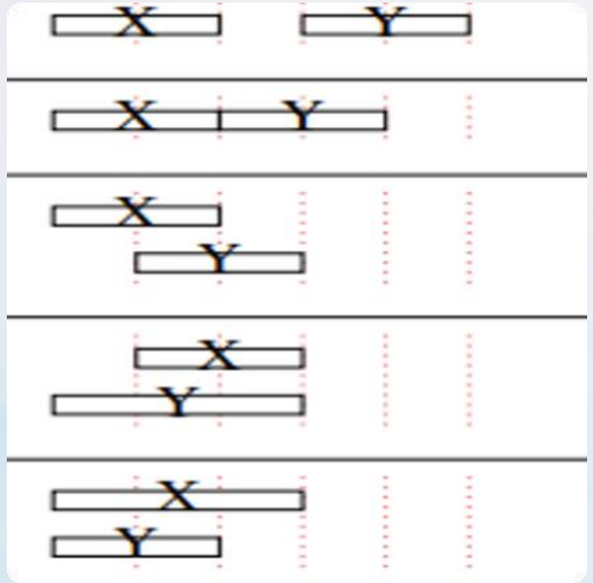
2. Basic Information

1. Framework
2. Benchmarks
3. LLM
4. Techniques

2.1 Framework:



2.2 three benchmarks



McTACO

Multi-choice Commonsense Reasoning Test for Adversarial Adversarial Reading Comprehension

CaTeRS

Causal and Temporal Relation Relation Scheme

TempEvalQA-Bi

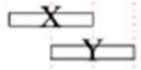
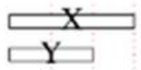
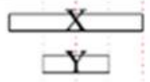
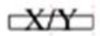
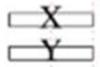
swap the temporal relation to its to its opposite to create negative negative Q&A pairs.

2.2.1 McTACO and EM&F1

- far behind human performance - 20%
- multiple correct answers
- duration
- temporal ordering
- typical time
- frequency
- stationarity

2.2.2 CaTeRS

- 320 five-sentence short stories sampled
- broad-coverage event entity： high inter annotator agreement
- semantic link： moderate agreement

Allen	Visualization	Allen - Inverse	TimeML	THYME	CaTeRS
X Before Y		Y After X	Before	Before	Before
X Meets Y		Y Is Met X	IBefore (Immediately)	-	-
X Overlaps Y		Y Is overlapped by X	-	Overlaps	Overlaps
X Finishes Y		Y Is finished by X	Ends	Ends-on	-
X Starts Y		Y Is started by X	Begins	Begins-on	-
X Contain Y		Y During X	During	Contains	Contains
X Equals Y		-	Identity	Identity	Identity
-		-	Simultaneous	-	-

2.2.3 TempEvalQA-Bi

- derived from TempEvalQA
- 448 humanannotated question–answer triples-only select some of them
- swap the temporal relation to its opposite to create negative question– answer pairs answer pairs

2.3 LLMs

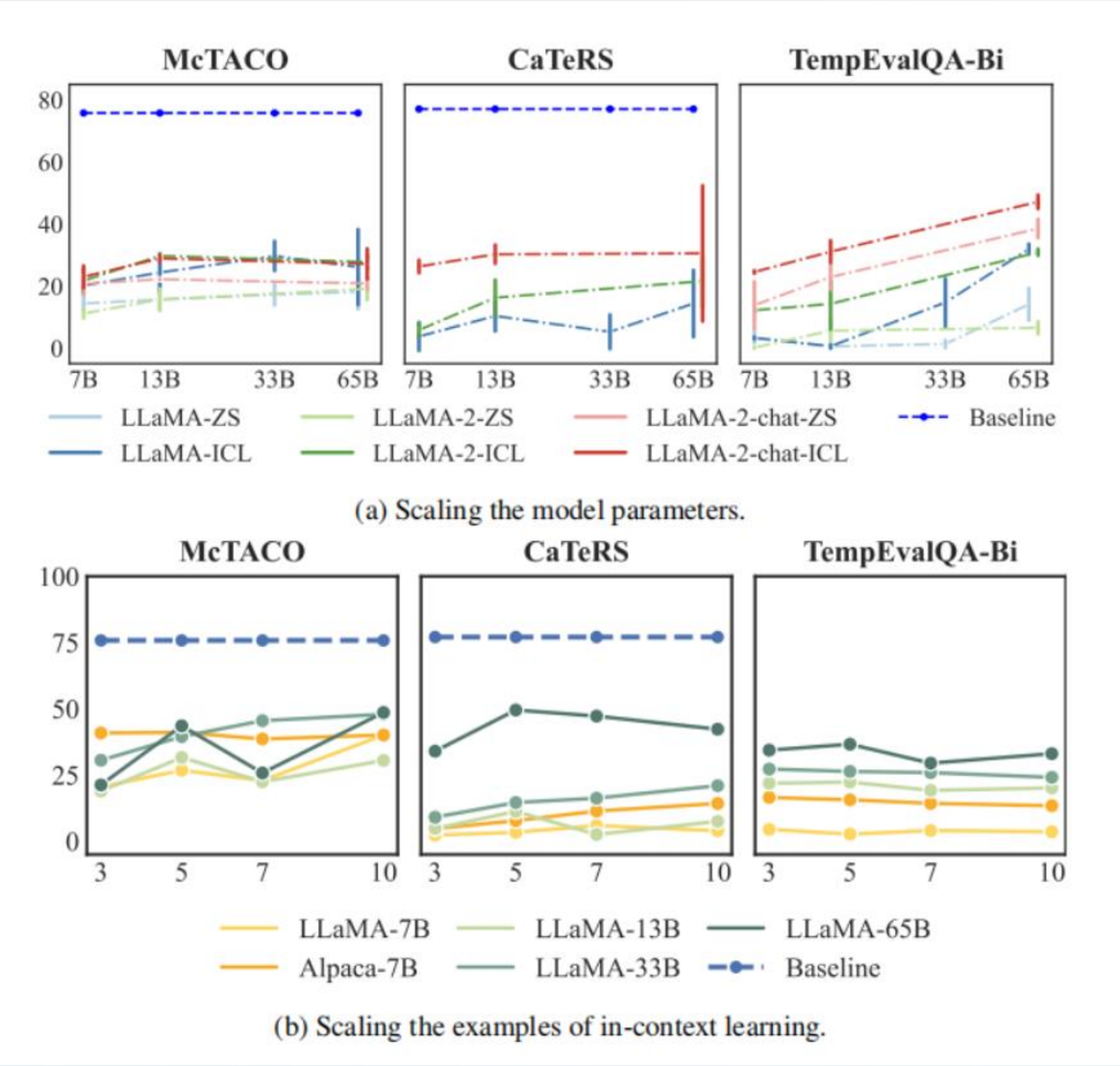
3.

3.1

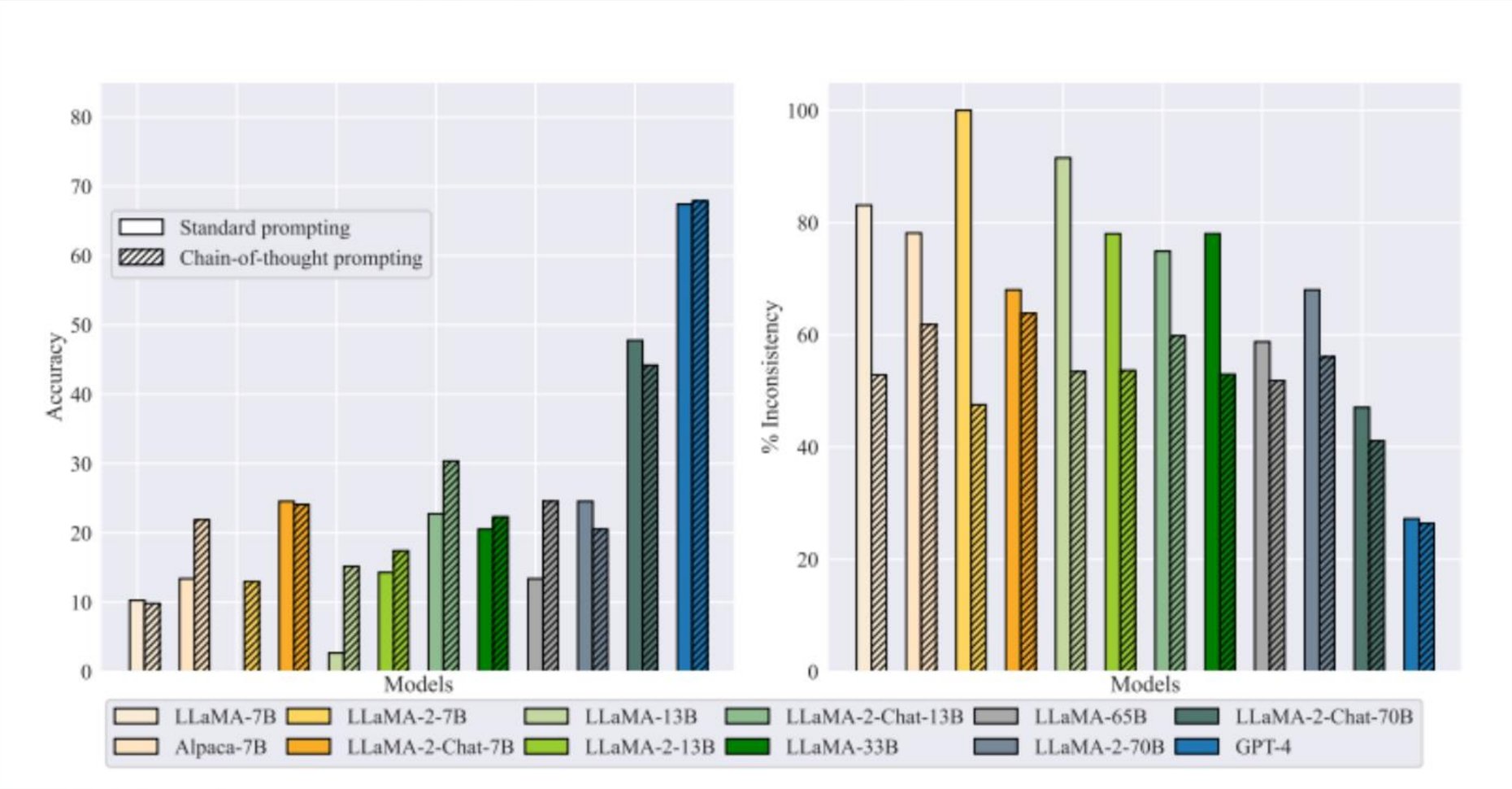
Models	McTACO				CaTeRS
	Zero-shot		Few-shot		
	Strict Acc.	F1	Strict Acc.	F1	Pair Acc.
RoBERTa	43.62	72.34			
TemporalBART					**77.06**
Human	**75.80**				
GPT-4	**28.45**	35.88	**50.15**	**65.27**	**60.51**
text-davinci-003	26.05	**48.30**	33.56	65.04	53.47
LLaMA-7B	$14.39_{2.82}$	$35.30_{15.18}$	$20.17_{2.46}$	$22.39_{5.07}$	$3.76_{4.58}$
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	$20.84_{2.08}$	$26.42_{5.98}$	$27.18_{4.9}$	$34.37_{7.75}$	$\mathbf{30.55}_{21.87}$

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	$\mathbf{38.54}_{3.04}$	$\mathbf{58.03}_{2.36}$	$\mathbf{46.42}_{1.18}$	$\mathbf{48.96}_{2.01}$

3.2 Would adding other technologies technologies improve it?



scaling experiments

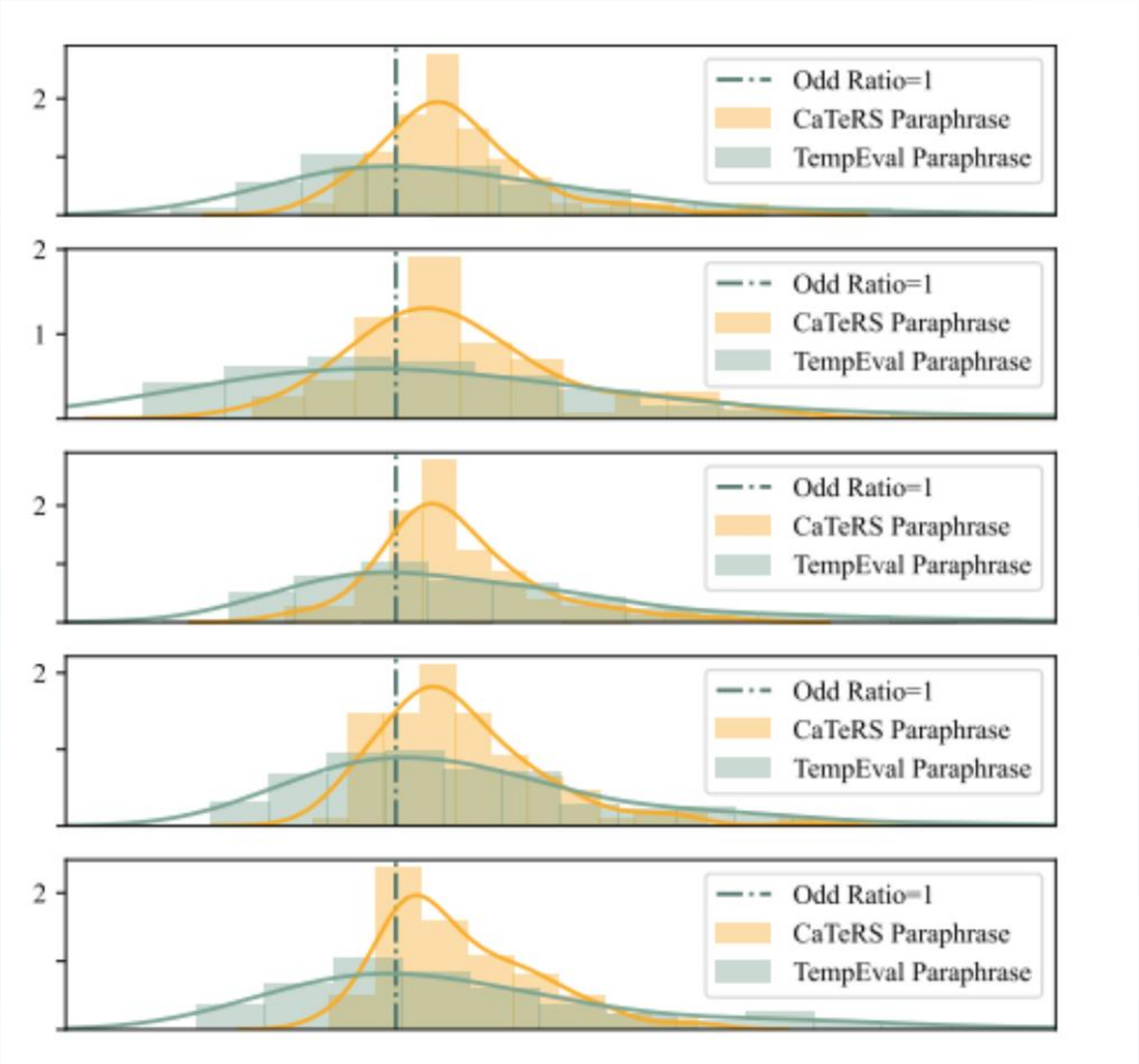


chain-of- thoughts prompting

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

different Prompts

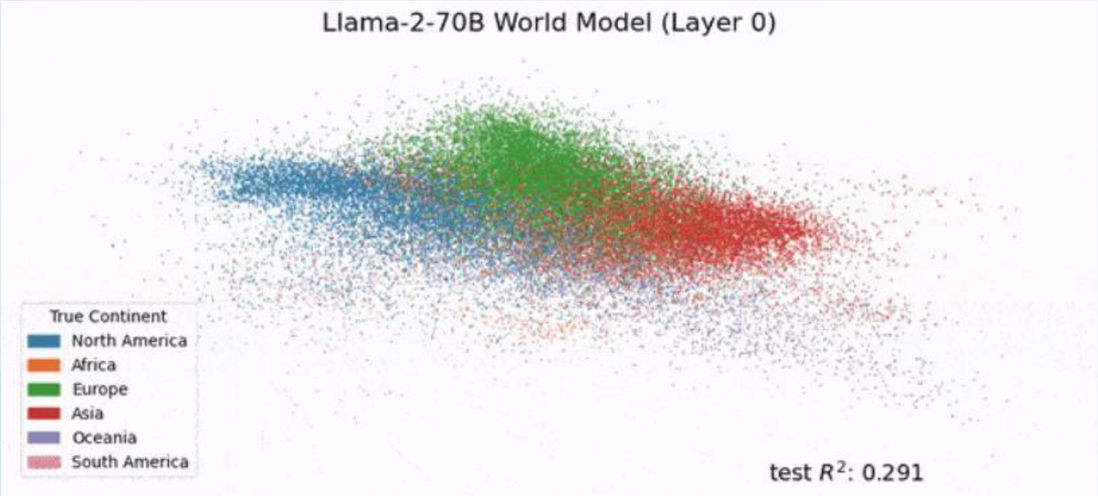
4. So, is there a lack of time?

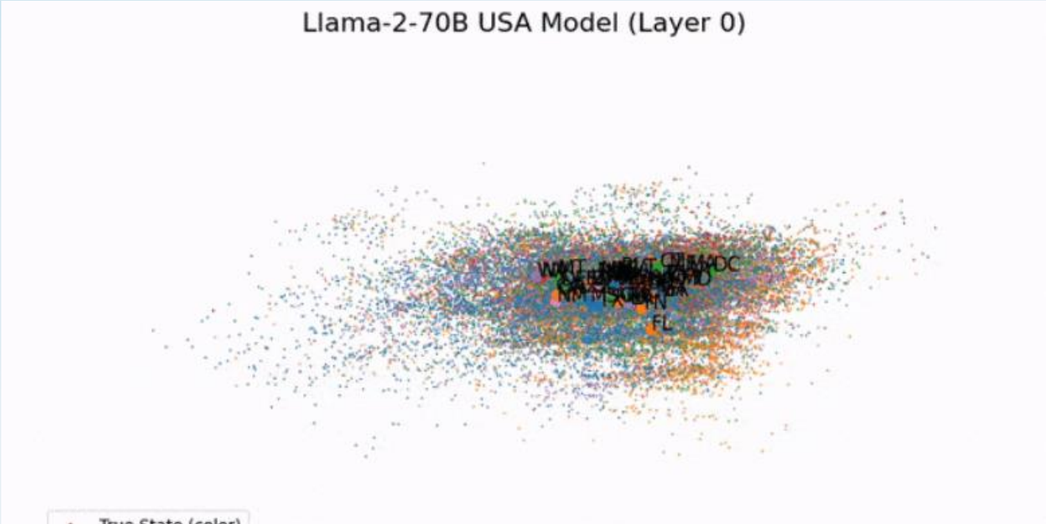

Odd Ratio=1
CaTeRS Paraphrase
TempEval Paraphrase
2
1

Further

Language Models Represent Space and Time

Wes Gurnee, Max Tegmark





Thanks for listening!

1. What do you think the next step should be to train the model to make it understand time better?
2. Which is better between a large model and a specialized model?