Perceive the Passage of Time: A Systematic Evaluation of Large Language Model in Temporal Relativity

Shuang Chen, Yining Zheng, Shimin Li, Qinyuan Cheng, Xipeng Qiu†
School of Computer Science, Fudan University
Correspondence: xpqiu@fudan.edu.cn

Abstract

Temporal perception is crucial for Large Language Models(LLMs) to effectively understand the world. However, current benchmarks primarily focus on temporal reasoning, falling short in understanding the temporal characteristics involving temporal perception, particularly in understanding temporal relativity. In this paper, we introduce TempBench, a comprehensive benchmark designed to evaluate the temporal-relative ability of LLMs. TempBench encompasses 4 distinct scenarios: Physiology, Psychology, Cognition and Mixture. We conduct an extensive experiments on GPT-4, a series of Llama and other popular LLMs. The experiment results demonstrate a significant performance gap between LLMs and humans in temporal-relative capability. Furthermore, the error types of temporal-relative ability in LLMs are proposed to thoroughly analyze the impact of multiple aspects and emphasize the associated challenges. We anticipate that TempBench will drive further advancements in enhancing the temporal-perceiving capabilities of LLMs.

1 Introduction

"*Time is what we want most, but what we use worst.*"

–William Penn

Temporal perception is an indispensable aspect of how humans perceive and comprehends the world, shaping our understanding of events and experiences over time. For instance, time seems to speed when a person is engaged in a pleasurable activity, whereas it appears to slow down during tedious or stressful situations (Zhou et al., 2019; Shi et al., 2024). This subjective sense of time is not only a reflection of human cognitive processes but also involves the integration of mathematical principles and a deep understanding of temporal concepts

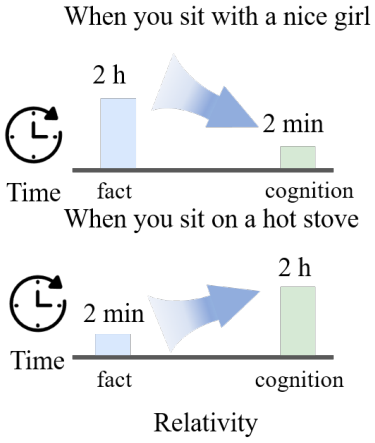

Figure 1: A temporal-relative scenario illustrating human temporal perception. When you sit with a nice girl, **2 hours** feel like **2 minutes**. When you sit on a hot stove, **2 minutes** feel like **2 hours.**

(Xiong et al., 2024; Yuan et al., 2024), as illustrated in Figure 1.

In recent years, Large Language Models (LLMs) have achieved remarkable success across various reasoning tasks, demonstrating their capacity to process and generate complex information with impressive accuracy (Zhao et al., 2023; Chang et al., 2024; Chowdhery et al., 2023; Zhang and Wan, 2024). However, despite these advancements, the domain of temporal perception, particularly in terms of temporal relativity, still remains relatively under-explored. Temporal relativity refers to the specific and varying perceptions of time experienced in different scenarios. For example, the way an individual perceives time during a high-pressure situation may differ dramatically from a leisurely activity. Although recent work has focused on temporal reasoning, addressing basic temporal concepts such as duration, intricate temporal relationships, and even computational tasks like arithmetic involving time (Su et al., 2024b; Fatemi et al., 2024; Su et al., 2024a; Wang and Zhao, 2023;

Proceedings of the 31st International Conference on Computational Linguistics, pages 8304–8313
January 19–24, 2025.

Category	Example
Physiological Relativity	Q: Children feel that one minute lasts much longer. How long does one minute feel to an elderly person? **A. One second** ✓ B. One minute ✗ C. One hour ✗ D. One day ✗
Psychological Relativity	Q: During a stressful exam, one hour feels like how long? A. One minute ✗ B. One day ✗ **C. Ten minutes** ✓ D. Two hours ✗
Cognitive Relativity	Q: When waiting in line for 10 minutes, how long does it feel? A. Five minutes ✗ B. Ten minutes ✗ **C. One hour** ✓ D. One day ✗
Mixed Relativity	Q: When experiencing a new and exciting event for one hour, how long does it feel? A. **One minute** ✓ B. One hour ✗ C. Two hours ✗ D. One day ✗

Figure 2: Examples of temporal-relative categories in TempBench.

Dhingra et al., 2022), these studies tend to overlook the more nuanced characteristics of temporal perception. The complexity of time perception extends beyond mere calculations and relations, encompassing subjective, context-dependent elements that pose unique challenges for LLMs.

To address the aforementioned challenges, we investigate a pivotal question: *Can LLMs effectively handle a variety of temporal relativity tasks?* To explore this question, we introduce the TEMPoral-relative BENCHmark (TempBench), a comprehensive evaluation dataset meticulously designed to assess fine-grained temporal perception. TempBench builds upon prior research into human perception and temporal understanding (Bardon, 2024; Wei et al., 2023; Tan et al., 2023; Yang et al., 2023; Son and Oh, 2023), providing a structured framework for evaluating LLMs' performance in temporal relativity. The benchmark is systematically categorized into 4 key aspects, reflecting the broad spectrum of higher-level cognitive processing. Each aspects within TempBench comprises one or more subtasks to evaluate the diversity and complexity of temporal understanding across multiple levels. Different from previous temporal reasoning tasks (Wang and Zhao, 2023; Chu et al., 2023), our benchmark, generated through a synthesized data pipeline, provides a fine-grained and precise evaluation of LLMs' temporal perception abilities. An illustration of TempBench is shown in Figure 2.

Overall, we make the following contribution in this paper:

- We publish TempBench, a novel temporal-relative benchmark grounded in a newly proposed taxonomy of temporal perception question types.
- A detailed methodology for the synthesized dataset pipeline is presented encompassing high-quality data construction and the implementation of quality check of temporal relativity.
- We conduct a comprehensive evaluation of LLMs on TempBench, offering valuable insights in addressing temporal-perceiving questions, particularly in the nuanced understanding of temporal relativity.

2 Temporal-relative Task

2.1 Definition

Temporal relativity refers to the phenomenon that the perception of time changes according to the different states, situations and reference frame of individuals. Typically, temporal during of perception t_i can be shorter than that t_j of fact when a person experiences a positive or concentrating event. Conversely, t_i can be longer than t_j when a person undergoes a negative or unfocused experience. For instance, as shown in Figure 1, the event(waiting in line for 10min) is a tedious action. The model requires selecting a time interval that is longer than 10 minutes and falls within the range of 10 minutes to one day. This involves not only identifying the appropriate time frame (e.g., one hour) but also classifying it correctly within the specified bounds, requiring careful consideration of how the event aligns with both subjective and objective measures of time.

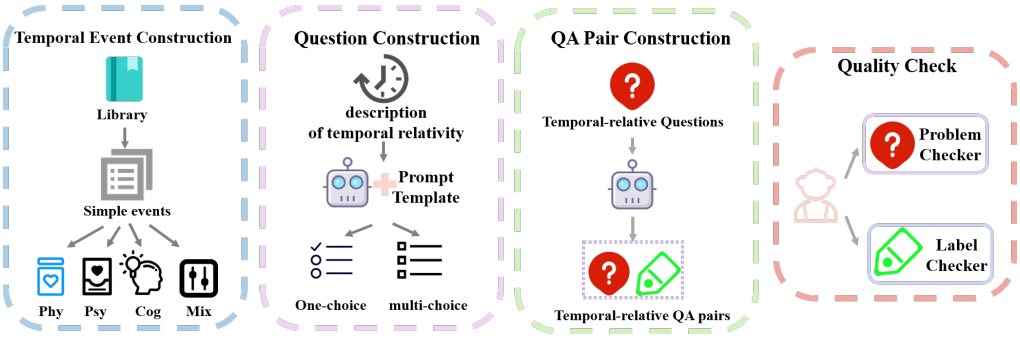

Figure 3: An illustration of TempBench construction. Firstly, we extract simple events with temporal relativity in Wikidata. Secondly, one-choice and multi-choice questions are constructed with description of temporal relativity by LLMs. Each question contains 4 suitable options. Thirdly, based on temporal-relativity questions, annotated answers for 4 different types are created by LLMs. Finally, temporal-relative QA pairs are checked by human annotation.

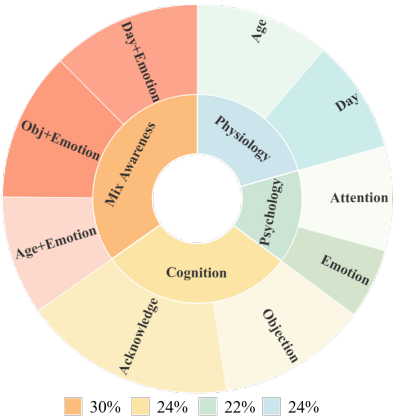

Figure 4: The distribution of our temporal-relative benchmark TempBench. It introduces the sub-tasks of each categories in TempBench.

2.2 Temporal-Relative Categories

Building on temporal relativity, TempBench encompasses 4 aspects from bodily rhythms (physiological) to emotional and mental states (psychological, cognitive and mixture) (Graziani et al., 2023; Buhusi and Meck, 2005; Hodroj et al., 2024; Bigg et al., 2024). The data distribution is shown in Figure 4):

- *Physiological relativity* represents the varying temporal interpretations of the same event by individuals with different identities. This set of tasks is crucial for evaluating the model's understanding of temporal relativity concerning core age and biological rhythms. For example, children typically perceive time as passing more slowly, while adults, especially the elderly, often feel time passes more quickly. This may be related to physiological changes and life experiences.

- *Psychological relativity* focuses on the temporal perception influenced by psychological states. These tasks primarily assess the model's understanding of different temporal perceptions influenced by emotions and attention. For instance, time seems to fly when people are in a positive mood or focused on a specific task.

- *Cognitive relativity* relies on the understanding of events from a human cognitive perspective. This set of tasks focuses on human cognition, where subjective time estimates may vary due to memory and different scenarios. For example, new experiences often make time feel slower because the brain processes more information and memories. When a person is waiting for a line, he tends to overestimate the time elapsed.

- *Mixed relativity* is a combination of the above types, presenting a particularly challenging scenario due to the complexity of real-world temporal relativity. A combination of the three types mentioned above. This category is particularly challenging due to the complexity across multiple tasks, requiring comprehensive temporal relativity reasoning. For instance, students perceive time more fast during holidays compared to adults, which relates to psychological relativity.

2.3 Dataset Construction

The overall process of TempBench construction is depicted in Figure 3. Following the pipeline of

Question Template
When you are [E] for [T], how long does you feel?
When you are [E], how long does [T] feel?
When you engrossed in [E] for [T], how long does you feel?
When you are [E] during [T], how long does you feel?

Figure 5: The illustration of simple templates.

prior work (Tan et al., 2023; Virgo et al., 2022), data construction is divided into 3 key steps:

- *Temporal Event Construction* We utilize Wikidata (Vrandečić and Krötzsch, 2014) as our knowledge source to extract simple events related to temporal relativity, such as children going to school or traveling. These events are manually classified into respective categories of temporal relativity.

- *Question Construction* Following Zhou et al. (2019) and Hoffman and Deffenbacher (1992), we structure time-relative events and compare its temporal relativity to develop questions. There are 2 primary approaches: (1) construct temporal-relative questions with simple templates and description of temporal relativity considering the temporal events, shown in Figure. 5. (2) directly classify questions into one-choice questions and multi-choice questions concerning the types of temporal relativity. For example, Mixture temporal relativity is a complex temporal perception that demonstrate the multiple aspects of temporal understanding. Therefore, we set the questions type of Mixture temporal relativity into multi-choice questions.

- *Question-Answer Pair Construction* We primarily employ GPT-4 with simple instruction to generate question-answer pairs. The correct answer strictly adhere to the rule of temporal comparison, while the incorrect answers are formed through random temporal during (1 minute, 10 minutes, 1 hour, etc.).

2.4 Quality Check

Following Fatemi et al. (2024), we conduct multiple rounds of benchmark quality checks to verify: (1) the accuracy of correct answer, and (2) the clarity of generated questions. The process is to confirm the generated questions are sufficient to produce results comparable to those obtained by humans through temporal perception.

3 Experiments and Analysis

3.1 Experiments Setting

To evaluate the performance of LLMs on temporal-relative tasks, we conducted experiments across multiple models to gain a deeper understanding of their performance on fine-grained temporal perception tasks. We selected an equal number of samples from each temporal perception task to ensure a balance in question type and complexity. In the experiment, we aim to address the following questions:

- How well do LLMs perform in answering questions related to temporal relativity?

- What types of temporal-relative questions are more difficult or easier for LLMs to answer?

Evaluated Models: We evaluate LLMs on the TempBench benchmark, categorized into 2 main groups: open-source models (Llama 3 8B (Dubey et al., 2024), Llama 2 7B (Touvron et al., 2023), Llama 2 13B (Touvron et al., 2023), Mistral 7B (Jiang et al., 2023) and Qwen 7B (Bai et al., 2023)) and closed-source models (GPT-4 (Achiam et al., 2023)). Each model is accessed using the appropriate API keys; GPT-4 is accessed via the OpenAI API, while the Llama models are accessed through Huggingface. Considering the constraints of API cost, we randomly selected 200 samples from each category. For categories with fewer than 200 samples, all available instances were used.

Metrics: Previous evaluations of temporal reasoning used Exact Match (EM) and token-level F1 scores (Rajpurkar et al., 2016; Kwiatkowski et al., 2019), which tend to overestimate accuracy by considering the highest score across all possible answers. In this work, we employ a stricter accuracy metric (Acc) (Zhong et al., 2022), which calculates correctness only when the predicted answer exactly matches the golden answer. This metric is applied to one-choice questions in TempBench. Additionally, the fixed accuracy are calculated for multi-choice questions, where one of the corrected choices can be selected.

3.2 Main Results

LLMs partially grasp temporal relativity Our analysis in Table 1 reveals that GPT-4 consistently

	Physiological		Psychological		Cognitive		Mix	Average
	Age	Day	Atten	Emo	Ack	Obj		
Human	0.9817	0.9579	1.0000	0.9241	0.9496	0.9738	0.9824	0.9671
GPT-4	**0.8049**	**0.7662**	**0.6871**	**0.7403**	**0.7887**	**0.7344**	**0.6251**	**0.7352**
Llama2 7B	0.2582	0.2237	0.0000	0.2979	0.7029	0.6937	0.2754	0.3503
Llama2 7B (5-shot)	0.2642	0.2061	0.2087	0.3208	0.7216	0.7095	0.2971	0.3897
Llama2 7B (5-shot+CoT)	0.3026	0.2583	0.2408	0.3727	0.7549	0.7324	0.3936	0.4365
Llama2 13B	0.5715	0.6082	0.0000	0.5495	0.5841	0.7178	0.5168	0.5068
Llama2 13B (5-shot)	0.5867	0.6492	0.3716	0.5628	0.5659	0.6927	0.5483	0.5682
Llama2 13B (5-shot+CoT)	0.6506	0.7031	0.4592	0.6037	0.6056	0.7129	0.5528	0.6126
Llama3 8B	0.2333	0.4342	0.2571	0.4681	0.5427	0.5014	0.4049	0.4060
Llama3 8B (5-shot)	0.2683	0.4427	0.2861	0.4418	0.5706	0.5360	0.4427	0.4269
Llama3 8B (5-shot+CoT)	0.3005	0.5037	0.3590	0.5036	0.6028	0.5824	0.4691	0.4744
Mistral 7B	0.7218	0.1316	0.5429	0.6157	0.5168	0.4082	0.3524	0.4699
Mistral 7B (5-shot)	0.7338	0.2084	0.5473	0.6352	0.5260	0.4718	0.3925	0.5021
Mistral 7B (5-shot+CoT)	0.7652	0.2764	0.5809	0.6728	0.5721	0.5384	0.4291	0.5478
Qwen2 7B	0.2667	0.5000	0.6857	0.7021	0.5879	0.2161	0.4708	0.4899
Qwen2 7B (5-shot)	0.2821	0.5375	0.6618	0.6943	0.6027	0.2561	0.4854	0.5028
Qwen2 7B (5-shot+CoT)	0.3168	0.5531	0.6992	0.7416	0.6481	0.2755	0.5028	0.5330

Table 1: Performance comparison of different LLMs on TempBench though accuracy metric. **Bold scores** indicate superior performance compared to others LLMs. The background colors of pink , yellow and blue represent the best score of each task in temporal relativity, respectively. The underlined values with various color are the sub-optimal results of temporal relativity tasks. Overall, human performance serves as an upper bound, while GPT-4 consistently outperforms other LLMs under both zero-shot and few-shot setting across temporal-relative tasks. The experimental results illustrate that there exists significant room for improvement on temporal perception.

surpasses other LLMs across 7 temporal-relative categories, maintaining a performance lead of over 10%. However, despite being the best-performing model among all LLMs, GPT-4 still lags behind human performance by 18%. The result indicates there has a significant room for improving the temporal-relative capabilities of LLMs. In the zero-shot setting, LLMs face the challenge of solving attention-relative task without any prior examples or context. Llama2 7B struggles significantly with an average score of 0.3503, underperforming in temporal-relative tasks like attention relativity (0.0000) and day relativity (0.2582). While securing the second-highest average score of 0.5068 in temporal relativity task, Llama2 13B falters notably in psychological relativity task. Specifically, Llama2 13B struggles to accurately understand human concepts of time perception related to attention. The results indicates Llama 2 7B and 13B exist a misalignment in understanding temporal relativity.

In mixed relativity scenarios, Llama2 7B and Mistral 7B still encounter difficulties in comprehending temporal perception across multiple contexts compared to single-scenario tasks, obtaining the scores of 0.2754 and 0.3524. Furthermore, Qwen with 7B parameters, having the same parameter size as Llama2 7B, performs significantly better, particularly in psychological relativity tasks. These experimental results demonstrate that vari-

ations in training data may enhance a model's capability to comprehend and process temporal perception. Interestingly, despite Llama2 7B has the smaller size, it outperforms the larger Llama 2 13B in cognitive awareness tasks. The observation highlights that a larger model size does not inherently equate to superior performance, indicating several factors could contribute to the outcome.

3.3 Simple Investigations for Improvement

In order to improve the performance of temporal relativity in LLMs, we utilize the prompting engineering incorporating both standard prompting and chain-of-thought prompting. These evaluations are performed in zero-shot and few-shot setting, enabling a comprehensive analysis of performance across varying levels of task exposure.

Standard Prompting Following Brown (2020) and Kojima et al. (2022), questions are presented directly without the need for additional steps in the prompt. Considering the following example from the exemplar answers (4 choices) are provided alongside the question from physiological relativity. For few-shot standard prompting, the 5 exemplar answers (one of 4 choices) are provided within the given question. The overall standard procedure across all tasks can be categorized into: (1) Direct Question-answering: Pose the question directly to the model without any intermediary steps or additional guidance. (2) Answer Solicitation: Request the model to choose and provide the most appropriate answer based on the information given.

Chain-of-Thought Prompting In contrast, zero-shot CoT learning takes inspiration from Wei et al. (2022) through the instruction "Choose the correct answer by thinking step by step". For few-shot CoT, we manually craft the step-by-step process for 5-shot exemplars in the development set. The process is as follows: (1) Understand the temporal information. (2) Identify key events from the question. (3) Compare the temporal during in choices with temporal information. (4) Conclusion. Given the temporal information and questions, the chosen answer is more plausible and contextually appropriate.

3.4 Further Analysis on TempBench

The further analysis based on 5-shot and 5-shot+CoT is shown in Table 1. In 5-shot settings, all LLMs benefit from receiving temporal-relative examples, indicating considerable importance compared to LLM without prompting. Llama2 13B emerges as the top performer and significantly outperform other LLMs by a large margin. The experimental results demonstrate that prompt engineering further improve the LLMs with more parameter size in temporal-relative tasks. It is noteworthy that open-source LLMs exhibit a large performance decline compared to proprietary LLMs when transitioning from few-shot, few-shot+CoT and zero-shot scenarios. Llama2 7B, Mistral 7B and Qwen2 7B show gains of 10%, 6.4%, 2.6%, respectively. We contribute the performance increasing to the quality of examples. Compared to the restriction of instruction-following capability in LLMs, open-source LLMs with few-shot prompting are better approach for stimulating their temporal-relative ability.

Previous research has found that chain-of-thought prompting can enhance the temporal reasoning of LLMs (Wei et al., 2023; Kojima et al., 2022). We aim to explore the following question: *How does CoT Prompting bring consistent improvement in temporal-relative tasks?* Considering the diversity of temporal relativity, the above question has not yet been definitively answered.

CoT reasoning is consistently effective As illustrated in Table 1, introducing few-shot CoT prompting results in consistent raising, with an overall increase of 7.2%. There is a 13.25% improvement in physiological temporal relativity, while a significant rising of 3.6% in cognitive temporal relativity. In psychological relativity, there is a slight improvement of 3.625%. In few-shot CoT setting, almost all models exhibit significant improvement in overall temporal-relative tasks. For physiological relativity, We contributes this to the temporal knowledge understanding in temporal relativity in LLMs. In psychological temporal relativity, improvements mainly stem from datasets involving step-by-step temporal reasoning, indicating that CoT is more effective for implicit relativity understanding. In summary, CoT has a positive impact on cognitive and complex temporal-relative tasks.

Impact of CoT prompting across temporal-relative tasks In order to thoroughly explore the impact of CoT on various temporal-relative tasks within each categories, we undertake a manual analysis of the models' erroneous or inappropriate responses. The errors in temporal relativity are classified into 4 types:

- *Refusal to Answer*. This issue predominantly

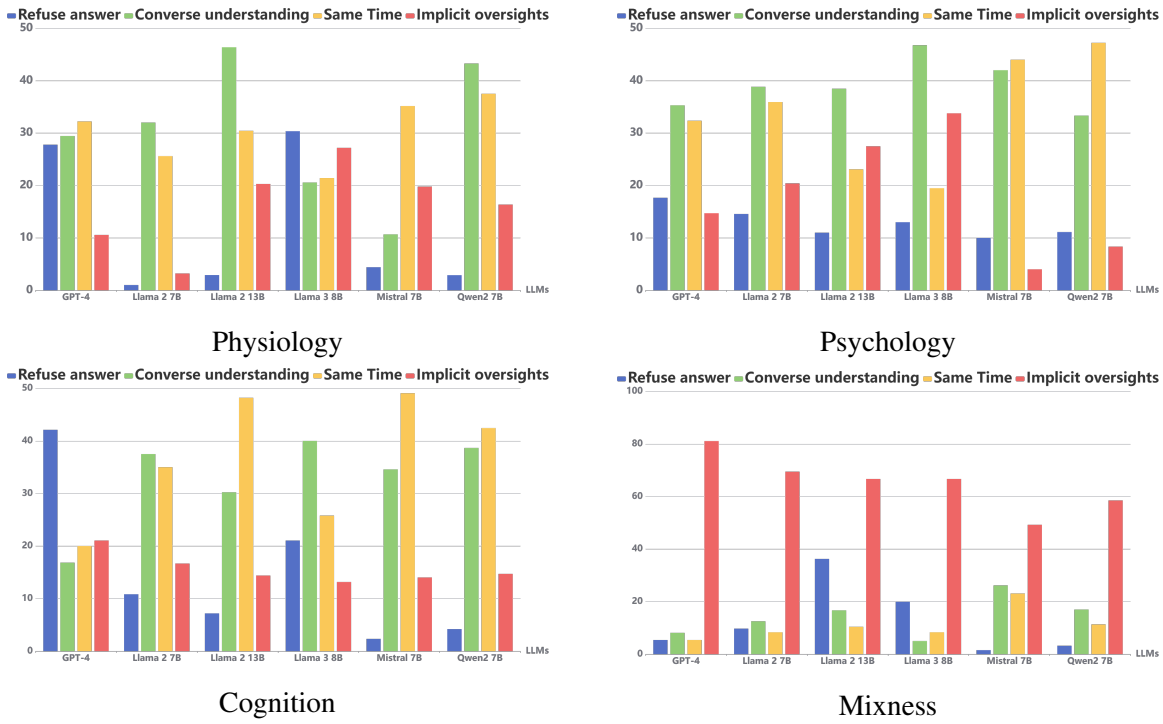

Figure 6: The error analysis of temporal-relative ability in LLMs. Overall, errors are concentrated in 2 categories: Converse Understanding and Same as the Original Time. The experimental results demonstrate that complex temporal-relative task integrates multiple aspects of temporal understanding are challenging for LLMs.

occurs in models with over 13B parameters, where responses are often laden with excessive explanations and fail to directly address the posed question.

- *Converse Understanding* Models incorrectly interpret temporal perception, confusing a shorter time duration with longer ones, and vice versa.

- *Same as the Original Time* Models demonstrate a lack of comprehension of temporal relativity by selecting the time mentioned in the question as the answer without adjustment.

- *Implicit Oversights* Models are unable to accurately understand temporal relativity in complex scenarios involving multiple temporal relativity.

Error Analysis Figure 6 demonstrates the common error types and their proportions at each task. Overall, errors are concentrated in 2 categories: Opposite Understanding and Same as the Original Time. Notably, GPT-4 and Llama3 8B frequently exhibit errors related to refusal to answer the question. LLMs with 7B parameters predominantly display Converse Understanding errors, which account for 30% of all errors in physiological relativity. Similar results are observed within psychological relativity. Additionally, in cognitive relativity, GPT-4 often opts to refuse answering these questions, whereas other models typically provide incorrect answers. In mixed relativity, all errors are concentrated in implicit oversights, representing 65.26% of total errors. For instance, *when a young person goes on vacation, their perception of time differs from that of older adults. Two distinct patterns emerge: younger individuals tend to perceive time more slowly compared to older adults, while vacations generally cause time to feel as though it passes more quickly. When these factors are combined, the accelerated perception of vacation time overrides the slower perception associated with youth, resulting in an overall perception of time speeding up.* However, most LLMs fail to recognize this nuanced interaction. The experimental results demonstrate that there remains considerable room to understand and process complex temporal relativity scenarios for LLMs.

4 Discussion

Scaling effect of model size In order to investigate how the parameter scale of models affects temporal-relative capabilities, we compare the per-

formance of a series of Llama2. As the scale of model increases, there is a notable improvement in performance. When the parameter size from 7B to 13B, Llama2 show improvements of 30%. Furthermore, when Llama2 scales up to 70B, the trend of performance follows a log-linear relationship with scale.

Challenges in Temporal-Relative Tasks The performance of all LLMs in temporal-relative tasks is unsatisfactory. A noticeable decrease is observed in mix relative task compared to other temporal tasks. This is because the mix relative task necessitates a multi-step temporal reasoning process. It first unifies multiple time units and event, and subsequently engages in time comparison, while other temporal-relative tasks can be completed with a single reasoning step. The complexity of these multi-step processes contributes to the observed performance drop in this category.

5 Relative Work

Temporal perception, as an integral component of temporal reasoning, has its foundations in the evolving techniques of temporal domain datasets. While existing work has significantly advanced the understanding of temporal reasoning in Large Language Models (LLMs), a notable gap remains in the exploration of how these models handle the intricacies of human temporal perception. Temporal perception goes beyond basic reasoning, requiring a model to understand subjective experiences of time across varying contexts. Current invaluable benchmarks do not fully capture this complex cognitive process, thus highlighting the need for a more refined evaluation of temporal perception capabilities in LLMs. Existing temporal reasoning datasets have also laid the groundwork for a detailed evaluation of temporal perception capabilities within the current paradigm of LLMs.

Temporal Reasoning Benchmarks In the field of temporal reasoning, previous datasets have emerged to address various challenges. As early as 2003, TimeBank (Pustejovsky et al., 2003) has focused primarily on temporal relationships. Relying on the TimeBank, TempEval-3 (UzZaman et al., 2012) has expanded this scope by introducing multiple tasks, including temporal entity extraction and relation extraction. TimeQA has established the first dataset aimed at studying the comprehension of time-sensitive facts. Recently, there has been a surge in the development of temporal reasoning QA datasets, such as MCTACO (Zhou et al., 2019), TEMPLAMA (Dhingra et al., 2022), TEMPREASON (Tan et al., 2023), and MenatQA (Wei et al., 2023). Each of these datasets addresses specific challenges in temporal reasoning, ranging from commonsense understanding to complex temporal logic.

Limitations of Existing Benchmarks However, these datasets are limited in their coverage of real-world temporal complexity. They do not sufficiently evaluate LLMs on tasks involving temporal relativity, where the perception of time can change depending on contextual factors, such as attention, stress, or cognitive load. This represents a critical gap, as human temporal perception is not merely a matter of understanding events in sequence, but also involves interpreting how those events feel over time. In contrast, our benchmark focuses on human temporal perception, addressing different dimensions of temporal understanding, and provide a more fine-grained and comprehensive evaluation of temporal reasoning than previous benchmarks.

Conclusion

In this paper, we introduce TempBench, a comprehensive dataset specifically designed to facilitate the exploration of temporal relativity in LLMs. It is the first dataset containing fine-grained temporal factors that can be used as an evaluation benchmark for assessing the time relativity. Extensive experiments have revealed a substantial gap between the performance of LLMs and that of humans. Moreover, the parameter size of LLMs substantially influences their capacity for temporal relativity. We also discover the types of Converse Understanding and Same as the Original Time are the crucial challenges in LLMs. We hope that TempBench will serve as a foundation for further advancements in enhancing temporal relativity in LLMs.

Limitation

Although TempBench provides a comprehensive standard for evaluating temporal perception, it has certain limitations. One of the primary issues is that the dataset construction mainly relies on seed events from Wikidata, which restricts the scope of coverage. In future iterations, we plan to expand the dataset and explore the post training of models. Additionally, effectively enhancing the temporal perception of LLMs is unfocused. Therefore, we intend to develop sophisticated prompts,

with the goal of improving the temporal perception of LLMs.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (No. 62441602).

Ethics Statement

To the best of our knowledge, our work does not pose any direct societal risks or ethics concerns. Our data is built upon Wikidata. There is no explicit detail that leaks an annotator's personal information. The dataset has low risks of containing sentences with toxicity and offensiveness. Additionally, we value the contributions of all individuals involved in the annotation process, providing them with appropriate recognition and ensuring fair compensation for their efforts.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Adrian Bardon. 2024. *A brief history of the philosophy of time*. Oxford University Press.

Anthony Bigg, Andrew J Latham, Kristie Miller, and Shira Yechimovitz. 2024. Episodic imagining, temporal experience, and beliefs about time. *Philosophy and Phenomenological Research*.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Catalin V Buhusi and Warren H Meck. 2005. What makes us tick? functional and neural mechanisms of interval timing. *Nature reviews neuroscience*, 6(10):755–765.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

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*.

Bhuwan Dhingra, Jeremy R Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W Cohen. 2022. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Bahare Fatemi, Mehran Kazemi, Anton Tsitsulin, Karishma Malkan, Jinyeong Yim, John Palowitch, Sungyong Seo, Jonathan Halcrow, and Bryan Perozzi. 2024. Test of time: A benchmark for evaluating llms on temporal reasoning. *arXiv preprint arXiv:2406.09170*.

Ernesto Graziani, Francesco Orilia, Elena Capitani, and Roberto Burro. 2023. Common-sense temporal ontology: an experimental study. *Synthese*, 202(6):193.

Batoul Hodroj, Andrew J Latham, and Kristie Miller. 2024. The moving open future, temporal phenomenology, and temporal passage. *Asian Journal of Philosophy*, 3(1):1–20.

Robert R Hoffman and Kenneth A Deffenbacher. 1992. A brief history of applied cognitive psychology. *Applied Cognitive Psychology*, 6(1):1–48.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. 2003. The timebank corpus. In *Corpus linguistics*, volume 2003, page 40. Lancaster, UK.

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

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, and Hao Wang. 2024. Continual learning of large language models: A comprehensive survey. *arXiv preprint arXiv:2404.16789*.

Jungbin Son and Alice Oh. 2023. Time-aware representation learning for time-sensitive question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 70–77.

Zhaochen Su, Juntao Li, Jun Zhang, Tong Zhu, Xiaoye Qu, Pan Zhou, Yan Bowen, Yu Cheng, et al. 2024a. Living in the moment: Can large language models grasp co-temporal reasoning? *arXiv preprint arXiv:2406.09072*.

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

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2023. Towards benchmarking and improving the temporal reasoning capability of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14820–14835.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Naushad UzZaman, Hector Llorens, James Allen, Leon Derczynski, Marc Verhagen, and James Pustejovsky. 2012. Tempeval-3: Evaluating events, time expressions, and temporal relations. *arXiv preprint arXiv:1206.5333*.

Felix Virgo, Fei Cheng, and Sadao Kurohashi. 2022. Improving event duration question answering by leveraging existing temporal information extraction data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4451–4457.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

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

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.

Yifan Wei, Yisong Su, Huanhuan Ma, Xiaoyan Yu, Fangyu Lei, Yuanzhe Zhang, Jun Zhao, and Kang Liu. 2023. Menatqa: A new dataset for testing the temporal comprehension and reasoning abilities of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1434–1447.

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

Sen Yang, Xin Li, Lidong Bing, and Wai Lam. 2023. Once upon a time in graph: Relative-time pretraining for complex temporal reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11879–11895.

Chenhan Yuan, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. Back to the future: Towards explainable temporal reasoning with large language models. In *Proceedings of the ACM on Web Conference 2024*, pages 1963–1974.

Yunxiang Zhang and Xiaojun Wan. 2024. Situatedgen: incorporating geographical and temporal contexts into generative commonsense reasoning. *Advances in Neural Information Processing Systems*, 36.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Victor Zhong, Weijia Shi, Wen-tau Yih, and Luke Zettlemoyer. 2022. Romqa: A benchmark for robust, multi-evidence, multi-answer question answering. *arXiv preprint arXiv:2210.14353*.

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. *arXiv preprint arXiv:1909.03065*.