Human-LLM Divergence in Temporal Reasoning under Mixed Time Expressions

Feifei Sun[1] Ziyi Tong[1]
Teeradaj Racharak[2] Minh Le Nguyen[1]
[1]Japan Advanced Institute of Science and Technology (JAIST)
[2]Advanced Institute of So-Go-Chi (Convergence Knowledge) Informatics, Tohoku University
{feifei.sun, ziyi.tong, nguyenml}@jaist.ac.jp
{racharak.teeradaj.c3}@tohoku.ac.jp

Abstract

Understanding temporal relations in natural language often requires integrating absolute and relative time expressions within a single narrative. While recent work has shown that mixed time expressions degrade large language model (LLM) performance, much less is known about how such conditions affect human temporal reasoning — and whether model behavior aligns with human judgments. In this work, we conduct a controlled human–LLM comparison on a sentence-level temporal event ordering task under two settings: absolute-time (AT) narratives with explicit timestamps, and mixed-time (MT) narratives where absolute references are partially replaced by relative or event-anchored expressions. Our results reveal a clear divergence between human and model reasoning. Although both humans and LLMs perform reliably under AT conditions, human performance drops more sharply under MT settings, while several frontier LLMs maintain relatively stable ordering consistency. This contrast suggests that mixed-time narratives impose substantial cognitive difficulty for humans, whereas LLMs may rely on alternative strategies that are less sensitive to temporal ambiguity. These findings highlight that strong performance in mixed-time temporal reasoning does not necessarily reflect human-like understanding, underscoring the importance of human baselines when interpreting model robustness in temporally underspecified narratives.

1 Introduction

Temporal reasoning is a fundamental aspect of language comprehension, enabling readers to reconstruct event sequences and interpret narratives in context. In natural discourse, however, temporal information is rarely expressed in a fully explicit or linear manner. Narratives frequently combine absolute time references (e.g., in 1945), relative expressions (e.g., two years later), and event-anchored cues (e.g., shortly after the war), resulting in timelines that are non-linear and partially implicit. Such mixed-time narratives pose challenges not only for computational models, but also for human readers, who must integrate distributed and often underspecified temporal cues across discourse.

Recent advances in LLMs have demonstrated promising performance on temporal reasoning tasks. However, most existing evaluations emphasize settings with explicit temporal anchors or simplified temporal relations, which do not fully reflect the structure of real-world narratives found in biographies, historical accounts, or cross-cultural texts. Although several benchmarks have begun to explore more complex temporal phenomena [1, 2, 3], comparatively little attention has been paid to how **human** temporal reasoning behaves under the same mixed-time conditions, or how closely model predictions align with human judgments. Prior work has identified anchoring effects and weaknesses in handling relative temporal expressions in LLMs [4, 5], yet it remains unclear whether strong model performance under mixed temporal cues reflects human-like reasoning or alternative strategies that diverge from human temporal interpretation.

In this work, we focus on this gap by conducting a controlled **human–LLM comparison** on a sentence-level temporal event ordering task. Using narratives that systematically vary between AT and MT conditions, we evaluate how humans and LLMs differ in their ability to re-

cover global chronological structure. Our results reveal a clear divergence: while both humans and LLMs perform reliably under AT settings, human performance degrades more sharply under mixed-time narratives, whereas several frontier LLMs maintain relatively stable ordering consistency. This contrast suggests that mixed-time temporal reasoning introduces cognitive difficulty for humans that is not mirrored by model behavior, highlighting fundamental differences in how temporal uncertainty is handled.

These findings underscore the importance of incorporating human baselines into temporal reasoning evaluation and caution against interpreting strong model performance in mixed-time settings as evidence of human-like temporal understanding.

2 Task and Experimental Setup

2.1 Task Definition

We study temporal reasoning through a sentence-level event ordering task. Given a small set of sentences describing events from a narrative, the task is to recover their correct chronological order. Unlike question answering or temporal relation classification, this formulation directly evaluates whether a system can reconstruct a global temporal structure from natural language descriptions.

Each instance consists of four event sentences sampled from a single narrative and presented in shuffled order. Systems are required to output a permutation of sentence indices corresponding to the inferred temporal order. This setup allows for a controlled comparison between human judgments and model predictions without introducing additional confounding factors such as question interpretation or answer generation.

2.2 Temporal Settings

We evaluate temporal reasoning under two conditions that differ in how temporal information is expressed:

Absolute-Time (AT). All event sentences contain explicit temporal anchors, such as calendar years or dates (e.g., in 1945, in March 2003). This setting represents narratives where chronological relations can be inferred primarily from surface-level timestamps.

Mixed-Time (MT). Some absolute time expressions are replaced with relative or event-anchored references (e.g., two years later, after the war), while preserving the underlying event order. As a result, temporal relations must be inferred by integrating multiple cues across the narrative rather than relying solely on explicit anchors.

By comparing performance across these two settings, we isolate the effect of mixed temporal expressions on temporal reasoning for both humans and language models.

2.3 Temporal Settings

Table 1 summarizes the characteristics of the 100-sample subset used in our experiments. The AT and MT settings contain a comparable number of events per instance, ensuring that differences in performance cannot be attributed to event count alone. At the same time, the two settings differ systematically in the distribution of temporal expression types and granularity. In particular, the MT setting contains a higher proportion of relative expressions and coarse-grained temporal cues, reflecting increased temporal underspecification. This design allows us to examine how humans and language models respond differently to mixed temporal signals under otherwise comparable narrative complexity.

Setting	AT	MT
#Samples	30	70
Avg Events	4.4	4.5
High Granularity (%)	53.3	42.9
Low Granularity (%)	46.7	57.1
Abs : Rel	53 : 47	43 : 57

Table 1: Statistics of the 100-sample subset for human annotation and probing experiments. This subset maintains temporal diversity while ensuring cognitive feasibility for human reasoning. *Avg Events* denotes the average number of annotated events per passage. *High Granularity* = expressions specifying *year+month+day* and *year+month*, and *Low Granularity* = expressions specifying only the *year*.

2.4 Human and Model Evaluation

To examine differences between human and model temporal reasoning, we evaluate both under identical task conditions. Human performance is measured using annotations from three trained annotators, each independently producing a complete event order for every instance. Model performance is evaluated using representative fron-

Setting	Model	EM	Kendall's τ
AT	DEEPSEEK-REASONER	**0.63**	0.65
	DEEPSEEK-V3	0.46	0.47
	GPT-4	0.60	0.69
	GPT-3.5-TURBO	0.13	0.30
	QWEN2.5-7B	0.03	0.12
	QWQ-32B	**0.63**	0.66
	Annotator 1	0.56	0.62
	Annotator 2	0.56	**0.73**
	Annotator 3	0.56	0.63
MT	DEEPSEEK-REASONER	**0.60**	**0.65**
	DEEPSEEK-V3	0.41	0.49
	GPT-4	0.42	0.46
	GPT-3.5-TURBO	0.16	0.08
	QWEN2.5-7B	0.15	0.20
	QWQ-32B	0.59	**0.65**
	Annotator 1	0.30	0.32
	Annotator 2	0.34	0.43
	Annotator 3	0.42	0.50

Table 2: Overall model and human performance on event ordering under AT and MT conditions. EM = exact match accuracy; Kendall's τ measures rank correlation between predicted and gold orders.

tier large language models, selected to reflect current high-performing systems rather than to provide exhaustive coverage.

All systems are evaluated using the same input format and output requirements. Performance is measured primarily using Kendall's τ, which captures the degree of agreement between predicted and gold-standard event orders by accounting for pairwise ordering consistency. This rank-based metric allows for meaningful comparison between humans and models even when full sequence agreement is not achieved.

3 Results: Human–LLM Comparison

3.1 Overall Performance under Absolute and Mixed Time

We first examine overall temporal ordering performance for humans and language models under AT and MT settings. Table 2 reports Kendall's τ between predicted and gold-standard event orders.

Under AT conditions, humans and frontier LLMs achieve comparably high performance, indicating that explicit temporal anchors enable reliable reconstruction of global event order. Human annotators show strong consistency with gold orders, and several LLMs reach similar levels of agreement.

When moving to MT settings, performance degrades across all systems, but the magnitude of this degradation differs substantially. Human performance exhibits a pronounced drop in Kendall's τ, whereas several frontier LLMs retain relatively stable ordering consistency. This contrast suggests that mixed temporal expressions introduce difficulty that disproportionately affects human temporal reasoning, even when overall narrative complexity remains comparable.

System	AT τ	MT τ
DEEPSEEK-REASONER	**0.84**	**0.77**
QwQ-32B	0.82	0.76
GPT-4	0.75	0.65
DEEPSEEK-V3	0.68	0.68
GPT-3.5	0.51	0.45
QWEN2.5-7B	0.37	0.33
Annotator 1	0.80	0.55
Annotator 2	0.75	0.62
Annotator 3	0.83	0.67

(a) Model–Gold and Human–Gold Kendall's τ

System	AT τ	MT τ
DEEPSEEK-REASONER	**0.84**	**0.59**
QwQ-32B	0.83	0.58
GPT-4	0.72	0.46
DEEPSEEK-V3	0.71	0.53
GPT-3.5	0.52	0.39
QWEN2.5-7B	0.40	0.28

(b) Direct Model–Human Kendall's τ

Table 3: Human–model agreement measured by Kendall's τ across AT and MT settings.

3.2 Divergent Sensitivity to Mixed-Time Expressions

Beyond overall performance, humans and LLMs also differ in how their predictions align under mixed temporal cues. Figure 1 illustrates human–LLM agreement patterns under AT and MT settings. While AT narratives exhibit substantial overlap between human and model predictions, MT narratives lead to a marked reduction in overlap and an increase in divergent judgments. This qualitative shift indicates that humans and models respond differently to temporal ambiguity introduced by mixed-time expressions.

This divergence is further supported by agreement statistics reported in Table 3. Although some LLMs maintain relatively high correlation with gold-standard orders under MT settings, direct human–model agreement decreases

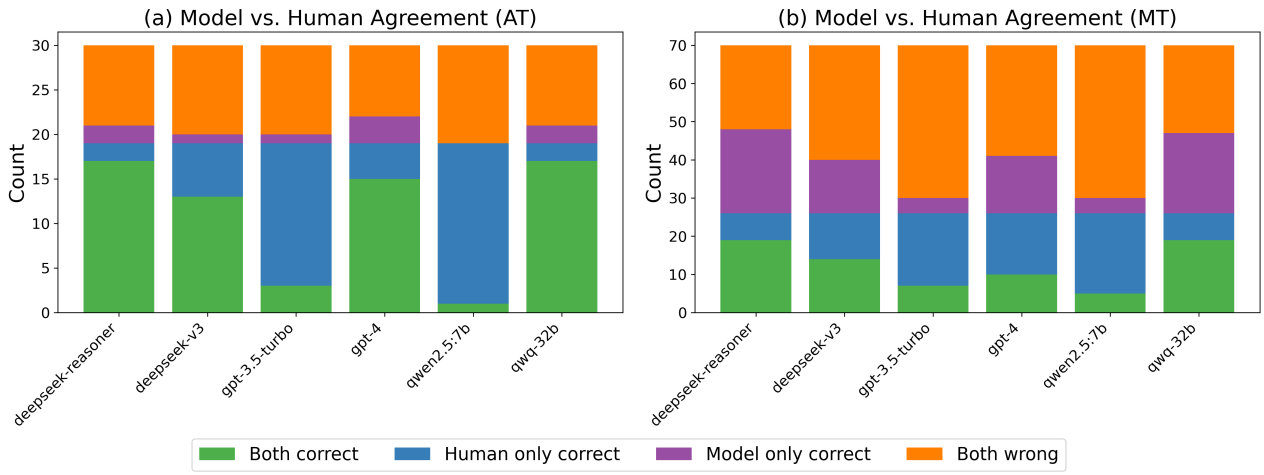

Figure 1: **Model vs. Human Agreement under AT and MT settings.** Stacked bar plots show the distribution of prediction outcomes for each model: *Both correct* (green), *Human only correct* (blue), *Model only correct* (purple), and *Both wrong* (orange). (a) **Absolute Time (AT):** Models such as DEEPSEEK-REASONER, GPT-4, and QWQ-32B achieve the highest human–model overlap, while DEEPSEEK-V3 and QWEN2.5-7B show more human-only correct cases, indicating weaker recovery of masked time expressions. (b) **Mixed Time (MT):** All models see a sharp drop in *Both correct* counts, with increased *Model only correct* and *Both wrong* cases. The gap between human and model judgments widens under ambiguous or relative time cues, especially for QWEN2.5-7B and DEEPSEEK-V3.

more noticeably. These results indicate that similar performance levels do not necessarily correspond to aligned temporal judgments. Instead, humans and LLMs appear to rely on different strategies when integrating relative or event-anchored temporal cues.

Taken together, these findings demonstrate that mixed-time temporal reasoning affects humans and LLMs in qualitatively different ways. Observed model robustness under mixed temporal conditions should therefore be interpreted with caution, as it may reflect alternative inference strategies rather than human-like temporal understanding.

4 Discussion and Conclusion

This study examined human and LLM temporal reasoning under narratives that mix absolute and relative time expressions. By directly comparing human judgments with model predictions on the same event ordering task, we revealed a clear divergence in how mixed-time temporal cues are processed.

A central finding is that mixed-time narratives introduce cognitive difficulty that disproportionately affects human reasoning. While humans perform reliably when explicit temporal anchors are available, their ability to maintain a coherent global timeline degrades more sharply once temporal relations must be inferred across relative or event-anchored expressions. This sensitivity likely reflects the reliance of human temporal reasoning on explicit anchors and discourse-level integration, which become more demanding under temporal underspecification.

In contrast, several frontier LLMs exhibit comparatively stable performance under mixed-time conditions. Importantly, this stability should not be interpreted as evidence of human-like temporal understanding. Our agreement analyses show that similar performance levels do not necessarily correspond to aligned temporal judgments between humans and models. Instead, LLMs may rely on alternative inference strategies, such as learned narrative regularities or surface-level temporal heuristics, that allow them to maintain ordering consistency without resolving temporal ambiguity in the same way humans do.

These findings have important implications for temporal reasoning evaluation. First, they highlight the necessity of incorporating human baselines when interpreting model robustness, particularly in settings where temporal information is implicit or underspecified. Second, they caution against equating strong model performance with cognitive plausibility, as divergent reasoning behaviors may underlie similar quantitative results.

References

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

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

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

[4] Yiming Huang, Biquan Bie, Zuqiu Na, Weilin Ruan, Songxin Lei, Yutao Yue, and Xinlei He. An empirical study of the anchoring effect in llms: Existence, mechanism, and potential mitigations. **arXiv preprint arXiv:2505.15392**, 2025.

[5] Shuang Chen, Yining Zheng, Shimin Li, Qinyuan Cheng, and Xipeng Qiu. Perceive the passage of time: A systematic evaluation of large language model in temporal relativity. In **Proceedings of the 31st International Conference on Computational Linguistics**, pages 8304–8313, 2025.