Temporal Reasoning Breaks under Mixed Time Expressions in Large Language Models

Feifei Sun[1] Ziyi Tong[1] Houjing Wei[1] Cheng Peng[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, houjing.w, cpeng, nguyenml}@jaist.ac.jp
{racharak.teeradaj.c3}@tohoku.ac.jp

Abstract

Temporal reasoning remains challenging for large language models (LLMs), especially when temporal cues are implicit or mixed. We investigate sentence-level event ordering under narratives that combine absolute and relative time expressions. Using a benchmark constructed from biographical texts, we systematically replace absolute timestamps with natural relative references to create mixed-time contexts. Through controlled comparisons, we observe consistent performance degradation across diverse LLMs in mixed-time settings, even for strong models. Further analysis shows that this degradation is amplified under coarser temporal granularity and longer event sequences. These results indicate that failures in mixed-time reasoning stem from weakened temporal anchoring rather than sequence length alone.

1 Introduction

Temporal reasoning is a fundamental component of natural language understanding, supporting tasks such as event sequencing and temporal inference. Accordingly, it has been extensively studied in natural language processing, particularly within question answering (QA) benchmarks that evaluate models' ability to reason over explicit and implicit temporal information. Early datasets such as TempQuestions [1] and TimeDial [2] focus on answering questions involving explicit, implicit, and ordinal temporal cues, while later work explores temporal alignment and drift between textual narratives and structured knowledge sources [3]. More recent benchmarks, including TempReason [4], further extend temporal QA to multi-level reasoning over time–time, time–event, and event–event relations.

Despite this progress, temporal reasoning is often evaluated as part of broader reasoning pipelines, where event ordering appears only as a latent intermediate step. As a result, it remains difficult to directly assess whether large language models (LLMs) can reliably recover global chronological structure, especially in nonlinear narratives with fragmented temporal cues. This limitation is particularly salient in naturalistic texts, where absolute timestamps and relative temporal expressions frequently coexist.

Recent comprehensive benchmarks aim to broaden evaluation coverage by integrating multiple temporal reasoning skills into unified frameworks [5, 6]. However, these benchmarks predominantly assume linear narratives or rely on explicit temporal anchors. Consequently, the challenges posed by mixed time expressions—including implicit anchoring, granularity mismatch, and nonlinearity—remain underexplored.

In this work, we isolate sentence-level event ordering as a standalone task to examine temporal reasoning under mixed time expressions. We construct a benchmark based on biographical narratives, systematically replacing absolute time expressions with natural relative references while preserving the underlying event order. This controlled design enables direct comparison between absolute-time and mixed-time settings.

Our experiments show that temporal reasoning consistently degrades under mixed time expressions across diverse LLMs, including strong instruction-tuned models. Further analysis reveals that this degradation is amplified under coarser temporal granularity and longer event sequences. These results suggest that the observed failures

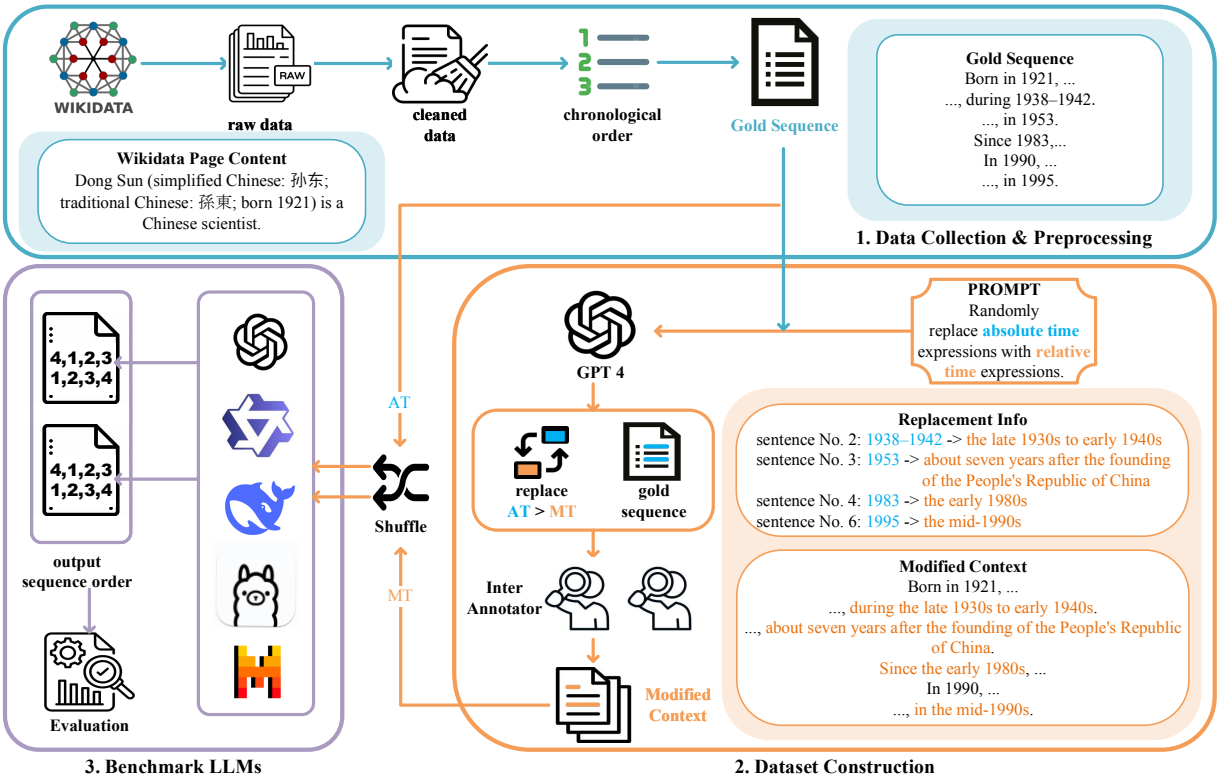


Figure 1: Framework for constructing mixed-time narratives and evaluating sentence-level event ordering. Biographical texts are converted into chronological event sequences, then partially rewritten by replacing absolute timestamps with relative expressions. Models predict the correct event order under absolute-time (**AT**) and mixed-time (**MT**) settings.

stem not from sequence length alone, but from weakened implicit temporal anchoring in mixed time systems.

2 Task and Framework Overview

This section introduces the sentence-level event ordering task studied in this work and provides an overview of the framework used to construct mixed-time narratives in a controlled manner.

2.1 Task Definition

We formulate temporal reasoning as a sentence-level event ordering task. Given a set of sentences describing events from a narrative, the model is required to predict the correct chronological order of these events. Each input consists of a shuffled sequence of event sentences, and the output is a permutation representing the inferred temporal order.

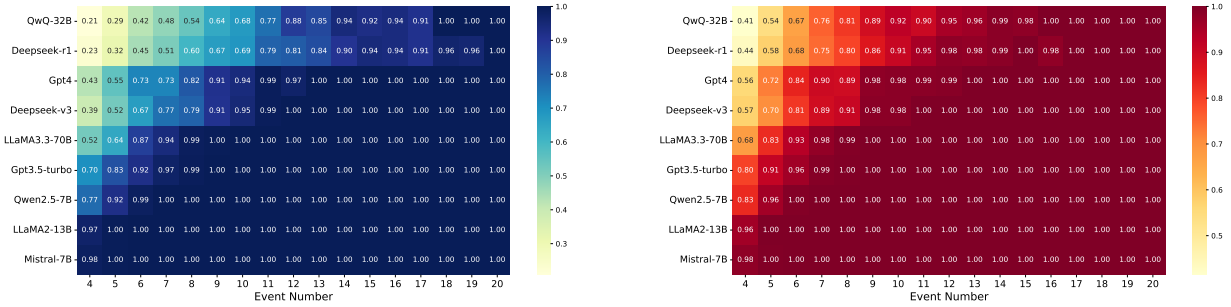
Unlike question answering formulations, this task directly evaluates a model's ability to recover global temporal structure from textual descriptions. By isolating event ordering as a standalone objective, we avoid confounding factors introduced by question interpretation or answer generation, enabling a more focused assessment of temporal reasoning ability.

2.2 Framework for Mixed-Time Construction

To study temporal reasoning under mixed time expressions, we construct narrative variants that combine absolute and relative temporal cues while preserving the underlying chronological order. Figure 1 provides an overview of the proposed framework.

Starting from biographical narratives with explicit temporal information, we first establish a gold chronological sequence based on absolute timestamps. This sequence serves as the ground truth ordering for all experimental settings.

We then generate mixed-time narratives by systematically replacing a subset of absolute time expressions with natural relative temporal references. Importantly, this transformation is performed under a controlled setting: while the surface realization of temporal cues is altered, the true event order remains unchanged. As a result, differences in model performance between absolute-time and



(a) AT Error Rates by Model and Event Number

(b) MT Error Rates by Model and Event Number

Figure 2: Comparison of AT and MT error rates across different models and event numbers. Error rate is defined as 1 - Exact Match (EM), representing the proportion of outputs that fail to exactly match the gold permutation.

mixed-time settings can be attributed to the presence of mixed temporal expressions rather than changes in narrative content or ordering.

The resulting narratives are used to evaluate large language models on the event ordering task. By comparing performance across absolute-time and mixed-time conditions, the framework enables a controlled investigation of how mixed temporal signals affect a model's ability to infer chronological structure.

3 Experimental Setup

This section describes the experimental setup used to evaluate large language models on the sentence-level event ordering task under absolute-time and mixed-time conditions.

3.1 Models

We evaluate a representative set of large language models that cover both frontier and open-source paradigms. Specifically, we include a strong proprietary model and two high-performing open-source models with different training and instruction-tuning characteristics. This selection allows us to examine whether the observed temporal reasoning failures under mixed time expressions are model-specific or systematic across model families.

All models are evaluated in a zero-shot or one-shot prompting setting, following a unified input–output format. No task-specific fine-tuning or additional supervision is applied.

3.2 Evaluation Settings and Metrics

For each narrative, models are given a shuffled list of event sentences and are required to output a predicted chronological order. We evaluate model performance under two settings: **absolute-time** (**AT**), where explicit temporal expressions are preserved, and **mixed-time** (**MT**), where a subset of absolute expressions is replaced with relative temporal references.

Model outputs are evaluated using exact match accuracy, which measures whether the predicted ordering exactly matches the gold chronological sequence. This metric provides a strict assessment of a model's ability to recover global temporal structure, as any local ordering error leads to an incorrect prediction.

To ensure fair comparison, all other aspects of the input—including narrative content and event order—are held constant across the two settings.

4 Results and Analysis

4.1 Sensitivity to Sequence Length

Figure 2 illustrates how error rates change with increasing event sequence length under AT and MT settings. For all models, error rates rise as the number of events increases, with substantially steeper growth under MT. This trend indicates that longer sequences amplify ordering errors, especially when temporal cues are implicit or relative. The results suggest that mixed time expressions reduce robustness to increased reasoning complexity.

Figure 3: MT-AT error rate increase under different time granularities for Qwen and DeepSeek models. Both show greater degradation with coarse-grained inputs, with QwQ-32B and DeepSeek-r1 most affected, suggesting reduced robustness to underspecified temporal cues.

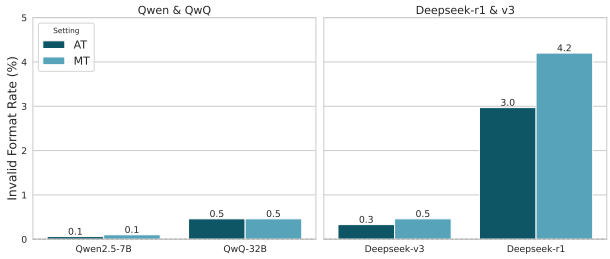


Figure 4: Invalid output rate (%) for Qwen and DeepSeek models under AT and MT. DeepSeek-r1 shows notably higher error rates, especially in MT, indicating reduced stability when processing relative time inputs.

4.2 Sensitivity to Temporal Granularity

Figure 3 shows that performance degradation under mixed-time settings is exacerbated as temporal expressions are coarsened from fine-grained to coarse-grained forms. This effect is consistently observed across model families. Notably, stronger models such as QwQ-32B and DeepSeek-r1 exhibit larger relative drops, suggesting a greater reliance on precise temporal anchoring. These results indicate that mixed time expressions undermine temporal reasoning particularly when explicit granularity cues are weakened.

4.3 Model Stability under Mixed-Time Inputs

Figure 4 reports invalid output rates for representative models under AT and MT settings. We observe that stronger reasoning variants within the same model family exhibit higher rates of format violations, particularly under mixed-time inputs. This pattern suggests a trade-off between deep semantic reasoning and adherence to structural constraints when temporal information is underspecified. The effect is more pronounced in the MT setting, indicating increased instruction-following instability induced by relative temporal expressions.

5 Discussion and Conclusion

This work examines temporal reasoning in large language models under mixed time expressions, using sentence-level event ordering as a controlled evaluation task. Across diverse models and metrics, we consistently observe significant performance degradation when absolute temporal cues are partially replaced with relative expressions, revealing a systematic limitation rather than isolated model failures.

Our analysis suggests that mixed time expressions disrupt temporal reasoning primarily by weakening implicit temporal anchoring. Without explicit anchors, models struggle to recover global chronological structure even when local semantic cues are available. This effect is further amplified under coarser temporal granularity and longer event sequences, leading to compounding ordering errors. Notably, stronger models often exhibit increased instability under mixed-time inputs, indicating difficulties in maintaining structural control when temporal information is underspecified.

These observations point to a broader challenge in how current LLMs represent and reason about time. While models can leverage surface temporal cues when they are explicit, they appear less capable of maintaining consistent temporal representations when temporal information must be inferred implicitly or integrated across multiple expressions. These findings highlight a gap in current temporal reasoning evaluations, which frequently rely on linear narratives or explicit timestamps and may overestimate model robustness in realistic settings. Our results suggest that improving temporal reasoning requires not only stronger inference abilities, but also better handling of implicit temporal cues and uncertainty.

In conclusion, we show that temporal reasoning in LLMs breaks in predictable ways under mixed time expressions. By isolating event ordering under hybrid temporal conditions, this study provides a focused view of where and why these failures arise, and motivates evaluation frameworks that better reflect the complexity of temporal expressions in natural language.

References

[1] Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. Tempquestions: A benchmark for temporal question answering. In **Companion Proceedings of the The Web Conference 2018**, pages 1057–1062, 2018.

[2] Lianhui Qin, Aditya Gupta, Shyam Upadhyay, Luheng He, Yejin Choi, and Manaal Faruqui. Timedial: Temporal commonsense reasoning in dialog. **arXiv preprint arXiv:2106.04571**, 2021.

[3] Wenhu Chen, Xinyi Wang, and William Yang Wang. A dataset for answering time-sensitive questions. **arXiv preprint arXiv:2108.06314**, 2021.

[4] 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.

[5] 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.

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