Benchmarking Temporal Reasoning: Can Large Language Models Navigate Time When Stories Refuse to Follow a Straight Line?

Anonymous ACL submission

Abstract

Temporal reasoning remains a challenging task for **L**arge **L**anguage **M**odels (LLMs), particularly when confronted with nonlinear narratives and mixed time systems, where events are presented out of chronological order. While human cognition effortlessly reconstructs temporal sequences in such narratives, LLMs often exhibit inconsistent reasoning and fail to infer the correct event order. In this paper, we present a comprehensive study on sentence-level event ordering to evaluate emerging frontier LLMs in temporal reasoning tasks. We contribute (i) a novel dataset derived from historical records, blending absolute and relative time expressions across varied granularities; (ii) a benchmark covering emerging frontier LLMs including GPT family, DeepSeek series, Qwen models, and open-source models; and (iii) an absolute-relative time conversion table to support future research on mixed time systems. [1] Our experiments reveal substantial limitations across current models, with a consistent performance decline when relative time disrupts chronological signals. We further provide a detailed benchmark analysis across multiple dimensions, including model types, sentence length, temporal granularity, and format violations. Our findings offer key insights and valuable resources to advance temporal reasoning research in LLMs.

1 Introduction

Temporal reasoning is a fundamental component of natural language understanding, underpinning applications such as question answering, narrative comprehension, and timeline construction. Despite rapid progress in **L**arge **L**anguage **M**odels (LLMs), reasoning over temporal sequences—especially within nonlinear narratives—remains a persistent challenge. Unlike humans, who can effortlessly reconstruct event orders from fragmented or non-chronological inputs, LLMs often struggle when faced with mixed time systems involving both absolute and relative time expressions.

Nonlinear narratives, characterized by disrupted temporal flow and interleaved time references, are common in historical texts, biographies, and storytelling. These contexts require models not only to interpret explicit time expressions but also to infer implicit event dependencies across varying temporal granularities (e.g., year, month, day). While existing benchmarks have explored temporal reasoning through question answering or multi-task datasets(Jia et al., 2018; Qin et al., 2021; Chu et al., 2023; Wang and Zhao, 2023; Tan et al., 2023), they often underrepresent event ordering as a standalone capability. As LLMs continue to advance, dedicated benchmarks for this fundamental yet fragile skill—particularly under naturalistic and temporally ambiguous conditions—are increasingly needed.

In this work, we address this gap by formulating sentence-level event ordering as a core temporal reasoning task under nonlinear narrative settings. We construct a benchmark derived from historical records sourced from Wikidata, where each sentence is temporally anchored and spans a range of granularities. To simulate realistic narrative complexity, we include both absolute and relative time expressions, capturing scenarios where temporal cues are implicit, vague, or mixed.

We evaluate a suite of leading frontiers LLMs, including models from the GPT, DeepSeek, Qwen, and LLaMA families, along with Mistral-7B, focusing on their ability to recover event order, recognize temporal dependencies, and reason effectively under disrupted chronological signals.

To support future research, we also release a curated table of over 6,000 absolute-to-relative time expression that links structured time expressions (e.g., “1945”) with natural references (e.g., “the end of World War II”), offering a reusable resource for investigating mixed-time systems.

[1] Anonymous Github: https://anonymous.4open.science/r/MTS-benchmark-3035/

Our work makes the following contributions: we propose sentence-level event ordering as a benchmark task for evaluating temporal reasoning in nonlinear narratives; we construct a novel dataset based on historical texts, enriched with both absolute and relative time annotations across varied temporal granularities; we present a comprehensive benchmark study involving both leading frontier models (e.g., GPT-4, Deepseek, QWQ) and strong open-source baselines (e.g., LLaMA 3.3, Mistral, LLaMA 2-13B), systematically evaluating their ability to reason over mixed time systems; and we release an absolute-relative time conversion table to support further research in temporal inference.

Guided by these contributions, we investigate the following research questions:

- *How do different model architectures perform in temporal reasoning tasks?*
- *How do temporal granularity and event sequence length influence reasoning accuracy?*
- *Is there an interaction between time type and reasoning complexity?*
- *To what extent do relative time expressions affect model performance?*

2 Related Work

Temporal Question Answering Temporal reasoning (TR) has long been recognized as a core challenge in natural language processing, essential for tasks involving event sequencing, duration inference, and causal understanding. Early QA-style benchmarks, such as TempQuestions(Jia et al., 2018) and TimeDial(Qin et al., 2021), focus on reasoning under explicit, implicit, and ordinal temporal constraints. Other datasets, like that of Chen et al. (Chen et al., 2021), explore temporal drift through Wikipedia–Wikidata alignment, revealing the sensitivity of language models to subtle time-based context changes. TempReason (Tan et al., 2023) expands the temporal QA paradigm to a multi-level framework, encompassing time-time, time-event, and event-event reasoning. This line of work demonstrates the increasing complexity of temporal understanding required by modern QA systems.

However, while these QA datasets reflect diverse forms of temporal reasoning, they often embed event ordering as a latent step within broader reasoning chains, making it difficult to isolate and evaluate this capability directly. In contrast, our work treats event ordering as a standalone task, enabling focused assessment of model performance under temporally ambiguous and nonlinear narrative conditions.

Comprehensive Temporal Benchmarks Recent benchmarks such as TimeBench(Chu et al., 2023) and TRAM(Wang and Zhao, 2023) evaluate a broad spectrum of temporal reasoning skills by combining multiple tasks—such as duration estimation, temporal arithmetic, frequency detection, and causal inference—into large-scale evaluation suites. TempReason (Tan et al., 2023) adopts a more structured design with three reasoning levels, but remains grounded in the question answering paradigm.

In contrast, we focus on sentence-level event ordering—an underexplored yet challenging subtask—under hybrid time conditions that mix absolute and relative expressions. This design enables a finer-grained evaluation of LLMs' ability to recover global temporal structure from fragmented, nonlinear narratives.

While existing work has addressed absolute or relative temporal reasoning in isolation, the distinct challenges of mixed time—such as implicit anchoring, granularity mismatch, and nonlinearity—remain underexplored. We outline these issues and their implications for benchmark construction in Section 3.2.

Instruction Sensitivity and Model Coverage Recent work has shown that instruction tuning alone may not ensure reliable execution of structured or temporally grounded tasks (Lou et al., 2024), especially in scenarios requiring compositional reasoning or strict output format adherence (Chia et al., 2023; Wang et al., 2022; Xu et al., 2023). Although instruction-tuned models demonstrate strong performance in QA and classification, they often struggle in tasks demanding sequence-level reasoning or alignment with latent structural constraints (Peng et al., 2023; Min et al., 2023).

Our benchmark contributes to this line of research by providing a comparative analysis of instruction-following behaviors across model families—including underexplored but high-performing models such as DeepSeek and Qwen—under temporally sensitive, zero-/one-shot prompting settings. While many prior studies focus on GPT-family models or open-domain QA tasks (Kimura et al., 2021; Chen et al., 2021; Saxena et al., 2021; Dhingra et al., 2022; Tan et al., 2023; Gupta et al., 2023; Jia et al., 2024; Xiong et al., 2024; Fatemi et al., 2024; Deroy and Maity, 2024; Su et al., 2024; Yuan et al., 2024; Zhang et al., 2024; Deng et al., 2024; Ruiz et al., 2025), recent open-source models like DeepSeek and Qwen—despite their strong reasoning capabilities—remain underexplored in temporal settings. Our benchmark fills this gap by providing targeted evaluations of instruction-following behavior across both frontier and open models under mixed-time conditions.

3 Benchmark Setup

3.1 Task Overview

We formulate temporal reasoning in nonlinear narratives as a sentence-level event ordering task. Given a short passage composed of n unordered sentences $P = \{s_1, s_2, \ldots, s_n\}$, where each s_i describes an event associated with a time expression t_i, the model is tasked with inferring the correct chronological order of the events. The time expressions can be absolute (e.g., "in 1923") or relative (e.g., "three years later"), or a combination of both.

The expected output is a permutation π over the indices $\{1, ..., n\}$ such that the reordered sequence $\{s_{\pi(1)}, s_{\pi(2)}, ..., s_{\pi(n)}\}$ respects the underlying temporal timeline implied by the input. This task requires interpreting time expressions, resolving references, and aligning events across possibly fragmented or non-chronological inputs.

3.2 Challenges of Mixed Temporal Reasoning

Temporal reasoning in mixed time systems introduces challenges beyond standard timeline inference. First, relative expressions (e.g., "the following year") require anchoring to implicit reference points, which are often unstated. Second, absolute and relative expressions may co-occur, requiring joint interpretation and temporal alignment. Third, varying temporal granularity—some events given as years, others as full dates—creates ambiguity in sequencing. Finally, nonlinear narratives frequently present events out of order, demanding global integration of dispersed time cues.

3.3 Experimental Factors

To systematically investigate how different aspects of temporal structure affect model performance, we design benchmark settings along the following dimensions:

Mixed time expressions: introducing temporal ambiguity by randomly replacing a subset of absolute time expressions with relative references using an LLM-based rewriting strategy. We allow minor imprecision or implicit temporal references—such as GPT-4 occasionally grounding expressions like "this year" as 2023 irrespective of narrative context—as long as they do not alter the overall event order. This design choice reflects the inherent ambiguity in mixed-time narratives and evaluates whether models can still recover global chronological structure under such conditions.

Temporal granularity: comparing passages with coarse-grained (year-only) versus fine-grained (month or day included) time annotations.

Event sequence length: varying the number of events from 4 to 40 to examine how model performance scales with narrative length, and whether reasoning abilities degrade as the temporal chain becomes longer.

These experimental factors enable a fine-grained analysis of model sensitivity to temporal complexity under diverse and naturalistic conditions.

3.4 Dataset Construction

We construct our dataset from Wikidata (Vrandečić and Krötzsch, 2014) by extracting 15,000 historical and contemporary figures born after 1900, focusing on occupations such as scientists, historians, and politicians to ensure temporal and professional diversity. For each entity, we retrieve the English Wikipedia page and extract time-anchored event sentences using regex-based patterns. Sentences are filtered for grammaticality, relevance, and valid absolute dates, then chronologically sorted to form gold-standard event sequences. We retain passages containing 4 to 40 events to balance sequence complexity and data coverage.

To simulate mixed-time narratives, we randomly convert a subset of absolute expressions into relative or descriptive forms using GPT-4o. A controlled prompt ensures the rewrites are semantically faithful and logically consistent with surrounding context. To assess the quality of these rewrites, two NLP expert annotators—also co-authors of this work—independently evaluate 200 randomly sampled passages on three dimensions: (i) *Info Accuracy*, (ii) *Context Logic*, and (iii) *Naturalness*. Agreement scores are high for accuracy (79.5%) and contextual coherence (71.5%), while naturalness exhibits moderate variance (quadratic

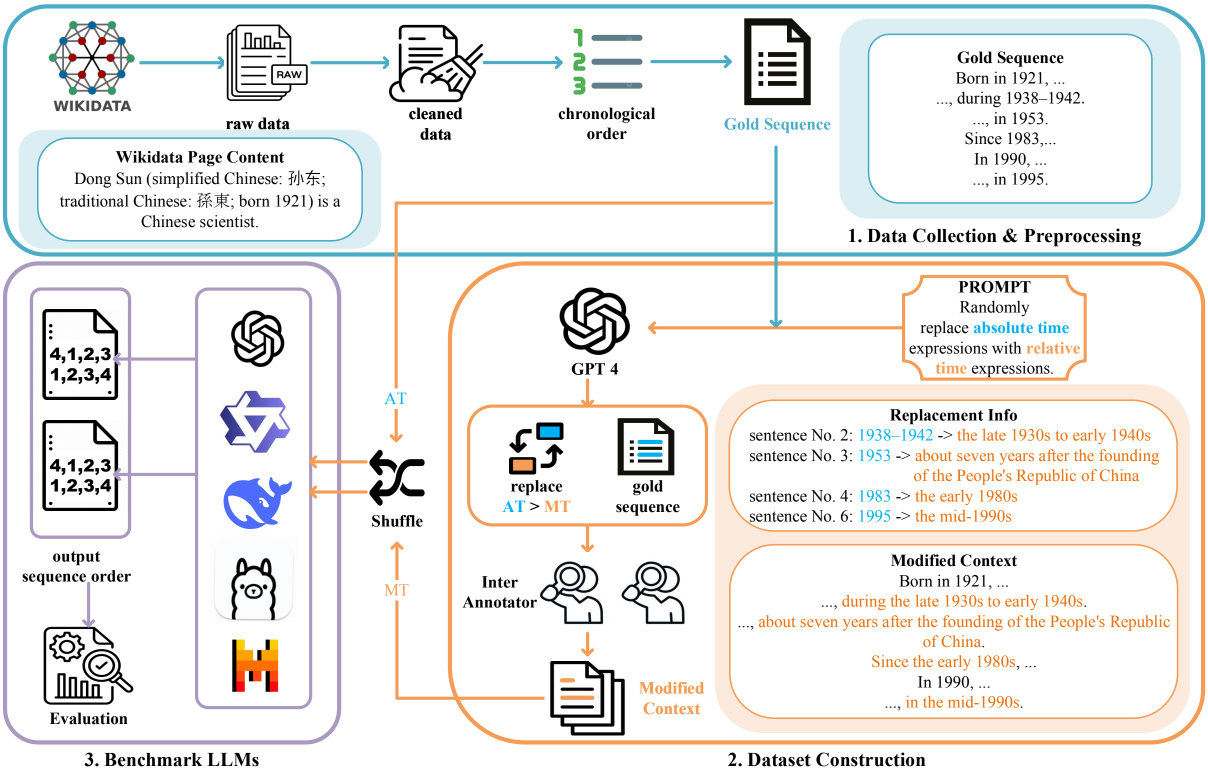


Figure 1: Overview of our benchmark construction pipeline. (1) We collect and clean biographical content from Wikidata and Wikipedia, extracting temporally anchored sentences to construct a gold-standard chronological sequence. (2) To simulate mixed-time scenarios, we use GPT-4o to rewrite a subset of absolute time expressions into natural relative expressions, producing both a modified context and a replacement mapping. Annotators then evaluate the quality of rewritten passages. (3) Multiple LLMs are benchmarked on sentence-level event ordering under both absolute-time (**AT**) and mixed-time (**MT**) settings. Models are required to output a comma-separated list of sentence indices (e.g., 2,1,4,3) to indicate the predicted event order.

weighted Cohen's $\kappa = 0.19$). These results confirm that most rewritten expressions are reliable for constructing mixed-time inputs.

The final dataset comprises 4,824 passages with an average of 8 events each. In the mixed-time setting, 56.7% of expressions are rewritten as relative forms. Distributions by event count and temporal granularity are shown in Table 2 and Table 3. We also release a time expression conversion table (e.g., "1945" → "the end of World War II") to support future work on temporal paraphrasing and normalization (see Appendix C).

3.5 Benchmark Settings and Models

We evaluate LLMs on a sentence-level temporal ordering task. Given a passage with shuffled event sentences, the model must predict the correct chronological order as a permutation of sentence indices. We define two task variants:

Absolute-Time Task (AT): Passages contain only absolute time expressions (e.g., "in 1945").

Mixed-Time Task (MT): Some absolute expressions are rewritten as natural relative references (e.g., "the end of World War II") using a GPT-based strategy. See Table 1 for the formal definition of time expression types.

All models are evaluated using a one-shot instruction-style prompt with a single illustrative example. We include both closed-source and open-source models spanning a range of training paradigms:

Closed-source Frontier Models: Including GPT-4, GPT-3.5, Deepseek-v3(Liu et al., 2024), Deepseek-r1(Guo et al., 2025), Qwen2.5-7B(Qianwen et al., 2024), and QwQ-32B(Team, 2025).

Open-source Models: Including LLaMA3.3-70B(Grattafiori et al., 2024), LLaMA2-13B(Touvron et al., 2023), and Mistral-7B(Jiang et al., 2023).

All models are tested using a consistent one-shot prompt setup that includes a single illustrative example and a standardized instruction format (see

Expression Type	Example
Absolute Time	"in 1945", "in March 2007", "on July 20, 1969"
Relative Time	"three years later", "shortly after the war"
Event-Anchored Time	"the end of World War II", "during the Great Depression"

Table 1: Time expression types used in our benchmark. The latter two categories are treated as *relative* for MT setting.

Appendix A) for details.

Statistic	Value
Total passages	4,824
Avg. events per passage	7.99
Temporal granularity — year	70.72%
Temporal granularity — month	23.46%
Temporal granularity — day	5.82%

(a) Absolute-time dataset before conversion.

Statistic	Value
Total passages	4,824
Avg. relative per passage	4.53
Avg. absolute per passage	3.46
Relative time ratio (relative / all)	56.73%

(b) Mixed-time dataset after relative replacement.

Table 2: Comparison of dataset statistics before and after the conversion from absolute-only to mixed-time representations.

3.6 Evaluation Metrics

We report the following evaluation metrics:

Exact Match (EM): The percentage of outputs that exactly match the gold-standard permutation, reflecting the model's ability to recover the *global temporal structure* of the passage. We additionally report the error rate, defined as 1 - EM, which captures the proportion of incorrect predictions.

Kendall's τ: Rank correlation between predicted and gold orders. This captures the *local temporal consistency* between event pairs.

Pairwise Accuracy: Fraction of correctly ordered sentence pairs.

We further apply:

McNemar's Test: For EM significance across AT and MT conditions.

Wilcoxon Signed-Rank Test: For Kendall's τ significance across AT and MT.

Malformed outputs are excluded. We also analyze EM and Kendall's τ scores by passage length and model family in Section 4.

Appendix E provides dataset visualizations, including event count distributions (Figure 8) and the temporal granularity of time expressions (Figure 7).

Event Count Range	Number of Passages	Percentage
4 – 9	3,817	79.13%
10 – 14	593	12.29%
15 – 19	184	3.81%
20 – 29	133	2.76%
30 – 39	56	1.16%
≥ 40	41	0.85%

Table 3: Distribution of passages by event count intervals. The majority of passages include no more than 10 events, aligning with the practical reasoning capacity of current LLMs. Longer passages are also retained to assess their ability to handle extended event sequences.

4 Results and Analysis

We analyze model performance on sentence-level event ordering under AT and MT conditions, covering nine models across proprietary and open-source families. Evaluation uses EM, Kendall's τ, and significance testing to assess sensitivity to temporal ambiguity. We further analyze model performance from three key perspectives—temporal granularity, event sequence length, and the presence of relative time expressions—to systematically address our four research questions.

4.1 Overall Model Performance

To address our first research question concerning the performance of different model architectures in temporal reasoning tasks, we begin by comparing overall accuracy across all evaluated models.

Significant Performance Gaps Between Frontier and Lightweight Models

The strongest overall performance is achieved by QwQ-32B and DeepSeek-R1, with EM scores of 0.54 and 0.52 in the AT setting, respectively, and high Kendall's τ values above 0.70. Notably, both models outperform GPT-4, which achieves an

Model	EM (AT)	EM (MT)	Kendall's τ (AT)	Kendall's τ (MT)
QwQ-32B	**0.54**	**0.33** ($\downarrow$39%)	**0.73**	**0.53** ($\downarrow$27%)
Deepseek-r1	0.52	0.32 ($\downarrow$38%)	0.70	**0.53** ($\downarrow$24%)
Deepseek-v3	0.33	0.21 ($\downarrow$36%)	0.51	0.38 ($\downarrow$25%)
GPT-4	0.31	0.15 ($\downarrow$52%)	0.50	0.34 ($\downarrow$32%)
LLaMA3.3-70B	0.21	0.13 ($\downarrow$38%)	0.40	0.30 ($\downarrow$25%)
GPT-3.5 turbo	0.12	0.07 ($\downarrow$42%)	0.21	0.17 ($\downarrow$19%)
Qwen2.5-7B	0.07	0.05 ($\downarrow$29%)	0.20	0.14 ($\downarrow$30%)
LLaMA2-13B	0.01	0.01 ($\downarrow$0%)	0.00	0.05 ($\uparrow$–)
Mistral-7B	0.00	0.01 ($\uparrow$–)	0.05	0.06 ($\uparrow$20%)

Table 4: Performance comparison across models under absolute-time (AT) and mixed-time (MT) conditions. Percentage change is calculated as: $\frac{\text{MT}-\text{AT}}{\text{AT}} \times 100\%$. Percentage changes from AT to MT are highlighted with color: red for performance drops, green for gains.

EM score of 0.50 and τ below 0.70 under the same setting.

High EM scores indicate strong reconstruction of the global event sequence, while high Kendall's τ reflects consistent pairwise ordering. These results align with prior findings from TimeBench (Chu et al., 2023), where GPT-family models excelled in structured temporal reasoning, while smaller models like Mistral-7B struggled with commonsense and relative time. This supports our observation that frontier models better preserve both global and local temporal structure. Table 4 reports EM and Kendall's τ under AT and MT settings, along with relative performance drops to assess robustness under temporal ambiguity.

Deeper Reasoning Comes at the Cost of Instruction Following

While both the Qwen and Deepseek model families achieve superior performance in temporal reasoning tasks compared to other models, we observe a notable divergence in output format adherence within each family. As shown in Figure 2, the stronger reasoning variants—QwQ-32B and Deepseek-r1—exhibit significantly higher rates of invalid format outputs than their smaller counterparts. This pattern is consistent with findings from instruction-following literature(Lou et al., 2024), which highlight that larger models, despite superior reasoning abilities, are more likely to deviate from strict output constraints—particularly in settings without strongly grounded demonstrations. An illustrative example of such a violation is provided in Appendix F.1.

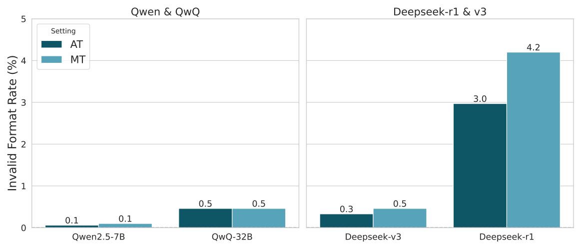


Figure 2: Invalid output rate (%) for Qwen and DeepSeek models under AT and MT. DeepSeek-r1 shows notably higher error rates, especially in MT, indicating reduced stability when processing relative time inputs.

These results reveal a trade-off between deep reasoning and strict instruction adherence. As models develop more complex inference capabilities, they may favor semantic interpretation over rigid output formatting, particularly under ambiguous prompts. This tension between interpretive depth and structural control is further evidenced by increased format violations, detailed in Appendix F.1.

4.2 Temporal Granularity Analysis

To address part of our second research question regarding the effect of temporal granularity on reasoning performance, we analyze how the granularity of time expressions influences model accuracy under both AT and MT conditions. Passages are grouped into two levels: those with only year-level expressions (*coarse-grained*) and those that include month or day annotations (*fine-grained*).

Fine-grained timestamps lead to more stable reasoning

Models perform more robustly on fine-grained passages, where temporal cues are more precise. These timestamps help disambiguate events that occur in the same year but at different times, enabling better alignment and control over sentence reordering.

To quantify this effect, we compare error rates between AT and MT across both granularity levels. As shown in Figure 10, the performance gap between AT and MT is consistently larger under coarse-grained inputs. For example, GPT-4 and QwQ-32B both show over 25% error rate increase when relative time replaces coarse absolute timestamps.

Stronger Models Within Families Are More Affected by Coarse-Grained Time Inputs

All models show performance degradation when temporal inputs are coarsened from day/month to year-level granularity. Notably, the strongest models—QwQ-32B and DeepSeek-r1—exhibit the largest MT–AT error increases under coarse-grained conditions (Figure 3), suggesting a reliance on fine-grained temporal cues. As specificity declines, these models may resort to overgeneralized reasoning, increasing deviation from the gold standard. This aligns with Yang et al.(Yang et al., 2024), who show that temporally aware embeddings enhance reasoning but amplify sensitivity to time granularity. In contrast, weaker models appear less affected, likely due to simpler, more conservative reasoning. Detailed results are in Appendix F.2.

Relative time expressions are *less harmful* when granularity is high

The negative impact of switching to relative time is most severe under vague or underspecified temporal conditions. When time granularity is higher, relative expressions carry more specific temporal meaning—mitigating ambiguity and supporting more stable reasoning.

These findings highlight the interaction between surface-level time granularity and deeper temporal reasoning ability. Improving model robustness to coarse-grained relative time may require explicit training on relational semantics and underspecified narratives.

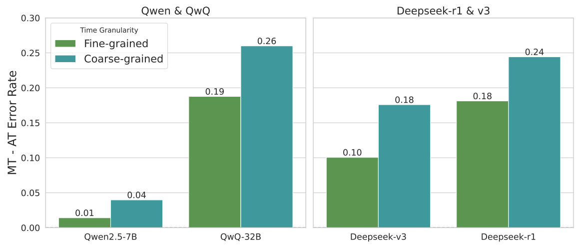


Figure 3: MT-AT error rate increase under different time granularities for Qwen and DeepSeek models. Both show greater degradation with coarse-grained inputs, with QwQ-32B and DeepSeek-r1 most affected, suggesting reduced robustness to underspecified temporal cues.

4.3 Event Sequence Length Analysis

To address our third research question regarding the interaction between time type and reasoning complexity, we now examine how event sequence length influences temporal reasoning performance.

Longer sequences sharply degrade model performance

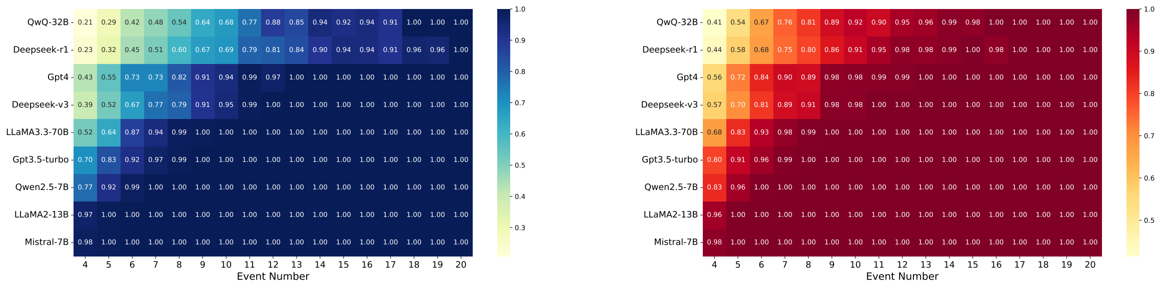
Figure 4 visualizes error rates for all evaluated models under both the AT and MT settings. We observe a clear trend: as the number of events rises, nearly all models experience a steady and often steep increase in error rate.

Most models begin with reasonably low error rates (e.g., 0.2–0.4) on short passages (4–6 events), particularly under the AT setting. However, accuracy degrades quickly, and by 12 events, even the best-performing models (e.g., GPT-4, Deepseek-R1, Qwen-32B) approach near-complete failure in the MT setting.

Relative time increases vulnerability to sequence length

The contrast between AT and MT is particularly striking: while AT error rates often increase more gradually, MT error rates rise faster and reach 1.0 earlier. This pattern reveals that relative time reasoning is disproportionately affected by sequence length—likely because models must track more implied temporal links without the support of explicit anchors.

Among all models, Qwen-32B and DeepSeek-R1 stand out for maintaining lower MT error rates in the 4–8 event range, while others such as Mistral and LLaMA variants fail almost immediately. The



(a) AT Error Rates by Model and Event Number

(b) MT Error Rates by Model and Event Number

Figure 4: 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.

robustness of these models may stem from better generalization over temporal language, or implicit pretraining biases favoring temporal coherence.

By 15–20 events, nearly all models saturate at an error rate of 1.0 in both AT and MT conditions. These results indicate that existing models struggle to maintain coherence in long event chains, and relative-time reasoning becomes brittle under increased temporal complexity.

4.4 Absolute vs. Mixed Time Comparison

To directly address our last research question, we compare model performance between passages with AT and MT.

Mixed-Time Settings Introduce Substantial Difficulty

All models exhibit performance drops under the MT setting, though the magnitude varies. GPT-4 and GPT-3.5 Turbo experience steep EM reductions of over 50%, suggesting a strong reliance on explicit absolute-time cues. In contrast, frontier models like QwQ-32B and Deep see k-R1 show more graceful degradation, with EM drops around 20%, and Kendall's τ remaining above 0.50.

Reliance on explicit timestamps amplifies degradation

Models like GPT-4 perform well under AT conditions but degrade sharply in MT, suggesting strong reliance on explicit date cues. In contrast, DeepSeek-R1 and QwQ maintain more stable performance, indicating better generalization to natural temporal variation. As shown in Table 4, color-coded drops in EM and Kendall's τ highlight that smaller models (e.g., Mistral, LLaMA2-13B) not only perform poorly overall, but also show minimal AT–MT difference—suggesting weak temporal sensitivity. These findings underscore the need to evaluate LLMs under both controlled and realistic temporal settings to fully assess their reasoning capabilities.

5 Conclusion

In this work, we present a novel benchmark for evaluating LLMs' temporal reasoning in complex narratives that combine absolute and relative time expressions across varied granularities. Unlike prior datasets limited to a single time type or simplified task settings, our sentence-level benchmark captures the hybrid temporal structures found in real-world biographies.

Through extensive analysis across time conditions, granularity levels, and sequence lengths, we find that even the strongest models (e.g., QwQ-32B, DeepSeek-R1) struggle to maintain temporal coherence under mixed-time settings—particularly with coarse-grained or long-range dependencies. These results highlight LLMs' reliance on surface-level cues and their limited capacity for relational temporal reasoning.

To support broader research on temporal modeling, we additionally release a large-scale conversion table of aligned absolute-to-relative time expressions—a novel resource for studying time normalization and contextual rewriting.

We hope our benchmark and accompanying resources encourage future work on time-aware inference, instruction-following under temporal ambiguity, and constraint-driven model alignment.

Future work will aim to improve model generalization in relative time settings and enhance instruction adherence through better prompting and constraint-aware training.

Limitations

While our benchmark offers a robust platform for evaluating temporal reasoning in LLMs, several limitations remain.

First, the dataset is built from Wikipedia-style biographies, which—though rich in timestamped events—do not cover all narrative types. Domains such as scientific writing or fiction may exhibit different temporal patterns.

Second, we adopt a sentence-level event abstraction, omitting finer discourse phenomena like simultaneity or intra-sentential shifts. Time expressions are automatically extracted and occasionally noisy, which may affect alignment.

Third, relative expressions are generated via GPT-4o rewrites. While this improves lexical diversity, it introduces ambiguity—e.g., "2004" may become "the following year," requiring prior context. Annotators observed occasional grounding errors (e.g., "this year" interpreted as 2023), but such cases are accepted if event order is preserved.

Fourth, our evaluation focuses on global metrics (EM, Kendall's τ), which may overlook partial correctness in passages with underspecified temporal cues.

Fifth, we evaluate models under zero- and one-shot prompting only, without fine-tuning or architectural changes (e.g., temporal embeddings), which may further improve performance.

We also observe frequent instruction-following failures in open-source models. Despite format constraints, models like Mistral-7B often produce verbose outputs. One-shot prompting improves compliance, but we do not compare prompting strategies systematically due to budget constraints.

Finally, performance collapses on very long passages (e.g., >30 events), likely due to compounded reasoning and context-length challenges. These cases are excluded from analysis and underscore the need for better long-context temporal reasoning.

References

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

Yew Ken Chia, Pengfei Hong, Lidong Bing, and Soujanya Poria. 2023. Instructeval: Towards holistic evaluation of instruction-tuned large language models. *arXiv preprint arXiv:2306.04757*.

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*.

Irwin Deng, Kushagra Dixit, Vivek Gupta, and Dan Roth. 2024. Enhancing temporal understanding in llms for semi-structured tables. *arXiv preprint arXiv:2407.16030*.

Aniket Deroy and Subhankar Maity. 2024. A short case study on understanding the capabilities of gpt for temporal reasoning tasks. *Authorea Preprints*.

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.

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*.

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

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Vivek Gupta, Pranshu Kandoi, Mahek Bhavesh Vora, Shuo Zhang, Yujie He, Ridho Reinanda, and Vivek Srikumar. 2023. Temptabqa: Temporal question answering for semi-structured tables. *arXiv preprint arXiv:2311.08002*.

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

Zhen Jia, Philipp Christmann, and Gerhard Weikum. 2024. Faithful temporal question answering over heterogeneous sources. In *Proceedings of the ACM Web Conference 2024*, pages 2052–2063.

Albert Q Jiang, A Sablayrolles, A Mensch, C Bamford, D Singh Chaplot, Ddl Casas, F Bressand, G Lengyel, G Lample, L Saulnier, et al. 2023. Mistral 7b. arxiv. *arXiv preprint arXiv:2310.06825*, 10.

Mayuko Kimura, Lis Kanashiro Pereira, and Ichiro Kobayashi. 2021. Towards a language model for temporal commonsense reasoning. In *Proceedings* of the Student Research Workshop Associated with RANLP 2021*, pages 78–84.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Renze Lou, Kai Zhang, and Wenpeng Yin. 2024. Large language model instruction following: A survey of progresses and challenges. *Computational Linguistics*, 50(3):1053–1095.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Peng Qianwen, Gao Yanzipeng, Li Xiaoqing, Min Fanke, Li Mingrui, Wang Zhichun, and Liu Tianyun. 2024. . In *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations)*, pages 294–301, Taiyuan, China. Chinese Information Processing Society of China.

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

Alfredo Garrachón Ruiz, Tomás de la Rosa, and Daniel Borrajo. 2025. On the temporal question-answering capabilities of large language models over anonymized data. *arXiv preprint arXiv:2504.07646*.

Apoorv Saxena, Soumen Chakrabarti, and Partha Talukdar. 2021. Question answering over temporal knowledge graphs. *arXiv preprint arXiv:2106.01515*.

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

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

Qwen Team. 2025. Qwq-32b: Embracing the power of reinforcement learning.

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*.

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

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

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

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

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Wanqi Yang, Yanda Li, Meng Fang, and Ling Chen. 2024. Enhancing temporal sensitivity and reasoning for time-sensitive question answering. *arXiv preprint arXiv:2409.16909*.

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 Web Conference 2024*, pages 1963–1974.

Xinliang Frederick Zhang, Nick Beauchamp, and Lu Wang. 2024. Narrative-of-thought: Improving temporal reasoning of large language models via recounted narratives. *arXiv preprint arXiv:2410.05558*.

A Prompt Templates

We provide below the two prompt templates used in our study: one for rewriting absolute time expressions into relative ones, and another for evaluating temporal reasoning via event ordering. Both prompts follow a standardized instruction style to ensure consistency across model families.

(1) Relative Time Conversion Prompt

Time Replacement Prompt

You are a time conversion assistant. Your task is to replace exactly {num_to_keep} absolute time expressions with relative time expressions.

- Absolute time refers to any date in year, month-year, or full-date format.
- Retain {num_to_keep} absolute times, convert the rest into natural relative references.
- Avoid repeating the same phrasing.
- Do not simply compute or state time differences.

Return:

- **Modified Context**: the rewritten passage.
- **Replacement Information**: lines showing original → relative expressions.

(2) Event Ordering Prompt (Benchmark)

One-Shot Benchmark Prompt

The following is a set of shuffled sentences. Please infer the correct order and return the sentence order as a sequence of numbers.

Instructions: - Only return a comma-separated sequence of numbers.
- Do not include any explanations, additional text, or line breaks.
- The sequence should reflect the correct order of the given sentences.

Example:
Input:
1. The sun rises in the east.
2. It is early morning.
3. The birds are singing.
Correct output:
2,1,3
Now, process the following sentences:
{context}
Please output only the sequence of numbers.

B Example of Relative Time Conversion

Below are three representative examples showing how absolute time expressions are converted into relative expressions using our prompt-based generation pipeline.

We acknowledge that certain time replacements (e.g., replacing "2015" with "a few years after joining MIT") may introduce implicit event dependencies, such as the need to infer the timing of the prior event (i.e., joining MIT). However, our task primarily evaluates whether models can recover the correct chronological order of events rather than verifying precise temporal anchoring of each individual expression.

Therefore, as long as the replacement does not alter the relative order of events in the passage, such substitutions are considered acceptable within our task framework. These more ambiguous or indirect expressions are intentionally included to simulate the diversity and complexity of naturally occurring narratives with mixed temporal expressions.

Example 1

Original Passage (Gold Sequence)

(1) His father, Babalyk, born in 1860, was the only child in the family.
(2) He studied at a Kazakh school, then in the Tatar language school, then in **1941–1943** he graduated from the gymnasium in the city of Tacheng.
(3) In **1943–1947**, while studying at the university in Ürümqi, he was arrested for nationalist actions and imprisoned.
(4) After the founding of the Communist State, he became governor of Ili Kazakh Autonomous Prefecture in June 1955, and held that office until 1958.

Converted Passage (Mixed Time Expressions)

(1) His father, Babalyk, born in 1860, was the only child in the family.
(2) He studied at a Kazakh school, then in the Tatar language school, then **during the early 1940s** he graduated from the gymnasium in the city of Tacheng.
(3) **Around the mid-1940s**, while studying at the university in Ürümqi, he was arrested for nationalist actions and imprisoned.
(4) After the founding of the Communist State, he became governor of Ili Kazakh Autonomous Prefecture in June 1955, and held that office until 1958.

Replacement Mapping

1. Sentence 2: `1941–1943` → *during the early 1940s*
2. Sentence 3: `1943–1947` → *around the mid-1940s*

Example 2

Original Passage (Gold Sequence)

(1) Zaharia was a gold medalist at the International Collegiate Programming Contest, where his team University of Waterloo placed fourth in the world and first in North America in 2005.
(2) While at University of California, Berkeley's AMPLab in **2009**, he created Apache Spark as a faster alternative to MapReduce.
(3) In 2013 Zaharia was one of the co-founders of Databricks where he is chief technology officer.
(4) He joined the faculty of MIT in **2015**, and then became an assistant professor of computer science at Stanford University in 2016.
(5) In 2019 he was spearheading MLflow at Databricks, while still teaching.

Converted Passage (Mixed Time Expressions)

(1) Zaharia was a gold medalist at the International Collegiate Programming Contest, where his team University of Waterloo placed fourth in the world and first in North America in 2005.
(2) While at University of California, Berkeley's AMPLab **several years later**, he created Apache Spark as a faster alternative to MapReduce.
(3) In 2013 Zaharia was one of the co-founders of Databricks where he is chief technology officer.
(4) **A few years after joining MIT**, he became an assistant professor of computer science at Stanford University in 2016.
(5) In 2019 he was spearheading MLflow at Databricks, while still teaching.

Replacement Mapping

1. Sentence 2: `2009` → *several years later*
2. Sentence 4: `2015` → *a few years after joining MIT*

Example 3

Original Passage (Gold Sequence)

(1) In 1937, Schulze moved to Peenemünde Army Research Center; in 1939, he was appointed chief of the Propulsion Unit, a position he held until **1945**.
(2) Classified as wards of the state, the seven men landed at Fort Strong on **September 29, 1945**; all but von Braun, Schulze included, were then transferred to Aberdeen Proving Ground to translate and catalog 14 tons of V-2 documents taken from Germany.
(3) By 1946, Schulze was among the Operation Paperclip scientists employed at Fort Bliss.
(4) He moved to Alabama, where he was naturalized in Birmingham on November 11, 1954.

Converted Passage (Mixed Time Expressions)

(1) In 1937, Schulze moved to Peenemünde Army Research Center; in 1939, he was appointed chief of the Propulsion Unit, a position he held until **the end of World War II**.
(2) Classified as wards of the state, the seven men landed at Fort Strong **during the late 1940s**; all but von Braun, Schulze included, were then transferred to Aberdeen Proving Ground to translate and catalog 14 tons of V-2 documents taken from Germany.
(3) By the year after World War II ended, Schulze was among the Operation Paperclip scientists employed at Fort Bliss.
(4) He moved to Alabama, where he was naturalized in Birmingham on November 11, 1954.

Replacement Mapping

1. Sentence 1: `1945` → *the end of World War II*
2. Sentence 2: `September 29, 1945` → *during the late 1940s*

C Time Expression Conversion Table

To support future research on temporal rewriting and normalization, we release a conversion table that records all absolute-to-relative time expression rewrites applied during the construction of the our dataset. Each entry in the table represents a single replacement performed by GPT-4o during the mixed-time generation process.

Table 5 presents representative examples. The rewrites range from grounded historical interpretations (e.g., "1945" → "the end of World War II") to relative references that depend on the surrounding narrative timeline (e.g., "2004" → "the following year").

Original Time	Rewritten Time
1970	early 1970s
1979	late 1970s
2000	the turn of the millennium
2004	the following year
1967	several decades ago
1976	a little over four decades ago
1993	Approximately three decades back
2009	Fourteen years ago
2012	eight years ago
1948	a little over 75 years ago
1949	about 74 years back
1951	early 1950s
1956	approximately mid-1950s
1989	the last decade of the 1980s
October 2, 2003	early October 2003
January 23, 2004	late January 2004
January 23, 2004	Soon after
January 28, 2004	at the end of January 2004
February 1, 2004	shortly after

Table 5: Sample entries from the time expression conversion table, covering grounded historical, approximate, and relative rewrites.

We caution that not all rewritten expressions are context-independent. While some rewrites refer to widely understood historical periods (e.g., "1945" → "the end of World War II"), others depend on the internal narrative timeline. For example, "2004" is sometimes rewritten as "the following year", which is contextually appropriate only if the previous sentence refers to "2003". Such replacements, though semantically coherent in context, may not be suitable for standalone use.

Moreover, during our double-annotation process (see Section 3.4 and Section D for details), we adopt a practical criterion: a rewritten time expression is considered valid as long as it preserves the overall temporal order of the passage, even if the substitution is not lexically precise. This design choice reflects our focus on evaluating temporal reasoning rather than surface-level rewriting fidelity.

We therefore encourage users to consult the context when applying this conversion table in downstream tasks such as generation, normalization, or rule extraction. The conversion table is best viewed as a supporting resource rather than a standalone ground truth.

D Annotation Protocol and Analysis

To evaluate the quality of GPT-generated relative time expressions, we conducted a double annotation study on 200 sampled passages. Each annotator was presented with the original passage (*Gold Sequence*), the GPT-modified passage (*Gpt Modified Context*), and a detailed list of substitutions (*Replacement Info*).

Although some relative expressions are not exact translations of the original absolute timestamps, we consider the replacement acceptable as long as the temporal sequence of events remains unaffected. This evaluation criterion was reflected in the annotation guidelines for the "Info Accuracy" dimension. This decision aligns with our task definition, where the primary goal is to evaluate models' ability to reconstruct the correct temporal order, rather than the surface accuracy of individual time expressions.

For each passage, annotators were instructed to evaluate the following three dimensions:

1. **Info Accuracy (Y/N)**: Whether the relative expression generated by GPT-4 accurately reflects the semantics of the original absolute timestamp.
    - **Y**: The relative time correctly corresponds to the absolute time and aligns with the provided substitution info.
    - **N**: The expression is semantically incorrect, overly vague, or omits critical temporal details.
2. **Context Logic (Y/N)**: Whether the modified relative expression fits logically and temporally within the surrounding passage.
    - **Y**: The expression is coherent in context and does not break the narrative or event sequence.
    - **N**: The expression introduces chronological contradictions or disrupts temporal flow.
3. **Naturalness Score (1–5)**: Fluency and readability of the modified sentence, regardless of correctness.

Field	Example Annotation
Original Passage	In March 2007 she was elected to the fellowship of the Royal Society of Edinburgh. In 2018 she was appointed Head of the School of Informatics at Edinburgh, taking over from Johanna Moore, until succeeded by Helen Hastie in 2023. In 2018, Hillston was elected the membership of the Academia Europaea. Hillston was elected a Fellow of the Royal Society in May 2022. Since January 1st 2023 Hillston has been Editor-in-Chief of Proceedings of the Royal Society A (the first female Editor-in-Chief in the journal's history).
Rewritten Passage	In March 2007 she was elected to the fellowship of the Royal Society of Edinburgh. In 2018 she was appointed Head of the School of Informatics at Edinburgh, taking over from Johanna Moore, until succeeded by Helen Hastie this year. In 2018, Hillston was elected the membership of the Academia Europaea. Hillston was elected a Fellow of the Royal Society in May of last year. Since January 1st 2023 Hillston has been Editor-in-Chief of Proceedings of the Royal Society A (the first female Editor-in-Chief in the journal's history).
Replacement Mapping	Sentence 2: 2023 → this year Sentence 4: May 2022 → May of last year
Accuracy(Y/N)	Y
Coherence(Y/N)	N
Naturalness Score(1-5)	3
Error Type	InfoLoss
Free-form Comment	The current year is assumed to be 2023, causing a disruption in contextual coherence.

Table 6: Full annotation example including rewritten passage and free-form comment.

- 5: Fully natural and indistinguishable from human-written text.
- 4: Mostly fluent with only minor disfluency.
- 3: Somewhat awkward but understandable.
- 2: Clearly unnatural with evident phrasing issues.
- 1: Machine-like and syntactically poor.

Annotators were also encouraged to optionally tag common issues using a predefined label set:

- `infoless`: Key temporal information is missing.
- `vague`: Time span is ambiguous (e.g., "many years later").
- `inconsistent`: Logical contradiction in event ordering.
- `HardUnderstand`: Converted sentence is semantically unclear.
- `Other`: Additional problems not captured by the above categories.

To ensure consistency, annotators jointly reviewed 5–10 initial examples and were encouraged to leave free-form comments for both high-quality and problematic samples. The estimated annotation time per passage ranged from 1–3 minutes.

Here, "free-form" refers to an open comment field in the annotation interface, where annotators could optionally write their reflections on the quality of time expression rewriting, such as naturalness, contextual alignment, or specific GPT-related issues.

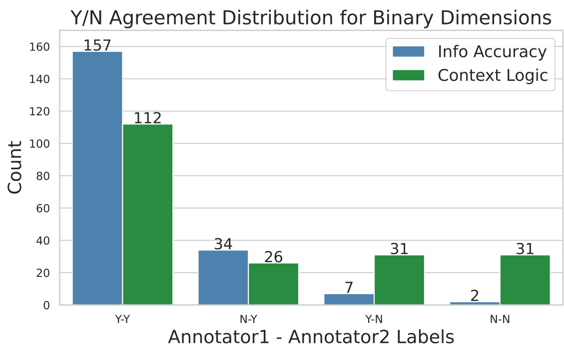


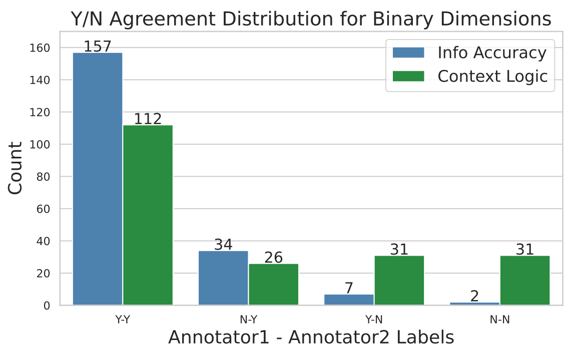
Figure 5: Absolute counts of issue types labeled by Annotator A and Annotator B.

Annotation Results and Agreement

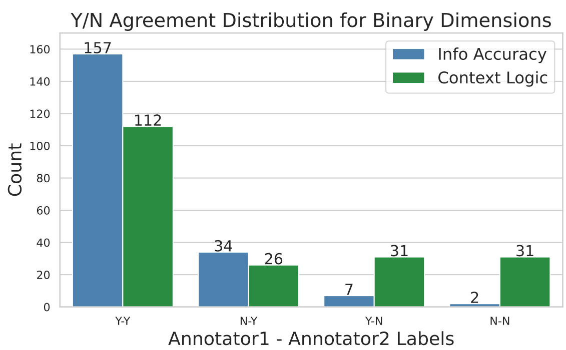
Figure 6 summarizes inter-annotator agreement patterns. For naturalness, Annotator 1 was generally more lenient, assigning the highest score (5) in 126 instances, while Annotator 2 gave more conservative, mid-range ratings. For binary categories, most passages received consistent "Y-Y" judgments, although moderate disagreement (e.g., "Y-N") remained, particularly in *Contextual Coherence*.

Overall, raw agreement reached 79.5% for *Replacement Accuracy* and 71.5% for *Contextual Coherence*. For naturalness, the quadratic-weighted Cohen's κ score was 0.19, indicating moderate agreement and highlighting the inherent subjectivity in fluency assessment.

These results confirm that the majority of GPT-generated relative time expressions are accurate, contextually appropriate, and linguistically natural to human readers. Despite some disagreements, the double-annotation protocol validates the reliability of our rewriting strategy and supports its use in constructing temporally ambiguous test sets for evaluating LLMs.



(a) Binary agreement (Y/N)



(b) Naturalness score distribution

Figure 6: Annotation agreement patterns across evaluation dimensions.

Error Type Distribution

To better understand annotator preferences and tendencies in error labeling, we compare the absolute count of common issue types annotated by each annotator. As shown in Figure 5, Annotator A overwhelmingly labeled vague expressions (62 instances), while Annotator B distributed their annotations more evenly across multiple categories.

Specifically, Annotator B marked 27 instances each of `InfoLoss` and `Incosistent`, as well as 18 instances of `HardUnderstand`, compared to Annotator A's respective counts of 6, 11, and 1. These differences suggest that Annotator A is particularly sensitive to ambiguity and imprecision in temporal phrasing, whereas Annotator B applies stricter standards in identifying information loss and logical inconsistency.

Despite these differences in emphasis, both annotators consistently identified problematic passages, reinforcing the value of error-type labels in guiding future improvements. The complementary nature of these annotation styles also offers useful insights into the diverse aspects of failure in GPT-based time rewriting.

Due to limited computational budget, we did not conduct adjudication to resolve annotation disagreements, which may leave some borderline cases open to interpretation.

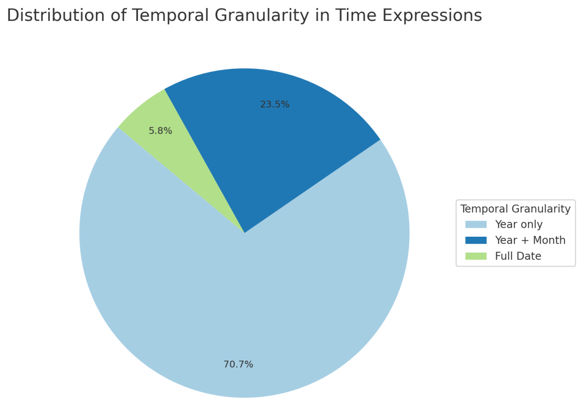
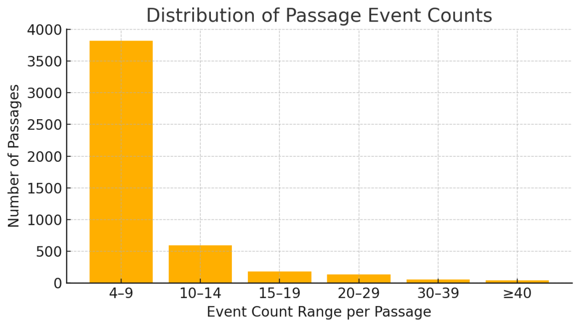


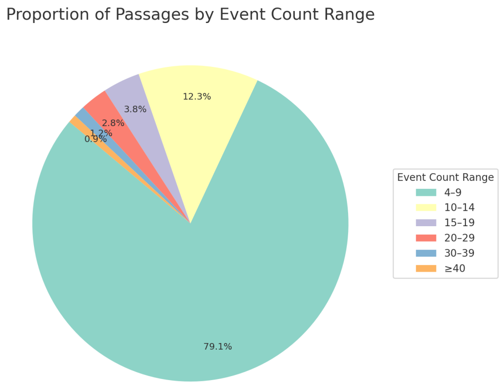
Figure 7: Distribution of temporal granularity among all absolute time expressions.

E Dataset Distribution Visualizations

To better illustrate the internal structure of our dataset, we present a set of visualizations that highlight the distribution of event counts and temporal granularity across all passages.



(a) Histogram of passages by event count. Most passages contain fewer than 10 events.



(b) Proportional breakdown by event count range.

Figure 8: Event count distribution across passages in the dataset.

Figure 7 shows the distribution of temporal granularity for all absolute time expressions in the dataset. The majority (70.7%) of expressions specify only the year (e.g., “in 1987”), while 23.5% include both year and month (e.g., “July 1987”), and only 5.8% provide a full date (e.g., “July 15, 1987”). This skew toward coarse-grained time references reflects common patterns in Wikipedia-style biographical writing and suggests that many temporal relations must be inferred from sparse cues.

Figure 8 presents the distribution of event counts across passages. Figure 8a displays a histogram showing that most passages contain fewer than 10 events, with a peak in the 4–9 range. Figure 8b provides a proportional breakdown, confirming that 79.1% of passages fall into the 4–9 event range. Passages with more than 20 events are relatively rare, accounting for less than 5% of the dataset.

This distribution suggests that the dataset is centered around passages with fewer than 10 events, maintaining a manageable level of complexity for most temporal reasoning tasks. At the same time, a small number of long-sequence passages (with 20 or more events) are included to support long-tail evaluation and stress-test models under extended temporal contexts.

F All model result

F.1 Format Violation Analysis

We first provide a representative example of a format violation. As shown in the box below, instead of returning a comma-separated list of sentence indices as instructed, the model outputs a verbose sequence of full event descriptions. This behavior constitutes a clear deviation from the expected format and illustrates a common failure mode among instruction-sensitive models. Such violations not only complicate automated evaluation but also indicate potential weaknesses in instruction comprehension, particularly when temporal reasoning is embedded in ambiguous inputs.

Example of Invalid Output Format

Gold Order: `[5,4,1,3,2]`
Expected Format: A comma-separated list of indices, e.g., `5,2,1,4,3`
Model Output:

> 1970, She was established what was for nearly a decade the only protein crystallography laboratory in Israel., In 1970, Her parents ... Then, from 1979 to 1984 she was a group leader ... On Saturday, 18 October 2014, Professor Yonath ... She was visiting professor at the University of Chicago ...

Violation Type: `Verbose explanation instead of index list`

Figure 9 shows format violation rates under AT and MT settings. Violations include missing indices, extraneous text, or malformed outputs.

While most models perform well (violation <1%), **Mistral-7B** (51.3% AT, 34.0% MT) and **LLaMA2-13B** (14.7% AT, 12.7% MT) show significant instability. In contrast, models like Qwen2.5-7B, Deepseek-v3, and GPT-3.5-turbo maintain consistently low violation rates.

MT settings generally increase format errors, highlighting the destabilizing effect of relative expressions on instruction-following. Notably,

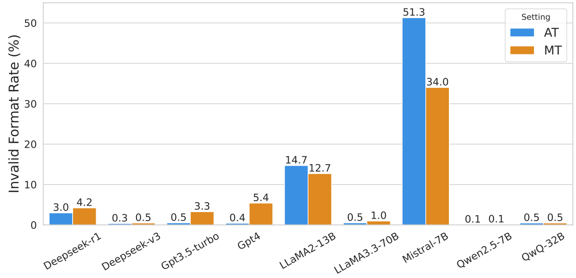


Figure 9: Prompt format violation rates across models in both AT and MT settings.

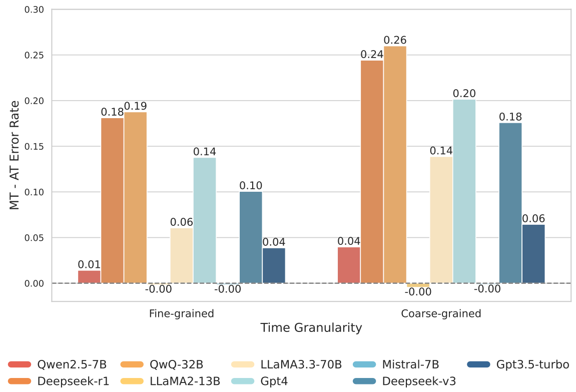


Figure 10: EM error rate increase (MT - AT) across time granularity levels. Coarse-grained (year-only) passages lead to stronger degradation under mixed-time input, especially for models like Deepseek-r1 and Qwen-32B.

Mistral-7B and LLaMA2-13B often generate verbose explanations instead of plain index lists.

These findings suggest that instruction adherence is not solely determined by model size or reasoning ability, and remains fragile under ambiguous temporal input.

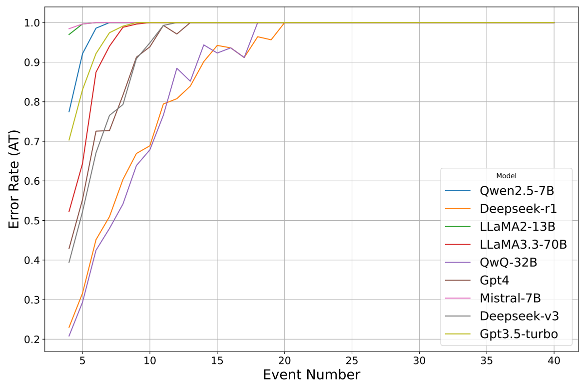
F.2 Granularity Analysis

Figure 10 reveals that when only year-level timestamps are present, models rely heavily on numerical comparison (e.g., 1995 vs. 2000) under AT. Once these cues are replaced with vague relative phrases like "a few years later," performance degrades sharply. The absence of fine-grained resolution compounds the difficulty of interpreting relative time.

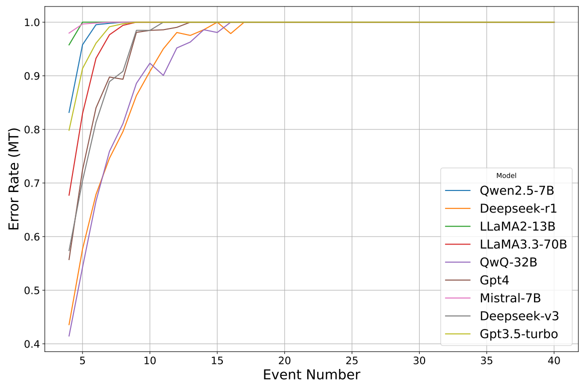
Interestingly, under fine-grained conditions, the performance gap between AT and MT narrows. While absolute timestamps are more complex (e.g., full dates), the corresponding relative phrases (e.g., "early that year," "a few months earlier") are often more informative. These naturalistic expressions provide additional linguistic cues that partially compensate for the loss of exact time, helping models maintain ordering accuracy.

G Full Error Rate Curves Across Event Numbers

Figure 11 provide a comprehensive view of model scalability when handling increasing event chains. While the main text focuses on results with up to 15 events (where most meaningful distinctions occur), we include these extended plots to show that beyond this point, most models saturate to an error rate of 1.0, suggesting a consistent upper bound on current models' capacity for temporal reasoning in complex narratives.



(a) Error Rate (AT) vs. Event Number for all evaluated models.



(b) Error Rate (MT) vs. Event Number for all evaluated models.

Figure 11: Full error rate trends under AT and MT; most models saturate at 1.0 beyond 15 events, indicating scalability limits.