Time Tells: Temporal Event Ordering in Frontier LLMs — Performance, Limitations, and Human Comparison

Feifei Sun[1], **Ziyi Tong**[1,],
Teeradaj Racharak[2], **Minh Le Nguyen**[1],

[1]Japan Advanced Institute of Science and Technology (JAIST), Japan
[2]Advanced Institute of So-Go-Chi (Convergence Knowledge) Informatics,
Tohoku University, Japan

Correspondence: racharak@tohoku.ac.jp, nguyenml@jaist.ac.jp

Abstract

Understanding temporal information is essential for natural language understanding, yet both humans and large language models (LLMs) face challenges when narratives mix absolute and relative time expressions. In this study, we systematically evaluate frontier LLMs on temporal event ordering tasks and compare their performance against human baselines under absolute-time (AT) and mixed-time (MT) conditions. Using a recently constructed temporal reasoning dataset [1], we analyze representative models including GPT-4, DeepSeek Reasoner, and QwQ-32B.

Our findings reveal three key insights: (1) Frontier LLMs can achieve near-human Kendall's τ value in AT settings, with GPT-4 and DeepSeek Reasoner performing competitively. (2) In MT scenarios, human performance drops more sharply than some LLMs, suggesting that frontier models can maintain stronger consistency in long-text temporal reasoning with mixed time expressions. (3) Targeted probing with time masking confirms that LLMs rely heavily on explicit temporal anchors, showing fragility when such cues are removed.

These results demonstrate that temporal reasoning remains a core challenge for both humans and LLMs, while also revealing conditions under which models can rival—or even approach—human performance. Our analysis provides actionable insights for improving the interpretability and robustness of LLMs in temporally grounded language tasks.

1 Introduction

Temporal reasoning is a fundamental aspect of language comprehension, enabling interlocutors to reconstruct event sequences and interpret narratives in context. In natural discourse, however, temporal information is rarely expressed in a fully explicit or linear fashion. Narratives across languages frequently weave together absolute time references (e.g., in 1945), relative expressions (e.g., two years later), and event-anchored cues (e.g., shortly after the war), producing timelines that are non-linear and partially implicit. This mixture poses challenges not only for computational models, but also for human readers—particularly when temporal cues are sparse, distributed, or dependent on discourse structure.

Recent advances in LLMs have shown promising results in temporal reasoning tasks. Yet, most evaluations have been conducted in settings dominated by explicit temporal anchors or simplified temporal relations. Such conditions do not fully reflect the complexity of real-world narratives, including those found in historical accounts, biographies, and cross-cultural storytelling, where temporal markers are often underspecified or require inferential bridging. However, existing benchmarks rarely capture this mixture of absolute and relative time, and few studies have systematically tested how LLMs behave under these more naturalistic, mixed-time conditions (Chu et al., 2023; Wang and Zhao, 2023; Tan et al., 2023). Moreover, while recent studies have identified anchoring biases in LLMs (Huang et al., 2025) and systematic weaknesses in handling relative temporal expressions (Chen et al., 2025), little is known about whether current evaluation methods can reveal the sources of model errors more comprehensively—specifically, whether they arise from difficulties in anchoring absolute references, integrating relative cues, or bridging across discourse gaps. From a computational linguistics perspective, understanding how models—and humans—navigate such mixed-time conditions is key to developing systems that are both robust and interpretable across languages.

Building on a recently introduced mixed-time temporal reasoning benchmark (Sun et al., 2025), this study makes three contributions:

[1] https://github.com/fantastic-Feifei/MTS-benchmark

- Human–model comparative evaluation: We establish a human baseline for temporal event ordering under absolute-time (AT) and mixed-time (MT) conditions, quantifying the inherent difficulty and ambiguity of hybrid-time narratives.
- Probing temporal robustness: We conduct targeted masking experiments that selectively remove explicit year expressions, revealing the extent to which LLMs rely on surface-level temporal anchors.

Our findings show that while frontier LLMs can match or even exceed human performance in explicit AT settings, both humans and models struggle when temporal anchors are obscured or replaced with relative expressions. Crucially, masking a single year often triggers a collapse in model ordering accuracy, underscoring the importance of context integration beyond explicit timestamps. These results have direct implications for the design of temporally aware NLP systems and contribute to the broader understanding of how temporal reasoning operates across varied linguistic and narrative structures.

2 Related Work

Temporal Reasoning in NLP Temporal reasoning has been a long-standing challenge in natural language processing, involving tasks such as temporal expression normalization (Verhagen et al., 2010), event ordering (Chambers and Jurafsky, 2008; Ning et al., 2019), and temporal question answering (Khot et al., 2020; Chen et al., 2021). Most existing benchmarks, including TimeBank, MATRES, and TORQUE, primarily focus on scenarios with explicit absolute time anchors or simplified temporal relations. However, real-world narratives often exhibit *mixed temporal structures* that combine absolute, relative, and event-anchored expressions, forming non-linear timelines. Deriving a coherent global event order from such narratives remains challenging, especially when temporal cues are implicit or distributed across long contexts.

Large Language Models for Temporal Understanding Recent research has explored the ability of LLMs to perform temporal reasoning in complex narratives. Although state-of-the-art LLMs such as GPT-4 and DeepSeek demonstrate strong capabilities in general reasoning tasks, their performance often degrades on event ordering when explicit date cues are removed or when relative temporal expressions dominate (Xiong et al., 2024; Ding and Wang, 2025; Yuan et al., 2024). LLMs typically rely on surface-level temporal signals to achieve partial ordering consistency, but they struggle with non-linear or mixed time settings, where reasoning requires integrating both explicit and implicit temporal cues. This gap motivates the need for systematic evaluations of LLM temporal robustness under both AT and MT conditions.

Probing Methods for Model Reasoning Probing techniques offer a principled approach for investigating the internal reasoning behavior of LLMs (Belinkov and Glass, 2019; Elazar et al., 2021). In the context of temporal reasoning, one widely-used strategy is to mask time expressions—whether absolute or relative—during inference or intermediate training, and then evaluate the model's ability to reconstruct event order purely from context.

For example, Cole et al. 2023 introduce Temporal Span Masking, where temporal expressions are selectively masked during intermediate training to enhance performance on downstream temporal tasks. Similarly, the TempoBERT model (Rosin et al., 2021) employs explicit time masking incorporated into the model's inputs, boosting accuracy in temporal prediction tasks (Rosin et al., 2022). More recently, Liu et al., 2025 propose the Time-R1 framework, which includes a "masked time entity completion" subtask—directly analogous to our design—in its multi-stage training to assess the model's reliance on narrative context rather than explicit temporal markers.

Our method similarly adopts a targeted masking strategy by selectively removing absolute or relative time expressions and evaluating whether models can still recover the correct event sequence. Unlike prior masking-based probing methods that primarily evaluate lexical recovery, our design masks *absolute* time anchors (e.g., "in 1995"), which are crucial chronological cues in narrative texts. This allows us to directly test whether models can maintain temporal coherence and reason based on commonsense or contextual inference when explicit anchors are missing. Such a probing setup enables fine-grained analysis of model reliance on temporal cues, highlights differences between human and model reasoning, and supports controlled experiments assessing temporal robustness in long-context event ordering tasks.

3 Dataset and Task Setup

3.1 Dataset Overview

Statistic	Value
Total passages	4,824
Avg. events per passage	7.99
Avg. relative per passage	4.53
Avg. absolute per passage	3.46
Relative time ratio (relative / all)	56.73%

Table 1: Full dataset statistics, showing event density and distribution of absolute vs. relative time expressions.

We conducted our experiments on the hybrid-time temporal reasoning dataset, which contains 4,824 Wikipedia-style biographical passages (Table 1) with time-stamped events and various temporal expressions (Table 2). The dataset can be accessed from our GitHub repository (linked in the abstract).

Terminology

Global event refers to an event's position in the complete chronological sequence of a passage (e.g., *1960 → 1980 → 1990 → 2000*).

Local context denotes the immediately surrounding sentences that help determine an event's position.

Temporal cues are explicit or implicit indicators of time, which in our dataset include: **Absolute** (e.g., "in 1945"), explicit chronological anchors; **Relative** (e.g., "two years later"), requiring contextual inference; **Event-anchored** (e.g., "the end of 47th Olympics"), dependent on prior events.

This *mixed-time* design reflects natural narratives where explicit dates are interwoven with relative and discourse-dependent cues, requiring models to combine surface-level anchors with contextual reasoning. The coexistence of multiple expression types enables controlled comparisons of reasoning strategies and fine-grained robustness analysis.

Each passage is annotated with normalized temporal values and aligned to a global event sequence, supporting two evaluation settings: **AT** and **MT**. This design probes model sensitivity to explicit vs. implicit temporal cues and offers reusability for cross-linguistic evaluation, temporal QA, timeline extraction, and discourse analysis.

3.2 Dataset Sampling

To facilitate human annotation and maintain clarity in subsequent analysis, we sampled 100 passages from the full 4,824-instance dataset (Table 3), each containing between 4 and 7 event sentences. This range strikes a balance between narrative richness and human annotator feasibility. Our sampling decision was also guided by evidence from cognitive psychology showing that human comprehension of long, complex texts is constrained by working memory and processing limitations (Sugawara et al., 2020; Kalyuga, 2011). Specifically, when discourse length increases, readers must concurrently integrate earlier information while processing new content, which strains limited memory resources. Therefore, to ensure high-quality human annotation in our probing tasks, we limit the sample size to 100 passages while retaining narrative richness. As later confirmed in Section 4.1, humans exhibit lower consistency than models when reasoning over longer passages.

The selected subset preserves a diverse mixture of time expressions (absolute, relative) and event ordering structures. All 100 passages were independently annotated by three trained annotators and subsequently used for both human-model comparison and fine-grained probing experiments.

All annotators are domain experts in Information Science, with research backgrounds in machine learning and strong proficiency in English. This expertise ensures both familiarity with the technical aspects of the task and the linguistic competence required to process the biographical passages, providing a reliable baseline for comparison with model predictions. In addition, one of the annotators is the co-author of this work.

3.3 Task Definition

We define a sentence ordering task that requires models to recover the global temporal order of events described in natural language. Each input consists of a set of four event sentences sampled from a temporally rich passage. The model is prompted to output the correct chronological order of sentence indices (e.g., "1,3,2,4").

We evaluate model performance under two settings:

- **Absolute-Time (AT)**: All original absolute time expressions (e.g., "in 1945") are preserved.

Table 2: Examples of temporal expressions in our dataset, categorized by expression type: Absolute, Relative, and Event-Anchored.

Type (Temporal Expression)	Example
Absolute	"in 1995", "on September, 1920"
Relative	"three years later", "shortly after the Expo"
Event-Anchored	"the end of 47th Olympics", "during the Great Depression"

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 3: 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*.

- **Mixed-Time (MT)**: Some absolute expressions are rewritten into relative forms (e.g., "eight years later") or event-anchored references (e.g., "August 2024" rewritten as "the end of 47th Olympics") to simulate hybrid time contexts.

To further probe model dependency on explicit time anchors, we introduce a masked-time variant of this task, in which one temporal expression is replaced with a [MASK] token (see Section 3.5). Models must still recover the correct sentence order, revealing their temporal inference capabilities under partial information.

3.4 Model Selection

To evaluate temporal reasoning capabilities across a diverse range of model families and architectural scales, we selected the following representative large language models for analysis:

GPT-4 (OpenAI) (Achiam et al., 2023): A frontier proprietary model known for its strong general reasoning and instruction-following abilities. It serves as a high-performance reference in our evaluations.

DeepSeek-Reasoner (DeepSeek) (Guo et al., 2025): A reasoning-optimized model designed for multi-hop and structured inference tasks. It represents a model explicitly trained with a focus on reasoning capabilities.

DeepSeek-v3 (DeepSeek) (Liu et al., 2024): A general-purpose instruction-tuned model from the same family, included to contrast with the reasoning-augmented variant.

QwQ-32B (Alibaba) (Team, 2025): A large-scale open-source model with competitive performance in general benchmarks.

Qwen2.5-7B (Alibaba) (Yang et al., 2024): A strong open-source base model with relatively smaller scale (7B), chosen to assess how compact models perform under temporal reasoning tasks.

These models span different training paradigms (instruction tuning, reasoning augmentation), sizes (7B–32B), and sources (open-source vs. proprietary), enabling a broad comparison of temporal reasoning performance across architectural and methodological dimensions.

3.5 Probing Design

Masked Time Prediction

To further examine the temporal reasoning ability of language models, we design a *masked time prediction* task that targets a model's capacity to infer missing temporal information from surrounding context (Table 4).

For each passage in both the AT and MT settings, we randomly select a single sentence containing an absolute time expression (e.g., "1945") and mask only the year component with a [MASK] token. The remainder of the sentence and surrounding context are preserved, enabling us to isolate the model's ability to recover or approximate the masked temporal anchor based on event order and discourse-level clues. In the MT setting, the masked sentence may contain a rewritten relative reference (e.g., "eight years later"), making the task more reliant on understanding inter-event relations rather than directly reading explicit anchors.

We deliberately mask only one temporal expression per passage for three reasons:

1. **Experimental control:** Masking a single anchor allows us to attribute any performance change directly to the removal of that cue, avoiding confounding effects from masking multiple references simultaneously.

2. **Linguistic realism:** In naturally occurring narratives, explicit temporal anchors are often sparse but not entirely absent; removing just one simulates this partial loss of explicit cues.

3. **Interpretability:** With only one masked expression, we can more clearly trace whether the missing anchor disrupts global ordering, making error patterns easier to analyze.

Rather than evaluating the exact lexical reconstruction of the masked expression—which may admit multiple valid rewrites—we assess the model's ability to recover the correct global event order when the masked sentence is included. This design enables us to measure whether the removal of a key temporal anchor significantly degrades downstream temporal reasoning, thereby revealing the extent to which models depend on explicit time cues versus contextual inference.

Prompt Construction

To assess whether language models can preserve temporal reasoning capabilities when explicit time anchors are partially removed, we formulate a masked time prediction task as a sentence reordering problem. Each instance consists of four event sentences, one of which has its year expression masked (e.g., in 1997" → in [MASK]").

To ensure consistent and well-structured outputs across models, we adopt a one-shot prompting strategy. Preliminary experiments in the zero-shot setting showed that only GPT-4 reliably followed the expected output format (i.e., returning a comma-separated list of sentence indices such as "1,2,3,4"). In contrast, open-source models such as DeepSeek and Qwen often produced verbose, unstructured, or incomplete responses. To mitigate this, we include a concrete in-context example at the beginning of each prompt.

Each prompt is composed of two parts: a worked example followed by a test instance. Both follow the same structure—numbered sentences and an instruction asking the model to reorder them chronologically using index notation.

This design allows us to probe the impact of masking a single temporal anchor on the model's ability to infer global event order. By applying this setup to both AT and MT settings, we can evaluate the degree to which models rely on explicit time expressions versus contextual temporal reasoning.

A detailed example of our prompt format is provided in Appendix A.

3.6 Evaluation Metrics

To comprehensively assess the temporal reasoning capability of language models, we employ the following evaluation metrics:

- **Exact Match (EM):** For event ordering tasks, EM is a strict metric that checks whether the predicted event sequence exactly matches the gold order. Unlike rank-based correlation measures (e.g., Kendall's τ), which provide partial credit for partially correct rankings, EM only assigns credit when the entire sequence is perfectly correct.

- **Kendall's τ:** For sentence reordering tasks, we compute Kendall's τ rank correlation coefficient between the predicted and gold-standard sentence orders. This metric captures the *pairwise consistency* of temporal relationships between events, providing a gradient of correctness even when the full order is not exact.

In addition to aggregate scores, we conduct fine-grained analyses along the following dimensions:

- **Time Expression Type:** We compare model performance on absolute (e.g., "in 1982") vs. relative (e.g., "eight years later") expressions, to evaluate their sensitivity to different temporal formats.

- **Time Granularity:** We define granularity as the level of specificity expressed in temporal anchors (year, year+month, year+month+day). This is distinct from the *type* of temporal expression (e.g., absolute vs. relative vs. event-anchored), which we treat as a separate factor.

- **Context Length:** We investigate whether the number of surrounding events in a passage influences the model's ability to infer temporal relations, shedding light on context sensitivity.

Together, these metrics provide a multi-faceted view of model behavior, balancing surface-form fidelity with structural reasoning competence.

Original Context (Before)	**Masked Context**
He graduated in **1998** and started working the following year.	He graduated in **[MASK]** and started working the following year.
He became president in **2009** after a decade in parliament.	He became president in **[MASK]** after a decade in parliament.
She left the ministry in **March 2003** and returned briefly in 2005.	She left the ministry in **[MASK]** and returned briefly in 2005.

Table 4: Examples of original and masked contexts used in the probing task.

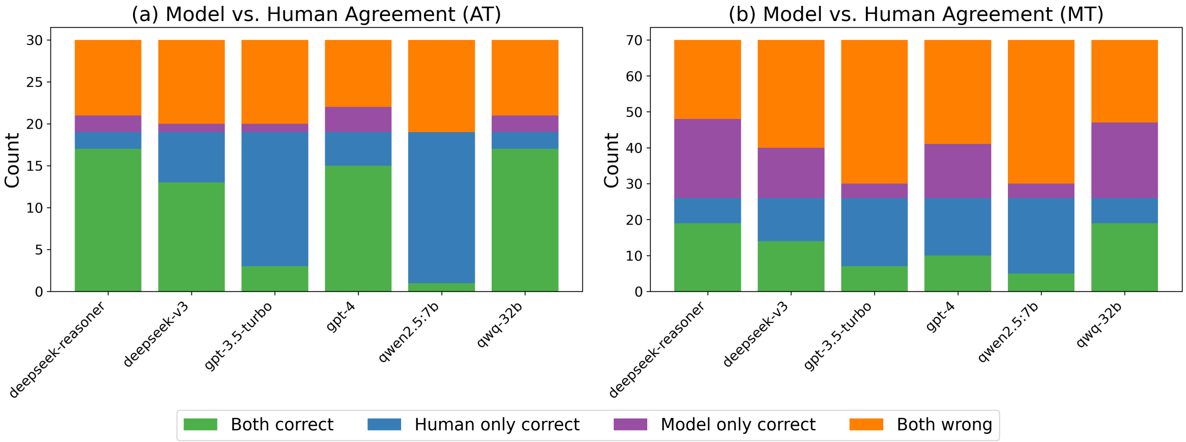

Figure 1: **Model vs. Human Agreement under AT and MT settings.** Stacked bar plots show the distribution of prediction outcomes for each model: *Both correct* (green), *Human only correct* (blue), *Model only correct* (purple), and *Both wrong* (orange). (a) **Absolute Time (AT):** Models such as DEEPSEEK-REASONER, 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.

4 Results and Analysis

4.1 Overall Model Performance

To establish a performance baseline, we first evaluate six language models on the temporal ordering task under two settings: AT and MT, using two metrics: exact match (EM) accuracy and Kendall's τ rank correlation. Table 5 summarizes the results for four representative models and three human annotators.

Performance under AT Under the AT setting, all models achieve relatively high accuracy, with DEEPSEEK-REASONER and QWQ-32B both reaching an EM score of 0.63. GPT-4 also performs well with an EM of 0.60 and the highest Kendall's τ of 0.69, suggesting strong global ordering consistency. Compared to human annotators, most models perform on par or slightly better in EM, though ANNOTATOR 2 achieves the highest τ score of 0.73, indicating the strongest ordering alignment with gold labels.

Reliable Yet Nuanced: Human Baseline Under AT Interestingly, all three annotators achieved an identical exact match (EM) score of 0.56 under the AT setting, despite differing moderately in their Kendall's τ values. This suggests that while their full-sequence predictions aligned with the gold order in the same proportion of cases, the extent to which their orderings agreed with the gold ranking varied. Such consistency in EM across annotators reinforces the reliability of the human baseline under AT, and highlights that partial sequence disagreements (captured by τ) may still occur even when EM scores coincide.

Performance under MT The MT setting introduces a pronounced performance drop for both human annotators and models. While DEEPSEEK-REASONER and QWQ-32B retain relatively strong performance (EM = 0.60 and 0.59, respectively), other models such as DEEPSEEK-V3 and GPT-4 show substantial degradation (EM = 0.41 and 0.42). Human performance declines even more noticeably: all three annotators exhibit reduced EM and Kendall's τ scores, with ANNOTATOR 1 dropping to an EM of just 0.30.

Humans Struggle, Models Endure under MT This gap between human and model performance may be attributed to the increased cognitive load imposed by long passages and temporally ambiguous references. Unlike the AT setting—where time expressions are explicit and reasoning is more straightforward—the MT condition requires interpreting implicit and relative temporal cues, often embedded within complex narratives. These challenges appear to hinder human consistency, whereas models like DEEPSEEK-REASONER and QWQ-32B demonstrate notable robustness, suggesting their superior ability to track and resolve temporal structure in long-context settings.

GPT-4 Mimics Human Temporal Reasoning Patterns Among all evaluated models, GPT-4 consistently demonstrated performance most closely aligned with that of human annotators. In the AT setting, its Kendall's τ of 0.69 is only marginally above the human range (0.62–0.73), and in the MT setting, its Kendall's τ of 0.46 remains within the variance of human annotators (0.32–0.50) (Table 5). This closeness in rank correlation suggests not only comparable accuracy but also similar ordering tendencies, indicating a reasoning style that aligns with human temporal judgments.

We hypothesize that GPT-4's relatively human-like behavior may stem from its training paradigm. Unlike open-source models such as DEEPSEEK-REASONER and QWQ-32B, which are often trained with strong emphasis on instruction tuning and retrieval-augmented generation, GPT-4 incorporates extensive reinforcement learning from human feedback (RLHF). This iterative alignment process likely encourages the model to mimic human preferences and inference styles, especially in ambiguous or underspecified contexts.

In contrast, DEEPSEEK-REASONER and QWQ-32B exhibit more decisive but less human-consistent behavior. Their superior performance under the MT condition suggests stronger capabilities in long-context tracking and structured reasoning, yet their predictions often diverge from human tendencies, possibly due to a more pattern-driven or memorization-based inference mechanism shaped by large-scale instruction-following pretraining.

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 5: 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.

Taken together, these findings suggest that GPT-4, while not always achieving the highest EM scores, may employ a reasoning process that is cognitively closer to human temporal understanding—a property valuable for downstream applications requiring interpretability or human-in-the-loop decision making.

Performance of Qwen2.5-7B and GPT-3.5-Turbo Both QWEN2.5-7B and GPT-3.5 underperform compared to larger frontier models. While QWEN2.5-7B yields near-random orderings (AT: EM=0.03, τ=0.12; MT: EM=0.15, τ=0.20), GPT-3.5 achieves slightly higher consistency (AT: EM=0.13, τ=0.30), yet still lags behind human annotators and fails to scale under MT (EM=0.16, τ=0.08). These results suggest that both models struggle with global event ordering, albeit for different reasons: QWEN2.5-7B appears particularly weak in leveraging absolute anchors, while GPT-3.5 shows instability in mixed-time reasoning.

Due to its consistently poor performance across both conditions, we exclude QWEN2.5-7B,GPT-3.5-TURBO, and DEEPSEEK-V3 from subsequent probing analyses.

4.2 Human-Model Agreement Patterns

To better understand the consistency between model predictions and human judgments, we analyze agreement patterns across six representative models under both AT and MT conditions. Each instance in our dataset was independently annotated by three human annotators. To enable consistent comparison with model outputs, we adopted a majority vote strategy to determine the gold-standard human response for each instance.

Figure 1 presents stacked bar plots that categorize each prediction outcome into one of four types: *Both correct*, *Human only correct*, *Model only correct*, and *Both wrong*, based on comparison with the majority-vote human answer.

In the AT setting (Figure 1a), most models—including DEEPSEEK-REASONER, GPT-4, and QWQ-32B—achieve high agreement with human annotators, with a substantial proportion of cases falling into the *Both correct* category. However, DEEPSEEK-V3 and QWEN2.5-7B exhibit relatively higher numbers of *Human only correct* cases, suggesting challenges in precise recovery of masked time expressions.

In contrast, under the MT setting (Figure 1b), all models show performance degradation. The number of *Both correct* cases drops notably, while *Model only correct* and *Both wrong* instances increase. This trend highlights the difficulty of temporal reasoning in contexts with ambiguous or relative time expressions. Models such as QWEN2.5-7B and DEEPSEEK-V3 especially struggle, with many instances correctly identified by humans but missed by the models.

These findings underscore the value of probing model behavior across both structured and ambiguous temporal settings, as performance under absolute time does not necessarily generalize to more naturalistic, mixed-time scenarios.

Quantifying Human–Model Alignment To complement the categorical outcome analysis, we further quantify the degree of alignment between models and human annotators using Kendall's τ. In addition to correlations with gold orders (Section 4.1), we directly compute the correlation between each model prediction and each individual annotator's sequence, averaging across annotators. This provides a finer-grained measure of human–model agreement beyond majority-vote correctness.

Table 6 summarizes the results under both AT and MT conditions. Table 6a reports Kendall's τ for model–gold and human–gold comparisons, while Table 6b presents direct model–human correlations. Several patterns emerge: (1) frontier models such as DEEPSEEK-REASONER, GPT-4, and QWQ-32B show strong alignment with both gold orders and human judgments in the AT setting; (2) QWEN2.5-7B consistently lags behind, particularly in the MT condition, reflecting its difficulty with ambiguous or relative time references; and (3) across most models, direct model–human correlations are slightly lower than model–gold ones (e.g., GPT-4, DEEPSEEK-V3), highlighting that high accuracy with respect to gold standards does not always translate into close alignment with human reasoning.

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

4.3 Probing Model Temporal Inference via Time Masking

To investigate whether LLMs can maintain temporal reasoning ability when explicit time anchors are removed, we conduct a masked time prediction experiment in both AT and MT settings. In each instance, a single year expression is masked in one of the event sentences, and models are prompted to output the correct chronological order of the events (as described in Section 3.5). This setup isolates the model's reliance on explicit temporal cues and evaluates its ability to infer event order from context alone.

Results. Table 7 summarizes model performance under the masked time probing task. In the AT setting, GPT-4 achieves an Exact Match (EM) of only 0.033, with a negative average Kendall's τ

of -0.059, indicating that its predicted orders are slightly worse than random. QwQ-32B performs comparably, while DeepSeek-Reasoner completely fails (EM = 0). The MT setting is even more challenging: all models fail to recover any correct orders (EM = 0), and Kendall's τ approaches zero or negative, reflecting near-random or even inversely correlated rankings.

Setting	Model	EM	Avg Kendall τ
AT	GPT-4	0.033	-0.059
	QwQ-32B	0.033	-0.056
	DeepSeek-Reasoner	0.000	-0.006
MT	GPT-4	0.000	-0.026
	QwQ-32B	0.000	-0.050
	DeepSeek-Reasoner	0.000	0.015

Table 7: Performance of LLMs under the masked time probing task. EM denotes the fraction of exact matches to the gold event order; Kendall's τ measures correlation between predicted and gold orders.

Analysis and Insights.

Our probing results reveal three key insights:

1. **Strong reliance on explicit temporal anchors.** Masking just a single year drastically degrades performance, indicating that models primarily leverage surface-level time expressions rather than robust event reasoning.

2. **Failure to generalize in mixed-time narratives.** In MT settings, where relative or vague temporal expressions dominate, masking any remaining anchor causes complete failure, with EM = 0 for all models.

3. **Temporal reasoning collapses without explicit cues.** Negative or near-zero Kendall's τ scores suggest that model predictions are effectively random, and sometimes even inversely correlated with the true order.

These findings demonstrate that current LLMs exhibit *shallow temporal reasoning*, heavily dependent on explicit timestamps. Our probing methodology thus exposes a critical weakness in LLM temporal inference, highlighting the necessity for models that can robustly infer event order in partially observed or naturally vague timelines.

5 Conclusion

This study examined the temporal reasoning abilities of frontier LLMs and human annotators under both AT and MT conditions, using a hybrid-time dataset that integrates absolute, relative, and event-anchored expressions. By combining controlled evaluation with a masked-time probing task, we identified clear performance gaps and reasoning patterns: while frontier models can approach human-level ordering accuracy in AT settings, both humans and models struggle when explicit temporal anchors are reduced or removed. The masking experiments further revealed a strong dependence on explicit cues, with ordering accuracy collapsing when even a single anchor is missing.

From a broader perspective, these findings highlight that temporal reasoning in naturally occurring narratives is shaped by more than surface-form date recognition: it requires resolving vague relative expressions, reconciling narrative sequencing with chronological order, and integrating long-distance discourse anchors. The hybrid-time design adopted here provides a reusable, language-agnostic framework for diagnosing such challenges, and can be readily adapted to other languages and narrative genres.

Future work will expand the dataset to multilingual settings, enabling cross-linguistic comparisons of temporal reasoning strategies. We also aim to explore model architectures and training objectives that promote robust integration of explicit and implicit temporal cues, moving towards temporally aware systems capable of handling the complexities of real-world discourse.

Acknowledgements

Feifei Sun acknowledges the support of the China Scholarship Council (CSC) under scholarship number 202408050023.
This work was also supported partly by JSPS KAKENHI Grant Number JP25H00459. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the author(s)' organization, JSPS or MEXT.
We thank all reviewers for their insightful and constructive feedback, which has helped us improve the clarity, precision, and completeness of this work.

Limitations

Our study has several limitations. First, the probing experiments are conducted on a small 100-sample subset to ensure human annotation feasibility and interpretability. While this allows for detailed human-model comparison, it limits the statistical generalizability of our findings to larger-scale temporal reasoning tasks.

Second, our current probing approach masks only one time expression per passage and evaluates its impact on event ordering. This simplified design highlights model reliance on explicit temporal cues but does not fully capture the complexity of real-world narratives with multiple missing or ambiguous temporal anchors.

Third, our evaluation focuses on surface-level ordering accuracy (EM and Kendall's τ) and does not analyze intermediate reasoning steps or latent temporal representations. As a result, our conclusions about model reasoning are inferential and may not fully reveal the internal mechanisms driving model predictions.

Fourth, although we include multiple model families (GPT-4, DeepSeek, QwQ), our selection omits smaller or domain-specialized models, and we exclude QWEN2.5-7B, GPT-3.5-TURBO and DEEPSEEK-V3 from probing due to its low baseline performance.

Fifth, a potential concern is that our passages originate from Wikipedia biographies, raising the possibility that models may have partially memorized specific dates or event sequences during pretraining. To mitigate this risk, during dataset construction, we performed data cleaning and randomization procedures to reduce direct overlap with seen text (details described in our companion work (Sun et al., 2025)). In particular, events were extracted and re-ordered into new narrative contexts, such that models could not rely on surface recall of document-level sequences.

Nevertheless, we acknowledge that memorization cannot be entirely excluded, as noted in our Limitations section. Importantly, our evaluation requires models to reconstruct global event orders across passages: even if individual dates were known, successful performance hinges on reasoning over relative and hybrid time references rather than verbatim recall.

Finally, our design of the Masked-Time probing task primarily aims to stress-test model robustness when crucial temporal anchors are missing. As the masked expression may admit multiple plausible human interpretations, constructing a unique human "gold" reference would be problematic. Therefore, we did not collect human annotations for this setting. We acknowledge this as a limitation, since direct human-model comparison could further illuminate the gap in reasoning strategies.

Future work should expand the model coverage and explore more fine-grained temporal reasoning diagnostics.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797.

Shuang Chen, Yining Zheng, Shimin Li, Qinyuan Cheng, and Xipeng Qiu. 2025. 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.

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

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*.

Jeremy R Cole, Aditi Chaudhary, Bhuwan Dhingra, and Partha Talukdar. 2023. Salient span masking for temporal understanding. *arXiv preprint arXiv:2303.12860*.

Xi Ding and Lei Wang. 2025. Do language models understand time? In *Companion Proceedings of the ACM on Web Conference 2025*, pages 1855–1868.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

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

Slava Kalyuga. 2011. Cognitive load aspects of text processing. *Cross-Disciplinary Advances in Applied Natural Language Processing: Issues and Approaches*, pages 114–132.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. Qasc: A dataset for question answering via sentence composition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8082–8090.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Zijia Liu, Peixuan Han, Haofei Yu, Haoru Li, and Jiaxuan You. 2025. Time-r1: Towards comprehensive temporal reasoning in llms. *arXiv preprint arXiv:2505.13508*.

Qiang Ning, Zhili Feng, and Dan Roth. 2019. A structured learning approach to temporal relation extraction. *arXiv preprint arXiv:1906.04943*.

Guy D Rosin, Ido Guy, and Kira Radinsky. 2022. Time masking for temporal language models. In *Proceedings of the fifteenth ACM international conference on Web search and data mining*, pages 833–841.

Saku Sugawara, Pontus Stenetorp, and Akiko Aizawa. 2020. Benchmarking machine reading comprehension: A psychological perspective. *arXiv preprint arXiv:2004.01912*.

Feifei Sun, Ziyi Tong, Houjing Wei, Cheng Peng, Teeradaj Racharak, and Le-Minh Nguyen. 2025. Benchmarking temporal reasoning: Can large language models navigate time when stories refuse to follow a straight line? In *First Workshop on Foundations of Reasoning in Language Models (FoRLM@NeurIPS)*.

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.

Marc Verhagen, Roser Sauri, Tommaso Caselli, and James Pustejovsky. 2010. Semeval-2010 task 13: Tempeval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 57–62.

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*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

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.

A Prompt Example for Masked Time Prediction

Below is a representative one-shot prompt used in our masked time prediction (sentence reordering) experiment. The example demonstrates the format provided to language models during inference:

> `Here is an example: Input Sentences:`
>
> 1. He was born in 1960.
>
> 2. He graduated in 1980.
>
> 3. He joined IBM in 1990.
>
> 4. He became a manager in 2000.
>
> `Answer:` 1,2,3,4
>
> `Now reorder the following sentences in chronological order. Respond only with the sentence indices in the correct order, separated by commas.`
>
> `Input Sentences:`
>
> A campaign, called Save the Ampelmännchen, was launched by the public and Ampelmännchen enthusiasts, resulting in the preservation of Peglau's Ampelmännchen in [MASK].
>
> Karl Peglau died in Berlin, Germany, on 29 November 2009, at the age of 82.
>
> Karl Peglau was a German traffic psychologist who invented the iconic Ampelmännchen traffic symbols used in the former East Germany in 1961.
>
> Peglau designed the glass human figures for the stop (red) and go (green) lights on the traffic signal in 1961, which became known as the Ampelmännchen.
>
> `Answer:` ___

In this example, the first sentence contains a masked year expression. The model must infer its correct position within the *global event timeline*—the ordered sequence of events in the passage, whether represented as explicit dates (e.g., *1960* → *1980* → *1990* → *2000*) or as major life milestones (e.g., *birth* → *graduation* → *career start* → *promotion*)—by leveraging both local context and surrounding temporal cues.