Structured yet Bounded Temporal Understanding in Large Language Models

Anonymous ACL submission

Abstract

Large language models (LLMs) increasingly show strong performance on temporally grounded tasks, such as timeline construction, temporal question answering, and event ordering. However, it remains unclear how their behavior depends on the way time is anchored in language. In this work, we study LLMs' temporal understanding through temporal frames of reference (t-FoRs), contrasting deictic framing (past-present-future) and sequential framing (before-after). Using a large-scale dataset of real-world events from Wikidata and similarity judgement task, we examine how LLMs' outputs vary with temporal distance, interval relations, and event duration. Our results show that LLMs systematically adapt to both t-FoRs, but the resulting similarity patterns differ significantly. Under deictic t-FoR, the similarity judgement scores form graded and asymmetric structures centered on the present, with sharper decline for future events and higher variance in the past. Under sequential t-FoR, similarity becomes strongly negative once events are temporally separated. Temporal judgements are also shaped by interval algebra and duration, with instability concentrated in overlap- and containment-based relations, and duration influencing only past events under deictic t-FoR. Overall, these findings characterize how LLMs organize temporal representation under different reference structures and identify the factors that most strongly shape their temporal understanding.

1 Introduction

Temporal understanding is a fundamental component of human intelligence, allowing us to represent, interpret, and reason about time and how events unfold in relation to it. Large language models (LLMs) have demonstrated strong performance on natural language understanding tasks, including temporally grounded tasks, such as temporal question answering (Zhou et al., 2019; Gurnee and Tegmark, 2023; Islakoglu and Kalo, 2025), timeline construction (Nylund et al., 2024; Bazaga et al., 2025), and event prediction (Dhingra et al., 2022), and narrative understanding (Min et al., 2023). However, temporal understanding still remains a challenge to LLMs (Zhou et al., 2019; Jain et al., 2023; Wallat et al., 2024; Qiu et al., 2024).

Various notions of time have been used in temporal understanding in the realm of natural language understanding. Some approaches adopt an absolute time where events are anchored to normalized timestamps and evaluate whether the systems are able to answer temporally grounded questions. Other works adopt a more relational time that models the relations between events. A mixed notion of time combines both views that align events to an absolute time and reason over their relative order. Among these notions, time is treated as a structural resource for ordering and constraining events.

From a different perspective, linguistic studies show that time is encoded through tense, aspect, temporal adverbials, and discourse structure, and cognitive science works have argued that the representation of time is metaphorical of space (McGlone and Harding, 1998; Lakoff et al., 1999; Gentner et al., 2002). As LLMs are trained on natural language, the data does not reflect the raw physical time, but linguistically and cognitively shaped representations of time. From this perspective, LLMs inherit the temporal abstractions encoded in natural language, including spatiotemporal representation (Gurnee and Tegmark, 2023) and temporal intervention (Nylund et al., 2024). Therefore, the appropriate question is more about how well LLMs capture the kind of temporality natural language itself encodes.

As human beings possess the ability to dynamically reason about temporal relations among events relative to a reference point under frames, we draw on the distinction between two temporal frames of reference (t-FoRs): deictic and sequential t-

FoRs (Evans, 2013). A t-FoR specifies how temporal relations are organized relative to a reference point. For example, in deictic t-FoR, the reference point is anchored in the experiencer's "now", or the "ego" in other terms, allowing the notions of past/present/future; in sequential t-FoR, it is anchored in a specific event to create the before/after notions. In this work, we attempt to explore LLMs' temporal understanding through the lens of t-FoRs. By manipulating reference point in each t-FoR, we can probe how LLMs adapt their temporal understanding and where their representations exhibit uncertainty. Inspired by recent work on language models' temporal cognition(Li et al., 2025), we incorporate similarity judgement task to observe the induced patterns. We aim to answer the following questions:

- **RQ1**: Do large language models adapt to different temporal frames of reference?
- **RQ2**: What factors affect the temporal understanding in large language models?

Our contributions can be summarized as follows:

- We introduce a frame-sensitive evaluation framework for temporal understanding in large language models, contrasting deictic and sequential t-FoRs through a similarity judgement task applied to real-world events.
- We propose a large-scale temporal dataset from Wikidata that preserves interval structure and duration, enabling systematic analysis of temporal distance, Allen's interval algebra, and event duration at scale.
- We provide comprehensive empirical evidence that LLMs' temporal behaviors are strongly frame-dependent and shaped by structural temporal factors, including past-future asymmetry under deictic t-FoR, rapid similarity collapse under sequential t-FoR, elevated variability for overlap-based relations, and duration effects retricted to past events.

2 Related Work

Research on temporal understanding in NLP has made substantial progress, but much of it implicitly targets specific levels of the above categories. A large body of works focuses on tasks such as temporal expression normalization (Bethard and Martin, 2007; Strötgen and Gertz, 2010), event ordering (Chambers and Jurafsky, 2008; Ning et al., 2018a,b), timeline construction (Do et al., 2012), and temporal question answering (Zhou et al., 2019; Chen et al., 2021). These tasks primarily evaluate a model's ability to represent and reason about temporal relations among events - that is, how events are positioned with respect to one another in time. More recent work explores time-aware embeddings, temporal pretraining objectives, and structured temporal representations, further strengthening this capability.

2.1 Temporal Reasoning Benchmarks

Question-answering datasets such as SQuAD (Rajpurkar et al., 2016, 2018) and Natural Questions (Kwiatkowski et al., 2019) construct temporal questions within a single, fixed time period. However, some questions may have different answers depending on when they are asked. To address this, several datasets include time-sensitive questions, such as TimeQA (Chen et al., 2021), TempLAMA (Dhingra et al., 2022), SituatedQA (Zhang and Choi, 2021), StreamingQA (Liska et al., 2022), RealtimeQA (Kasai et al., 2023), TempReason (Tan et al., 2023), and MenatQA (Wei et al., 2023).

Temporal commonsense knowledge, such as a fact that a leap year generally occurs every four years, is another important aspect of temporal reasoning. Datasets like MCTACO (Zhou et al., 2019) and TimeDial (Qin et al., 2021) focus specifically on evaluating models' understanding of temporal commonsense.

Beyond datasets targeting specific temporal properties or sequential relations, comprehensive benchmarks such as TRAM (Wang and Zhao, 2024) and TimeBench (Chu et al., 2024) evaluate temporal reasoning across multiple tasks. Additionally, datasets such as TimeQuestions (Jia et al., 2021), ExpTime (Yuan et al., 2024), TGQA (Xiong et al., 2024), and ToT (Fatemi et al., 2024) leverage knowledge graphs to assess models' understanding of time-event and event-event relations. Datasets such as ChronoSense (Islakoglu and Kalo, 2025) utilized Allen's interval algebra (Allen, 1983) to assess LLMs' temporal understanding capability.

2.2 Frame of Reference

A frame of reference (FoR) is a coordinate system used to determine the position of a figure relative to a ground from a particular perspective (Talmy, 2003), typically corresponding to the viewpoint of

an observer.

From a philosophical perspective, time can be conceptualized as either an A-series or a B-series (McTaggart, 1908). The A-series represents a tensed and dynamic view of time, where events are located relative to the observer's subjective "now" (the deictic center) as past, present, or future. In this view, a future event successively becomes present and eventually shifts into the past. In contrast, the B-series reflects a tenseless and static ordering of events, where each event is positioned as earlier than or later than another, independent of any observer's temporal standpoint.

From a cognitive perspective, time is commonly described using two conceptual metaphors: moving ego (ME) and moving time (MT) (McGlone and Harding, 1998; Boroditsky, 2000; Gentner et al., 2002). In the ME perspective, time is stationary while the observer (ego) moves forward through it, such as "We are approaching Tuesday". In the MT perspective, the observer remains stationary while time itself flows toward them, for example, "Tuesday is approaching".

A temporal frame of reference (t-FoR) typically involves three components: a target event (TE), a reference point (RP), and an origo (O). The TE denotes the event being fixed, the RP provides the point of comparison relative to the TE, and the O serves as the anchoring perspective. As on the of the first attempts to integrate temporal perspectives within a FoR framework, Moore distinguishes between ego-centered MT and non-deictic MT, proposing an ego-based vs. field-based taxonomy to capture these differences. Extending this approach, Moore groups ME and ego-centered MT as "ego-perspective", corresponding to the A-series, while field-based perspectives correspond to the B-series. The reference-point metaphors framework splits the MT perspective into two categories: ego-based and field-based, where former merges with the ME perspective under a deictic clasification (Núñez and Sweetser, 2006). The temporal framework models taxonomy integrates reference-point metaphors with descriptions of time across all three spatial frames of reference (s-FoRs) (Kranjec, 2006). Building on these developments, Evans aligns the A-series with ego-based frames (Moore, 2004) and ego-RP metaphors (Núñez and Sweetser, 2006), and the B-series with field-based frames (Moore, 2004) and time-RP metaphors (Núñez and Sweetser, 2006). Other revisions classify ME and MT as ego-relative temporal motion constructions, while sequence-based conceptions are termed positional time constructions (Sinha et al., 2011). Tenbrink characterizes time as possessing an "inbuilt asymmetry", which can be conceptualized either as a vector from past to future or as a vector expressing anteriority/posteriority in event sequences.

2.3 Temporal Reasoning in Language Models

Although Large Language Models (LLMs) have demonstrated impressive reasoning capabilities through in-context learning and post-training techniques, understanding and reasoning about temporal information remains challenging (Jain et al., 2023). Recent studies have highlighted that LLMs particularly struggle with relative time references, randomized time references, and typical temporal questions (Wallat et al., 2024).

Prior work has sought to enhance LMs' temporal understanding through several approaches, including post-training with external knowledge (Yuan et al., 2024; Tan et al., 2024; Xiong et al., 2024), leveraging code-format representations (Li et al., 2023; Zhu et al., 2023), and improving generalization across diverse temporal tasks (Su et al., 2024).

However, these approaches leave two issues underspecified. First, the frame of reference governing temporal judgements is frequently implicit, making it difficult to distinguish deictic from sequential reasoning. Second, temporal relations are often treated symmetrically or discretely, obscuring graded uncertainty and relation-specific difficulty.

3 What Temporal Understanding Means for Language Models

It is necessary to clarify what is reasonable to expect from language models (LMs) in terms of temporal understanding. In this work, we use the term 'event" as a temporally bounded occurrence that can be described and related to other occurrences. Our goal is to study how language models organize such events in time.

LMs do not encounter time as human beings do, but only through what is encoded in natural language. Therefore, whatever temporal understanding LMs may acquire is necessarily mediated by linguistic and cognitive representation of time that is present in human discourse. This makes the temporal structure available to LMs being perspectival, different from physical or metaphysical notions of time, that LLMs do not have access to absolute time nor do they experience lived temporality.

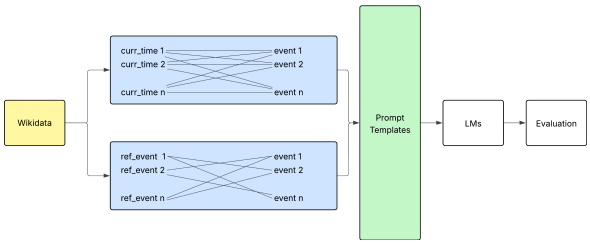

Figure 1: Overall benchmark steps.

This places LLMs in an intermediate place between the philosophical and cognitive views. On one hand, LLMs model temporal structure as stable relations between events, resembling what McTaggart (McTaggart, 1908) classified as B-series structure: the "earlier than" and "later than" relations between events, as prior works have shown (Xiong et al., 2024). On the other hand, natural language also encodes so called A-series structure that distinguishes positions from past to present to future, relative to an observer's viewpoint. These perspectival distinctions arise from how the model is situated within it. Though prior works have shown success on temporally grounded tasks, but they rarely test whether LLMs can adopt and update temporal perspectives.

Temporal frames of reference (t-FoRs) provide a way to probe this capacity. In a deictic t-FoR, temporal relations are anchored through experiencer-based reference strategy, allowing distinctions between past, present, and future. In a sequential t-FoR, the reference strategy is changed to event-based, where the reference point is anther event, creating before and after relations. Both frames are pervasive in language, linguistically grounded, and cognitively motivated (Evans, 2013).

We propose a frame-dependent benchmark that aims to probe temporal understanding in LLMs. Intead of evaluating factual correctness from multiple choices format, we measure graded similarity judgements as a function of temporal relations, temporal distance, and reference-point manipulation. As illustrated in Figure 1, the benchmark adopts an experiencer-based reference strategy from the deictic t-FoR and event-based strategy from the sequential t-FoR.

3.1 Event Extraction

We constructed a larger scale dataset by extracting real events from Wikidata (Vrandečić and Krötzsch, 2014). The extracted finite set of events $E = e_1, e_2, ..., e_n$, where each event $e_i = (t_i^{start}, t_i^{end}, l_i)$ is represented by a start time, end time, and name. We end up with 49,956 events dated from "0001-01-01" to "2100-12-31".

To avoid that models rely on memorized knowledge of specific historical events, we anonymized all event names using a list of rare animal names, while keeping start and end timestamps unchanged. The mapping from original to anonymized names is stored for prompt construction. Additionally, because events are temporally extended rather than instantaneous, we leveraged Allen's Interval Algebra (Allen, 1983) to characterize the full range of possible relations between two events. For later analysis and for balanced coverage across relations, we randomly sample 321 anonymized events such that all 13 Allen relations are represented.

3.2 Temporal Frames of Reference

We evaluate temporal understanding under two distinct temporal frames of reference (t-FoR), following Evan's taxonomy (Evans, 2013).

In the deictic t-FoR, temporal relations are established relative to a reference time $\tau \in T$, which is treated as the experiencer's "now". A sequence of reference times of "now", $T_{now} = \tau_1, \tau_2, ..., \tau_m$, is created by enumerating from the earliest event start time to the latest event end time. Therefore, the data pairs are constructed as (τ, e_i) for $\tau \in T$ and $e_i \in E$.

In the sequential t-FoR, the reference point is not an external "now" but an event itself. For each event $e_j \in E$, the temporal relations between e_j, the reference event, and all other events e_i are computed. The resulted data pairs are (e_i, e_j) for $e_i, e_j \in E$ and $e_i \neq e_j$.

For each data pair, both temporal relation and distance are calculated. The deictic t-FoR yields past/present/future relations and before/after for sequential t-FoR. The temporal distance is calculated in the unit of a day:

$$\Delta(r, e_i) = \begin{cases} t_i^{end} - t_r^{start}, & \text{if } t_r^{start} >= t_i^{end} \\ t_i^{start} - t_r^{end}, & \text{otherwise} \end{cases}$$

where r stands for the reference time/event and e_i refers to the target event. The temporal distance is signed that negative values correspond to past/before relations; 0 indicates present relation; and positive values means future/after relations.

3.3 Similarity Judgement

Inspired by recent works (Marjieh et al., 2022; Li et al., 2025), we leveraged similarity judgement

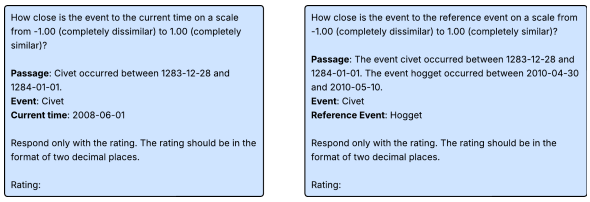

Figure 2: Prompt examples for similarity judgement. Left prompt is for deictic temporal frame of reference. Right prompt is for sequential temporal frame of reference.

task to probe LLMs to rate how similar the reference time/event is to the target event. Specifically, for each data pair (r, e_i), where r is the reference time/event, the tested LLMs are prompted to output a similarity score:

$$s_{LLM}(r, e_i) \in [-1.00, 1.00]$$

where 1.00 indicates mostly similar and -1.00 is completely dissimilar. The two prompt templates are shown in Figure 2.

Overall, the dataset contains 2,916,927 data entries for deictic t-FoR and 51,360 data entried for sequential t-FoR.

4 Experiment Setup

In this work, we focused on evaluating two open-source large language models (LLMs) with different sizes: Qwen3-4B (Qwen3-4b-instruct) and Qwen3-30B (Qwen3-30b-a3b-instruct-2507) (QwenTeam, 2025). The models can generate at most 64 new tokens for answers and scores are extracted using regex rules.

4.1 Measurement

For each signed temporal distance $\Delta = i$, the mean similarity score is computed by aggregating all outputs associated with that distance:

$$\overline{s}^i = \frac{1}{n_i} \sum_{j=1}^{n_i} s_{LLM}^{(j)}$$

where n_i is the total number of data pairs whose temporal distance equals i.

Visualizing the resulting mean similarity values as a function of temporal distance yields a characteristic response pattern that reflects the model's temporal cognition. Under a deictic t-FoR, it is expected to manifest as a maximum around the present reference point, with similarity gradually decresing as events move further into the past or future. Deviations from this pattern indicate asymmetries or instability in how the model interprets temporal relations as the deictic anchor shifts.

Under a sequential t-FoR, a different pattern is expected as the before/after relations are symmetric with respct to temporal ordering, so that similarity scores should roughly result in a symmetric distribution centered around zero temporal distance.

4.2 Allen's Interval Algebra Analysis

In addition to signed temporal distance, we further analyze LLMs' behavior patterns by leveraging Allen's Interval Algebra (Allen, 1983), which defines 13 mutually exclusive temporal interval relations between two intervals: *equals*, *before*, *after*, *overlaps*, *overlapped-by*, *contains*, *during*, *started-by*, *starts*, *finished-by*, *finishes*, *meets*, and *met-by*.

For each data pair, we computed the interval relations using the corresponding start and end times, allowing us to investigate whether LLMs certain interval relation introduce higher representational uncertainty. During evaluation, we grouped similarity scores by interval relation and computed the variance within each group. Higher variance indicates that the model has less consistent temporal assessments for that specific interval relation, even when temporal distance is comparable. This analysis allows us to identify where LLM judgements are most and least stable across the full structural space.

5 Results

5.1 Function of Temporal Distance

Figure 3 shows the mean similarity score (red dots) against signed temporal distance for each data pair, together with the standard deviation across instances at the same distance (blue vertical bars). The left two plots correspond to the deictic t-FoR, and the right two plots correspond to the sequential t-FoR.

Under the deictic t-FoR, we observe a clear asymmetric distance effect for both LLMs. Similarity ratings peak near zero temporal distance and drop sharply for future events (positive distances), where scores rapidly converge to -1.00 with comparatively low variance. In contrast, past events (negative distances) exhibit a more gradual decline in similarity, accompanied by substantially higher variance, especially for events occurring tens of thousands of days before the reference point. This indicates that the models' treatment of the future

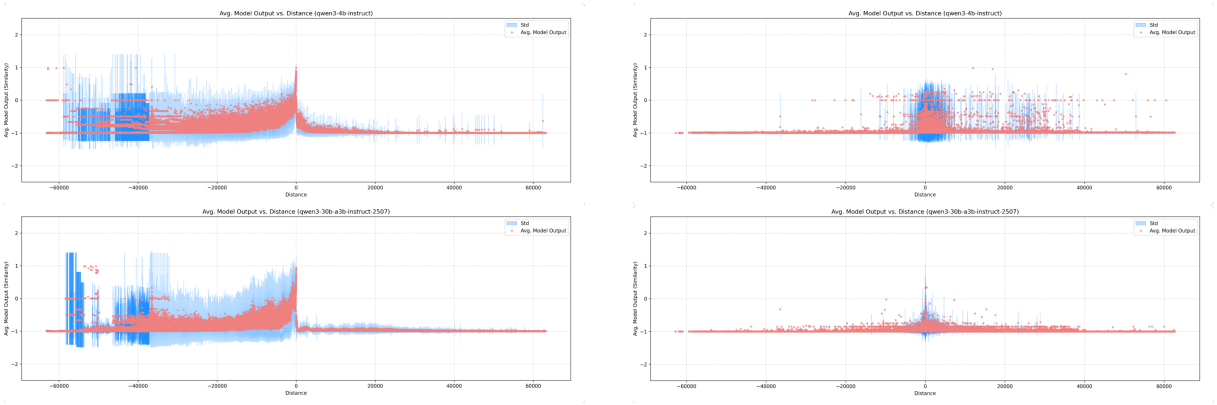

Figure 3: Mean similarity ratings vs. temporal distance. **Left**: in deictic t-FoR. **Right**: in sequential t-FoR. **Top**: qwen3-4b. **Bottom**: qwen3-30b. The blue shades indicate the variance of the ratings. Red dots correspond to mean similarity rating at a specific temporal distance.

is more uniform and compressed, whereas their representation of the past is richer but less stable. The pattern is more pronounced in the Qwen3-4b model, while the 30B model shows reduced variance near the present but preserves the same overall asymmetry.

Under the sequential t-FoR, the response curves differ significantly. Because the temporal relations are computed using event-based reference strategy, similarity scores remain strongly polarized for most non-zero distances, with values quickly saturating near the lower similarity bound. Variance is highest only in a narrow band around zero temporal distance: at temporal distance $\Delta = 0$ this reflects the coexistence of multiple interval relations, such as *overlaps*, *meets*, or *during*, while small non-zero distances still yield weakly separated *before/after* cases. Beyond this region, similarity ratings are highly stable and strongly negative, indicating that models treat most non-overlapping event pairs as mostly dissimilar in the sequential setting.

5.2 Interval-Specific Variance

Additionally, we leveraged Allen's Interval Algebra to better understand LLMs' behavior in special cases, as temporal distance $\Delta = 0$ contains different interval realtions, for example, "overlaps" or "meets". We reported the mean similarity score and standard deviation for each relation type, as shown in Figure 4. The left two plots correspond to the deictic t-FoR, while the right two plots are in sequential t-FoR.

Across both LLMs and both t-FoRs, we observe strong polarity effects for interval relations that are strictly ordered: intervals labeled as *after* consistently receive similarity ratings close to -1.00, followed by *before*, indicating that events occurring entirely after or entirely before the reference event are treated as mostly dissimilar. Oppositely, interval relations implying coincidence or direct boundary contact yield the highest positive similarity scores, reflecting strong alignment when events coincide or touch at their endpoints, particularly in *equal*, *meets*, and *met-by*.

More nuanced behavior is observed for containment and partial-overlap relations, such as *contains*, *during*, *overlaps*, and *overlapped-by*. These interval relations tend to have mid-range similarity scores but substantially higher variance, indicating less stable temporal assessments when interval topology is more complex. This variability is more pronounced under the sequential t-FoR, where temporal judgements depend on event-based reference strategy. Under the deictic t-FoR, variance is still elevated for partial-overlap relations, but similarity ratings remain more polarized, consistent with previously observed asymmetry between past and future.

The pattern holds across model sizes, although the Qwen3-30b model generally exhibits reduced dispersion relative to the 4B model, suggesting improved but still imperfect discrimination of structurally ambiguous temporal relations at larger scale.

5.3 Interaction between Temporal Distance and Event Duration

We further analyze whether temporal similarity depends jointly on temporal distance and event duration by averaging model outpus over (distance, duration) bins. Figure 5 shows the results for both LLMs.

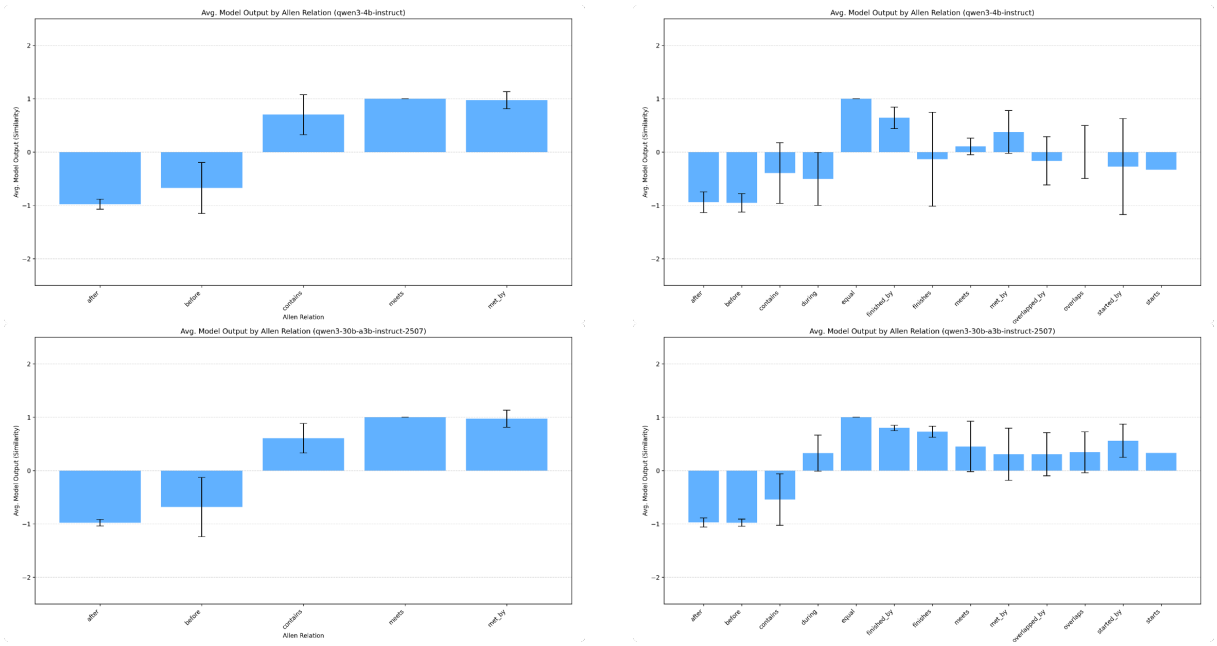

Figure 4: Mean similarity ratings by Allen's Interval Algebra. **Left**: in deictic t-FoR. **Right**: in sequential t-FoR. **Top**: qwen3-4b. **Bottom**: qwen3-30b. Bars indicate mean similarity score for each relation, and error bars denote the standard deviation across instances.

Under the deictic t-FoR, a clear pattern emerges for past events, that similarity increases with duration. Long-duration past events receive higher similarity ratings than short-duration ones at comparable distances, and in many cases distant long-duration past events are rated as more similar than short-duration events occurring much closer to the reference point. No comparable duration effect appears for future events, whose ratings rapidly converge to strongly negative values regardless of duration. The effect is qualitatively consistent across model sizes.

On the other hand, no systematic duration-distance structure is visible under the sequential t-FoR. Similarity ratings remain strongly polarized once events are temporally separated, and duration does not meaningfully modulate similarity.

6 Discussion

6.1 Adaptation to Temporal Frames of Reference

Based on the results, we observe that model behavior depends strongly on the choice of temporal frame of reference (t-FoR). Under the deictic t-FoR, similarity ratings vary smoothly with temporal distance and show a graded structure centered on the reference point. In contrast, under sequential t-FoR, similarity stabilizes into strongly negative values once events are temporally separated, with variability concentrated only near zero distance. This frame-specific divergence demonstrates that temporal understanding in large language models (LLMs) is frame-dependent, that they adapt their temporal judgements to the way time is anchored, whether to the experiencer's "now" or to another event. This adaptation is systematic rather than random, suggesting that temporal framing is encoded at a representational level, not only at the surface form of the task.

6.2 Effects of Temporal Distance and Interval Structure

The Allen's interval algebra result shows that similarity judgements are most stable for strictly ordered relations (*before*/*after*) and boundary-aligned or coincident relations (*meets*/*met-by*/*equal*), and least stable for overlap- and containment-based relations. This reveals that temporal understanding in LLMs is not merely a function of linear distance, but also of how intervals relate structurally.

The reduced stability for overlap relations suggests that LLMs handle ambiguous intervals less consistently than cleanly ordered ones. This aligns with the higher variance observed at small temporal distances, where multiple Allen's interval relations remain possible. These findings indicate that interval geometry is a key determinant of model behavior, that temporal understanding is sensitive

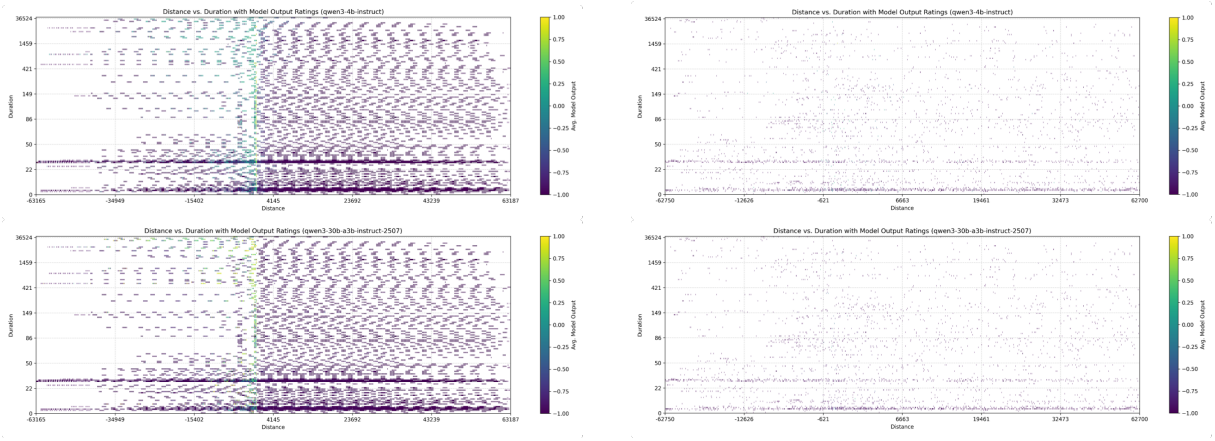

Figure 5: Heatmap of mean similarity ratings vs. (distance, duration). **Left**: in deictic t-FoR. **Right**: in sequential t-FoR. **Top**: qwen3-4b. **Bottom**: qwen3-30b.

to the structural complexity of temporal relations.

6.3 Effects of Event Duration and Past-Future Asymmetry

A second factor shaping temporal understanding is event duration, but only under specific conditions. Under the deictic t-FoR, duration systematically modulates similarity for past events: long-duration past events are judged as more similar than short-duration ones at comparable distances, and even more similar than short-duration events that occur much closer to the reference point. Crucially, the same effect does not appear for future events or under sequential t-FoR.

This selective retrospective duration sensitivity suggests that time is not represented symmetrically in LLMs. The past is encoded in a richer and more graded structure, while the future collapses toward uniform low similarity. The fact that duration effects disappear under sequential t-FoR reinforces that this is not purely a semantic property of events themselves, but a property of how LLMs anchor temporal meaning.

Therefore, temporal understanding in LLMs is influenced not only by temporal distance and interval structure, but also by duration. These effects are frame-dependent and directionally asymmetric.

6.4 Summary of Findings

Taken together, these findings indicate that LLMs demonstrate structured but bounded temporal understanding. They respond coherently to different frames of reference and their behavior is systematically modulated by temporal distance, interval relations, and duration. However, these effects are not uniform: temporal representations are richer in the past than in the future, and more stable for cleanly ordered relations than for overlapping intervals.

This suggests that LLMs' temporal understanding is shaped by linguistically acquired regularities rather than explicit temporal inference mechanisms. Temporal representation appears to be encoded as a graded, context-anchored similarity space, that adapts to framing, but does not collapse into a single global temporal metric.

7 Conclusion

We introduced a frame-sensitive evaluation of temporal understanding in large language models (LLMs), comparing behavior under deictic and sequential temporal frames of reference (t-FoRs). Our results show that LLMs adapt systematically to the chosen t-FoR: under deictic t-FoR, similarity judgements form graded and asymmetric patterns centered on the present, whereas under sequential t-FoR, similarity rapidly collapses to strongly negative values once events are temporally separated. Temporal judgements are additionally shaped by interval structure and duration, with instability concentrated in overlap-based relations and with duration affecting only past events under deictic anchoring. Altogether, these findings indicate that LLMs exhibit a structured yet bounded form of temporal understanding.

8 Limitations

This work currently focuses only on English text, leaving extension to other languages for future explorations. Our choice of t-FoR taxonomies (Evans, 2013) is based on the integration of notion of tran-

science which complements prior t-FoR taxonomies, but there is no common agreement of selecting which temporal frame of reference taxonomies to describe relations between events.

References

James F Allen. 1983. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843.

Adrián Bazaga, Rexhina Blloshmi, Bill Byrne, and Adrià de Gispert. 2025. Learning to reason over time: Timeline self-reflection for improved temporal reasoning in language models. pages 28014–28033. Association for Computational Linguistics.

Steven Bethard and James H Martin. 2007. Cu-tmp: Temporal relation classification using syntactic and semantic features. In *Proceedings of the fourth international workshop on semantic evaluations (SemEval-2007)*, pages 129–132.

Lera Boroditsky. 2000. Metaphoric structuring: Understanding time through spatial metaphors. *Cognition*, 75(1):1–28.

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

Wenhu Chen, Xinyi Wang, William Yang Wang, and William Yang Wang. 2021. A dataset for answering time-sensitive questions. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Haotian Wang, Ming Liu, and Bing Qin. 2024. Timebench: A comprehensive evaluation of temporal reasoning abilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1204–1228.

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.

Quang Do, Wei Lu, and Dan Roth. 2012. Joint inference for event timeline construction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 677–687.

Vyvyan Evans. 2013. Temporal frames of reference. *Cognitive linguistics*, 24(3).

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*.

Dedre Gentner, Mutsumi Imai, and Lera Boroditsky. 2002. As time goes by: Evidence for two systems in processing space→ time metaphors. *Language and cognitive processes*, 17(5):537–565.

Wes Gurnee and Max Tegmark. 2023. Language models represent space and time. *arXiv preprint arXiv:2310.02207*.

Duygu Sezen Islakoglu and Jan-Christoph Kalo. 2025. Chronosense: Exploring temporal understanding in large language models with time intervals of events. *arXiv preprint arXiv:2501.03040*.

Raghav Jain, Daivik Sojitra, Arkadeep Acharya, Sriparna Saha, Adam Jatowt, and Sandipan Dandapat. 2023. Do language models have a common sense regarding time? revisiting temporal commonsense reasoning in the era of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6750–6774.

Zhen Jia, Soumajit Pramanik, Rishiraj Saha Roy, and Gerhard Weikum. 2021. Complex temporal question answering on knowledge graphs. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 792–802.

Jungo Kasai, Keisuke Sakaguchi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A Smith, Yejin Choi, Kentaro Inui, and 1 others. 2023. Realtime qa: What's the answer right now? *Advances in neural information processing systems*, 36:49025–49043.

Alexander Kranjec. 2006. Extending spatial frames of reference to temporal concepts. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 28.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

George Lakoff, Mark Johnson, and John F Sowa. 1999. Review of philosophy in the flesh: The embodied mind and its challenge to western thought. *Computational Linguistics*, 25(4):631–634.

Lingyu Li, Yang Yao, Yixu Wang, Chubo Li, Yan Teng, and Yingchun Wang. 2025. The other mind: How language models exhibit human temporal cognition. *arXiv preprint arXiv:2507.15851*.

Xingxuan Li, Liying Cheng, Qingyu Tan, Hwee Tou Ng, Shafiq Joty, and Lidong Bing. 2023. Unlocking temporal question answering for large language models with tailor-made reasoning logic. *arXiv preprint arXiv:2305.15014*.

Adam Liska, Tomas Kocisky, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, Cyprien De Masson D'Autume, Tim Scholtes, Manzil Zaheer, Susannah Young, and 1 others. 2022. Streamingqa: A benchmark for adaptation to new knowledge over time in question answering models. In *International Conference on Machine Learning*, pages 13604–13622. PMLR.

Raja Marjieh, Ilia Sucholutsky, Theodore R Sumers, Nori Jacoby, and Thomas L Griffiths. 2022. Predicting human similarity judgments using large language models. *arXiv preprint arXiv:2202.04728*.

Matthew S McGlone and Jennifer L Harding. 1998. Back (or forward?) to the future: The role of perspective in temporal language comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24(5):1211.

J Ellis McTaggart. 1908. The unreality of time. *Mind*, 17(68):457–474.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40.

Kevin Ezra Moore. 2004. Ego-based and field-based frames of reference in space to time metaphors. *Language, culture, and mind*, pages 151–165.

Kevin Ezra Moore. 2011. Ego-perspective and field-based frames of reference: Temporal meanings of front in japanese, wolof, and aymara. *Journal of Pragmatics*, 43(3):759–776.

Qiang Ning, Hao Wu, Haoruo Peng, and Dan Roth. 2018a. Improving temporal relation extraction with a globally acquired statistical resource. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 841–851.

Qiang Ning, Hao Wu, and Dan Roth. 2018b. A multi-axis annotation scheme for event temporal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1318–1328.

Rafael E Núñez and Eve Sweetser. 2006. With the future behind them: Convergent evidence from aymara language and gesture in the crosslinguistic comparison of spatial construals of time. *Cognitive science*, 30(3):401–450.

Kai Nylund, Suchin Gururangan, and Noah A Smith. 2024. Time is encoded in the weights of finetuned language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2571–2587.

Lianhui Qin, Aditya Gupta, Shyam Upadhyay, Luheng He, Yejin Choi, and Manaal Faruqui. 2021. TIMEDIAL: Temporal commonsense reasoning in dialog. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7066–7076.

Yifu Qiu, Zheng Zhao, Yftah Ziser, Anna Korhonen, Edoardo Ponti, and Shay B Cohen. 2024. Are large language model temporally grounded? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7064–7083.

QwenTeam. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789. Association for Computational Linguistics.

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

Chris Sinha, Vera Da Silva Sinha, Jörg Zinken, and Wany Sampaio. 2011. When time is not space: The social and linguistic construction of time intervals and temporal event relations in an amazonian culture. *Language and Cognition*, 3(1):137–169.

Jannik Strötgen and Michael Gertz. 2010. Heideltime: High quality rule-based extraction and normalization of temporal expressions. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 321–324.

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

Leonard Talmy. 2003. *Toward a cognitive semantics, volume 1: Concept structuring systems*, volume 1. MIT press.

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

Long Papers), pages 14820–14835. Association for Computational Linguistics.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2024. Towards robust temporal reasoning of large language models via a multi-hop QA dataset and pseudo-instruction tuning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6272–6286.

Thora Tenbrink. 2011. Reference frames of space and time in language. *Journal of pragmatics*, 43(3):704–722.

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

Jonas Wallat, Adam Jatowt, and Avishek Anand. 2024. Temporal blind spots in large language models. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 683–692.

Yuqing Wang and Yun Zhao. 2024. TRAM: Benchmarking temporal reasoning for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6389–6415.

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

Siheng Xiong, Ali Payani, Ramana Kompella, and Faramarz Fekri. 2024. Large language models can learn temporal reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10452–10470.

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.

Michael Zhang and Eunsol Choi. 2021. SituatedQA: Incorporating extra-linguistic contexts into QA. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7371–7387.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. pages 3363–3369. Association for Computational Linguistics.

Xinyu Zhu, Cheng Yang, Bei Chen, Siheng Li, Jian-Guang Lou, and Yujiu Yang. 2023. Question answering as programming for solving time-sensitive questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12775–12790.