ChronoSense: Exploring Temporal Understanding in Large Language Models with Time Intervals of Events

Anonymous ACL submission

Abstract

Large Language Models (LLMs) still face significant challenges in reasoning and arithmetic. Although temporal reasoning has raised increasing research attention, comprehensive testing of Allen's interval relations (e.g., before, after, during) -a fundamental framework for temporal relationships- remains underexplored. To fill this gap, we present *ChronoSense*, a new benchmark for evaluating LLMs' temporal understanding. It includes 16 tasks, identifying the Allen relation between two temporal events and temporal arithmetic. We assess the performance of seven recent LLMs. The results indicate that models handle Allen relations, even symmetrical ones, quite differently. Moreover, the findings suggest that the models may rely on memorization to answer time-related questions. Overall, the models' low performance highlights the need for improved temporal understanding in LLMs. Our dataset and the source code are available at https://bit.ly/chronosense

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable proficiency across various tasks in NLP. Despite these advancements, significant challenges persist in areas such as reasoning, arithmetic (BIG-bench authors, 2023), and working with numerical values (Wei et al., 2022). These limitations affect their performance in temporal reasoning and numerical arithmetic.

Recent research has shown a growing interest in evaluating the temporal reasoning capabilities of LLMs. Efforts have focused on event ordering, comparing temporal events, temporal question answering, and event forecasting (Chu et al., 2023). However, a notable gap remains: the comprehensive testing of Allen's intervals, one of the most fundamental temporal reasoning frameworks that have been in use for over 30 years (Allen, 1989).

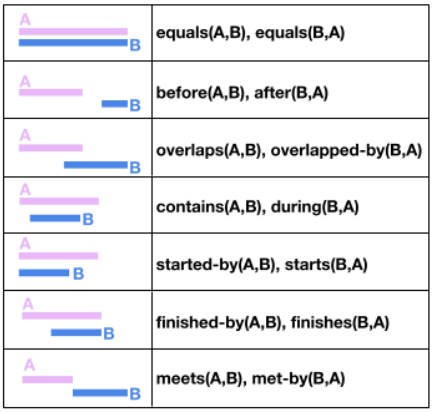

Figure 1: 13 Allen relations between two intervals, covering all combinations.

Allen's intervals provide a formal structure for representing temporal relationships between events, defining thirteen possible relations between time intervals. Despite its importance, existing benchmarks cover only subsets of these relations. We demonstrate these 13 relations in Figure 1.

To illustrate our task, consider the following example: In Figure 2, the first event is the *fourth cholera pandemic* which occurred between 1863 and 1875, while *World War II* occurred between 1939 and 1945. In our prompt, we list these two events with their names and respective start and end years and then ask a *True/False* question about one of the 13 Allen relations. For example, we ask the LLM whether the *fourth cholera pandemic* happened "before" *World War II*.

While such tasks are straightforward for humans, they pose considerable difficulty for LLMs due to the need to compare numerical values accurately. Our research focuses on reasoning about time intervals, and assessing how models perform on temporal understanding tasks. We also incorporate three time arithmetic tasks to challenge the models further.

Our contributions can be summarized as follows:

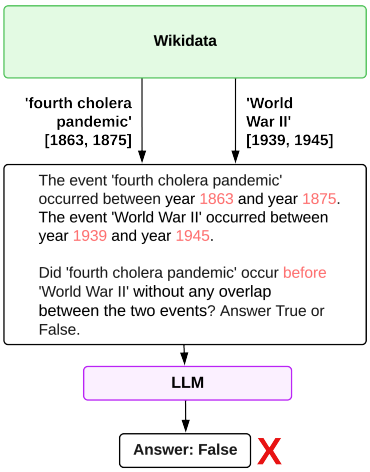

Figure 2: An example for comparing two temporal events with LLMs.

- We present a comprehensive evaluation of LLMs' performance on temporal reasoning tasks using our *ChronoSense* benchmark. Our evaluation spans Allen relations and temporal arithmetic tasks across 0-shot, few-shot, and chain-of-thought (CoT) prompting scenarios.
- We demonstrate the effectiveness of few-shot and CoT prompting in improving LLM performance, especially on temporal arithmetic tasks that require step-by-step reasoning.
- We investigate the influence of memorization on LLMs' ability to perform temporal reasoning tasks, especially when models encounter real-world event names that might have been part of pre-training data.

2 Preliminaries

Allen's Interval Algebra. Allen's interval algebra (IA) (Allen, 1989) provides 13 different relations between two intervals. As illustrated in Figure 1, these relations are "Equals", "Before", "After", "Overlaps", "Overlapped-by", "Contains", "During", "Started-by", "Starts", "Finished-by", "Finishes" "Meets", and "Met-by". These relations are mutually exclusive and cover all possible temporal relationships between two intervals. IA serves as a base for artificial intelligence and has been used in many applications (Janhunen and Sioutis, 2019). Although it is not the focus of this study, it allows deriving new facts. For instance, through transitivity, if Event e_1 happens before Event e_2, and Event e_2 happens before Event e_3, then Event e_1 happens before Event e_3. Therefore, correctly identifying the relationships between intervals is essential to support this type of reasoning.

3 ChronoSense Dataset

We create an event-centric dataset, named *ChronoSense*[1]. This dataset is designed to diagnose how well LLMs comprehend temporal events and the relationships between them, as illustrated in Figure 2. ChronoSense contains True/False questions that include different temporal dimensions. It features two types of questions: (1) Allen questions (requiring models to determine the Allen relation of two time intervals) and (2) temporal arithmetic tasks focused on a single event (challenging models to draw conclusions based on explicit time information). We set the time granularity to years for both question types. The prompts used in ChronoSense can be seen in Table 3 in Appendix A.

Question Type 1: Comparing Two Temporal Events with Allen Relations. We extract real event pairs from the Wikidata (Vrandečić and Krötzsch, 2014) (Section A.1). Similar to (Yang et al., 2023), every test instance in our dataset is in (Context, Hypothesis, Correctness) format. *Context* introduces the events and explicitly states the time periods when the events have occurred (e.g. *The event 'fourth cholera pandemic' occurred between year 1863 and year 1875. The event 'World War II' occurred between year 1939 and year 1945.*). *Hypothesis* verbalizes an Allen relation in natural language (e.g. *Did 'fourth cholera pandemic' occur before 'World War II' without any overlap between the two events? Answer True or False.*). *Correctness* is True if *Hypothesis* describes the temporal relationship between these two events correctly and False otherwise (e.g. *True* for the example above.).

Question Type 2: Temporal Arithmetic With A Single Event. To get insights into models' ability to perform temporal arithmetic, we also include temporal arithmetic questions in *ChronoSense*. *Context* introduces a single event and explicitly states the time information and optionally a temporal feature such as its duration or frequency (e.g. *'Event A' first occurred in year 1909. 'Event A' occurs every 12 years.*). *Hypothesis* is a statement that is not covered in *Context* and requires arithmetic calculations to verify (e.g. *Did 'Event A' occur again in the year 1921? Answer True or False.*). *Correctness* is True if *Hypothesis* matches with the calculations based on the *Context* and False otherwise (e.g. *True* for the example above).

[1]The dataset will be released under the CC BY 4.0 license.

The temporal arithmetic questions cover three different aspects. *End Timepoint* focuses on the duration of an event and requires models to determine the end time based on the given start time and duration. *Next Occurrence* focuses on the frequency of events and challenges models to calculate when an event occurs again based on a given frequency. *Intermediate Timepoint*, which is novel to this work, challenges models to infer whether an event was happening between its start and end time by asking if it happened at a certain year in time. Due to the limited number of events with frequency from Wikidata, we synthetically create these questions. Therefore, the events do not have event names, but rather we name them as "Event A". For each question, we create a negative sample by creating a wrong *Hypothesis* (e.g. by changing the next occurrence year in the previous example from 1921 to 1950.).

Different event abstraction levels. For Allen questions, we have an abstract version of each question where we hide the names of the events by replacing them with letters such as "Event A" and "Event B". This setting allows us to see how the memorization affects LLM's performance by comparing the abstract versions with the original versions (where we have event names).

Different prompts for questions. There are multiple ways to ask a question so we create two different additional prompts for each question to understand the effect of the prompt. All prompts can be seen in Table 3 and Table 10 in the Section A.

Negative samples. To evaluate the robustness of the LLM's predictions, we generate negative examples for each data instance (detailed in A.1.1). Therefore, the *Correctness* value is "*True*" in 50% of the data instances, and "*False*" in the other half.

Dataset statistics. For each Allen relation and each temporal arithmetic question, ChronoSense has 4,000 training samples, 500 validation samples, and 50 test samples to ensure reproducibility.

4 Experiments

We evaluate the performance of various LLMs on a task framed as binary classification. Specifically, the models are tasked with answering *True* or *False* to a set of prompts on temporal reasoning. We evaluate the accuracy of the models, where we have a random chance accuracy of 50%. We compare the following LLMs in our experiments: Gemma2-9, GPT-4o, GPT-4o-mini, Meta-Llama-3.1-8B-Instruct, Mistral-7B-Instruct-v0.2 (Jiang et al., 2023), Mixtral-8x7B-Instruct-v0.1 (Jiang et al., 2024), Phi-3-mini-128k-instruct. Each model can generate up to 64 new tokens for an answer; however, in the chain-of-thought (CoT) setting, the maximum token limit is increased to 512 to provide more space for reasoning. For both question types (Allen and temporal arithmetic), we report on different settings: 0-shot, 1-shot, 3-shot, Chain-of-Thought (CoT) prompting. For Allen questions, we also report on abstract versions in which we remove the real event names. As mentioned in Section 3, the temporal arithmetic questions are all in the abstract setting. We report the averaged results in Table 1. The complete experimental results, including the experiments on individual Allen relations, can be found in A.2. Moreover, in Table 2, we zoom in and report the 0-shot performance on individual Allen relations for three models.

General Findings. (1) The models exhibit low performance and lack consistency on ChronoSense questions across the experiments, given the fact that the random prediction would lead to 0.50 accuracy. This suggests the need for improvements in temporal understanding in LLMs. (2) Arithmetic questions are typically more challenging than Allen relations in both zero-shot and few-shot settings. (3) Few-shot setting is helpful for most models for Allen questions, although CoT has no consistent effect. Despite these improvements, the tasks remain challenging, as several models still have an accuracy below 0.60. (4) CoT prompting helps all models in arithmetic questions. This is expected as these questions require step-by-step reasoning. (5) When averaged over models, some Allen relations are easier and some are more challenging for the models. First, "Before" and "After" are easier than other relations in all experiments. This is expected as these relations are the most frequently used phrases among others. This may also indicate that the models are better at detecting relations that do not contain any overlap. Second, "Equals" is the hardest relation in zero-shot and abstract settings, and "Finishes" is the hardest for few-shot and "Overlapped-by" for the CoT setting. (6) The models do not perform similarly for symmetrical Allen relations. For instance, despite their symmetric nature, averaged model performance for "Before" is higher than "After". Similarly, "Contains" and "Finished-by" are easier than their symmetrical relations "During" and "Finishes" (except for one tie case). (7) The abstract versions are more

Type	Setting	Gemma2-9B	GPT-4o	GPT-4o-mini	Llama3.1-8B	Mistral-7B	Mixtral-8x7B	Phi-3-mini
Allen	0-shot	0.08*	0.86	0.73	0.18*	0.49	0.54	0.45
	1-shot	0.74	0.95	0.73	0.54	0.47	0.57	0.59
	3-shot	0.79	0.95	0.74	0.60	0.48	0.58	0.70
	CoT	0.75	0.64	0.70	0.59	0.52	0.54	0.80
	Abstract	0.04*	0.83	0.64	0.06*	0.21	0.36	0.38
Arithmetic	0-shot	0.82	0.62	0.62	0.18*	0.35	0.36	0.34
	1-shot	0.68	0.60	0.50	0.01*	0.38	0.45	0.0*
	3-shot	0.66	0.64	0.37	0.0*	0.44	0.66	0.0*
	CoT	0.96	0.98	0.99	0.72	0.72	0.73	0.98

Table 1: The average performance comparison between different settings on two different question types in ChronoSense. (*) indicates the models that perform poorly due to a high number of answers that did not follow the instruction e.g. the model did not answer with "True" or "False".

Allen Relation	GPT-4o	Mixtral-8x7B	Phi-3-mini
Equals	0.76	0.4	0.44
Before	0.88	0.88	0.84
After	0.98	0.74	0.6
Overlaps	0.9	0.38	0.48
Overlapped-By	0.76	0.52	0.46
Contains	0.88	0.44	0.34
During	0.86	0.5	0.42
Started-By	0.86	0.58	0.32
Starts	0.84	0.48	0.46
Finished-By	0.9	0.4	0.32
Finishes	0.94	0.48	0.32
Meets	0.9	0.76	0.46
Met-By	0.84	0.58	0.44

Table 2: 0-shot setting results for GPT-4o, Mixtral-8x7B, and Phi-3-mini on 13 Allen relations.

challenging for the models. Models may rely on memorization to answer temporal understanding questions for the events included in the pre-training data. In other words, the implicit knowledge from pre-training can influence their performance on temporal understanding. (8) The different ways of asking the same question affect the performance, but not significantly (Table 11 in Section A.3). This validates our decision to choose one prompt variant to report performances for all the experiments.

5 Related Work

Temporal reasoning has been extensively studied in NLP (Terenziani, 2009; Sanampudi and Kumari, 2010) and QA over temporal knowledge graphs (Dhingra et al., 2022; Zhao et al., 2022; Saxena et al., 2021; Chen et al., 2021; Jia et al., 2018a,b, 2021). A new line of work focuses on LLMs' temporal knowledge and reasoning. TimeBench (Chu et al., 2023) covers abstract temporal expressions, commonsense reasoning, and event relationships. Other benchmarks include those by (Jain et al., 2023) for commonsense-based temporal tasks and TimeLlama (Yuan et al., 2023) for event forecasting. TGQA (Xiong et al., 2024) evaluates synthetic temporal QA but only covers three simple event relations. TRACIE (Zhou et al., 2021) assesses reasoning over implicit events, while TEMPREASON (Tan et al., 2023a) probes three levels of temporal understanding but primarily focuses on factual recall. TRAM (Wang and Zhao, 2023) includes event relations from (UzZaman et al., 2013) but lacks explicit events. (Tan et al., 2023b) has temporal arithmetic but it is event-independent. LTLBench (Tang and Belle, 2024) uses linear temporal logic to model the temporal relationships between events. Test of Time (Fatemi et al., 2024) creates a synthetic dataset to isolate temporal reasoning. Recent works on event ordering include TDDiscourse (Naik et al., 2019), which classifies implicit event relations overlapping with Allen's framework. Datasets from (Vashishtha et al., 2020) focus on event ordering and duration, while TORQUE (Ning et al., 2020) presents a reading comprehension dataset to investigate the temporal ordering of events but lacks explicit start and end times. Despite the variety of benchmarks, none covers all 13 of Allen's interval relations.

6 Conclusion

We introduce ChronoSense, a diagnostic dataset designed to assess LLMs' ability to compare event timelines using Allen relations and perform temporal arithmetic. We show that models frequently struggle with these tasks and may rely on memorization rather than reasoning. This raises critical concerns about their reliability in applications such as historical analysis, legal AI, and medical timelines. Future research should focus on improving LLMs' temporal reasoning capabilities, integrating temporal constraint-based reasoning, and analyzing multi-event comparisons.

7 Limitations

Our work has some limitations regarding the dataset and the evaluation. Concerning the dataset, we limit the size of the test set due to reproducibility concerns and the computational cost of large models. This may affect the generalizability of the results. Moreover, some Wikidata events have ambiguous names that may mislead the model, e.g., an exhibition event named after a painter, which may not clearly indicate a temporal event to the models. On the evaluation side, our study involves a relatively small selection of models and some closed-source models (e.g. GPT-4o). Moreover, although we test 3 different prompt versions per task, we acknowledge that the prompt content may influence the model performance. Lastly, we truncate the LLM outputs when they exceed the maximum token lengths. This potentially omits some of the correct answers and leads to lower accuracy scores for the respective models.

8 Ethics Statement

Our dataset, which sources events from Wikidata, inherently carries the risk of containing incorrect information. This could unintentionally propagate misinformation. While our script filters out data points containing certain triggering keywords, some event names may still include inappropriate or harmful content. This does not reflect the views or opinions of the authors. Moreover, the data points in ChronoSense do not represent individuals but rather events categorized as instances or subclasses of "occurrence" [2]. However, some events include the names of individuals, such as exhibitions named after artists. Furthermore, we acknowledge the environmental impact associated with LLMs. Although our study only utilizes pretrained models, inference with these models still demands significant computational resources.

References

James F. Allen. 1989. *Maintaining Knowledge about Temporal Intervals*, page 361–372.

BIG-bench authors. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Wenhu Chen, Xinyi Wang, and William Yang Wang. 2021. A dataset for answering time-sensitive questions. *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*, abs/2311.17667.

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. *Preprint*, arXiv:2406.09170.

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.

Tomi Janhunen and Michael Sioutis. 2019. Allen's interval algebra makes the difference. *Preprint*, arXiv:1909.01128.

Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. 2018a. Tempquestions: A benchmark for temporal question answering. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 1057–1062.

Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. 2018b. Tequila: Temporal question answering over knowledge bases. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 1807–1810.

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.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris

[2] https://www.wikidata.org/wiki/Q1190554

Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Aakanksha Naik, Luke Breitfeller, and Carolyn Rose. 2019. TDDiscourse: A dataset for discourse-level temporal ordering of events. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 239–249.

Qiang Ning, Hao Wu, Rujun Han, Nanyun Peng, Matt Gardner, and Dan Roth. 2020. TORQUE: A reading comprehension dataset of temporal ordering questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1158–1172.

Suresh Kumar Sanampudi and G Vijaya Kumari. 2010. Temporal reasoning in natural language processing: A survey. *International Journal of Computer Applications*, 1(4):68–72.

Apoorv Saxena, Soumen Chakrabarti, and Partha Talukdar. 2021. Question answering over temporal knowledge graphs. 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 6663–6676.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2023a. Towards benchmarking and improving the temporal reasoning capability of large language models. In *Annual Meeting of the Association for Computational Linguistics*.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2023b. 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.

Weizhi Tang and Vaishak Belle. 2024. Ltlbench: Towards benchmarks for evaluating temporal logic reasoning in large language models. *ArXiv*, abs/2407.05434.

Paolo Terenziani. 2009. *Qualitative Temporal Reasoning*, pages 2225–2229.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. SemEval-2013 task 1: TempEval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9.

Siddharth Vashishtha, Adam Poliak, Yash Kumar Lal, Benjamin Van Durme, and Aaron Steven White. 2020. Temporal reasoning in natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4070–4078.

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

Yuqing Wang and Yun Zhao. 2023. Tram: Benchmarking temporal reasoning for large language models. *ArXiv*, abs/2310.00835.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

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

Zonglin Yang, Xinya Du, Rui Mao, Jinjie Ni, and Erik Cambria. 2023. Logical reasoning over natural language as knowledge representation: A survey. *Preprint*, arXiv:2303.12023.

Chenhan Yuan, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2023. Back to the future: Towards explainable temporal reasoning with large language models. *Preprint*, arXiv:2310.01074.

Ruilin Zhao, Feng Zhao, Guandong Xu, Sixiao Zhang, and Hai Jin. 2022. Can language models serve as temporal knowledge bases? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2024–2037.

Ben Zhou, Kyle Richardson, Qiang Ning, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2021. Temporal reasoning on implicit events from distant supervision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1361–1371.

A Appendix

A.1 Allen Question Generation

To generate the Allen questions, we take the following steps:

1. We extract real-world event pairs from Wikidata (Vrandečić and Krötzsch, 2014) via SPARQL. The used Wikidata content is licenced under CC0 [3].
2. We determine the valid Allen relation for this event pair by comparing the time intervals of these events.
3. In order to map these relations into text, we verbalize each Allen relation using the prompts as depicted in Table 3.

A.1.1 Negative Samples For Allen Questions

For the positive samples, we put the correct Allen relations to the *Hypothesis* and set the *Correctness* as *True*. However, for negative samples, we choose another Allen relation (e.g. choosing the "Meets" relation instead of "Before") and set the *Correctness* to *False*. However, since we set the time granularity as years instead of days, generating negative samples for Allen relations presents certain challenges. For example, the "Equals" relation requires that both the start and end points of two events match exactly. When we create a negative sample for "Equals", we cannot use the "Contains" relation. This is because the second event could start later and end earlier than the first event, even if the years are the same. Since the exact days/dates of the events are not known, the information provided in the context will be ambiguous. To address this issue, we exclude such problematic relations from the pool of candidate relations during negative sampling.

Below we provide a list of Allen relations along with the Allen relations that are excluded from its negative sample candidates to avoid such inconclusive cases.

- "Equals": "Overlaps", "Contains", "During", "Overlapped-By", "Started-By", "Starts", "Finished-By", "Finishes"
- "Started-By": "Contains", "Overlapped-By"
- "Starts": "Overlaps", "During"
- "Finished-By": "Overlaps", "Contains"
- "Finishes": "During", "Overlapped-By"
- "Meets": "Before", "Overlaps"
- "Met-By": "Overlapped-By", "After"

[3] https://www.wikidata.org/wiki/Wikidata:Licensing

A.2 Detailed Results

For Allen questions, we report the 0-shot, 1-shot, 3-shot, and Chain-of-Thought results in Table 4, Table 5, Table 6, and Table 7. Moreover, Table 8 includes the results for the abstract setting, where we replace the actual event names with abstract names such as "Event A" and "Event B".

Table 9 reports the results of the 0-shot, few-shot, chain-of-thought for temporal arithmetic questions (*End Timepoint*, *Intermediate Timepoint* and *Next Occurrence*).

A.3 Different prompts

ChronoSense has different prompt variants for each question type. The templates for prompt variants can be seen in Table 10. In order to show the effect of different prompts, we report the average accuracy values with standard deviation across three prompt variants in Table 11. Although there are cases with high standard deviation, we do not observe a relation that has consistently high values.

A.4 Computational Budget

We ran all experiments using HuggingFace on a single Nvidia H100 - 80GB or via the OpenAI API. None of the experiments per model took longer than 24 hours. The experiments via the OpenAI API caused costs of less than 100$.

Type	Question	Template
Allen	Equals	Did ‘Event A’ begin in the same year as ‘Event B’ and end in the same year as ‘Event B’? Answer True or False.
Allen	Before	Did ‘Event A’ occur before ‘Event B’ without any overlap between the two events? Answer True or False.
Allen	After	Did ‘Event A’ occur after ‘Event B’ without any overlap between the two events? Answer True or False.
Allen	Overlaps	Did ‘Event A’ begin before ‘Event B’ and end before ‘Event B’ ended, with some overlap between the two events? Answer True or False.
Allen	Overlapped-By	Did ‘Event B’ begin before ‘Event A’ and end before ‘Event A’ ended, with some overlap between the two events? Answer True or False.
Allen	Contains	Did ‘Event A’ begin before ‘Event B’ began and end after ‘Event B’ ended, entirely containing ‘Event B’? Answer True or False.
Allen	During	Did ‘Event A’ begin after ‘Event B’ began and end before ‘Event B’ ended, being entirely contained within ‘Event B’? Answer True or False.
Allen	Started-By	Did ‘Event B’ begin in the same year as ‘Event A’, but end before ‘Event A’ ended? Answer True or False.
Allen	Starts	Did ‘Event A’ begin in the same year as ‘Event B’, but end before ‘Event B’ ended? Answer True or False.
Allen	Finished-By	Did ‘Event B’ begin after ‘Event A’ began and end in the same year as ‘Event A’? Answer True or False.
Allen	Finishes	Did ‘Event A’ begin after ‘Event B’ began and end in the same year as ‘Event B’? Answer True or False.
Allen	Meets	Did ‘Event A’ end in the same year as ‘Event B’ began? Answer True or False.
Allen	Met-by	Did ‘Event B’ end in the same year as ‘Event A’ began? Answer True or False.
Arithmetic	End timepoint	Did ‘Event A’ end in the year [start+duration]? Answer True or False.
Arithmetic	Next occurrence	Did ‘Event A’ occur again in the year [next-occurrence]? Answer True or False.
Arithmetic	Intermediate timepoint	Was ‘Event A’ happening in the year [intermediate]? Answer True or False.

Table 3: Templates used in ChronoSense.

	Gemma2-9B	GPT-4o	GPT-4o-mini	Llama3.1-8B	Mistral-7B	Mixtral-8x7B	Phi-3-mini	Average
Equals	0.04	0.76	0.58	0.14	0.18	0.4	0.44	0.36
Before	0.52	0.88	0.82	0.04	0.88	0.88	0.84	0.69
After	0.18	0.98	0.82	0.26	0.9	0.74	0.6	0.63
Overlaps	0.02	0.9	0.8	0.2	0.46	0.38	0.48	0.46
Overlapped-By	0.04	0.76	0.72	0.3	0.42	0.52	0.46	0.46
Contains	0.04	0.88	0.8	0.1	0.48	0.44	0.34	0.44
During	0.02	0.86	0.6	0.38	0.42	0.5	0.42	0.45
Started-By	0.02	0.86	0.76	0.18	0.38	0.58	0.32	0.44
Starts	0.0	0.84	0.74	0.24	0.48	0.48	0.46	0.46
Finished-By	0.02	0.9	0.78	0.14	0.46	0.4	0.32	0.43
Finishes	0.06	0.94	0.64	0.16	0.42	0.48	0.32	0.43
Meets	0.14	0.9	0.76	0.06	0.5	0.76	0.46	0.51
Met-By	0.0	0.84	0.7	0.14	0.5	0.58	0.44	0.45
Average	0.08	0.86	0.73	0.18	0.49	0.54	0.45	

Table 4: 0-shot setting results on 13 Allen questions with explicit event names.

	Gemma2-9B	GPT-4o	GPT-4o-mini	Llama3.1-8B	Mistral-7B	Mixtral-8x7B	Phi-3-mini	Average
Equals	0.64	0.96	0.7	0.62	0.36	0.62	0.58	0.63
Before	1.0	0.96	0.92	0.74	0.92	0.94	0.88	0.90
After	0.86	1.0	0.94	0.76	0.78	0.78	0.86	0.85
Overlaps	0.82	0.96	0.58	0.46	0.42	0.52	0.66	0.63
Overlapped-By	0.68	0.84	0.56	0.5	0.36	0.52	0.66	0.58
Contains	0.8	1.0	0.7	0.46	0.48	0.52	0.56	0.64
During	0.7	0.94	0.7	0.66	0.44	0.58	0.42	0.63
Started-By	0.54	1.0	0.86	0.42	0.3	0.5	0.34	0.56
Starts	0.86	0.98	0.8	0.56	0.44	0.48	0.7	0.68
Finished-By	0.7	1.0	0.76	0.5	0.4	0.48	0.74	0.65
Finishes	0.5	0.94	0.52	0.46	0.42	0.52	0.44	0.54
Meets	0.84	0.92	0.86	0.46	0.42	0.46	0.42	0.62
Met-By	0.72	0.88	0.6	0.46	0.48	0.5	0.44	0.58
Average	0.74	0.95	0.73	0.54	0.47	0.57	0.59	

Table 5: 1-shot setting results on Allen questions with explicit event names.

	Gemma2-9B	GPT-4o	GPT-4o-mini	Llama3.1-8B	Mistral-7B	Mixtral-8x7B	Phi-3-mini	Average
Equals	0.76	0.98	0.76	0.64	0.44	0.58	0.78	0.70
Before	0.98	0.98	0.94	0.78	0.68	0.94	0.98	0.89
After	0.86	0.98	0.9	0.8	0.74	0.8	0.88	0.85
Overlaps	0.86	0.94	0.54	0.54	0.44	0.5	0.8	0.66
Overlapped-By	0.68	0.76	0.52	0.54	0.48	0.36	0.8	0.59
Contains	0.86	0.98	0.78	0.54	0.46	0.54	0.52	0.66
During	0.82	1.0	0.78	0.6	0.42	0.56	0.48	0.66
Started-By	0.62	1.0	0.88	0.44	0.42	0.52	0.54	0.63
Starts	0.92	0.98	0.96	0.64	0.44	0.68	0.92	0.79
Finished-By	0.86	0.98	0.56	0.54	0.46	0.44	0.68	0.64
Finishes	0.5	0.98	0.54	0.56	0.44	0.42	0.54	0.56
Meets	0.9	0.96	0.84	0.62	0.46	0.66	0.68	0.73
Met-By	0.66	0.88	0.74	0.66	0.48	0.64	0.56	0.66
Average	0.79	0.95	0.74	0.60	0.48	0.58	0.70	

Table 6: 3-shot setting results on Allen questions with explicit event names.

	Gemma2-9B	GPT-4o	GPT-4o-mini	Llama3.1-8B	Mistral-7B	Mixtral-8x7B	Phi-3-mini	Average
Equals	0.46	0.6	0.76	0.8	0.3	0.4	0.76	0.58
Before	1.0	0.78	0.82	0.56	0.9	0.92	0.9	0.84
After	0.96	0.8	0.8	0.64	0.94	0.68	0.86	0.81
Overlaps	0.68	0.58	0.66	0.6	0.64	0.48	0.62	0.60
Overlapped-By	0.66	0.5	0.72	0.44	0.4	0.46	0.62	0.54
Contains	0.72	0.58	0.78	0.52	0.42	0.66	0.82	0.64
During	0.8	0.58	0.78	0.58	0.46	0.64	0.8	0.66
Started-By	0.78	0.68	0.48	0.64	0.32	0.58	0.9	0.62
Starts	0.8	0.58	0.58	0.58	0.4	0.46	0.92	0.61
Finished-By	0.76	0.56	0.54	0.54	0.46	0.4	0.82	0.58
Finishes	0.62	0.5	0.56	0.48	0.54	0.42	0.76	0.55
Meets	0.82	0.84	0.86	0.74	0.42	0.54	0.8	0.71
Met-By	0.78	0.86	0.8	0.62	0.56	0.48	0.94	0.72
Average	0.75	0.64	0.70	0.59	0.52	0.54	0.80	

Table 7: Chain-of-Thought setting results on Allen questions with explicit event names.

	Gemma2-9B	GPT-4o	GPT-4o-mini	Llama3.1-8B	Mistral-7B	Mixtral-8x7B	Phi-3-mini	Average
Equals	0.02	0.42	0.36	0.1	0.02	0.02	0.62	0.22
Before	0.04	0.92	0.84	0.02	0.46	0.88	0.52	0.52
After	0.0	0.98	0.94	0.02	0.38	0.62	0.4	0.47
Overlaps	0.02	0.9	0.5	0.04	0.14	0.3	0.34	0.32
Overlapped-By	0.04	0.64	0.48	0.06	0.14	0.2	0.36	0.27
Contains	0.06	0.92	0.72	0.06	0.28	0.42	0.34	0.39
During	0.26	0.8	0.52	0.06	0.18	0.28	0.38	0.35
Started-By	0.0	0.92	0.66	0.08	0.06	0.28	0.24	0.32
Starts	0.04	0.92	0.64	0.1	0.02	0.24	0.42	0.34
Finished-By	0.0	0.9	0.76	0.08	0.22	0.36	0.34	0.38
Finishes	0.06	0.78	0.58	0.06	0.06	0.26	0.28	0.29
Meets	0.08	0.94	0.72	0.1	0.42	0.42	0.42	0.44
Met-By	0.0	0.82	0.66	0.06	0.42	0.46	0.38	0.39
Average	0.04	0.83	0.64	0.06	0.21	0.36	0.38	

Table 8: 0-shot setting results on Allen questions with the abstract event names.

	Gemma2-9B	GPT-4o	GPT-4o-mini	Llama3.1-8B	Mistral-7B	Mixtral-8x7B	Phi-3-mini	Average
End-Timepoint (0-shot)	0.8	0.64	0.6	0.18	0.48	0.42	0.04	0.45
Next-Occurrence (0-shot)	0.88	0.24	0.28	0.1	0.12	0.12	0.08	0.26
Intermediate-Timepoint (0-shot)	0.78	1.0	1.0	0.26	0.46	0.54	0.92	0.70
Average (0-shot)	0.82	0.62	0.62	0.18	0.35	0.36	0.34	
End-Timepoint (1-shot)	0.54	0.64	0.36	0.02	0.34	0.4	0.0	0.32
Next-Occurrence (1-shot)	0.52	0.18	0.16	0.02	0.36	0.32	0.0	0.22
Intermediate-Timepoint (1-shot)	1.0	1.0	1.0	0.0	0.46	0.64	0.0	0.58
Average (1-shot)	0.68	0.60	0.50	0.01	0.38	0.45	0.0	
End-Timepoint (3-shot)	0.46	0.7	0.08	0.0	0.26	0.64	0.0	0.30
Next-Occurrence (3-shot)	0.58	0.24	0.04	0.0	0.44	0.54	0.0	0.26
Intermediate-Timepoint (3-shot)	0.94	1.0	1.0	0.0	0.62	0.82	0.0	0.62
Average (3-shot)	0.66	0.64	0.37	0.0	0.44	0.66	0.0	
End-Timepoint (CoT)	1.0	0.98	0.98	0.78	0.82	0.92	1.0	0.92
Next-Occurrence (CoT)	0.92	0.96	1.0	0.74	0.8	0.74	0.96	0.87
Intermediate-Timepoint (CoT)	0.98	1.0	1.0	0.64	0.54	0.54	1.0	0.81
Average (CoT)	0.96	0.98	0.99	0.72	0.72	0.73	0.98	

Table 9: The results on all temporal arithmetic questions in 0-, 1-, and 3-shot settings, as well as using CoT prompting.

Question	Prompt alternative 1	Prompt alternative 2
Equals	Does ‘Event A’ have identical start and end years as ‘Event B’? Answer True or False.	Are the starting and ending years of ‘Event A’ and ‘Event B’ the same? Answer True or False.
Before	Is it true that ‘Event A’ took place completely before ‘Event B’? Answer True or False.	Can it be confirmed that ‘Event A’ completely preceded ‘Event B’? Answer True or False.
After	Is it true that ‘Event A’ took place completely after ‘Event B’? Answer True or False.	Can it be confirmed that ‘Event A’ completely succeeded ‘Event B’? Answer True or False.
Overlaps	Does ‘Event A’ overlap with ‘Event B’ by starting before and ending during it? Answer True or False.	Is there a period where ‘Event A’ and ‘Event B’ overlapped, with ‘Event A’ starting and ending first? Answer True or False.
Overlapped-By	Does ‘Event A’ overlap with ‘Event B’ by starting after and ending after it? Answer True or False.	Is there a period where ‘Event A’ and ‘Event B’ overlapped, with ‘Event A’ starting and ending last? Answer True or False.
Contains	Does ‘Event A’ fully enclose ‘Event A’, starting before and ending after ‘Event B’? Answer True or False.	Does the time interval of ‘Event A’ contain the time interval of ‘Event B’? Answer True or False.
During	Is ‘Event A’ fully enclosed by ‘Event B’, starting and ending within ‘Event B’s duration? Answer True or False.	Can ‘Event A’ be considered to occur entirely during ‘Event B’, from start to finish? Answer True or False.
Started-By	Does ‘Event A’ have the same starting year as ‘Event B’ but finish later? Answer True or False.	Did ‘Event A’ initiated in the same year as ‘Event B’ yet end later? Answer True or False.
Starts	Does ‘Event A’ have the same starting year as ‘Event B’ but finish earlier? Answer True or False.	Did ‘Event A’ initiated in the same year as ‘Event B’ yet end sooner? Answer True or False.
Finished-By	Does ‘Event A’ initiate before the start of ‘Event B’ and finish in the same calendar year? Answer True or False.	Is ‘Event A’ starting earlier than ‘Event A’ and concluding within the same year? Answer True or False.
Finishes	Does ‘Event A’ initiate after the start of ‘Event B’ and finish in the same calendar year? Answer True or False.	Is ‘Event A’ starting later than ‘Event B’ and concluding within the same year? Answer True or False.
Meets	Is the end of ‘Event A’ coinciding with the start of ‘Event B’ in the same year? Answer True or False.	Does the end of ‘Event A’ align with the beginning of ‘Event B’ within the same year? Answer True or False.
Met-by	Is the start of ‘Event A’ coinciding with the end of ‘Event B’ in the same year? Answer True or False.	Does the beginning of ‘Event A’ align with the end of ‘Event B’ within the same year? Answer True or False.
End timepoint	Is the conclusion of ‘Event A’ marked within the year [start+duration]? Answer True or False.	Can it be confirmed that ‘Event A’ finished in the year [start+duration]? Answer True or False.
Next occurrence	Is a recurrence of ‘Event A’ expected in the year [next-occurrence]? Answer True or False.	Can we anticipate another instance of ‘Event A’ in the year [next-occurrence]? Answer True or False.
Intermediate timepoint	During the year [intermediate], was ‘Event A’ in progress? Answer True or False.	In the year [intermediate], can it be verified that ‘Event A’ was active? Answer True or False.

Table 10: The different prompts used in ChronoSense.

	Gemma2-9B	GPT-4o	GPT-4o-mini	Llama3.1-8B	Mistral-7B	Mixtral-8x7B	Phi-3-mini
Equals	0.15 ± 0.15	0.85 ± 0.08	0.47 ± 0.13	0.23 ± 0.08	0.12 ± 0.05	0.44 ± 0.07	0.43 ± 0.06
Before	0.74 ± 0.21	0.93 ± 0.06	0.92 ± 0.09	0.17 ± 0.14	0.84 ± 0.11	0.93 ± 0.05	0.83 ± 0.12
After	0.32 ± 0.12	0.97 ± 0.01	0.78 ± 0.20	0.29 ± 0.06	0.64 ± 0.23	0.64 ± 0.16	0.56 ± 0.14
Overlaps	0.05 ± 0.05	0.85 ± 0.06	0.83 ± 0.05	0.19 ± 0.06	0.49 ± 0.06	0.46 ± 0.07	0.45 ± 0.08
Overlapped-By	0.05 ± 0.01	0.66 ± 0.23	0.65 ± 0.16	0.22 ± 0.07	0.45 ± 0.03	0.43 ± 0.09	0.40 ± 0.06
Contains	0.11 ± 0.07	0.93 ± 0.04	0.63 ± 0.16	0.15 ± 0.05	0.47 ± 0.01	0.49 ± 0.06	0.45 ± 0.13
During	0.05 ± 0.03	0.83 ± 0.03	0.46 ± 0.14	0.33 ± 0.04	0.45 ± 0.03	0.48 ± 0.02	0.41 ± 0.03
Started-By	0.08 ± 0.07	0.89 ± 0.03	0.72 ± 0.09	0.21 ± 0.03	0.39 ± 0.01	0.51 ± 0.11	0.46 ± 0.13
Starts	0.08 ± 0.07	0.86 ± 0.03	0.74 ± 0.06	0.29 ± 0.06	0.44 ± 0.05	0.49 ± 0.01	0.48 ± 0.03
Finished-By	0.09 ± 0.08	0.73 ± 0.27	0.66 ± 0.28	0.21 ± 0.07	0.47 ± 0.06	0.43 ± 0.05	0.35 ± 0.03
Finishes	0.05 ± 0.03	0.91 ± 0.03	0.65 ± 0.05	0.21 ± 0.06	0.47 ± 0.04	0.52 ± 0.09	0.39 ± 0.06
Meets	0.17 ± 0.03	0.94 ± 0.04	0.79 ± 0.06	0.09 ± 0.02	0.49 ± 0.01	0.57 ± 0.17	0.42 ± 0.03
Met-By	0.01 ± 0.01	0.79 ± 0.05	0.69 ± 0.01	0.20 ± 0.05	0.50 ± 0.00	0.52 ± 0.06	0.33 ± 0.09
End Timepoint	0.56 ± 0.22	0.59 ± 0.15	0.53 ± 0.09	0.16 ± 0.09	0.35 ± 0.25	0.45 ± 0.03	0.05 ± 0.01
Next Occurrence	0.77 ± 0.10	0.34 ± 0.09	0.27 ± 0.04	0.14 ± 0.04	0.19 ± 0.07	0.18 ± 0.06	0.13 ± 0.05
Intermediate	0.51 ± 0.31	1.00 ± 0.00	0.96 ± 0.05	0.25 ± 0.01	0.28 ± 0.22	0.26 ± 0.25	0.75 ± 0.16

Table 11: The mean accuracy and standard deviation values for three prompt variants.