TRAM: Benchmarking Temporal Reasoning for Large Language Models

Anonymous ACL submission

Abstract

Reasoning about time is essential for understanding the nuances of events described in natural language. Previous research on this topic has been limited in scope, characterized by a lack of standardized benchmarks that would allow for consistent evaluations across different studies. In this paper, we introduce TRAM, a temporal reasoning benchmark composed of ten datasets, encompassing various temporal aspects of events such as order, arithmetic, frequency, and duration, designed to facilitate a comprehensive evaluation of the TeR capabilities of large language models (LLMs). We evaluate popular LLMs like GPT-4 and Llama2 in zero-shot and few-shot scenarios, and establish baselines with BERT-based and domain-specific models. Our findings indicate that the best-performing model lags significantly behind human performance. It is our aspiration that TRAM will spur further progress in enhancing the TeR capabilities of LLMs.

1 Introduction

Temporal reasoning is fundamental for humans to understand the world and distinguish between everyday events. For instance, when given the activities "watching a movie" and "watching a sunset", we intuitively recognize that, though both are time-bound, a movie typically lasts longer than a sunset. Moreover, while movies can be watched repeatedly, sunsets transpire once a day. Such innate comprehension is not just about sequencing events or understanding durations; it extends to more intricate aspects of time, allowing us to make sense of complex narratives and the causality of events. Despite advancements in natural language processing (NLP) and the advent of large language models (Min et al., 2021; Zhao et al., 2023; Wang et al., 2023), mastering temporal reasoning remains a significant challenge due to its intricate nature, the variability of temporal expressions, and the need for contextual understanding.

Recent work in temporal reasoning (TeR) has primarily focused on time-sensitive question-answering (Zhou et al., 2019; Chen et al., 2021; Dhingra et al., 2022; Tan et al., 2023), demonstrating that despite significant advancements in NLP, current language models have yet to reach human-level performance in this domain. Furthermore, these studies, while addressing explicit temporal elements such as order, duration, and time-event relations, overlook more complex aspects of TeR, like temporal narratives and causality. Importantly, the establishment of a unified framework including broad facets of TeR has not yet been achieved.

Frequency (Commonsense)	**Q:** It is also a love story , between Ace and Tobio, a trans woman. How often do they break up? A. Once ✓ B. Always ✗ C. Once per week ✗
Ambiguity Resolution (Interpretation)	**Q:** A historic event is documented to have happened 'before you know it'. When did it take place? A. The next day ✗ B. Without hesitation ✗ C. Before long ✓
Temporal Causality (Cause)	**Q:** She noticed that all the wall clocks in the store were set to ten past ten. What's the more plausible CAUSE? A. It is a common display setting for clocks and watches. ✓ B. B. It was ten minutes past ten at that moment. ✗
Temporal Storytelling	**Q:** I woke up so late this morning. I was panicked when I saw what time it was. I had to be at work on time. I threw myself together quickly. Which of the two endings is the most plausible correct ending to the story? A. I was able to get a job at a local restaurant. ✗ B. I was still thirty minutes late. ✓
Arithmetic (24-hour Adjustment)	**Q:** What is 00:18 - 23:50? A. 0:28 ✓ B. 1:44 ✗ C. 22:15 ✗ D. 1:35 ✗

Figure 1: Example questions in TRAM.

To facilitate research in this direction, we present the **T**emporal **R**easoning for large l**A**nguage **M**odel benchmark (or TRAM for short), a collection of ten temporal reasoning tasks. These tasks range from foundational understanding (e.g., duration, frequency) to advanced temporal interpretations and computations (e.g., arithmetic, causality). Each task consists of one or more subtasks, all of which are specifically crafted to assess a model's TeR capabilities across varying levels of understanding and difficulty. In total, our benchmark includes 38 distinct subtasks, comprising a total of 526.7k questions. Answers have been derived through a combination of expert annotations and programmatic

generation. Diverging from prior TeR research and in line with (Hendrycks et al., 2020), our questions are formatted as straightforward multiple-choice tests rather than generative tasks, thereby more appropriately evaluating LLMs. Example questions in TRAM are shown in Figure 1.

To gain deeper insight into the TeR challenges posed by TRAM, we extensively evaluate several prominent language models, including BERT (Kenton and Toutanova, 2019), RoBERTa (Liu et al., 2019), the domain-specific TeR model RST (Yuan and Liu, 2022), and recent LLMs including Llama2 (Touvron et al., 2023), Gemini Pro (Team et al., 2023), GPT-3.5, and GPT-4 (OpenAI, 2023). We use limited training data to fine-tune BERT-style and RST models. LLMs are evaluated using standard and chain-of-thought prompting under zero-shot and few-shot learning paradigms. Our results indicate that GPT-4 excels in most tasks, achieving an average accuracy of up to 84.4%. Moreover, we observe notable performance disparities across tasks among the models. Despite the impressive performance of GPT-4, it falls short of human proficiency by over 10%, highlighting significant room for LLMs to improve their TeR capabilities. Manual error analysis shows that models struggle with nuanced understanding and interpreting implicit cues across all task categories.

In summary, our contributions are threefold:

(1) We introduce TRAM, a comprehensive collection of ten distinct TeR tasks featuring 526.7k questions presented in a multiple-choice format. Ranging from foundational temporal concepts to intricate temporal interpretations, TRAM serves as a unified framework for assessing the TeR capabilities of LLMs.

(2) We conduct extensive experiments on TRAM, evaluating leading language models including BERT-style models, a TeR-specific model, and LLMs such as Llama2 and GPT-4. Our results reveal that even the best-performing model notably falls short of human-level performance, underscoring the opportunities for continued research in this area.

(3) Manual error analysis reveals consistent TeR challenges for current LLMs, particularly in nuanced comprehension and decoding implicit temporal cues, highlighting the need for further research to enhance LLM capabilities in addressing these specific errors.

2 Related Work

Our proposal for a comprehensive TeR benchmark builds on the evolution of datasets in this domain while addressing the lack of a unified system for evaluation. The modern NLP landscape sets the stage for a nuanced evaluation of both pretrained models and LLM paradigms.

Temporal Reasoning Benchmarks In the realm of TeR, several datasets have emerged to address distinct challenges. Early benchmarks, such as TimeBank (Pustejovsky et al., 2003), predominantly focused on temporal relations. TempEval-3 (UzZaman et al., 2013) broadened the scope by introducing multiple tasks, including temporal entity extraction and temporal relation extraction. Recently, there has been a surge in the development of time-sensitive question-answering datasets like MCTACO (Zhou et al., 2019), Time-sensitive QA (Chen et al., 2021), TEMPLAMA (Dhingra et al., 2022), TEMPREASON (Tan et al., 2023), and MenatQA (Wei et al., 2023). However, these datasets often specialize in narrower aspects of TeR, such as duration, frequency, or event-time relations. In contrast, our benchmark offers a comprehensive scope of TeR, addressing diverse levels and dimensions of understanding about time, aiming to provide a more complete representation of TeR challenges than previously available datasets.

Training Paradigms in LLMs In NLP research, pretraining language models on vast amounts of diverse texts has become standard practice. Through this process, the models encapsulate a broad spectrum of information across various domains. BERT-based models like BERT (Kenton and Toutanova, 2019) and RoBERTa (Liu et al., 2019) are representative examples. These models have been applied to a diverse set of tasks, including disease prediction (Zhao et al., 2021), text classification (Wang et al., 2022b), time series analysis (Wang et al., 2022c), and more. However, the advent of GPT-3 (Brown et al., 2020) shifted the focus towards minimal fine-tuning approaches, such as zero-shot and few-shot learning, allowing models to adapt to new tasks with only a few training examples (Brown et al., 2020). This transition has spurred the development of advanced prompting techniques aimed at enhancing the understanding and reasoning capabilities of LLMs. Some representative prompting methods include CoT prompting (Wei et al., 2022), self-consistency (Wang et al., 2022a), tree-of-thought prompting (Yao et al.,

2023), and metacognitive prompting (Wang and Zhao, 2023). In this work, we establish baseline evaluations by considering traditional BERT-based models, a domain-specific TeR model, and recent LLMs such as Llama2 and GPT-4 to provide a comprehensive understanding of their strengths and limitations in diverse TeR tasks.

3 Tasks and Datasets

TRAM encompasses ten distinct tasks, presented as multiple-choice questions (MCQs) across a range of time-related domains. For clarity and directness, we ensure that each question has only one correct answer. The main purpose of TRAM is to spur further research into the advanced TeR capabilities of LLMs. Overall, these tasks fall under three distinct groups. (1) *Foundational Temporal Understanding Tasks:* Covering basic temporal comprehension, this group incorporates tasks such as ordering, frequency, duration, and typical time. (2) *Temporal Interpretation and Computation Tasks:* Centered on the interpretative and computational aspects of time, this group includes tasks like ambiguity resolution and arithmetic. (3) *Advanced Temporal and Conceptual Understanding Tasks:* Dedicated to exploring intricate temporal relationships and narratives, this category features tasks like relation, temporal NLI, causality, and storytelling. In this work, certain task names, such as 'relation' and 'causality', can have varied interpretations across different contexts. However, they are specifically emphasized for their temporal aspects in this work. Although we might occasionally omit the term 'temporal' for brevity, readers should note that the tasks are centered on time-related elements.

In TRAM, each task is designed with one or more problem types, ensuring diverse representation across data attributes, complexities, and domains. The benchmark includes 526,668 problems in total. For each dataset, we introduce a few-shot development set, with five questions per category, and a separate test set for evaluation. Table 1 provides an overview of these tasks, and more details can be found in Appendix A. The majority of tasks employ accuracy as the evaluation metric due to their straightforward MCQ structure. For tasks like 'relation' and 'temporal NLI', which exhibit an imbalanced label distribution inherent to their fixed class structure, both accuracy and the F1 score are utilized, even when they are presented as MCQs.

3.1 Foundational Temporal Understanding Tasks

This group of tasks is fundamental for assessing a model's proficiency in core temporal concepts. For the tasks below, data from the Multiple Choice TemporAl COmmon-sense (MCTACO) dataset incorporates both validation and test sets, while data from the Stanford Question Answering Dataset (SQuAD) dataset includes both training and validation sets. Unless otherwise mentioned, the options for each dataset are generated through a blend of human curation and algorithmic processes, tailored to each specific task. For instance, in the ordering task, correct answers strictly adhere to the accurate chronological sequence of events, while incorrect choices are formed through random permutations.

Ordering The temporal ordering task evaluates a model's ability to understand the sequence in which events occur. This task is divided into two primary problem types. For *commonsense* problems, we mainly source questions from the MCTACO dataset (Zhou et al., 2019), specifically targeting subcategories related to temporal ordering. For each individual question selected from this dataset, we pose questions in the format, "Is {candidate answer} possible?" While MCTACO's expected answers are "true" or "false", we introduce another layer of complexity by also including an "undetermined" option. Additionally, we curate another set of commonsense questions by manually structuring event sequences logically, followed by programmatic question generation. Concurrently, recognizing the significance of tasks rooted in real-world events, we introduce *facts* problems. These focus on major historical events, spanning from ancient to contemporary times, and are sourced from Wikipedia timelines. Models are posed with challenges such as sequencing: "Arrange the following events in chronological order" and verification queries like, "Is the following sequence of events in the correct chronological order?".

Frequency The frequency task assesses a model's ability to understand how often events take place over time and comprises six distinct categories of problems. For the *commonsense* category, we source questions from the MCTACO dataset related to frequency. Each selected category ensures the presence of at least two incorrect options and one correct one. To prevent models from memorizing specific answer orders, we randomize the placement of the correct answers. In the *reading*

Table 1: Overview of ten tasks included in TRAM. The "Data Size" column aggregates totals from both the development and test sets. "*K*-Way MC" signifies a multiple-choice response format with *K* options. *Amb. Res.* denotes Ambiguity Resolution. *NLI* stands for natural language inference. "Same" indicates the text source is the same as the row above.

Task	Data Size	# Problem Types	Metrics	Answer Type	Text Sources
Foundational Temporal Understanding Tasks					
Ordering	29,462	2	Acc.	3-Way MC	MCTACO[1], Wikipedia, Misc.
Frequency	4,658	6	Acc.	3-Way MC	MCTACO[1], SQuAD[2], Misc.
Duration	7,232	7	Acc.	3-Way MC	Same
Typical Time	13,018	4	Acc.	3-Way MC	Same
Temporal Interpretation and Computation Tasks					
Amb. Res.	3,649	5	Acc.	3-Way MC	Misc.
Arithmetic	15,629	9	Acc.	4-Way MC	Same
Advanced Temporal and Conceptual Understanding Tasks					
Relation	102,462	1	Acc./F1	3-Way MC	TempEval-3[3]
Temporal NLI	282,144	1	Acc./F1	3-Way MC	MNLI[4], SNLI[5]
Causality	1,200	2	Acc.	2-Way MC	COPA[6], Misc.
Storytelling	67,214	1	Acc.	2-Way MC	ROC[7], SCT[8]

[1] (Zhou et al., 2019), [2] (Rajpurkar et al., 2016), [3] (UzZaman et al., 2013), [4] (Williams et al., 2018), [5] (Bowman et al., 2015), [6] (Roemmele et al., 2011), [7] (Mostafazadeh et al., 2016), [8] (Mostafazadeh et al., 2017)

comprehension category, questions are chosen from the SQuAD dataset (Rajpurkar et al., 2016) based on frequency-oriented keywords like "how often", "how many times", and "how frequent". The *application* and *computation* categories are mainly made up of human-curated templates that test the model's ability to infer time intervals and compute either previous or subsequent occurrences. The *comparison* problems blend real and artificially conceived events, challenging the model's ability to differentiate frequency nuances. Lastly, the *facts* category draws questions from various sources, with Wikipedia being the primary one, centering on queries related to events that are known to happen regularly or periodically in either historical or contemporary settings.

Duration The duration task is designed to assess a model's capability to comprehend the length of events or periods of time and encompasses seven distinct categories of problems. The *commonsense* problems are derived from the MCTACO dataset, probing the model's fundamental understanding of event durations grounded in everyday scenarios. The extraction methods mirror those used for the "frequency" task. The *reading comprehension* category sources questions from the SQuAD dataset, selecting those with duration-oriented keywords like "how long", "how many years", and "how much time". Apart from the aforementioned subtasks, all other categories consist of human-curated templates or problems. The *analogy inference* category assesses the model's ability to discern durations through analogical reasoning. The *computation* category tests mathematical precision, where problems often require arithmetic operations to determine event durations. Comparative analysis is examined in two subtasks: *direct comparison*, which demands straightforward judgments of event durations involving both real and artificial events; and *multi-step comparison*, which challenges the model to infer and integrate information across sequential statements. Lastly, the *facts* category primarily draws from Wikipedia, furnishing questions anchored in well-documented historical or contemporary durations.

Typical Time The typical time task is constructed to evaluate a model's understanding of when events or activities typically occur, segmented into four distinct categories. The *commonsense* category draws problems from the MCTACO dataset, exploring the model's innate comprehension of event timings as they manifest in daily scenarios. The extraction method for this subtask is similar to that used for the "frequency" task. The *comparison* category, comprising human-curated statements and problems, delves into relative timing. This cate-

gory involves determining which of two presented scenarios is more temporally typical or discerning which event customarily precedes the other. The *facts* category, primarily sourced from Wikipedia timelines spanning ancient history to the 21st century, provides the model with specific historical or established events and expects it to identify the precise times or periods associated with them. Lastly, the *reading comprehension* problem sets source questions from the SQuAD dataset. This category filters problems based on keywords like "at what time", "when did", and "in what year", challenging the model to extract specific temporal data from passages.

3.2 Temporal Interpretation and Computation Tasks

This group of tasks is fundamental in gauging a model's adeptness at deciphering, processing, and computing temporal information.

Ambiguity Resolution The temporal ambiguity resolution task aims to gauge a model's ability to decipher and resolve uncertainties related to temporal expressions, divided into five subtasks. The *interpretation* category evaluates the model's comprehension of ambiguous time-related phrases commonly used in everyday language. The *calendar shift* subtask necessitates the conversion between different calendar systems, such as the Julian and Gregorian. The *long-term shift*, *mid-term shift*, and *short-term shift* categories challenge the model's capacity to adjust dates over long (i.e., years), medium (i.e., months, weeks, days), and short (i.e., hours, minutes, seconds) timeframes, respectively. All questions across these categories originate from carefully crafted human templates.

Arithmetic The temporal arithmetic task evaluates a model's capacity to manage calculations related to time, organized into nine distinct subtasks. The *application* category presents real-world scenarios such as time calculations involving schooling, vacations, homework, and flights. *Date computation* sets focus on adding or subtracting days from specified dates to determine a new date. *hour adjustment* subtasks, divided into *12-hour* and *24-hour* formats, require the model to calculate time differences or additions. The *month shift* subtask examines the model's ability to pinpoint a month that is a certain number of months away from a specified month. The *week identification* problems determine the exact week number within a year based on a given date. In *year shift*, the model discerns a year a certain number of years relative to a provided year. *time computation* evaluates the model's proficiency in converting various time units, especially over shorter durations like days, hours, minutes, and seconds. Lastly, the *time zone conversion* category requires the model to convert times between different zones. Both the question templates and the programs used to formulate answers derive from human expertise.

3.3 Advanced Temporal and Conceptual Understanding Tasks

This group of tasks is fundamental in assessing a model's depth of comprehension in time-oriented narratives and in discerning complex conceptual relationships.

Relation The temporal relation task seeks to assess a model's ability to identify the relationship between two entities involving time, categorized as either an *event-to-time* or an *event-to-event* association. Questions are crafted based on the TempEval-3 Silver dataset (UzZaman et al., 2013). The context sentences, which contain the two entities in question, are directly extracted from the original passages. One inherent challenge of this task lies in the subtle nuances among the fixed set of relations. For instance, distinguishing between relations like "BEFORE" and "IMMEDIATELY BEFORE" can be particularly demanding, as they require fine-grained comprehension of temporal sequences. With the predetermined relations from the dataset, the correct relation option is randomized in its placement, while distractor options are chosen from the pool of remaining relations.

Temporal NLI The temporal NLI task is designed to evaluate a model's ability in *natural language inference*, with a particular emphasis on statements that involve temporal elements. We source questions from prevalent NLI datasets, including Stanford Natural Language Inference datasets (SNLI) (Bowman et al., 2015) and Multi-Genre Natural Language Inference (MNLI) (Williams et al., 2018). Data from the MNLI dataset includes training and validation sets, while data from the SNLI dataset includes training, validation, and test sets. We select problems based on keywords that capture a range of temporal nuances, such as explicit references (e.g., 'tomorrow', 'later'), months (e.g., 'May', 'October'), seasons (e.g., 'summer', 'winter'), and temporal actions (e.g., 'in advance', 'postpone'). Consistent with the original task, the three response options for all questions are: "En-

tailment", "Neutral", and "Contradiction".

Causality The temporal causality task assesses a model's capability to discern cause-and-effect relationships within scenarios influenced by time. Drawing inspiration from the Choice of Plausible Alternatives (COPA) dataset (Roemmele et al., 2011), we select questions that naturally contain temporal elements such as 'postpone', 'tomorrow', 'summer', and 'clock'. Additionally, we manually craft problems to highlight the temporal nature of COPA-style questions. Each problem presents a situation that revolves around time, followed by a question pinpointing either the most plausible *cause* or *effect* of that situation. Both options for these problems are carefully created by hand. For augmentation purposes, we create additional, mirrored instances for each original sample. This approach ensures that for a given question with two options, each option is supported by a uniquely tailored premise, effectively creating a distinct and relevant context for both choices.

Storytelling The *temporal storytelling* task is designed to assess a model's ability to predict the appropriate ending of stories that emphasize temporal elements. We source questions from the ROCStories (ROC) (Mostafazadeh et al., 2016) and Story Cloze Test (SCT) (Mostafazadeh et al., 2017) datasets. We identify and select stories that contain notable temporal components by filtering them using keywords such as 'now', 'tomorrow', 'future', 'always', and 'postpone', among others. The typical format of the task presents a story comprising four sentences, followed by two potential endings. The model is required to choose the most appropriate conclusion for the story. In the case of SCT, which inherently provides two endings for each story, our focus remains on selecting stories with evident temporal aspects. To further enrich our dataset, we take the initial four sentences from the ROC and employ GPT-2 (Radford et al., 2019) to produce an alternate, incorrect ending, initiated with the prompt "unexpectedly". Subsequently, we filter this augmented data to ensure that stories emphasize the desired temporal themes.

4 Experiments

In our evaluation, we compare the performance of prevalent LLMs across all datasets and analyze the mistakes they make. We report the best results after multiple runs for each experimental setting.

4.1 Experimental Setup

We evaluate the performance of several well-known language models on the TRAM benchmark, which is organized into two main categories. In the first category, we employ four popular LLMs: Llama-2-70b-chat (Touvron et al., 2023), Gemini Pro (Anil et al., 2023), GPT-3.5 Turbo, and GPT-4 Turbo (OpenAI, 2023). Each of these models is accessed using its corresponding API key. Specifically, we query Gemini through Google Vertex AI, the GPT models through the OpenAI API, and Llama2 via DeepInfra. Following (Tan et al., 2023) and considering API cost constraints, we evaluate model performance on 300 randomly selected examples per category from the test set, using all available examples for categories with fewer than 300. For all evaluations, greedy decoding (i.e., temperature = 0) is applied during model response generation. We evaluate each model using two prompting strategies: standard prompting (SP) (Brown et al., 2020; Kojima et al., 2022) and CoT (Wei et al., 2022) prompting. Under both strategies, the models undergo tests in zero-shot and 5-shot settings. In the 5-shot scenario, exemplars are consistently drawn from the development set. Step-by-step answers associated with CoT prompting are obtained through human annotation. More details about prompts can be found in Appendix B.

In the second category, we consider minimal supervision as opposed to traditional fully supervised learning to establish baseline evaluations. The rationale behind this decision is driven by the intention to leverage the inherent world knowledge of the models and to ensure an equitable comparison with the previously mentioned LLMs. We employ four representative BERT-style models, including BERT-base, BERT-large (Kenton and Toutanova, 2019), RoBERTa-base, and RoBERTa-large (Liu et al., 2019). For temporal NLI, we employ the Sequence Classification versions of BERT and RoBERTa from Huggingface, which align with the task demands. For other tasks, we use their Multiple Choice configurations. Additionally, we include the RST (Yuan and Liu, 2022), a domain-specific TeR model, to benchmark against the generalist models. The data sampling strategy for minimal supervision is structured based on the size of the original dataset. For datasets with around 1k samples, we randomly select 50% of the remaining data after setting aside the test data used for LLM evaluation. For datasets with sizes between 3k and

Table 2: Performance comparison of each model across ten tasks in TRAM. GPT-4 consistently outperforms other models under both zero-shot (0S) and 5-shot (5S) settings across the majority of tasks. Interestingly, the RoBERTa-large model achieves a higher average performance than models with larger architectures, such as Llama2. Human expert performance serves as an upper bound, illustrating that there still exists room for improvement in LLMs on TeR tasks. The abbreviations *Freq., Dur., Arith., Rel., Caus.* refer to frequency, duration, arithmetic, relation, and causality, respectively. All values are percentages. Best model results are highlighted in bold.

Model	**Order** *Acc.*	**Freq.** *Acc.*	**Dur.** *Acc.*	**Typical Time** *Acc.*	**Amb. Res.** *Acc.*	**Arith.** *Acc.*	**Rel.** *Acc./F1*	**NLI** *Acc./F1*	**Caus.** *Acc.*	**Story** *Acc.*	**Average**
Random	33.3	33.3	33.3	33.3	33.3	25.0	33.3/33.3	33.3/33.3	50.0	50.0	35.4
Llama2 (0S, SP)	51.3	73.5	64.9	74.1	46.9	52.6	35.2/33.1	64.4/63.9	90.5	86.7	61.4
Llama2 (0S, CoT)	52.9	75.4	66.3	75.5	49.4	55.6	40.1/38.5	67.7/67.4	92.0	88.2	64.1
Llama2 (5S, SP)	52.2	74.1	65.7	74.6	48.0	53.9	38.1/36.6	65.2/64.7	92.0	87.3	62.7
Llama2 (5S, CoT)	53.8	76.3	67.1	75.9	50.7	57.8	43.0/41.3	69.8/69.2	93.6	88.5	65.6
Gemini (0S, SP)	55.4	86.2	83.9	82.7	75.1	69.8	60.5/60.1	69.5/70.7	92.5	91.2	74.8
Gemini (0S, CoT)	56.9	87.6	84.2	83.6	76.9	71.8	64.2/63.6	70.9/71.8	94.0	92.0	76.5
Gemini (5S, SP)	56.4	86.5	84.5	82.9	75.8	70.4	62.8/62.3	70.4/71.0	94.2	91.5	75.7
Gemini (5S, CoT)	57.4	88.2	86.3	83.8	77.4	72.5	65.1/64.9	72.3/73.1	**95.3**	92.2	77.4
GPT-3.5 (0S, SP)	52.5	76.3	70.8	77.8	71.6	72.8	40.5/39.1	73.8/74.2	93.4	90.5	69.4
GPT-3.5 (0S, CoT)	53.7	78.3	72.3	78.7	74.6	74.8	44.1/42.9	75.2/75.7	94.5	91.7	71.4
GPT-3.5 (5S, SP)	53.2	77.8	71.6	79.2	73.4	73.7	42.5/41.3	74.5/75.0	94.5	91.0	70.6
GPT-3.5 (5S, CoT)	54.8	79.2	72.7	80.3	75.2	75.0	45.9/45.2	76.3/76.9	94.8	91.7	72.3
GPT-4 (0S, SP)	64.7	85.2	86.1	84.6	82.3	87.1	60.6/58.8	82.9/85.3	92.6	91.0	80.1
GPT-4 (0S, CoT)	66.2	87.7	86.4	85.5	84.1	88.9	63.6/62.9	85.4/87.2	92.9	93.2	82.0
GPT-4 (5S, SP)	65.8	86.3	87.3	84.8	83.6	88.3	62.0/61.5	83.7/86.4	92.6	91.6	81.2
GPT-4 (5S, CoT)	**69.5**	**90.7**	**89.2**	**87.2**	**87.1**	**91.2**	66.5/65.2	**87.7/89.6**	95.0	**93.6**	**84.4**
BERT-base	50.0	47.3	50.0	53.0	36.6	25.9	86.5/86.6	53.0/53.4	81.0	79.0	58.5
BERT-large	52.5	53.1	53.3	56.8	37.4	28.3	89.5/89.5	59.5/60.1	85.0	81.3	62.2
RoBERTa-base	50.8	54.5	51.8	55.3	37.4	26.4	87.0/86.8	64.5/64.9	82.3	81.3	61.9
RoBERTa-large	55.5	57.7	55.4	60.0	41.0	29.1	90.0/90.0	70.0/70.3	88.0	84.0	65.9
RST	54.5	56.2	52.3	58.7	39.8	31.6	**91.5/91.6**	68.2/68.7	87.5	82.2	65.2
Human Experts	86.0	96.3	97.7	94.5	94.8	98.7	96.0/96.0	92.0/92.4	100.0	98.0	95.2

10k, we select 10%. For those with sizes between 10k and 100k, we sample 2.5%, and for datasets with more than 100k examples, we take 1%. This limited training data is used for model fine-tuning. The same test set is used consistently with LLMs.

In addition to evaluating model performance, multiple expert annotators worked on each problem type for every task in TRAM to better understand human performance. Each expert answered a subset of the 50 questions from each category of every task, which were randomly selected from the test set. Collectively, they tackled about 1,900 questions across TRAM. Further details on human expert annotators and human non-specialists are provided in Appendix C.

4.2 Overall Performance Comparison

We compare the performance of different models across ten tasks, as shown in Table 2. There are several key takeaways. First, GPT-4 consistently outperforms other models across the majority of tasks, demonstrating a performance advantage of over 9% compared to the second-best model, Gemini, under 5-shot CoT. Second, CoT often results in performance enhancements, corroborating the findings from (Wei et al., 2022) and emphasizing the efficacy of step-by-step prompting in augmenting LLMs' performance in intricate reasoning tasks. Third, it is notable that RoBERTa-large, despite its size, surpasses the larger Llama2 in average performance. This observation underscores that sheer model size does not always equate to superior performance. Several factors might contribute to this outcome. RoBERTa-large may utilize optimization strategies particularly beneficial for these tasks. Additionally, inherent features or efficiencies in its architecture might enhance its ability to understand and process temporal cues. Delving deeper into task-specific performance, certain tasks such as ambiguity resolution and arithmetic show considerable variance across models. For LLMs, performance on the arithmetic task varies significantly, ranging from 52.6% to 91.2%. Moreover, BERT and RoBERTa exhibit exceptional performance in the temporal relation task, potentially due to their bidirectional contextual processing and emphasis

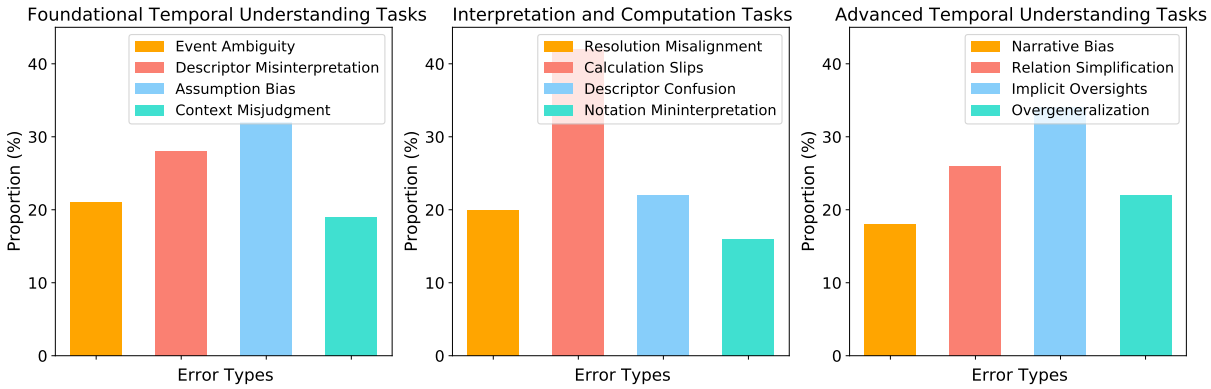

Figure 2: Error type distribution for three groups of tasks in TRAM. Models often struggle with subtle details and hidden clues across all categories.

on token-level relationships. This contrasts sharply with their average or below-average performance in other tasks. For the specialized RST model, we observe comparable average performance with RoBERTa-large, indicating the benefits of tailored training for domain-specific tasks. The discrepancies in performance among models suggest that certain architectures or training methodologies are better suited for specific types of reasoning or tasks, highlighting the need for tailored approaches. Finally, despite the lead of GPT-4, it remains 12.9% behind human performance, underscoring the potential for further LLM enhancements.

4.3 Error Analysis

To better understand the mistakes made by models, we manually analyze instances where a model has made incorrect choices or provided inappropriate answers, focusing exclusively on LLMs. Figure 2 illustrates the common error types and their proportions for each task group. In foundational temporal understanding tasks, "assumption bias" was the most frequent error, accounting for 32% of all mistakes. In the interpretation and computation tasks, "calculation slips" dominated, making up 42% of the errors. "Implicit oversights" led in the advanced temporal understanding tasks with a representation of 34%. Detailed descriptions of each error type can be found in Appendix D.

5 Discussion

We introduce TRAM, a comprehensive benchmark spanning ten diverse tasks, to evaluate the temporal reasoning of LLMs. The contrasting performances across models emphasize the significance of experimental strategies and shed light on the intrinsic challenges. This benchmark serves as a tool for researchers to identify model limitations and guide further advancements in this domain.

Implications of TRAM The introduction of TRAM establishes a new paradigm for probing the temporal reasoning capabilities of LLMs. Unlike previous benchmarks, which often offered fragmented insights into temporal tasks, TRAM provides a comprehensive system. This allows for a unified evaluation of how models comprehend both rudimentary temporal concepts and complex temporal narratives. The differentiation in task complexity within TRAM elucidates the various stages of temporal understanding. In particular, TRAM underscores challenges like decoding implicit temporal cues and navigating intricate temporal relationships, providing a roadmap for future improvements in LLMs in this area.

Future Directions TRAM has initiated a step towards evaluating LLMs' temporal reasoning capabilities, but there are further avenues to explore. Going forward, we will experiment with more test data and refine tailored prompting techniques for each task through iterative testing. Moreover, we plan to expand the benchmark to include varied question formats. For generative tasks, this might encompass short answers and summarization. Even within MCQs, we intend to incorporate questions that may have one or more correct answers, allowing for a more comprehensive evaluation. We also plan to fine-tune existing open-source LLMs on these tasks, such as Llama2. These efforts aim to create tailored LLMs that can better understand and reason about time across various contexts.

6 Limitations

While TRAM sets a holistic standard for TeR assessment, we acknowledge its limitations. One pri-

mary concern is the subset evaluation of the test set, which may not reflect the full spectrum of LLMs' TeR capabilities. Furthermore, the MCQ format may allow LLMs to guess randomly, skewing performance evaluations. Moreover, textual questions may not capture the entire complexity of TeR tasks, as real-world scenarios often integrate multi-modal cues such as images and videos.

References

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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

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.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2021. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849.

Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James Allen. 2017. Lsdsem 2017 shared task: The story cloze test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 46–51.

OpenAI. 2023. Gpt-4 technical report.

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. 2003. The timebank corpus. In *Corpus linguistics*, volume 2003, page 40. Lancaster, UK.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners.

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.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

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.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022a. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Yuqing Wang and Yun Zhao. 2023. Metacognitive prompting improves understanding in large language models. *arXiv preprint arXiv:2308.05342*.

Yuqing Wang, Yun Zhao, Rachael Callcut, and Linda Petzold. 2022b. Integrating physiological time series and clinical notes with transformer for early prediction of sepsis. *arXiv preprint arXiv:2203.14469*.

Yuqing Wang, Yun Zhao, and Linda Petzold. 2022c. Enhancing transformer efficiency for multivariate time series classification. *arXiv preprint arXiv:2203.14472*.

Yuqing Wang, Yun Zhao, and Linda Petzold. 2023. Are large language models ready for healthcare? a comparative study on clinical language understanding. *arXiv preprint arXiv:2304.05368*.

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.

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.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

Weizhe Yuan and Pengfei Liu. 2022. restructured pretraining. *arXiv preprint arXiv:2206.11147*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Yun Zhao, Yuqing Wang, Junfeng Liu, Haotian Xia, Zhenni Xu, Qinghang Hong, Zhiyang Zhou, and Linda Petzold. 2021. Empirical quantitative analysis of covid-19 forecasting models. In *2021 International Conference on Data Mining Workshops (ICDMW)*, pages 517–526. IEEE.

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. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3363–3369.

A Datasets

This section presents the datasheet for TRAM, a glossary of definitions for all subtasks, and details the dataset construction process, including human-crafted templates for programmatic question generation and the use of temporal keywords to filter questions from existing datasets. Additional example questions for each task, as well as those sourced from existing datasets, are provided. Furthermore, for tasks comprising multiple subtasks, we provide their distribution. Note that the following templates do not represent the full spectrum of templates we used when constructing the datasets.

A.1 Datasheet for TRAM

OVERVIEW

Motivation and Intended Uses.

1. What are the intended purposes for this benchmark?

The benchmark is designed to establish a standard for evaluating temporal reasoning in large language models. It focuses on three key areas: Foundational Temporal Understanding (such as Duration and Frequency), Temporal Interpretation and Computation (including Ambiguity Resolution and Arithmetic), and Advanced Temporal and Conceptual Understanding (encompassing areas like Causality and Storytelling).

2. Was it designed to address a specific task or fill a particular gap in research or application?

The benchmark is curated to address the need for a robust and comprehensive tool, specifically designed to evaluate temporal reasoning in large language models. It provides a diverse set of tasks that

challenge models in the more intricate aspects of temporal reasoning.

Limitations and Inappropriate Uses.

3. Are there any specific tasks or applications for which this benchmark should not be used?

The focus of the benchmark is on understanding and interpreting time-related concepts. Therefore, it may not be suitable for evaluations that significantly diverge from temporal reasoning, such as tasks involving texts that require contextual emotional intelligence, or domain-specific applications in medical or legal document analysis.

DETAILS

Composition.

4. What do the instances that comprise the benchmark represent?

The instances consist of multiple-choice questions, created from a combination of existing datasets and human-curated problems, with a focus on temporal reasoning tasks. Each instance is specifically designed to assess a language model's ability to process and reason about time in natural language.

5. How many instances are there in total (of each type, if appropriate)?

There are a total of 526,668 problems. Specifically, the dataset comprises 10 main tasks and 38 subtasks. The number of problems for each main task is as follows: Ordering (29,462), Frequency (4,658), Duration (7,232), Typical Time (13,018), Ambiguity Resolution (3,649), Arithmetic (15,629), Relation (102,462), Temporal NLI (282,144), Causality (1,200), and Storytelling (67,214).

6. Does the benchmark contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

Part of the benchmark comprises a curated selection of instances, representing a comprehensive but not exhaustive collection of temporal reasoning problems. Specifically, it includes problems selectively sourced from existing datasets that exemplify a wide array of temporal reasoning scenarios. Human expertise has verified and determined the representativeness of the selected problems.

7. Is there a label or target associated with each instance?

Yes, the label for each instance is the correct answer to the multiple-choice question, indicated as either A, B, C, or D, and this varies by task.

8. Is the benchmark self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?

The benchmark is partially self-contained. Problems derived from existing datasets have been integrated into TRAM in a way that makes them standalone. This integration includes manually adding distracting or confusing options, filtering out irrelevant questions for relevance, and reformulating problems. For transparency, references are provided for problems that originated from existing data. The remaining questions are heavily driven by human curation, supplemented by programmatic generation.

9. Does the benchmark contain data that might be considered sensitive in any way?

The benchmark does not contain any sensitive data.

Data Quality.

10. Is there any missing information in the benchmark?

Everything is included. No data is missing.

11. What errors, sources of noise, or redundancies are important for benchmark users to be aware of?

Firstly, some problems in the benchmark might contain contextual ambiguities leading to multiple plausible interpretations. The benchmark is designed to have one correct answer per question, with the final unique correct answer determined or verified by a group of professionals. Secondly, within the same main task, there may be similar problems with nuanced differences. While complete redundancy of problems across the entire benchmark is avoided, the presence of similar problems is not. Finally, for problems sourced from existing datasets, irrelevant or diverging options may occur during reformulation due to issues with the source data. Further verification checks will be conducted to minimize any errors or noise that may arise in the benchmark.

12. How was the data validated/verified?

The benchmark was initially verified by multiple professionals possessing advanced degrees (M.S. or Ph.D.) in cognitive science and psychology, who provided insights into the nuances of human temporal cognition, as well as in statistics, mathematics, and computer science, for their expertise in analytical rigor required by many tasks. They reviewed the problems for relevance and common errors, such as formatting inconsistencies or logical discrepancies in questions and answers. The

final review of the benchmark was conducted by the authors of the TRAM paper, who checked for relevance and removed any obvious noise and redundancies.

Pre-Processing, Cleaning, and Labeling.

13. What pre-processing, cleaning, and/or labeling was done on this benchmark?

In the preparation of the benchmark, several key steps were undertaken to ensure its overall quality and relevance:

1) Pre-processing: This step involved standardizing the format of problems sourced from relevant existing datasets to align with the TRAM benchmark's structure. It included unifying the formats of questions and answers, normalizing temporal expressions, and ensuring consistency in language and style. Additionally, over 100k problems in the benchmark were manually crafted, supplemented by program generation.

2) Cleaning: A thorough review was conducted to identify and correct any obvious errors in the data. This process involved resolving typos, rectifying factual inaccuracies, and eliminating ambiguous or misleading phrasing in both questions and options. However, nuanced errors such as acceptable bias in multiple interpretations of the same problem and subtle logical errors might be overlooked and could still be present in the current version of the benchmark.

3) Labeling: Each problem in the benchmark was carefully labeled with the correct answer. In the case of multiple-choice questions, plausible distractors were also manually created and added. Labels were verified for accuracy by subject matter experts to ensure that they correctly represented the intended temporal reasoning challenge.

14. Provide a link to the code used to preprocess/clean/label the data, if available.

The code for data pre-processing is available on the official GitHub page.

15. If there are any recommended data splits (e.g., training, validation, testing), please explain.

For each main task, there is a few-shot development set, with 5 questions per category (subtask), and a separate test set for evaluation.

ADDITIONAL DETAILS ON DISTRIBUTION AND MAINTENANCE

Distribution.

16. Will the benchmark be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?

Yes, the benchmark will be publicly available on the Internet.

17. How will the benchmark be distributed (e.g., tarball on website, API, GitHub)?

The benchmark is distributed via the official GitHub page.

18. When will the benchmark be distributed?

The benchmark was first released in September 2023.

Maintenance.

19. Who will be supporting/hosting/maintaining the benchmark?

The first author of the TRAM paper will be supporting and maintaining the benchmark.

20. Will the benchmark be updated (e.g., to correct labeling errors, add new instances, delete instances)?

Updates to question sets, error corrections, and results will be shared on the official GitHub page.

21. Will older versions of the benchmark continue to be supported/hosted/maintained?

Given any updates to the benchmark, older versions will be retained for consistency.

22. If others want to extend/augment/build on/contribute to the benchmark, is there a mechanism for them to do so?

Others wishing to do so should contact the original authors of TRAM about incorporating fixes or extensions.

A.2 Task Glossary Definitions

We provide a glossary with definitions of all tasks and subtasks encompassed within our TRAM benchmark for clarity. In our actual dataset formatting, the subcategory (if a task comprises multiple subtasks) or source (if a single subtask is sourced from an existing dataset) is marked for verification and convenient lookup.

Ordering: Chronological arrangement of events.

- *Commonsense*: Logical sequencing of events based on general knowledge.

- *Facts*: Accurate ordering of historical events based on factual information.

Frequency: Determination of how often events occur over time.

- *Commonsense*: Assessment of event occurrence rates based on general knowledge.
- *Reading Comprehension*: Frequency information extraction from passages.
- *Application*: Inference of time intervals and event frequencies.
- *Computation*: Calculation of event occurrences and intervals.
- *Comparison*: Differentiation of event frequencies in various contexts.
- *Facts*: Identification of periodically occurring events.

Duration: Determination of the length of events or time periods.

- *Commonsense*: Evaluation of time spans in everyday life scenarios.
- *Reading Comprehension*: Duration information extraction from passages.
- *Analogy Inference*: Discernment of relative time spans through contextual comparison.
- *Computation*: Calculation of event lengths.
- *Direct Comparison*: Straightforward assessment of event durations in a given set.
- *Multi-step Comparison*: Analysis of relative durations using layered information.
- *Facts*: Identification of length of factual events.

Typical Time: Determination of when events or activities typically occur.

- *Commonsense*: Analysis of usual event timings in daily life scenarios.
- *Comparison*: Assessment of relative event timings and typical sequences.
- *Facts*: Identification of historical times or periods from established events.
- *Reading Comprehension*: Specific time information extraction from passages.

Ambiguity Resolution: Resolution of uncertainties in temporal expressions.

- *Interpretation*: Understanding of ambiguous time-related phrases.
- *Calendar shift*: Conversion between different calendar systems.
- *Long-term shift*: Adjustment of dates over extended periods (years).
- *Mid-term shift*: Date adjustments over intermediate periods (months, weeks, days).
- *Short-term shift*: Time adjustments over brief periods (hours, minutes, seconds).

Arithmetic: Execution of time-related calculations.

- *Application*: Real-world time calculation scenarios (schooling, vacations, etc.).
- *Date Computation*: Addition or subtraction of days to find new dates.
- *12-hour Adjustment*: Time difference calculations in 12-hour format.
- *24-hour Adjustment*: Time difference calculations in 24-hour format.
- *Month Shift*: Identification of a future or past month from a given date.
- *Week Identification*: Determination of week numbers within a year.
- *Year Shift*: Calculation of future or past years from a specified year.
- *Time Computation*: Calculating future or past years from a specified year.
- *Time Zone Conversion*: Conversion of times between different time zones.

Relation: Identification of the temporal relationship between two entities, either as an event-to-time or event-to-event association.

Temporal NLI: Assessment of a 'hypothesis' as true (entailment), false (contradiction), or undetermined (neutral) relative to a 'premise' with temporal elements.

Causality: Analysis of cause-and-effect relationships in time-related scenarios.

- *Cause*: Identification of the initiator or reason leading to a particular event.
- *Effect*: Determination of the outcome or consequence resulting from a specific cause.

Storytelling: Prediction of appropriate story endings, with an emphasis on temporal elements.

A.3 Data Construction

Ordering For our ordering dataset, the *facts* problems were derived from actual events extracted from historical timelines on Wikipedia. Specifically, pages such as https://en.wikipedia.org/wiki/Timeline_of_the_18th_century served as our primary data sources. These timelines cover events ranging from ancient history to the 21st century, offering a rich foundation for our dataset. We explored dedicated pages for each available century, ensuring a diverse collection of events across various epochs.

Frequency For the frequency task, three main subtasks are generated based on templates: comparison, computation, and applications. Each template contains placeholders, denoted by {}, to represent both events and times. Table 3 outlines some representative templates for each subtask. The construction processes for other subtasks are detailed in the main paper.

Duration For the duration task, five main subtasks are generated based on templates: multi-step comparison, analogy inference, computation, direct comparison, and facts. Each template contains placeholders, denoted by {}, to represent both events and times. Table 4 outlines some representative templates for each subtask. The construction processes for other subtasks of the dataset are described in the main paper.

Typical Time For the typical time task, we crafted pairs of time-related events to test the model's proficiency in determining "Which statement is more typical in terms of time?" For instance, when presented with statements such as "People often have dinner in the early to late evening" and "People often have dinner in the mid to late afternoon", the model is prompted to recognize which one is more aligned with a conventional behavior. Similarly, it might evaluate statements like "Bars are typically busiest on Friday and Saturday nights" in comparison to "Bars are typically busiest on Sunday and Monday nights". Through these examples, we aim to assess the model's aptitude in discerning standard temporal practices.

Ambiguity Resolution For the ambiguity resolution task, we introduced templates to test the model's proficiency in resolving temporal ambiguities. Additionally, we manually gathered both common and uncommon temporal expressions that might perplex individuals and the model alike, such as "for a coon's age", "when pigs fly", and "in the nick of time". Table 5 presents representative templates for each subtask. Each template contains placeholders, denoted by {}, to represent both events and times.

Arithmetic We mainly adopted a programmatic generation approach, grounded in meticulously designed templates that focus on specific temporal calculations. These templates encompass a variety of temporal arithmetic tasks, ranging from basic time adjustments to more complex calculations like week identifications and real-world applications. Table 6 shows the major templates we use for constructing the arithmetic datasets. The variable values, denoted by {}, are randomly generated by programs. Through these templates, we can generate diverse questions that test a model's proficiency in handling different temporal arithmetic scenarios.

Relation To derive temporal relation questions from the TempEval-3 Silver dataset, we iterated through each temporal link (*tlink*) to extract the relationship type (*relType*) and relevant event and time IDs. For each *tlink*, the associated *eventInstanceID* provided the *eventID*, either directly or via the *MAKEINSTANCE* tag. We then identified the sentence containing this event as its contextual background. Using the gathered data, we crafted questions such as "What is the relationship between the event '$event_1$' and the event '$event_2$'?" or analogous questions pertaining to event-time relationships. The context, encompassing both events, was attached to the resulting question to ensure clarity.

Temporal NLI To construct our temporal NLI dataset, we adopted a keyword-based filtering approach from SNLI and MNLI datasets. Recognizing that NLI tasks can often hinge on nuanced temporal cues, we curated a comprehensive set of temporal keywords, as shown in Table 7. This selection was designed to capture a broader range of temporal relationships and nuances. Instances containing at least one term from this extended list were considered to possess temporal elements and were thus included for further analysis.

Causality Inspired directly by the style of the COPA dataset, our goal was to capture the intricate

Table 3: Major templates are used for constructing the frequency subtasks: comparison, computation, and applications. The symbols {} serve as placeholders for variable inputs, which can represent both events and times.

Category	Template
Comparison	Compare the frequency of {} and {}.
Computation	If {} happens {}, how many times will it occur in {} years? {} appears {}. If it was last seen in {}, when will it next appear? {} appears {}. If it took place in {}, when did it previously occur?
Application	If a person's job contract has a renewal every {} years, and they started working in {} and renewed it {} times without gaps, until what year is their current contract valid? A solar eclipse happens at least {} times a year. If the first one in {} is in {}, in which month can we expect the next one? If a plant blooms every {} days and it last bloomed on January 1, on what date will it next bloom? A comet passes Earth every {} years. If its last appearance was in {}, when will it next appear? If a magazine publishes a special edition every {} months and the last one was in January, in which month will the next special edition be? A company holds a general meeting every {} quarters. If the last one was in Q1 of a year, which quarter will the next meeting be? A species of cicada emerges every {} years. If they last emerged in {}, when will they next emerge? If a leap year occurs every 4 years and the last one was in {}, when is the next leap year? A festival is celebrated every {} years. If it was last celebrated in {}, when will it next be celebrated? If a building undergoes maintenance every {} months and the last maintenance was in January, which month will the next maintenance be?

weave of cause-and-effect relationships shaped by temporal elements. To this end, we prioritized the inclusion of diverse temporal factors in our dataset, encompassing aspects like seasons, specific times on clocks, special occasions, as well as both long-term and short-term causes and impacts. By meticulously crafting problems with these considerations, we have crafted a rich collection that illuminates the nuanced interplay between time and causality.

Storytelling To identify stories with temporal nuances from the ROCStories and SCT datasets, we employed a keyword-based filtering approach. The choice of our keyword set, as shown in Table 8, was shaped by the distinctive nature of the datasets and the contexts they encompass. In ROCStories, for instance, storytelling often employs varied and colloquial temporal expressions, necessitating a specific focus in our keyword selection. Stories containing at least one term from the list were considered to have temporal aspects and were subsequently selected for further processing.

A.4 Example Questions

For additional examples of various tasks, refer to the following figures: Figure 3 for the ordering task, Figure 4 for the frequency task, Figure 5 for the duration task, Figure 6 for the typical time task, Figure 7 for the ambiguity resolution task, and Figure 8 for the arithmetic task. The advanced temporal understanding group, comprising relation, temporal NLI, causality, and storytelling tasks, which have relatively fewer subtasks, are collectively presented in Figure 9. The correct choices are bolded.

A.5 Comparison of Source vs. Curated Datasets

We provide several representative examples sourced from existing datasets, allowing for a comparison between the original sources and our curated datasets. Specifically, Table 9 and Table 10 demonstrate the transformation of original Yes/No binary questions from the MCTACO dataset into our frequency and ordering tasks in MCQ for-

Commonsense	**Q:** Mike started his first business, a bakery. Then Mike launched his online cake delivery service. -True/False? A. Undetermined B. False **C. True**
Commonsense	**Q:** Arrange the following events in chronological order: (1) Sarah spoke her first words. (2) Sarah learned to ride a bicycle. (3) Sarah took her first steps. (4) Sarah started kindergarten. **A. (3), (1), (4), (2)** B. (1), (3), (4), (2) C. (4), (3), (2), (1)
Facts	**Q:** Is the following sequence of events in the correct chronological order? (1) The Periplus of the Erythrean Sea, a Graeco-Roman manuscript is written. It describes an established Indian Ocean Trade route (2) War of the Polish Succession. (3) East India Company starts operations in Bengal to smuggle opium into China. (4) Viking state in Russia founded under Rurik, first at Novgorod, then Kiev. (5) China conquers the Kingdom of Tungning and annexes Taiwan. A. True **B. False** C. Undetermined

Figure 3: Example questions on the temporal ordering task.

Facts	**Q:** How often does ICC Cricket World Cup occur? **A. Every 4 years** B. Every 5 years C. Once a year
Comparison	**Q:** Compare the frequency of 'Veterans Day' and 'Solar eclipse'. A. Veterans Day is more frequent **B. Solar eclipse is more frequent** C. Both events are equally frequent
Computation	**Q:** If 'Annual invisibility cloak fashion show' happens yearly, how many times will it occur in 100 years? **A. It will occur 100 times** B. It will occur 103 times C. It will occur 99 times
Application	**Q:** A species of cicada emerges every 22 years. If they last emerged in 1914, when will they next emerge? **A. 1936** B. 1939 C. 1934
Reading Comprehension	Q: There have been six instances as of 2009 in which the exemption process was initiated. Of these six, one was granted, one was partially granted, one was denied and three were withdrawn. Donald Baur, in The Endangered Species Act: law, policy, and perspectives, concluded," ... the exemption provision is basically a nonfactor in the administration of the ESA. A major reason, of course, is that so few consultations result in jeopardy opinions, and those that do almost always result in the identification of reasonable and prudent alternatives to avoid jeopardy." How many times has the exemption process been used, as of 2009? **A. Six** B. Eight C. Five

Figure 4: Example questions on the frequency task.

mats, respectively. Meanwhile, Table 11 shows the transformation of original short-answer questions from the SQuAD dataset into our duration task. Our benchmark combines the strengths of existing benchmarks with extensive manual effort, including the addition of distracting or confusing options, the filtering out of irrelevant questions for quality control, and the reformulation of problems, thereby setting a new standard for assessing temporal reasoning in LLMs.

A.6 Subtask Distributions

As shown in Table 1, if *Problem Types* count exceeds 1, then we consider it a task involving multiple subtasks. Figure 10 illustrates the distribution of subtasks for each temporal reasoning task. In the case of causality, two problem types are evenly distributed, each accounting for 50%.

B Prompts

We utilize both SP and CoT in our experiments with LLMs. For SP, questions are presented directly without the need for additional steps in the prompt. Consider the following example from the storytelling dataset:

"When I was a boy, my parents used to take my brother and me to the park. We would play, have lunch, and just walk around. One day, when all the picnic benches at the park were occupied, we had one. Two police officers approached and asked if they could join us. Which of the two endings is the most plausible correct ending to the story?

(A) They were there to take my brother and me to the police station.
(B) They let us operate the police car lights and siren."

For zero-shot SP, the model is simply prompted with the question: "Given the story 'When I was

Facts	**Q:** How long did California Gold Rush last?		
	A. 3 years	**B. 7 years**	C. 10 years
Computation	**Q:** For a conference, planning lasts for 9 months. If preparation is double that duration minus 15% of planning and keynote is the sum of planning and preparation divided by 2, what's the entire duration?		
	A. 40.5 months	B. 44.5 months	**C. 38.5 months**
Direct Comparison	**Q:** Which event lasted the longest: World War II, U.S. Woman Suffrage Movement, or British Raj in India?		
	A. World War II	B. U.S. Woman Suffrage Movement	**C. British Raj in India**
Multi-Step Comparison	**Q:** Art Exhibition has a duration of 2 months. Wine Tasting lasts as long as Art Exhibition and Tech Conference combined, where Tech Conference is triple of Art Exhibition. Which event has the shortest duration?		
	A. Tech Conference	**B. Art Exhibition**	C. Wine Tasting
Commonsense	**Q:** Lennon accuses his father of leaving him again, and then leaves, after telling his father that he won't live with him anymore . How long does this conversation between Lennon and his father take?		
	A. 10 minutes	B. 10 months	C. 6 weeks
Reading Comprehension	Q: In Canada, "college" generally refers to a two-year, non-degree-granting institution, while "university" connotes a four-year, degree-granting institution. Universities may be sub-classified (as in the Macleans rankings) into large research universities with many PhD granting programs and medical schools (for example, McGill University); "comprehensive" universities that have some PhDs but aren't geared toward research (such as Waterloo); and smaller, primarily undergraduate universities (such as St. Francis Xavier). How many years does a degree-granting university in Canada spend teaching students?		
	A. Four	B. Five	C. Six

Figure 5: Example questions on the duration task.

Facts	**Q:** In what year(s) did "The Phoenician alphabet is created" occur?		
	A. 1006 BCE	B. 1096 BCE	**C. 1050 BCE**
Commonsense	**Q:** Then, he pretended he was his father and pretended that he was driving the tractor. What time did he pretend to drive the tractor?		
	A. 1:00 PM	B. at midnight	C. 1:00 AM
Reading Comprehension	Q: In 1978 Aboriginal writer Kevin Gilbert received the National Book Council award for his book Living Black: Blacks Talk to Kevin Gilbert, a collection of Aboriginal people's stories, and in 1998 was awarded (but refused to accept) the Human Rights Award for Literature for Inside Black Australia, a poetry anthology and exhibition of Aboriginal photography. In contrast to previous definitions based solely on the degree of Aboriginal ancestry, in 1990 the Government changed the legal definition of Aboriginal to include any: What year was Gilbert awarded for his efforts?		
	A. 1960	**B. 1978**	C. 2017

Figure 6: Example questions on the typical time task.

a boy ... they could join us.' Which of the two endings is the most plausible correct ending to the story? (A) They were... or (B) They let us... The answer (A or B) is: { }." For few-shot SP, exemplar answers (A or B) are provided alongside the questions.

The overall SP procedure across all tasks can be summarized in three steps: (1) *Context Provision (if any):* Provide any necessary background information or context that may aid the model in understanding the scenario presented in the question. (2) *Direct Questioning:* Pose the question directly to the model without any intermediary steps or additional guidance. (3) *Answer Solicitation:* Request the model to choose and provide the most appropriate answer based on the information given.

In contrast, for CoT, zero-shot learning takes inspiration from (Kojima et al., 2022) by instructing the model to "Answer the question step by step". For few-shot CoT, we manually craft the step-by-step process for 5-shot exemplars in the development set. The procedure to approach this problem is as follows:

(1) *Read the Story Carefully:* Understand the main theme, setting, and characters introduced in the story. The dominant theme appears to be a nostalgic recollection of a family day out at a park.

(2) *Identify Key Elements from the Story:* The

Short-term Shift	**Q:** Your train's regular schedule is 10:53 AM. However, today it's running 58 minutes behind. When will it depart?		
	A. 11:51 AM	B. 11:23 AM	C. 11:38 AM
Mid-term Shift	**Q:** A marathon was supposed to happen this coming Wednesday, but got shifted three days earlier. When will it occur?		
	A: Thursday	**B. Sunday**	C. Tuesday
Long-term Shift	**Q:** The dynasty which fell in 1830 had risen to power roughly 90 years earlier. When was its establishment?		
	A: 1742	B. 1745	**C. 1740**
Facts	**Q:** If the date is 9/7/1872 in the Julian, what is the date in the Gregorian?		
	A. 6/6/1871	**B. 9/19/1872**	C. 5/26/1872

Figure 7: Example questions on the ambiguity resolution task.

12-hour Adjustment	**Q:** What is 08:24 AM - 07:42?			
	A. 12:42 AM	B. 1:56 AM	C. 10:31 PM	D. 11:34 PM
Year Shift	**Q:** Which year comes 11 years after 1718?			
	A. 1731	B. 1707	C. 1764	**D. 1729**
Month Shift	**Q:** Which month comes 2 months after December?			
	A. June	**B. February**	C. January	D. September
Date Computation	**Q:** What will be the time 16 years and 8 months after August 1412?			
	A. June 1430	B. May 1430	**C. April 1429**	D. June 1431
Week Identification	**Q:** In which week of year 2007 does the date 10-12-2007 occur?			
	A. Week 41	B. Week 28	C. Week 5	D. Week 10
Time Zone Conversion	**Q:** If it's 12 PM on May 4, 1904 in Asia/Kolkata, what's the date and time in US/Eastern?			
	A. 6 AM on May 4, 1904	B. 12 PM on May 4, 1904	**C. 1 AM on May 4, 1904**	D. 11 AM on May 4, 1904
Time Computation	Q: Subtract 1 minute 32 seconds from 1 hour 22 minutes.			
	A. 77 minutes 25 seconds	B. 90 minutes 38 seconds	C. 70 minutes 18 seconds	**D. 80 minutes 28 seconds**
Application	Q: If a girl is advised to take medicine every 139 minutes, how many times will she take the medicine in a day?			
	A. 12	B. 11	C. 8	**D. 10**

Figure 8: Example questions on the arithmetic task.

protagonist recalls a childhood memory. The primary setting is a park. The mood is both casual and reminiscent. Despite the park being crowded, they have a picnic spot. Subsequently, two police officers approach the family.

(3) *Evaluate Each Proposed Ending:* For the first ending, a sudden and unexpected twist is introduced that deviates from the story's initial light-hearted narrative. This ending lacks context about why they'd be taken to the police station. The second ending maintains the story's casual and friendly tone, presenting a scenario where the police officers engage positively with the family.

(4) *Comparison of the Two Endings:* Both endings involve the police officers, but the first one introduces a jarring twist without adequate prior context. The second ending aligns more consistently with the story's overarching mood and theme.

(5) *Conclusion:* Given the story's tone, setting, and characters, the second ending appears more plausible and contextually appropriate.

After defining the step-by-step procedure, we employ it to steer the model's thought process. This structured methodology better prepares the model to reason through the question and formulate a well-considered answer, thereby providing a distinct advantage over the SP method. We structure our prompt as follows: "Begin by reading the story carefully, ensuring you fully understand its main

Temporal Relation	**Q:** It added that the Ministry of Economic Affairs and Finance was assigned to draw up practical procedure for the ceding, while the Ministry of Welfare and Social Security would be responsible for identifying the beneficiaries in two months. What is the relationship between the event 'added' and the event 'ceding'? A. IS_INCLUDED B. SIMULTANEOUS **C. AFTER**
Temporal NLI	**Q:** Premise: Two guys playing football on a campus green. Hypothesis: They are practicing before the big game tomorrow A. Entailment **B. Neutral** C. Contradiction
Temporal Causality (Effect)	**Q:** The seasons changed from summer to autumn. What's the more plausible RESULT? A. People evacuated their homes. **B. Leaves fell from the trees.**
Temporal Storytelling	**Q:** There is a huge clock in my living room. I turned the clock back one hour for daylight savings. My wife also turned the clock back one hour for daylight savings. Our 2 kids each turned the clock back one hour for daylight savings. Which of the two endings is the most plausible correct ending to the story? **A. Then we wondered why it got so dark so early.** B. The kids were not happy.

Figure 9: Example questions on advanced temporal reasoning tasks, including relation, temporal NLI, causality, and storytelling.

theme, setting, and the characters. {Immediate analysis}. Subsequently, identify the key elements of the story. {Immediate analysis}. Assess each proposed ending within the context of the narrative. {Immediate analysis}. Compare the two endings, highlighting any thematic or tonal discrepancies. {Immediate analysis}. Conclude by determining which ending appears more plausible, offering a rationale for this selection {Immediate analysis}."

In general, the CoT procedure across all temporal reasoning tasks is as follows: (1) *Understanding Context:* Begin by reading the provided data, statement, or story attentively. Understand the overarching theme, objectives, or the problem's primary ask. (2) *Key Elements Extraction:* Identify and highlight crucial elements, specifics, or characters. This could mean different things for different tasks - key events in a story, terms in a mathematical problem, or clauses in a statement. (3) *Evaluation:* Assess the core objective of the problem in its context. This could be understanding the chronology for ordering, assessing frequency, gauging durations, or even understanding the logical or causal flow in more complex problems. (4) *Analysis and Comparison:* If there are multiple options or scenarios presented, conduct a deep analysis. Compare, contrast, and evaluate based on the preceding steps. (5) *Reasoned Conclusion:* Conclude with a structured answer or resolution to the problem, ensuring that the decision-making process aligns with the evidence or data presented. In practice, the procedure varies for each task to account for the diverse nature of temporal reasoning tasks.

C Human Assessment

In this section, we provide additional details on human participation in our benchmark, including the selection process for experts, verification of their capabilities, and a performance comparison with non-specialists.

Selection of Expert Annotators Our selection criteria for expert annotators emphasized a balanced proficiency in both temporal reasoning and quantitative analysis. We included professionals with advanced degrees (M.S. or Ph.D.) in disciplines that offer distinct perspectives on our tasks. This included cognitive science and psychology for qualitative understanding of human temporal cognition, crucial for interpreting more subjective aspects of the tasks. We also involved experts in statistics, mathematics, and computer science to address the quantitative complexities inherent in many of our benchmark tasks. This diverse expertise ensured a comprehensive evaluation of the problems within the TRAM dataset from both qualitative and quantitative angles.

Expertise Verification Process To ensure the high caliber of our expert panel, we implemented a robust screening process. This involved a thorough validation of their educational qualifications and a careful review of their professional and research experience, particularly focusing on time perception and quantitative problem-solving. Additionally, we administered a preliminary assessment composed of one random problem from each subtask, totaling 37 problems. The passing criterion for this assessment was set at an average accuracy rate of more than 92%, allowing a maximum of three incorrect responses. This stringent benchmark was estab-

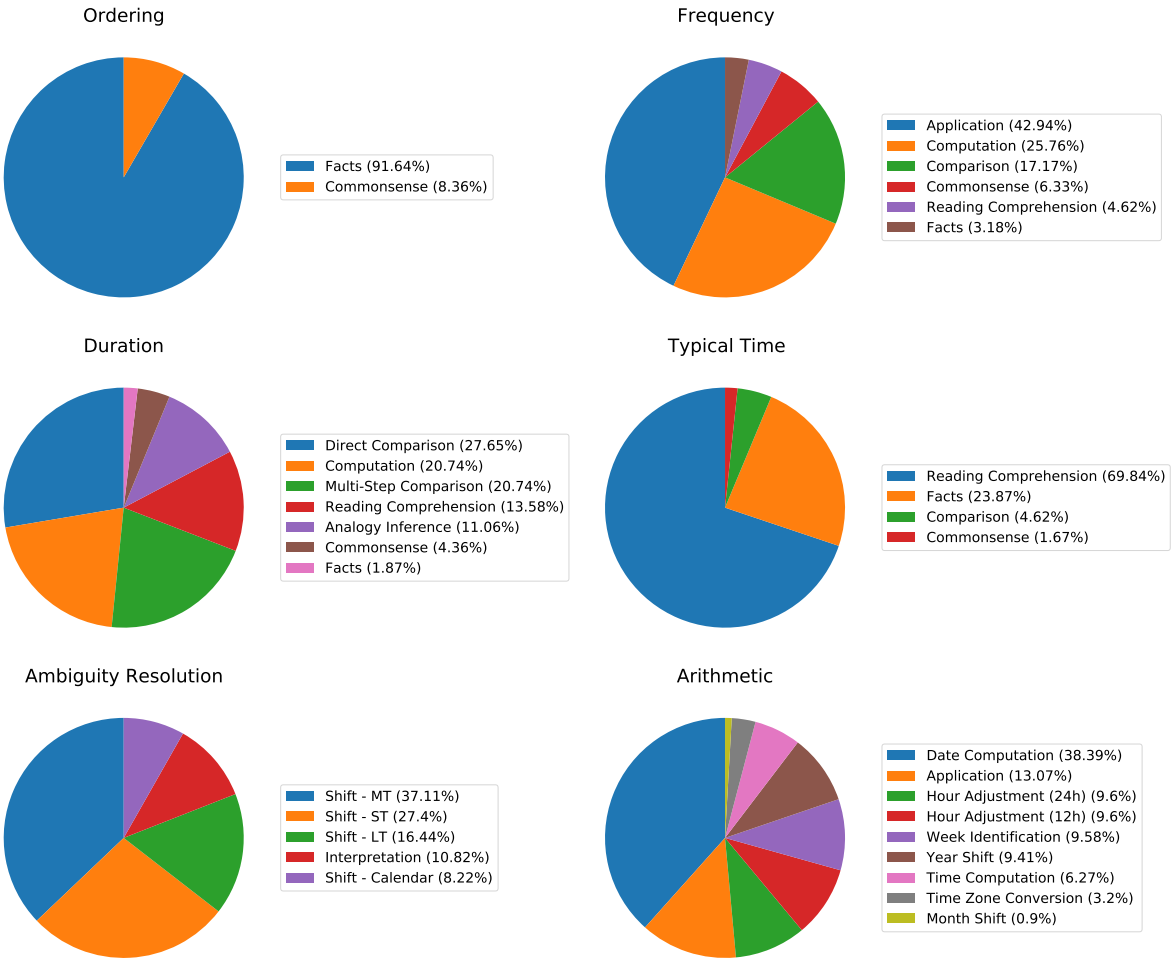

Figure 10: Distribution of subtasks for each distinct temporal reasoning task.

lished to guarantee the experts' capability in accurately addressing the complex problems in TRAM.

Comparison with Unspecialized Individuals In addition to expert assessments, we conducted a comparative analysis with human non-specialists to provide a broader perspective on human performance. These non-specialists, sourced from Amazon Mechanical Turk, consisted of individuals without specialized training in temporal reasoning or related fields. They were tasked with responding to the same set of 1,900 questions as the experts. This group achieved an overall accuracy rate of 63.5% across all tasks. This comparison not only underlines the proficiency of our expert panel but also offers insights into the general human ability to tackle TeR challenges, providing a baseline for non-expert performance in this area.

D Error Types

In this section, we delve into each specific error that LLMs commonly encounter in temporal reasoning tasks, as illustrated in Figure 2.

Foundational Temporal Understanding Tasks In foundational temporal understanding, LLMs encounter several distinct challenges. Firstly, *Assumption Bias* is evident when models over-rely on patterns from their training, often neglecting cultural or individual variations. Next, *Temporal Descriptor Misinterpretation* occurs when models misinterpret terms, such as perceiving "often" as a daily event instead of a possible weekly occurrence. *Event Ambiguity* presents another challenge, where events can be described in ways that allow for multiple interpretations, requiring models to select the most suitable one based on context. Lastly, *Contextual Misjudgment* is when models either miss or misinterpret explicit temporal clues, leading to errors in their reasoning.

Temporal Interpretation and Computation Tasks In computational and interpretable temporal reasoning, LLMs encounter various challenges. Firstly, *Calculation Slips* highlight instances where models often make calculation mistakes like inappropriate handling of time carries. Following this, *Descriptor Confusion* arises when models misalign qualitative terms such as "seldom" or "frequently" with their quantitative meanings. *Resolution Misalignment* represents the struggle models face with vague time references, such as deciphering the exact duration from terms like "in a while". Lastly, *Temporal Notation Misinterpretation* occurs when models confuse time formats, for example, mixing up AM with PM or not differentiating between 24-hour and 12-hour representations.

Advanced Temporal and Conceptual Understanding Tasks In advanced temporal reasoning tasks, LLMs frequently encounter certain pitfalls. Among the most prevalent is *Implicit Oversights*, where models overlook subtle but crucial temporal indications, resulting in inaccurate conclusions. Also, they may face *Relation Simplification*, wherein complex temporal interplays between events are either misunderstood or overly simplified. LLMs might also fall into the trap of *Narrative Bias*, where they overly depend on familiar story patterns, prioritizing recognized sequences over fresh interpretations. Lastly, *Overgeneralization* becomes evident when models incorrectly apply broad temporal conventions to specific situations, leading to misunderstandings when scenarios diverge from the norm.

Table 4: Major templates for constructing the duration subtasks:multi-step comparison, analogy inference, computation, direct comparison, and facts. The symbols {} serve as placeholders for variable inputs, which can represent both events and times.

Type	**Template**
Multi-Step Comparison	{} goes on for {}. {} is a third of {}, and {} is as long as {} and {} combined. Which event lasts the longest? Between {} that lasts {}, {} that is four times longer, and {} that's half the total duration of the two, which is the shortest? {} spans {}. {} is double that, but {} is only a third of {}. Which has the most extended duration? If {} lasts {}, {} is twice as long, and {} is half of {}, which event has the medium duration? {} lasts for {}. {} is half of {}'s duration, and {} is triple the combined length of both {} and {}. Which event has the shortest duration?
Analogy Inference	During {}, the audience had a chance to enjoy a long opera, while {} showcased just one act, and {} played only an overture. Which event was the shortest? People could indulge in a seven-course meal during {}, a three-course meal in {}, but only an appetizer during {}. Which event was in the middle in terms of duration? {} felt like watching an epic trilogy, {} was more of a feature-length film, and {} was just a brief trailer. Which event was probably the longest? Participants at {} went through an entire yoga session, {} allowed for a short warm-up, while {} was only a few stretches. Which event was the shortest? During {}, attendees could finish a whole board game, in {} they played just a few rounds, and in {} merely set up the pieces. Which event was likely the longest?
Computation	The duration of {} is {}. If {} is a quarter shorter than {} and {} is four times the length of {} for {}, how long do all the activities last? For {}, {} takes {}. If {} is twice that duration minus 10% of {}, and {} is half of the sum of {} and {}, how long is the whole event? The total duration of {} is four times the time of {} which is {}. If {} is half of {} minus 5% of {} and {} is twice {} plus 15% of {}, how long do the {} and {} together take? In {}, {} is twice as long as {} which takes {}. If {} is the difference between {} and {}, how long in total? For {}, {} lasts for {}. If {} is double that duration minus 15% of {} and {} is the sum of {} and {} divided by 2, what's the entire duration?
Direct Comparison	Which event lasted longer: {} or {}? Which event lasted the longest: {}, {}, or {}?
Facts	How long did {} last?

Table 5: Major templates used for constructing the ambiguity resolution dataset. The symbols {} serve as placeholders for variable inputs, which can represent both events and times.

Type	**Template**
Short-term Shift	Your plane is supposed to depart at {}. If it's preponed by {}, when is the revised departure? The meal was promised to be on the table at {}. If it's going to be {} postponed, when can you expect to dine? You have an exciting date at {}. If you're lagging by {}, when will you probably meet your date?
Mid-term Shift	The match initially set for {} has now been advanced by {}. Which day is it on now? Your usual spa day on {} of every week has been postponed {}. When will it be next week? The weekly town hall usually on {} is delayed by {}. When will it happen? The town carnival usually during the {} week of {} will now be {}. About which date is it now? The music fest during the {} week of {} will be held {}. Around which date will it likely be? The product launch in the {} week of {} has been shifted {}. Around when will it likely be?
Long-term Shift	The star, predicted to explode in {}, has its explosion postponed by {} years. When is the new prediction for the explosion? The dynasty which fell in {} had risen to power roughly {} years earlier. When was its establishment?
Calendar Shift	If the date is {}/{}/{} in the {}, what is the date in the {}?
Interpretation	You receive a memo with the timestamp {}. When should you be prepared? A festival is being organized {}. When would that be? A note suggests meeting {}. When is this suggesting?

Table 6: Major templates used for constructing the arithmetic dataset. The symbols {} serve as placeholders for variable inputs, which are randomly generated by programs.

Category	**Template**
24-hour Adjustment	What is {}:{} +/- {}:{}?
12-hour Adjustment	What is {}:{} AM/PM +/- {}:{}?
Year Shift	Which year comes {} years after {}? Which year was {} years before {}?
Month Shift	Which month comes {} months after {}? Which month was {} months before {}?
Date Computation	What will be the time {} years and {} months after month {}? If you add/subtract {} days to the date {}, what will be the new date? If you add/subtract {} months and {} days to the date {}, what will be the new date? If you add/subtract {} weeks and {} days to the date {}, what will be the new date?
Week Identification	In which week of year {} does the date {} occur?
Time Zone Conversion	If it's {} in the source zone, what's the date and time in target zone?
Time Computation	Convert {} days into minutes. Convert {} minutes into hours. Convert {} days into hours. Convert {} seconds into hours. Add {} minutes {} seconds and {} minutes {} seconds. Subtract {} minutes {} seconds from {} hours {} minutes.
Application	If a person takes a leave of {} days starting from start_date, on which day may the leave end? If a person was {} years {} month(s) old when he joined school and now he is {} years {} month(s) old, for how long has he been in school? If a person is advised to take medicine every {} minutes, how many times will she take the medicine in a day? If a person starts doing homework at {} and finishes at {} PM, how many hours did he spend on homework? If a flight takes off at {} and the duration of the flight is {} hours, at what time will it land? If a person walks at a speed of {} km/hr and after every km, she takes a rest for {} minutes, how many minutes will it take her to cover {} km? How long will it take to travel a distance of {} kilometers in minutes?

Table 7: Keywords used for filtering SNLI and MNLI datasets that contain temporal aspects.

Category	**Keywords**
Explicit References	today, tomorrow, yesterday, now, soon, later, before, after, day, week, month, year, hour, minute, second, morning, evening, night, noon, midnight, anniversary
Days of the Week	Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday
Months	January, February, March, April, May, June, July, August, September, October, November, December
Seasons	spring, summer, fall, autumn, winter
Periods and Eras	decade, century, millennium, epoch, era
General Terms	annual, biannual, quarterly, hourly, daily, weekly, quarter, monthly, fortnight, biweekly, bimonthly, semester, trimester
Relative Terms	past, future, current, upcoming, recent, lately, ago, in advance, later, previous, next, moment, time, when, while, duration, period, early, earlier
Implicit Temporal Actions	wait, postpone, delay, reschedule, expire, due, schedule, begin, start, end, finish, commence, conclude, last, extend
Temporal Transitions and Connectors	until, by the time, as soon as, whenever, since, during, whilst
Other Temporal Entities	sunset, sunrise, dusk, dawn, midday, eve, annually, eventually, seldom, often, always, never, sometimes, usually, frequently, occasionally, rarely, just, once, still

Table 8: Keywords used for filtering ROCStories and SCT datasets that contain temporal aspects.

Category	Keywords
Time References	before, after, recently, now, then, earlier, later, today, tonight, yesterday, tomorrow
Temporal Intervals	soon, nowadays, currently, presently, eventually, ultimately, suddenly, immediately, momentarily, previously, formerly
Recurring Time Periods	periodically, seasonally, daily, weekly, monthly, annually, biennially
Fixed Time Periods	century, decade, millennium, year, minute, hour, day, week, month
Parts of the Day	morning, noon, evening, night
Duration & Frequency	duration, instant, temporarily, intermittently, frequently, always, never, sometimes, often, rarely, usually
Starting Actions	begin/begins/began, start/starts/started, commence/commences/commenced
Ending Actions	end/ends/ended, finish/finishes/finished, cease/ceases/ceased, expire/expires/expired, elapse/elapses/elapsed
Continuing & Delaying	last/lasts/lasted, continue/continues/continued, resume/resumes/resumed, linger/lingers/lingered, postpone/postpones/postponed, procrastinate/procrastinates/procrastinated

Table 9: Comparison of source (MCTACO) and curated question in TRAM for the Frequency task.

Source Dataset (MCTACO)	Curated Dataset (TRAM)
Question: Allan crouched over his desk once more, pen in hand and mind blank. How often does Allan crouch over his desk?	**Question:** Allan crouched over his desk once more, pen in hand and mind blank. How often does Allan crouch over his desk?
Options/Answers: • Once a second - No • Once two years ago - No • Every day - Yes • Several times per second - No • Daily - Yes	**Options:** • (A) Every day • (B) Several times per second • (C) Once a second **Answer:** • (A) Every day
Commentary: Binary Yes/No format, simple frequency assessment.	**Commentary:** Transition to an MCQ format enriches the question's complexity by offering closely related alternatives.

Table 10: Comparison of source (MCTACO) and curated question in TRAM for the Ordering task.

Source Dataset (MCTAC0)	**Curated Dataset (TRAM)**
Question: Church is brought back to life, but is an evil shell of himself. What did Church do next?	**Question:** Church is brought back to life, but is an evil shell of himself. What did Church do next? Is "took a nap" possible?
Options/Answers: • "took a nap" - No	**Options:** • (A) Undetermined • (B) TRUE • (C) FALSE **Answer:** • (B) Two months
Commentary: Binary Yes/No format, simple ordering assessment.	**Commentary:** Transition to an MCQ format introduces additional ambiguity and uncertainty into the question.

Table 11: Comparison of source (SQuAD) and curated question in TRAM for the Duration task.

Source Dataset (SQuAD)	**Curated Dataset (TRAM)**
Question: It was not until January 1518 that friends of Luther translated the 95 Theses ... Within two weeks, copies of the theses had spread throughout Germany; within two months, they had spread throughout Europe. How long did it take for the Theses to spread through Europe?	**Question:** It was not until January 1518 that friends of Luther translated the 95 Theses ... Within two weeks, copies of the theses had spread throughout Germany; within two months, they had spread throughout Europe. How long did it take for the Theses to spread through Europe?
Options/Answers: • Short answer: Two months	**Options:** • (A) 45 days • (B) Two months • (C) 2 days **Answer:** • (B) Two months
Commentary: Short-answer format, simple duration assessment.	**Commentary:** Transition to an MCQ format introduces additional numerical ambiguity in problems involving multiple numbers.