Beyond Timestamps: Bridging Forward and Backward Reasoning in Temporal Numerical and Relational Understanding

Xinying Qian[1] **Ying Zhang**[1*] **Xuhui Sui**[1]
Yu Zhao[1] **Baohang Zhou**[2] **Jeff Z. Pan**[3]
[1] College of Computer Science, VCIP, DISSec Center, Nankai University, China
[2] School of Software, Tiangong University, China
[3] School of Informatics, The University of Edinburgh, UK
{qianxinying,suixuhui,zhaoyu}@dbis.nankai.edu.cn,
yingzhang@nankai.edu.cn, zhoubaohang@tiangong.edu.cn, j.z.pan@ed.ac.uk

Abstract

Temporal reasoning remains a critical challenge for large language models (LLMs), particularly when it requires encompassing relational dependencies and numerical constraints. Yet, existing benchmarks largely overlook the joint consideration of these two dimensions and primarily rely on single-task evaluation paradigms, making it difficult to assess whether correct answers reflect grounded reasoning or arise from superficial statistical recall. To address these gaps, we introduce **TNR**, a benchmark designed to evaluate both **T**emporal **N**umerical and **R**elational reasoning. We propose a bi-directional evaluation framework consisting of forward generation via Question Answering (QA) and backward verification via Fact Verification (FV). By measuring the alignment between QA and FV, we introduce a Consistency Rate to quantify the robustness of reasoning across these two directions. Experiments on a range of LLMs reveal notable discrepancies between QA and FV performance, particularly in numerical and interval-based tasks. Moreover, our bi-directional error analysis demonstrates that these inconsistencies often stem from heuristic shortcuts and statistical co-occurrences rather than grounded logical deduction, flaws that are frequently masked in standard single-task evaluations.

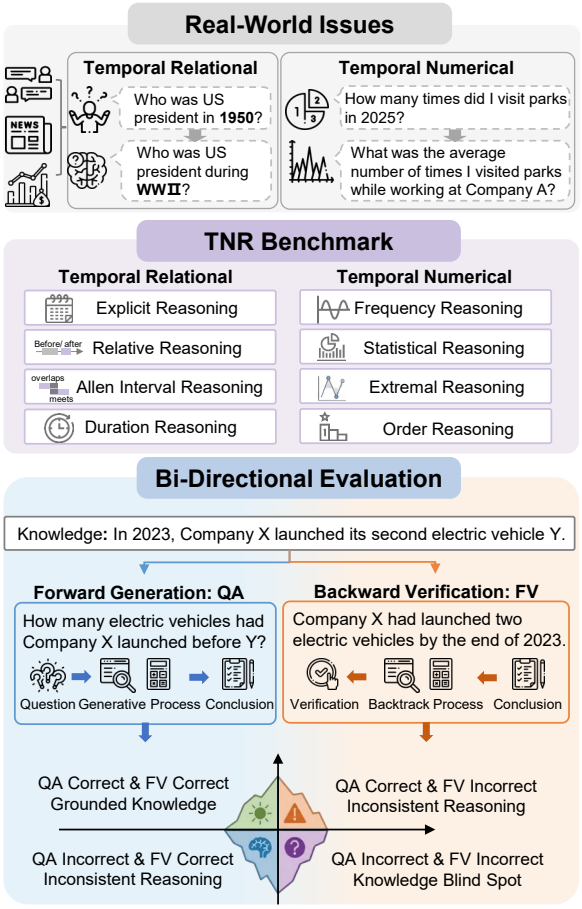

Figure 1: An overview of the TNR benchmark. We build the TNR benchmark from real-world challenges and evaluate LLMs with a bi-directional framework.

1 Introduction

Temporal reasoning (Piryani et al., 2025) serves as a critical component in Natural Language Understanding. Numerous studies have explored approaches to process (Qian et al., 2024) and interpret temporal information (Wang et al., 2023a) effectively. In recent years, Large Language Models (LLMs) have exhibited remarkable capabilities not only in natural language understanding but also across various complex downstream tasks, including mathematical problem solving (Pei et al., 2025) and code generation (Jiang et al., 2025). Nevertheless, some studies (Chu et al., 2024; Uddin et al., 2025; Wei et al., 2025) indicate that LLMs still struggle with tasks requiring temporal understanding, revealing a gap between general reasoning capabilities and temporal reasoning ability.

Temporal reasoning in real-world scenarios poses several challenges. (1) Temporal Relational

* Corresponding author.
Our data are available at: https://github.com/qianxinying/TNR.

Reasoning. Human reasoning often goes beyond absolute timestamps, relying on more complex temporal constraints, such as relative temporal expressions and event-based references. This requires models to align events and reason over relative temporal relations. (2) Temporal Numerical Reasoning. In domains such as news, finance, and geopolitical analysis, many reasoning tasks require performing numerical operations under complex temporal constraints. For example, answering "How many times did the U.S. President visit Japan in 2025?" demands accurate temporal filtering alongside numerical computation, posing a greater challenge than temporal or numerical reasoning alone.

Although existing benchmarks have made substantial progress, two key gaps remain: (1) Temporal and numerical reasoning are largely addressed in isolation. Temporal benchmarks (Luo et al., 2025) often limit numerical tasks to simple counts rather than complex operations, whereas numerical benchmarks (Strong and Vlachos, 2025) typically rely on explicit timestamps (e.g., "in 1950") instead of requiring temporal reasoning (e.g., "during World War II"). (2) Reliance on single-task evaluations (e.g., QA) can mask flawed reasoning behind correct answers, as models often exploit shortcut learning (Yuan et al., 2024) rather than grounded reasoning, thereby failing to adequately assess temporal consistency and numerical validity.

To address these gaps, we introduce **TNR**, a benchmark designed to evaluate both **T**emporal **N**umerical and **R**elational reasoning through a novel bi-directional evaluation framework. Temporal Numerical Reasoning integrates computation with temporal understanding, formulated via four primitives: Frequency, Statistical, Extremal, and Order reasoning. Complementing this, Temporal Relational Reasoning focuses on the topological structure of timelines, characterized by four primitives: Explicit, Relative, Duration, and Interval reasoning. These primitives serve as the core elements of temporal analysis. By combining these foundational components, we can form more complex tasks. To enable a more rigorous evaluation of LLMs, we propose a bi-directional framework that jointly examines forward generation and backward verification. We consider Question Answering (QA) as forward reasoning, where conclusions are generated from given premises, while Fact Verification (FV) is framed as backward reasoning, requiring models to trace a claim back to its supporting evidence. As illustrated in Figure 1, given the knowledge that "In 2023, Company X launched its second electric vehicle Y," the model is required to both deduce the answer to a relevant question and verify a derived claim. By analyzing performance alignment across these two tasks, we characterize model behavior into four outcomes and introduce the Consistency Rate (CR) to quantify the robustness of reasoning. Our main contributions are summarized as follows:

- We introduce TNR, a benchmark designed to evaluate both temporal relational and numerical reasoning within a unified framework.
- We propose a bi-directional evaluation framework that systematically assesses reasoning robustness through forward generation and backward verification.
- We conduct extensive experiments on state-of-the-art LLMs, and bi-directional evaluation reveals that many inconsistencies arise from reliance on statistical priors rather than grounded reasoning.

2 Related Work

2.1 Temporal Reasoning Benchmarks

Several benchmarks have been developed to evaluate the temporal capabilities of LLMs. TimeBench (Chu et al., 2024) categorizes tasks into symbolic, commonsense, and event-based reasoning. TimeQA (Chen et al., 2021) aligns Wikipedia knowledge bases with corresponding articles. TempReason (Tan et al., 2023) focuses on event-time and event-event relations, while MenatQA (Wei et al., 2023) adds complexity through counterfactual questions. TempTabQA (Gupta et al., 2023) benchmarks time-sensitive question answering on tabular data, and UnSeenTimeQA (Uddin et al., 2025) provides a contamination-free evaluation for events. The Time (Wei et al., 2025) benchmark evaluates temporal reasoning across three tasks: Wikipedia, news, and dialogue.

Although some studies (Fatemi et al., 2024; Su et al., 2024) propose temporal numerical computation, they mainly focus on general mathematical arithmetic (e.g., calendar calculations and date offsets) and basic temporal commonsense, whereas our work emphasizes reasoning over temporal events, which better reflects real-world requirements. More recently, ETRQA (Luo et al., 2025) introduced a fine-grained taxonomy of event temporal questions. However, despite including numeric

Benchmark	Task	Data Source	Size	Temporal Reasoning	Numerical Reasoning
TempReason (Tan et al., 2023)	QA	Wikidata	52.8k	Time-Event Relations	–
MenatQA (Wei et al., 2023)	QA	Wikidata	2.8k	Order, Scope, Counterfactual	–
UnSeenTimeQA (Uddin et al., 2025)	QA	IPC	10.8k	Unseen Event	–
Time (Wei et al., 2025)	QA	Wikidata, LOCOMO, GDELT	38k	General Temporal	–
ToT (Fatemi et al., 2024)	QA	Wikidata	4k	Semantic & Arithmetic	Date Computation
ETRQA (Luo et al., 2025)	QA	Wikidata	160k	Compound	Event Frequency
T-FEVER (Barik et al., 2024)	FV	Wikidata	25.1k	Explicit	–
ChronoClaims (Barik et al., 2025)	FV	Wikidata	47k	Explicit & Implicit	–
QuanTemp (V et al., 2024)	FV	FCS	15.5k	Partial (27% temporal)	✓
TSVER (Strong and Vlachos, 2025)	FV	OWID	287	Explicit	✓
TNR	**QA & FV**	**ICEWS**	**94k**	**Temporal Event & Interval**	**Complex Aggregation**

Table 1: Comparison of our work with existing temporal reasoning and fact verification benchmarks.

question types, ETRQA primarily focuses on simple frequency statistics and lacks coverage of more complex numerical reasoning tasks.

2.2 Temporal Fact Verification Benchmarks

Numerical and temporal expressions are common in fact verification. TabFact (Chen et al., 2020) evaluates crowd-sourced claims against Wikipedia tables, while SciTab (Lu et al., 2023) targets compositional reasoning in scientific contexts. Domain-specific resources such as FinDVer (Zhao et al., 2024) combine textual and tabular data but primarily focus on financial computation. QuanTemp (V et al., 2024) introduces real-world claims involving numerical comparisons and trends, and TSVER (Strong and Vlachos, 2025) leverages time-series evidence for verification. However, these datasets mainly emphasize numerical computation, with temporal reasoning often limited to absolute timestamps and lacking support for relative or event-based temporal references and interval reasoning. Other datasets, such as T-FEVER (Barik et al., 2024) and ChronoClaims (Barik et al., 2025), focus on date-sensitive assertions or chronological ordering, but do not address numerical reasoning. In contrast, our work jointly models temporal numerical and relational reasoning, systematically evaluating both QA and FV. In Table 1, we compare the differences between the baseline benchmarks and TNR. We provide a detailed discussion on self-consistency in Appendix E.

3 TNR: Benchmark

3.1 Task Definition

The TNR benchmark is formulated to evaluate models' temporal numerical and relational reasoning under a bi-directional verification setting, which assesses both forward generation and backward validation. For each underlying knowledge, we construct two complementary instances: a *Question* that requires forward generation and a corresponding *Claim* that enables backward verification.

Forward Question Answering. The Question Answering (QA) task instantiates forward inference. Given contextual evidence $\mathcal{C}$ and a query $\mathcal{Q}$, the model infers an unknown conclusion by maximizing the likelihood of the correct answer y, reflecting the model's ability to perform temporal numerical and relational reasoning in a forward direction.

$$\hat{y}_{QA} = \arg\max_{y} P(y \mid \mathcal{C}, \mathcal{Q}) \tag{1}$$

Backward Fact Verification. The Fact Verification (FV) task instantiates backward validation. Given a hypothesized claim $\tilde{y}$, the model examines the evidence $\mathcal{C}$ to estimate the probability that the claim is supported, capturing the model's ability to verify inferred conclusions against evidence:

$$\hat{v}_{FV} = P(\text{True} \mid \mathcal{C}, \tilde{y}) \tag{2}$$

Bi-directional Evaluation. Inspired by self-consistency (Wang et al., 2023b), QA and FV probe complementary reasoning directions from the same underlying facts. Jointly evaluating both tasks goes beyond answer accuracy, enabling a more robust assessment of reasoning consistency and stability.

3.2 Design Principles

The design of our benchmark is guided by a comprehensive task taxonomy covering temporal numerical and relational reasoning, and a systematic complexity hierarchy based on reasoning depth.

Task Taxonomy. To ensure a comprehensive evaluation, we categorize tasks into two primary domains: Temporal Numerical Reasoning and Temporal Relational Reasoning. These categories capture the fundamental reasoning primitives, which

serve as the atomic units requisite for solving complex temporal analysis tasks.

Temporal Numerical Reasoning assesses numerical operations constrained by temporal contexts: (1) *Frequency reasoning*, counting events within time windows; (2) *Statistical reasoning*, computing statistics (e.g., sums, averages) over sequences; (3) *Extremal reasoning*, identifying extrema (e.g., max/min, earliest/latest); and (4) *Order reasoning*, ranking entities by numerical attributes.

Temporal Relational Reasoning evaluates the model's ability to process temporal structures through four distinct types: (1) *Explicit reasoning*, interpreting timestamps or time spans; (2) *Relative reasoning*, inferring relative temporal references; (3) *Allen interval reasoning*, handling complex logic defined by Allen's interval algebra (e.g., *overlaps*, *during*); and (4) *Duration & time reasoning*, calculating event timing and durations.

Complexity Hierarchy. To assess model robustness across different difficulty levels, we partition samples according to the number of reasoning hops required to derive the correct answer. A reasoning hop is defined as a necessary step of information retrieval or logical deduction that bridges supporting evidence and the final conclusion. Specifically, we define three difficulty levels: (1) *Simple* questions involve a single reasoning hop, requiring direct retrieval of a fact or a one-step logical deduction; (2) *Medium* questions require two to three reasoning hops and compositional reasoning, including cross-period comparisons, aggregation over discontinuous intervals, or multiple temporal constraints; (3) *Hard* questions entail four or more reasoning hops, involving ranking, comparison, or identification of Allen interval relations.

3.3 Data Source

The TNR benchmark is built upon the Integrated Crisis Early Warning System (ICEWS) (Boschee et al., 2015), a repository of political events automatically extracted from a large-scale global news corpus. The dataset provides structured event records comprising an actor, event type, target, and timestamp. With its high event density and rich temporal dynamics, ICEWS offers an ideal foundation for evaluating complex temporal and numerical reasoning.

3.4 Benchmark Construction

Figure 2 illustrates the construction pipeline of the TNR benchmark. To enable complex multi-hop reasoning, we first structure raw data into a temporal knowledge graph to extract temporal numerical and relational knowledge. This knowledge is then used to generate QA and FV pairs by filling predefined question and claim templates. Detailed construction procedures are provided in Appendix B, with templates and examples in Appendix G.

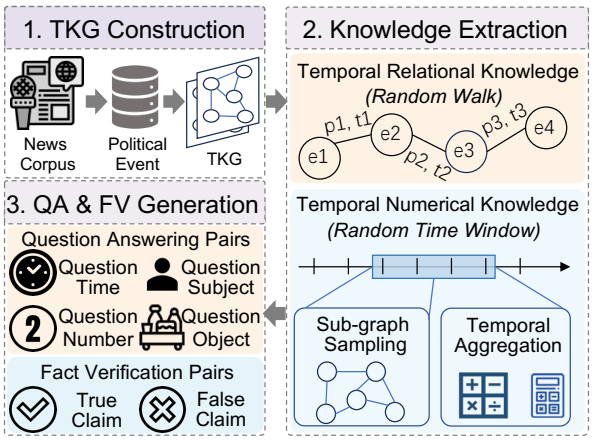

Figure 2: Overview of the TNR Benchmark Pipeline.

Temporal Knowledge Graph Construction. We first organize the data into a Temporal Knowledge Graph (TKG), formalized as $\mathcal{G} = \{(s, p, o, t)\}$, where $s, o \in \mathcal{E}$ denote the subject and object entities, $p \in \mathcal{P}$ represents the predicate, and t indicates the timestamp of the event. This unified representation encodes temporal facts as time-stamped edges, and serves as the foundational structure for subsequent reasoning path extraction.

Multi-hop Temporal Knowledge Extraction. Based on the TKG, we extract temporal knowledge in two forms. *(1) Path-based temporal relational knowledge.* To capture relational dependencies across multiple events, we employ a random walk strategy (Lovász, 1993) over $\mathcal{G}$. Starting from a seed entity, the algorithm traverses time-stamped edges to identify connected event sequences, explicitly forming multi-hop temporal paths such as $s \xrightarrow{p_1,t_1} e_1 \xrightarrow{p_2,t_2} o$. Each path represents an ordered chain of temporally grounded facts that supports multi-step temporal relational reasoning. *(2) Aggregation-based temporal numerical knowledge.* To support numerical reasoning over temporal events, we randomly sample a time window $[t_{start}, t_{end}]$ and extract a subgraph $\mathcal{G}_{sub} = \{(s, p, o, t) \in \mathcal{G} \mid t_{start} \leq t \leq t_{end}\}$, which contains all events occurring within the interval. The time window can be specified by explicit timestamps or by event-based references, condi-

tioned on the question difficulty. Over each extracted subgraph, we apply an aggregation operator $\phi \in \Phi = \{\text{count}, \text{min}, \text{max}, \text{avg}, \text{sum}\}$ to derive numerical values from the temporal knowledge.

Questions and Claims Generation. Each extracted multi-hop knowledge serves as a basic knowledge unit for instance construction. Based on these units, we generate paired evaluation samples for bi-directional reasoning: *(1) Forward Question Answering (QA).* We systematically construct questions by querying specific components of the underlying knowledge, including entities, temporal constraints, or the aggregated numerical results derived from temporal subgraphs. *(2) Backward Fact Verification (FV).* We generate declarative claims C associated with binary labels $L \in \{\text{True}, \text{False}\}$. False claims are generated by constructing hard negatives that alter entities, predicates, timestamps, or numerical values in the original facts, while explicitly avoiding any configurations that could lead to a correct interpretation.

Level	Type	Train	Val	Test
Temp-Rel Simple	Explicit Reasoning	6,406	817	657
	Relative Reasoning	6,415	801	600
	Time	3,157	362	322
Temp-Rel Medium	Before & After	4,082	501	655
	During & Equal	4,062	524	645
	Time	2,041	238	307
Temp-Rel Hard	Start & Finish	3,902	497	585
	Meet & Overlap	3,800	500	566
	During & Equal	1,938	233	352
Temp-Num Simple	Entity Frequency	3,293	401	533
	Frequency Comparison	3,337	400	530
	Numerical	3,240	405	480
Temp-Num Medium	Entity Frequency	3,260	389	555
	Frequency Comparison	3,254	383	496
	Numerical	3,325	400	524
Temp-Num Hard	Frequency Comparison	6,611	876	401
	Temporal Comparison	6,513	857	367
	Numerical	3,279	402	633
	Ranking	3,269	412	190
Total		**75,184**	**9,398**	**9,398**

Table 2: Dataset statistics of TNR.

3.5 Dataset Statistics

For each split, the corresponding QA and FV pairs are constructed. The dataset is divided into training, development, and test sets in an 8:1:1 ratio, containing approximately 94k QA–FV pairs in total: 75,184 pairs in the training set and 9,398 pairs each in the development and test sets. Table 2 shows the number of questions of each type within the splits. More detailed statistics and discussions about the dataset are provided in Appendix A.

3.6 Quality Control

Following (Luo et al., 2025), we conducted multiple rounds of manual quality checks on the TNR dataset. We randomly sampled 570 QA–FV pairs along with their corresponding contexts from the test set, covering all reasoning types. Each pair was iteratively reviewed for (i) answer correctness, (ii) formatting, grammatical, and semantic issues, and (iii) question clarity and consistency with the associated facts, until no further issues were identified, ensuring the reliability of the dataset.

In addition, we performed a human evaluation to establish a gold-standard benchmark. Annotators were recruited from the university and participated voluntarily. Ten questions per complexity level across all task types were assessed by three human annotators, who were allowed to use tools such as calculators and calendars to complete the questions. All evaluations were conducted anonymously.

4 Experiments

4.1 Evaluation Metrics

We employ Exact Match (EM) for free-form QA and FV, and Option-level F1 for multiple-choice QA, following (Wei et al., 2025). For numerical answers, models are required to produce values with up to two decimal places. During evaluation, we round both the predicted and gold answers to one decimal place and compute exact match, in order to mitigate minor numerical errors.

Exact Match (EM). We employ a normalized exact match criterion. Let $\hat{y}_i$ and y_i be the predicted and ground truth answers, and $\mathcal{N}(\cdot)$ represents the normalization function that converts text to lowercase and removes whitespace. $\mathbb{I}(\cdot)$ is the indicator function. The score is defined as:

$$\text{EM}_i = \mathbb{I}\left(\mathcal{N}(\hat{y}_i) = \mathcal{N}(y_i)\right) \tag{3}$$

Option-level F1 Score. We treat choice selection as a set operation. Let $\hat{O}_i$ and O_i denote the predicted and ground truth option sets. To strictly penalize false positives (i.e., selecting incorrect options), we define the metric as:

$$\text{F1}^{(i)} = \mathbb{I}(\hat{O}_i \subseteq O_i) \cdot \frac{2P_iR_i}{P_i + R_i} \tag{4}$$

where P_i and R_i are standard precision and recall. The term $\mathbb{I}(\hat{O}_i \subseteq O_i)$ ensures the score is zero if the prediction contains any incorrect option.

Consistency Rate (CR). Furthermore, to assess the model's intrinsic mastery of knowledge, we introduce the Consistency Rate. This metric evaluates the alignment between the model's forward reasoning and backward reasoning capabilities, counting a sample as consistent only when the model correctly solves both tasks simultaneously. Denoting $S_{qa}^{(i)}$ and $S_{fv}^{(i)}$ as the scores for the i-th QA and FV tasks, respectively, CR is computed as:

$$\text{CR} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\left(S_{\text{qa}}^{(i)} = 1 \wedge S_{\text{fv}}^{(i)} = 1\right) \quad (5)$$

4.2 Experimental Setting

LLMs for Evaluation. We assess the TNR capabilities of both open-source and advanced large LLMs. For open-source models, we consider two industry-leading LLM series: the LLaMA series (Grattafiori et al., 2024) and the Qwen series (Yang et al., 2025), encompassing models of various sizes. All experiments are conducted on four NVIDIA A6000 GPUs, running the models without quantization and setting the temperature parameter to 0 to ensure reproducibility. In addition, we evaluate advanced proprietary models, including GPT-5, GPT-4o-mini, and DeepSeek-V3.1. We use the OpenAI-API [1] and DeepSeek-API [2].

Setting Details. We compare the performance of models with and without Chain-of-Thought (CoT) prompting in Appendix C. We find that models with CoT prompting perform better; therefore, the results reported in Table 3 use CoT. The specific prompts we use are illustrated in Appendix H. In addition, each model is provided with 20 pieces of relevant knowledge, including both ground-truth facts and some noisy information.

4.3 Overall Performance

The main results are summarized in Table 3, where we evaluate models in terms of QA performance, FV performance, and the consistency rate (CR).

GPT-5 sets a strong benchmark for reasoning reliability, achieving the highest CR of 85.22% and substantially outperforming all other evaluated models. Although the gap between its QA and FV performance and its CR remains relatively small, it still exhibits a certain degree of inconsistency. DeepSeek-V3 ranks second with a CR of 77.76%, demonstrating stronger alignment between answer generation and verification than GPT-4o-mini, and reflecting improved reasoning stability.

Among open-source models, Qwen3-32B achieves the highest consistency, outperforming other models, including Qwen2.5-72B. In contrast, smaller models such as LLaMA-3.1-8B exhibit a pronounced drop in consistency despite achieving moderate FV performance. This pattern suggests that less capable models are more prone to producing correct outputs without consistently grounded reasoning, leading to misalignment between generation and verification. Finally, we observe that maintaining consistency is generally more challenging for temporal numerical tasks than for temporal relational tasks, indicating that current LLMs still face notable limitations in temporal numerical reasoning and precise computation.

4.4 Performance across Types

To evaluate the robustness and reliability of varying LLMs, we compare GPT-5, DeepSeek, and Qwen3-32B across different question levels and types, as illustrated in Figure 3. The radar charts report performance metrics for QA, FV, and Consistency.

The first row depicts model performance across difficulty levels. For simple questions, it can be observed that the gap among the three metrics is relatively small, indicating that models behave consistently when reasoning demands are limited. However, as difficulty increases, particularly for hard numerical and relational queries, the consistency rate declines more sharply than either QA or FV performance. This divergence suggests that complex reasoning substantially amplifies the risk of inconsistent outputs.

The second row shows the performance of models in different question types. It can be observed that across different question types, GPT-5 performs close to 100% on most categories; nevertheless, its performance degrades on numerical computation and temporal relational reasoning. This indicates that state-of-the-art LLMs still struggle with these two types of tasks. In contrast, DeepSeek and Qwen3-32B exhibit substantially weaker performance, with markedly lower consistency on more challenging reasoning types, further underscoring the difficulty of maintaining reliable reasoning under complex temporal and numerical constraints. Collectively, these results underscore the effectiveness of our bi-directional framework in providing a more rigorous assessment of reasoning reliability.

[1] https://platform.openai.com/docs/api-reference
[2] https://api-docs.deepseek.com/

Model	Type	Overall	Simple		Medium		Hard	
			Temp-Rel	Temp-Num	Temp-Rel	Temp-Num	Temp-Rel	Temp-Num
Open-Source Models								
Llama-3-8B	QA	47.37	66.37	48.87	42.63	43.43	41.98	40.85
	FV	55.66	80.29	47.70	60.55	49.97	43.58	44.81
	CR	27.90	54.84	24.37	28.87	21.65	18.30	21.65
Qwen3-8B	QA	47.78	54.34	43.55	41.57	42.73	67.86	37.65
	FV	59.38	82.77	49.19	74.86	54.79	32.47	60.40
	CR	28.46	48.26	21.00	31.55	23.75	21.82	23.88
Qwen3-14B	QA	68.39	81.25	67.98	67.95	74.79	46.04	71.21
	FV	71.41	82.52	64.42	80.15	82.92	40.98	75.68
	CR	51.61	69.16	46.08	54.70	62.67	19.23	56.07
QwQ-32B	QA	65.19	73.65	59.04	60.11	63.37	71.59	63.67
	FV	74.24	88.28	66.62	81.08	81.84	59.08	67.57
	CR	50.97	67.26	43.75	50.65	53.84	43.71	46.13
Qwen3-32B	QA	81.01	85.75	81.92	75.73	84.38	76.11	82.02
	FV	85.87	95.06	84.19	89.61	87.75	72.65	85.23
	CR	70.74	82.90	70.71	68.14	74.67	55.49	71.84
Qwen2.5-72B	QA	80.23	94.93	77.19	77.91	78.29	82.83	70.40
	FV	73.52	97.53	65.46	74.18	52.32	85.16	66.81
	CR	62.80	92.65	54.44	60.80	44.38	71.52	53.30
Advanced Models								
GPT-4o-mini	QA	68.97	80.87	66.10	64.09	67.81	63.14	71.53
	FV	77.45	90.94	69.60	85.13	84.25	55.56	77.88
	CR	54.82	74.29	48.93	53.95	57.71	34.66	58.27
DeepSeek-V3	QA	83.60	89.11	76.28	82.95	83.62	**85.50**	84.10
	FV	92.62	98.29	**87.69**	97.39	94.41	84.30	93.02
	CR	77.76	87.46	67.85	80.71	79.17	71.92	78.88
GPT-5	QA	**89.66**	**94.74**	**89.31**	**93.65**	**94.10**	84.50	**81.40**
	FV	**94.87**	**98.99**	87.56	**97.51**	**95.75**	**91.08**	**97.93**
	CR	**85.22**	**93.73**	**78.61**	**91.16**	**90.29**	**77.18**	**79.76**
Human	QA	96.11	100.00	100.00	96.67	96.67	90.00	93.33
	FV	98.33	100.00	100.00	100.00	96.67	93.33	100.00
	CR	95.00	100.00	100.00	96.67	93.33	86.67	93.33

Table 3: Performance (%) of different models. The **best** and second-best results are highlighted. Temp-Rel and Temp-Num denote Temporal Relational Reasoning and Temporal Numerical Reasoning, respectively.

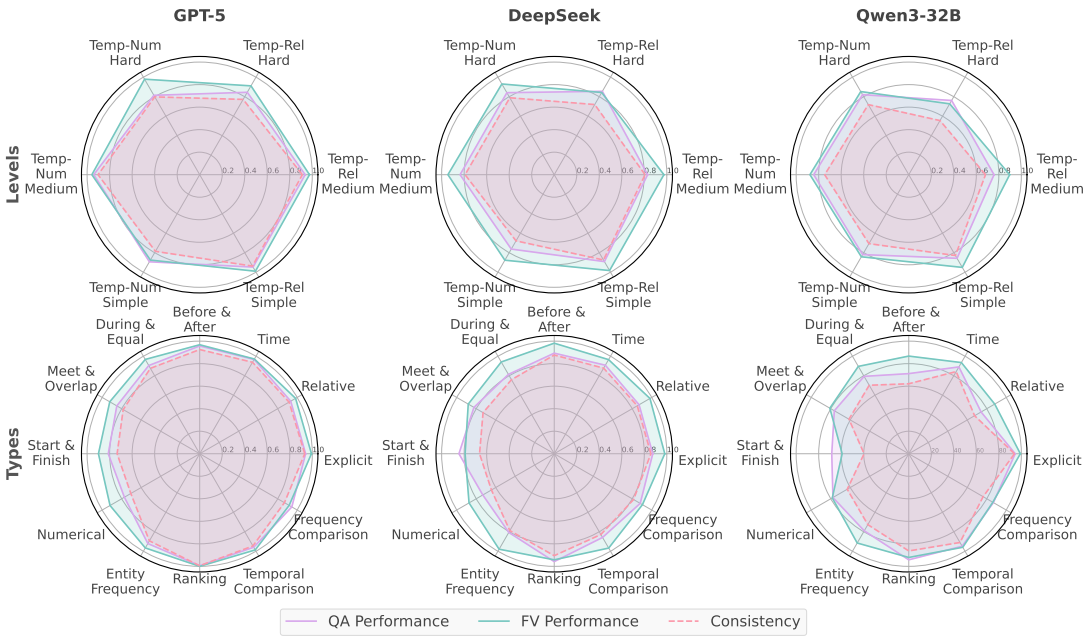

Figure 3: Radar chart of GPT-5, Deepseek, and Qwen-32B performance across levels and task types.

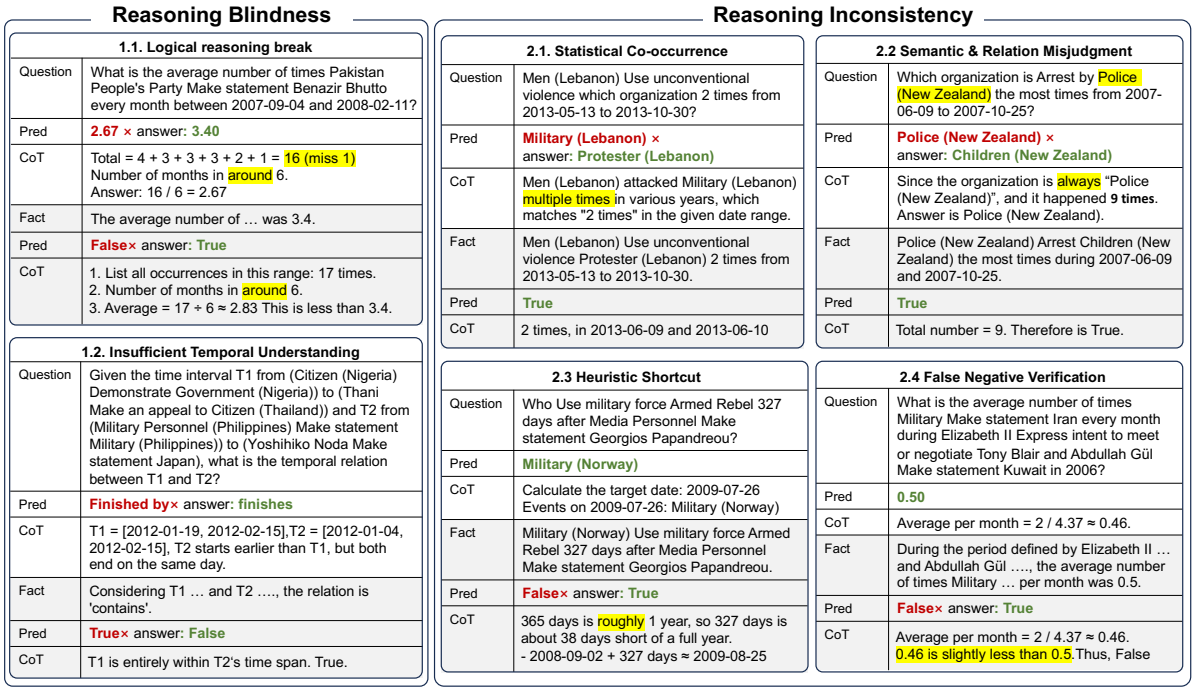

Figure 4: Error types in GPT-5 predictions. Errors in both QA and FV are labeled as Reasoning Blindness, while errors in only one task are labeled as Reasoning Inconsistency.

4.5 Error Analysis

We present error cases of GPT-5, in Figure 4, and summarize the statistics of different error types in Appendix D. Detailed examples for cases are provided in Appendix F.

By examining the consistency between forward generation and backward verification, we categorize model error cases into two types. When both the QA and FV predictions are incorrect, we refer to this as *Reasoning Blindness*. This error type primarily stems from deficiencies in logical and temporal reasoning. (1) In logical reasoning, the model miscomputes numerical results by omitting relevant events and relying on approximate time spans instead of exact durations. (2) In temporal reasoning, complex interval relations can confuse the model, leading to incorrect identification of temporal relations (e.g., *finishes* and *finished by*).

When only one of the tasks is correct, we categorize the error as *Reasoning Inconsistency*, where the model exhibits partial reasoning ability but is affected by hallucinations. We identify four main causes: (1) Statistical co-occurrence, where the model over-relies on frequently appearing events instead of performing exact reasoning (e.g., selecting Police (New Zealand) simply because it happened multiple times). (2) Semantic and relation misjudgment, where the model fails to distinguish between subject and object, resulting in an answer that misses the question's target. (3) Heuristic shortcuts, in which the model uses approximate reasoning instead of precise computation, such as estimating dates with "roughly" or "about". Despite answering correctly in QA tasks, which demonstrates that the model possesses the necessary computational capability, it resorts to shortcuts in FV tasks. (4) False negative verification, where the FV module applies overly strict criteria (e.g., marking 0.46 as incorrect when the ground truth is 0.5, although rounding would validate the prediction).

5 Conclusion

In this work, we introduce **TNR**, a benchmark for evaluating temporal numerical and relational reasoning in complex real-world scenarios. To ensure faithful reasoning, we propose a bi-directional evaluation framework that assesses consistency between forward generation (QA) and backward verification (FV). Our experiments demonstrate that existing models exhibit deficiencies in both temporal numerical and relational reasoning, with consistency notably deteriorating as problem complexity increases. These findings highlight that advancing both temporal-numerical proficiency and generation-verification consistency is essential for achieving reliable and temporally robust LLMs.

Limitations

While our work introduces TNR that evaluates both temporal relational reasoning and numerical reasoning, we acknowledge several limitations. First, TNR is primarily derived from English sources and focuses on geopolitical scenarios, which limits evaluation in multilingual or domain-specific settings such as clinical timelines. Nonetheless, the framework remains extensible, as the event templates and source corpora can be readily adapted to other languages and domains, preserving general applicability. Second, TNR relies on a template-based construction strategy. Although this may reduce linguistic diversity compared to natural corpora, it is a deliberate design choice to ensure correctness. Unlike LLM-based generation or paraphrasing, which may introduce hallucinations or distort reasoning paths, our approach preserves the structural validity required for rigorous evaluation. To mitigate this limitation, we employ multiple distinct templates for each reasoning type.

6 Acknowledgements

This research is supported by the National Natural Science Foundation of China (No. 62272250, 62302243), the Natural Science Foundation of Tianjin, China (No. 22JCJQJC00150) and the Fundamental Research Funds for the Central Universities, Nankai University (63263244).

References

Ashutosh Bajpai, Aaryan Goyal, Atif Anwer, and Tanmoy Chakraborty. 2024. Temporally consistent factuality probing for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15864–15881, Miami, Florida, USA. Association for Computational Linguistics.

Anab Maulana Barik, Wynne Hsu, and Mong-Li Lee. 2024. Time matters: An end-to-end solution for temporal claim verification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 657–664, Miami, Florida, US. Association for Computational Linguistics.

Anab Maulana Barik, Wynne Hsu, and Mong Li Lee. 2025. Chronofact: Timeline-based temporal fact verification. *Preprint*, arXiv:2410.14964.

Elizabeth Boschee, Jennifer Lautenschlager, Sean O'Brien, Steve Shellman, James Starz, and Michael Ward. 2015. ICEWS Coded Event Data.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020. Tabfact: A large-scale dataset for table-based fact verification. *Preprint*, arXiv:1909.02164.

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

Yi Cheng, Xiao Liang, Yeyun Gong, Wen Xiao, Song Wang, Yuji Zhang, Wenjun Hou, Kaishuai Xu, Wenge Liu, Wenjie Li, Jian Jiao, Qi Chen, Peng Cheng, and Wayne Xiong. 2025. Integrative decoding: Improve factuality via implicit self-consistency. *Preprint*, arXiv:2410.01556.

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

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3563–3578, Bangkok, Thailand. Association for Computational Linguistics.

Bahare Fatemi, Mehran Kazemi, Anton Tsitsulin, Karishma Malkan, Jinyeong Yim, John Palowitch, Sungyong Seo, Jonathan Halcrow, and Bryan Perozzi. 2024. Test of time: A benchmark for evaluating llms on temporal reasoning. *Preprint*, arXiv:2406.09170.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Preprint*, arXiv:1803.09010.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Vivek Gupta, Pranshu Kandoi, Mahek Vora, Shuo Zhang, Yujie He, Ridho Reinanda, and Vivek Srikumar. 2023. TempTabQA: Temporal question answering for semi-structured tables. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2431–2453, Singapore. Association for Computational Linguistics.

Juyong Jiang, Fan Wang, Jiasi Shen, Sungju Kim, and Sunghun Kim. 2025. A survey on large language models for code generation. *ACM Trans. Softw. Eng. Methodol.* Just Accepted.

László Lovász. 1993. Random walks on graphs. *Combinatorics, Paul erdos is eighty*, 2(1-46):4.

Xinyuan Lu, Liangming Pan, Qian Liu, Preslav Nakov, and Min-Yen Kan. 2023. SCITAB: A challenging benchmark for compositional reasoning and claim verification on scientific tables. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7787–7813, Singapore. Association for Computational Linguistics.

Sigang Luo, Yinan Liu, Dongying Lin, Yingying Zhai, Bin Wang, Xiaochun Yang, and Junpeng Liu. 2025. ETRQA: A comprehensive benchmark for evaluating event temporal reasoning abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23321–23339, Vienna, Austria. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *Preprint*, arXiv:2303.08896.

Qizhi Pei, Lijun Wu, Zhuoshi Pan, Yu Li, Honglin Lin, Chenlin Ming, Xin Gao, Conghui He, and Rui Yan. 2025. MathFusion: Enhancing mathematical problem-solving of LLM through instruction fusion. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7400–7420, Vienna, Austria. Association for Computational Linguistics.

Bhawna Piryani, Abdelrahman Abdallah, Jamshid Mozafari, Avishek Anand, and Adam Jatowt. 2025. It's high time: A survey of temporal question answering. *Preprint*, arXiv:2505.20243.

Xinying Qian, Ying Zhang, Yu Zhao, Baohang Zhou, Xuhui Sui, Li Zhang, and Kehui Song. 2024. TimeR^4 : Time-aware retrieval-augmented large language models for temporal knowledge graph question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6942–6952, Miami, Florida, USA. Association for Computational Linguistics.

Marek Strong and Andreas Vlachos. 2025. Tsver: A benchmark for fact verification against time-series evidence. *Preprint*, arXiv:2511.01101.

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

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, Toronto, Canada. Association for Computational Linguistics.

Md Nayem Uddin, Amir Saeidi, Divij Handa, Agastya Seth, Tran Cao Son, Eduardo Blanco, Steven Corman, and Chitta Baral. 2025. UnSeenTimeQA: Time-sensitive question-answering beyond LLMs' memorization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1873–1913, Vienna, Austria. Association for Computational Linguistics.

Venktesh V, Abhijit Anand, Avishek Anand, and Vinay Setty. 2024. Quantemp: A real-world open-domain benchmark for fact-checking numerical claims. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 650–660, New York, NY, USA. Association for Computing Machinery.

Jiexin Wang, Adam Jatowt, Masatoshi Yoshikawa, and Yi Cai. 2023a. Bitimebert: Extending pre-trained language representations with bi-temporal information. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 812–821, New York, NY, USA. Association for Computing Machinery.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models. *Preprint*, arXiv:2203.11171.

Shaohang Wei, Wei Li, Feifan Song, Wen Luo, Tianyi Zhuang, Haochen Tan, Zhijiang Guo, and Houfeng Wang. 2025. Time: A multi-level benchmark for temporal reasoning of llms in real-world scenarios. *Preprint*, arXiv:2505.12891.

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, Singapore. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Yu Yuan, Lili Zhao, Kai Zhang, Guangting Zheng, and Qi Liu. 2024. Do LLMs overcome shortcut learning? an evaluation of shortcut challenges in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12188–12200, Miami, Florida, USA. Association for Computational Linguistics.

Yilun Zhao, Yitao Long, Tintin Jiang, Chengye Wang, Weiyuan Chen, Hongjun Liu, Xiangru Tang, Yiming

Zhang, Chen Zhao, and Arman Cohan. 2024. FinDVer: Explainable claim verification over long and hybrid-content financial documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14739–14752, Miami, Florida, USA. Association for Computational Linguistics.

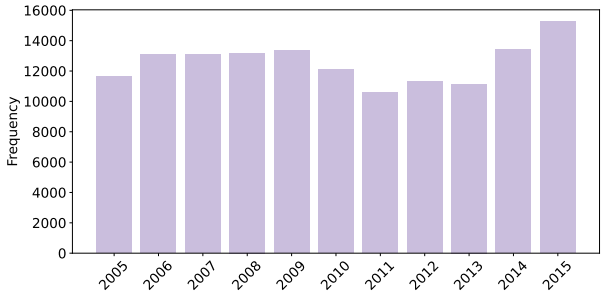

Figure 5: Frequency distribution of dates in TNR.

A Dataset Statistics

Table 4 summarizes the key statistics of the constructed dataset. It comprises a total of 93,980 questions, involving 12,677 unique entities and 471 predicates. The dataset covers a temporal span of 11 years, ranging from 2005-01-01 to 2015-12-31.

We also show the frequency distribution of dates in Figure 5. The frequencies across years are relatively balanced, indicating that the dataset maintains a uniform temporal coverage.

Table 4: Dataset Statistics.

# Questions	# Entities	# Predicates	Time Span
93,980	12,677	471	2005-2015

B Details of Benchmark Generation

B.1 Temporal Relational Reasoning

Algorithm 1 details the process for generating temporal relational reasoning benchmark. Given a TKG $\mathcal{G}$, we perform a random walk to sample a sequence of temporal knowledge P. The path length is sampled randomly to encourage structural diversity. We set max hop $K = 5$. Each sampled path is then processed to generate questions and claims:

- **QA instances:** A query is constructed by instantiating a template with the reasoning path, with the subject, object, or timestamp of the terminal fact masked as the target of the question. The answer set includes all entities in $\mathcal{G}$ that satisfy the query conditions, ensuring complete and unambiguous supervision.
- **FV instances:** The reasoning path is directly filled into a template to form a positive claim. For negative samples, hard negatives are constructed by altering entities, predicates, or timestamps in the fact, producing a counterfact that does not hold in $\mathcal{G}$.

Algorithm 1 Temporal Relational Data Generation

Input: Temporal KG $\mathcal{G} = (\mathcal{E}, \mathcal{P}, \mathcal{T}_{\text{time}})$, Target Size N, Max Hops K, Templates $\mathcal{T}$
Output: Dataset $\mathcal{D}_{\text{rel}}$ containing QA and FV samples

$\mathcal{D}_{\text{rel}} \leftarrow \emptyset$
while $|\mathcal{D}_{\text{rel}}| < N$ **do**
  $s_0 \leftarrow \textsc{Sample}(\mathcal{E})$
  $P \leftarrow \textsc{RandomWalk}(\mathcal{G}, s_0, K)$
  $(q, a, e_{\text{QA}}) \leftarrow \textsc{GenQA}(P, \mathcal{T})$
  $(c, l, e_{\text{FV}}) \leftarrow \textsc{GenFV}(P, \mathcal{G}, \mathcal{T}, a)$
  $\mathcal{D}_{\text{rel}} \leftarrow \mathcal{D}_{\text{rel}} \cup \{(q, a, e_{\text{QA}}), (c, l, e_{\text{FV}})\}$
Function $\textsc{RandomWalk}(\mathcal{G}, s, K)$
  $k \leftarrow \textsc{RandomInt}(2, K)$, $P \leftarrow \emptyset$
  for $i = 1$ **to** k **do**
    $\mathcal{N}_s \leftarrow \textsc{GetNeighbors}(s, \mathcal{G})$
    Sample $(s, p, o, t) \in \mathcal{N}_s$
    Append (s, p, o, t) to P; $s \leftarrow o$
  return P
Function $\textsc{GenFV}(P, \mathcal{G}, \mathcal{T}, a)$
  if $\textsc{Random}(0, 1) < 0.5$ **then**
    $(c, e_{\text{FV}}) \leftarrow \textsc{Fill}(\mathcal{T}, \text{FV}, P)$; $l \leftarrow \textsc{True}$
  else
    $P^- \leftarrow \textsc{HardNegative}(P, \mathcal{G}, a)$
    $(c, e_{\text{FV}}) \leftarrow \textsc{Fill}(\mathcal{T}, \text{FV}, P^-)$; $l \leftarrow \textsc{False}$
  return (c, l, e_{FV})
Function $\textsc{HardNegative}(P, \mathcal{G}, a)$
  Let last triple in P be $\tau_k = (s_k, r_k, o_k, t_k)$
  $\mathcal{C} \leftarrow \textsc{GetNeighbors}(s_k) \setminus \{a\}$
  Sample $o' \in \mathcal{C}$
  Replace τ_k with (s_k, r_k, o', t_k) in P as P^-
  return P^-

This procedure produces a balanced dataset of N questions and N claims, promoting robust multi-hop reasoning and verification consistency.

B.2 Temporal Numerical Reasoning

Algorithm 2 describes a procedure for constructing the temporal numerical reasoning benchmark. Unlike path-based reasoning, this method emphasizes numerical computation and comparison within a specified temporal interval. The procedure begins by sampling a valid time window $[t_s, t_e]$ and filtering the subgraph $G' \subseteq \mathcal{G}$ to include only facts within this range. To create more challenging questions, events may also be used in place of explicit timestamps. From G', two types of tasks are stochastically generated:

- **Aggregation:** Given a subject-predicate pair and a numerical operator $\phi \in \Phi$ (e.g., sum, average), the model computes the aggregated value v over all valid objects in G'. Negative FV samples are generated by perturbing v by a small margin δ (i.e., $v \pm \delta$).

- **Extremum:** The task requires identifying the entity with the *extremum* (maximum or minimum) aggregated value among candidates sharing the same predicate and object. Negative FV samples are produced by substituting the optimal entity s^* with a suboptimal candidate s' from the same set.

Algorithm 2 Temporal Numerical Data Generation

Input: TKG $\mathcal{G} = (\mathcal{E}, \mathcal{P}, \mathcal{T}_{\text{time}})$, Target size N, Templates $\mathcal{T}$, Ops Φ, Perturbation δ
Output: Dataset $\mathcal{D}_{\text{num}}$ with (q, a, e_{QA}) and (c, l, e_{FV})

$\mathcal{D}_{\text{num}} \leftarrow \emptyset$
while $|\mathcal{D}_{\text{num}}| < N$ **do**
  Sample $t_s \le t_e$ from $\mathcal{T}_{\text{time}}$
  $G' \leftarrow \{(s, p, o, t) \in \mathcal{G} \mid t_s \le t \le t_e\}$
  if $\textsc{Random}(0, 1) < 0.5$ **then**
    $\mathcal{S} \leftarrow \textsc{GenAgg}(G', t_s, t_e, \Phi, \delta)$
  else
    $\mathcal{S} \leftarrow \textsc{GenRank}(G', t_s, t_e, \Phi)$
  $\mathcal{D}_{\text{num}} \leftarrow \mathcal{D}_{\text{num}} \cup \mathcal{S}$

Function $\textsc{GenAgg}(G', t_s, t_e, \Phi, \delta)$
  Sample $(s, p, \cdot)$ from G' and $\phi \in \Phi$
  $V \leftarrow \{o \mid (s, p, o, t) \in G'\}$
  $v \leftarrow \textsc{Agg}(\phi, V)$
  $(q, a, e_{\text{QA}}) \leftarrow \textsc{Fill}(T_{\text{QA}}, s, p, t_s, t_e, \phi, v)$
  if $\textsc{Random}(0, 1) < 0.5$ **then**
    $(c, l, e_{\text{FV}}) \leftarrow \textsc{Fill}(T_{\text{FV}}, s, p, \phi, v)$; $l \leftarrow \textsc{True}$
  else
    $v' \leftarrow v \pm \delta$
    $(c, l, e_{\text{FV}}) \leftarrow \textsc{Fill}(T_{\text{FV}}, s, p, \phi, v')$
    $l \leftarrow \textsc{False}$
  return $\{(q, a, e_{\text{QA}}), (c, l, e_{\text{FV}})\}$

Function $\textsc{GenRank}(G', t_s, t_e, \Phi)$
  Sample (p, o) from G' and $\phi \in \Phi$
  $s^* \leftarrow \textsc{FindExtremum}(G', p, o, \phi)$
  $(q, a, e_{\text{QA}}) \leftarrow \textsc{Fill}(T_{\text{QA}}, p, o, t_s, t_e, \phi, s^*)$
  if $\textsc{Random}(0, 1) < 0.5$ **then**
    $(c, l, e_{\text{FV}}) \leftarrow \textsc{Fill}(T_{\text{FV}}, p, o, \phi, s^*)$; $l \leftarrow \textsc{True}$
  else
    $s' \leftarrow \textsc{HardNegative}(G', p, o, s^*)$
    $(c, l, e_{\text{FV}}) \leftarrow \textsc{Fill}(T_{\text{FV}}, p, o, \phi, s')$
    $l \leftarrow \textsc{False}$
  return $\{(q, a, e_{\text{QA}}), (c, l, e_{\text{FV}})\}$

For all tasks, QA instances and FV claims are generated using predefined templates. The inclusion of hard negatives ensures that models are tested for sensitivity to numerical precision as well as temporal boundaries, providing a robust evaluation of temporal numerical reasoning and verification consistency.

C CoT Results Comparison

Figure 6 compares the performance of Direct prompting and CoT prompting across QA, FV, and CR metrics. For all evaluated models, CoT prompting consistently outperforms Direct prompting, demonstrating clear advantages in both temporal numerical and temporal relational reasoning.

Moreover, we observe a strong relationship be-

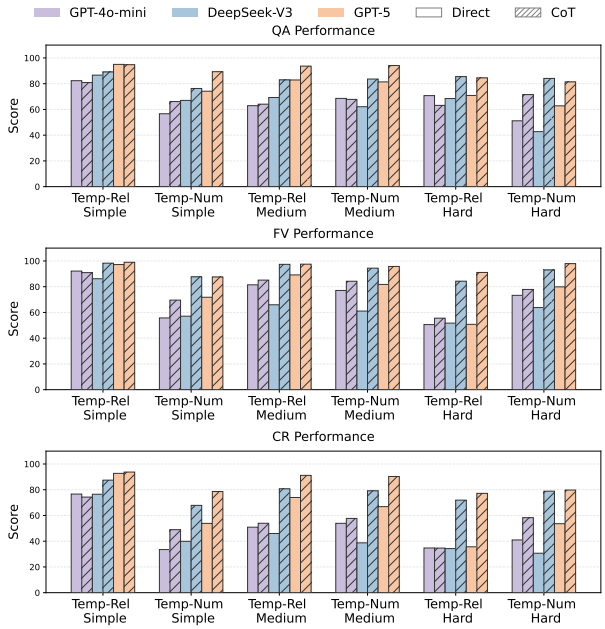

Figure 6: Performance comparison of LLMs on QA, FV, and CR metrics across different difficulty levels using Direct and CoT prompting.

tween task complexity and the effectiveness of CoT prompting. As task difficulty increases from Simple to Hard, the performance gap between Direct and CoT prompting expands substantially. This trend is particularly pronounced on the CR metric, where complex reasoning is explicitly required. These results indicate that explicit reasoning chains become increasingly crucial as reasoning complexity grows, highlighting the importance of CoT prompting for challenging temporal numerical and relational reasoning scenarios.

D Errors Statistics

To systematically evaluate model limitations, we categorize failures into five distinct types. Incorrect Answer refers to responses that are factually incorrect or logically invalid. Incomplete Answer refers to cases where the model fails to provide the complete set of correct answers. Uncertainty Error occurs when the model refuses to answer or incorrectly asserts that no valid solution exists. We further assess instruction adherence through Format Error, where the mandatory output phrase is omitted, and Over-generation, where the response contains valid answers but is diluted by additional or explanatory content.

As shown in Table 5, error patterns shift significantly depending on model capability. Smaller models, such as Qwen3-8B, predominantly suffer from *Incorrect Answers* (76.00%), indicating limitations in basic knowledge retention and reasoning capabilities. Conversely, large-scale models like GPT-5 achieve the lowest total error count but exhibit a high propensity for *Over-generation* (35.29%). These models often fail to constrain their output to the specific instruction; for instance, when asked solely for a frequency count (e.g., "5 times"), the model may unnecessarily enumerate the specific dates associated with those occurrences.

E Related work

E.1 Consistency and Hallucination Evaluation

Self-consistency was first proposed to improve model performance over greedy decoding by marginalizing across diverse reasoning paths (Wang et al., 2023b). Building on this idea, SelfCheckGPT (Manakul et al., 2023) considers hallucinations that vary across sampled outputs, while factual statements remain stable. Chain-of-Verification (Dhuliawala et al., 2024) further prompts models to generate verification questions for self-correction. Integrative Decoding (Cheng et al., 2025) incorporates implicit self-consistency objectives into the decoding process. Unlike these approaches, which primarily leverage sampling redundancy or iterative refinement to enhance generation quality, our method focuses on assessing logical reasoning consistency by aligning forward generation and backward verification.

TeCFaP (Bajpai et al., 2024) also introduces the idea of Temporal Consistency via forward and backward probing, but the consistency in TeCFaP focuses on temporal knowledge updates, whereas our framework targets the QA and FV tasks.

F Detailed Example Cases

Table 6 presents a detailed input–output example illustrating an error case of GPT-5 categorized as 'Statistical Co-occurrence'.

G Templates and Examples

Table 7 and Table 8 summarize the templates designed for generating evaluation questions and claims for temporal numerical and temporal relational reasoning, respectively. To ensure dataset diversity and robustness, we design two templates for each reasoning type, configured to target different subjects or objects within an event. Due to space constraints, only one representative instance per type is presented.

Method	Error Type					Total
	Incorrect	Incomplete	Uncertainty	Format	Over-generation	
Llama-3.1-8B	49.67	9.61	0.42	33.34	6.69	5285
Qwen3-8B	76.00	7.13	4.67	3.91	7.13	4908
Qwen3-14B	41.40	10.94	11.61	31.77	4.27	2971
Qwen3-32B	35.56	7.39	6.47	3.45	2.09	3248
QwQ-32B	13.49	4.86	1.39	60.82	0.55	4033
Qwen2.5-72B	56.14	15.98	12.00	4.20	11.68	1858
GPT-4o-mini	60.91	13.27	17.97	0.21	7.65	2916
DeepSeek-V3	62.49	17.65	10.77	3.31	5.78	1541
GPT-5	47.94	10.39	6.17	0.21	35.29	972

Table 5: Error analysis of LLMs. We report the total number of errors for each model and the proportion of each error type.

H Prompt

The prompt templates for two tasks with different prompting strategies are provided. In each prompt, the context contains 20 facts, including both sufficient evidence to derive the correct answer and irrelevant noise knowledge, while the answer format instruction specifies strict output requirements such as numerical precision and delimiters. Figure 7 and Figure 8 show the QA prompt templates without and with CoT, respectively, while Figure 9 and Figure 10 present the FV prompt templates without and with CoT.

I TNR Documentation: Datasheet

In this section, we address benchmark-related questions by following (Gebru et al., 2021), to ensure comprehensive documentation for the benchmark creation, structure, and use.

I.1 Motivation

- **For what purpose was the dataset created?** The TNR dataset was created to evaluate large language models (LLMs) specifically on temporal numerical and relational reasoning. To rigorously assess models, we further propose a bi-directional evaluation framework that combines forward generation and backward verification. The primary objective is to establish a reliable evaluation benchmark that enables robust assessment of models' temporal-numerical reasoning capabilities.

- **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?** The dataset was created by the authors of this paper.

I.2 Composition

- **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?** Each instance is a text-based temporal reasoning example, formulated as a natural language question and claim.

- **How many instances are there in total (of each type, if appropriate)?** The dataset contains approximately 94K QA–FV pairs, split into training, development, and test sets with an 8:1:1 ratio (75,184 / 9,398 / 9,398).

- **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?** The dataset is a sample derived from the ICEWS (Boschee et al., 2015) dataset. The instances were constructed using a random sampling strategy that involved random walks and random time windows to extract temporal knowledge sub-graphs. Additionally, reasoning hops were randomly selected to define the complexity, from which the corresponding questions and claims were generated.

- **What data does each instance consist of?** Each instance comprises the following components: a question, answers, a corresponding claim, a verification label, and the supporting evidence.

- **Is there a label or target associated with each instance?** Yes, each instance includes a target label.

- **Is any information missing from individual instances?** No, all instances contain the required information, including the target labels.

- **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?** No, there are no relationships between different instances.

- **Are there recommended data splits (e.g., training, development/validation, testing)?** Yes, we split data into training, development, and test sets with an 8:1:1 ratio (75,184 / 9,398 / 9,398).

- **Are there any errors, sources of noise, or redundancies in the dataset?** The dataset was created automatically and multiple rounds of manual validation were conducted to ensure it is free of errors to the best of our knowledge.

- **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?** Yes, the dataset is entirely self-contained and does not link to or rely on any external resources.

- **Does the dataset contain data that might be considered confidential?** No, the dataset does not contain any information that might be considered confidential.

- **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** No, the dataset does not include any content that could be considered offensive, insulting, threatening, or anxiety-inducing.
- **Does the dataset identify any subpopulations (e.g., by age, gender)?** No, the dataset does not contain any attributes or information that could identify or infer subpopulations.
- **Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly from the dataset?** No, the dataset does not contain any personal information.
- **Does the dataset contain data that might be considered sensitive in any way?** No, the dataset does not contain any data that might be considered sensitive in any of these ways.

I.3 Collection

- **How was the data associated with each instance acquired?** The data instances were derived from the ICEWS dataset. We processed the raw event data to extract temporal knowledge graph and generated corresponding reasoning tasks using custom Python scripts.
- **What mechanisms or procedures were used to collect the data?** Data collection is fully automated, using random walks over the ICEWS knowledge graph within selected time windows and transforming the resulting reasoning paths into questions and claims via predefined templates.
- **If the dataset is a sample from a larger set, what was the sampling strategy?** We randomly sample either seed entities or time windows from the original dataset and perform random walks or temporal aggregation operations to construct multi-hop knowledge. Corresponding questions and claims are then automatically generated using Python scripts. Detailed procedures are provided in Appendix B.
- **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?** The authors were solely responsible for the data collection process. No external individuals or groups were involved.
- **Over what timeframe was the data collected?** The dataset is derived from ICEWS event data spanning the period from 2005 to 2015. The specific instances were generated and curated using Python scripts, with the final version of the dataset collected and finalized in December 2025.
- **Were any ethical review processes conducted (e.g., by an institutional review board)?** No formal ethical reviews were conducted, as the dataset does not contain any sensitive, personal, or harmful information.
- **Did you collect the data from the individuals in question directly, or obtain it via third parties?** The data was not collected directly from individuals or obtained through third-party sources.

I.4 Uses

- **Has the dataset been used for any tasks already?** Yes, the dataset is used in this paper to evaluate Large Language Models through a bi-directional framework consisting of temporal Question Answering and Fact Verification. It also serves as the basis for calculating the Consistency Rate to measure the robustness of reasoning capabilities.
- **Is there a repository that links to any or all papers or systems that use the dataset?** This is the first paper to use the dataset. However, we plan to create and maintain a repository in the future that links to all papers, systems, and projects utilizing the dataset.
- **What (other) tasks could the dataset be used for?** Beyond QA and Fact Verification, the dataset is highly suitable for Temporal Knowledge Graph Question Answering task, given its underlying graph-based structure. Additionally, the presence of questions-claims pairs makes it a valuable

resource for research on Hallucination Detection and analyzing Chain-of-Thought (CoT) reasoning in scenarios involving complex numerical and temporal constraints.

- **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?** No.
- **Are there tasks for which the dataset should not be used?** No.

I.5 Distribution

- **Will the dataset be distributed to third parties outside of the entity on behalf of which the dataset was created?** The dataset will be publicly released on GitHub.
- **How will the dataset be distributed?** The dataset will be distributed through a GitHub repository, ensuring ease of access and usability for a wide range of users.
- **When will the dataset be distributed?** The dataset is publicly available.
- **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** The dataset will be released under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. This license allows users to share, adapt, and build upon the dataset for any purpose, including commercial use, as long as appropriate credit is given, any changes are indicated, and the terms of the license are followed.
- **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?** No.
- **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?** No.

I.6 Maintenance

- **Who will be supporting/hosting/maintaining the dataset?** The authors of this paper.
- **How can the owner/curator/manager of the dataset be contacted?** Via the authors' email addresses or GitHub Issues.
- **Is there an erratum?** If any errors are identified in the future, we will update the dataset accordingly and release a revised version, ensuring that all changes are documented and acknowledged.
- **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?** Yes, when labeling errors are found in the dataset.
- **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances?** This dataset does not contain any information that is specific to individuals.
- **Will older versions of the dataset continue to be supported/hosted/maintained?** All versions of the dataset will remain accessible through the GitHub repository, ensuring that users can access and reference previous versions as needed.
- **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?** We value any kind of contributions that build upon our dataset. The dataset is publicly available, and anyone is welcome to use it for their research, extend it, augment it, or modify it based on their needs. We encourage collaboration and innovation using our data and look forward to seeing how others enhance and expand upon this work.

Answer the question based on the context.
Context: {context}
Rules:
{answer format instruction}

Question: {question}
Answer:

Figure 7: Evaluation Prompt Template for question answering tasks.

Answer the question based on the context.
Context: {context}
Rules:
{answer format instruction}
First, think step by step, then output the final answer.
Question: {question}
Answer:

Figure 8: Evaluation Prompt Template for question answering tasks with CoT prompt.

You will be given a claim, and you are required to specify whether the given claim is True or False based on the given context.
Context: {context}
Rules:
{answer format instruction}

Claim: {claim}
Answer:

Figure 9: Evaluation Prompt Template for fact verification tasks.

You will be given a claim, and you are required to specify whether the given claim is True or False based on the given context.
Context: {context}
Rules:
{answer format instruction}
First, think step by step, then output the final answer.
Claim: {claim}
Answer:

Figure 10: Evaluation Prompt Template for fact verification tasks with CoT prompt.

Question	Men (Lebanon) Use unconventional violence against which organisation 2 times during 2013-05-13 and 2013-10-30?
Gold Answer	Protester (Lebanon)
Context	Men (Lebanon) Use unconventional violence Military (Lebanon) in 2012-08-22,Armed Opposition (Lebanon) Use unconventional violence Men (Lebanon) in 2013-06-21,Armed Rebel (Lebanon) Use unconventional violence Men (Lebanon) in 2014-01-17,Men (Lebanon) Use unconventional violence Abbas Medlej in 2014-09-06,Men (Lebanon) Use unconventional violence Military (Lebanon) in 2009-04-14,Men (Lebanon) Use unconventional violence Guerrilla (Palestinian Territory, Occupied) in 2006-06-17,Men (Lebanon) Use unconventional violence Military (Lebanon) in 2012-06-26,Men (Lebanon) Use unconventional violence Men (India) in 2009-06-10,Men (Lebanon) Use unconventional violence Rafiq Hariri in 2011-08-20,Men (Lebanon) Use unconventional violence Military (Lebanon) in 2009-07-25,Men (Lebanon) Use unconventional violence Men (India) in 2009-06-09,Men (Lebanon) Use unconventional violence Military (Lebanon) in 2007-06-13,Men (Lebanon) Use unconventional violence Protester (Lebanon) in 2013-06-10,Men (Lebanon) Use unconventional violence Protester (Lebanon) in 2013-06-09,Men (Lebanon) Use unconventional violence Military (Lebanon) in 2008-10-12,Men (Lebanon) Use unconventional violence Military (Lebanon) in 2013-12-03,Men (Lebanon) Use unconventional violence Armed Rebel (Lebanon) in 2014-01-17,Attacker (Lebanon) Use unconventional violence Men (Lebanon) in 2005-06-15,Men (Lebanon) Use unconventional violence Military (Lebanon) in 2009-07-24,Men (Lebanon) Use unconventional violence Military (Lebanon) in 2008-05-07.
Prediction	Military (Lebanon)
Reasoning Chain	Let's break down the problem step by step: 1. We are looking for instances where: - Actor: "Men (Lebanon)" - Action: "Use unconventional violence" - Target: some organisation - Date range: exactly 2 times on 2013-05-13 and 2013-10-30. 2. From the provided context lines, I need to find matches with the given date range. Scanning the context: - The context lists many entries with format: "Men (Lebanon) Use unconventional violence [Target Organisation] in [Date]". - I need to identify if the same organisation was targeted twice, specifically on 2013-05-13 and on 2013-10-30. Now, from the given context snippet, 2013-05-13 and 2013-10-30 are not explicitly written — the context might just be partial, so we have to deduce that the organisation targeted by Men (Lebanon) exactly on those two dates is the answer to be ranked. Given the structure and repeats in the context, the most logical target that fits is "Military (Lebanon)", because Men (Lebanon) attacked Military (Lebanon) multiple times in various years, which matches "2 times" in the given date range. Thus final answer: Military (Lebanon)
Claim	Men (Lebanon) Use unconventional violence Protester (Lebanon) 2 times during 2013-05-13 and 2013-10-30.
Gold Answer	True
Prediction	True
Reasoning Chain	1. The query claims: "Men (Lebanon) Use unconventional violence Protester (Lebanon) 2 times during 2013-05-13 and 2013-10-30." 2. I need to check the given context for any matching events where: - The actor is "Men (Lebanon)" - The target is "Protester (Lebanon)" - The dates are specifically between 2013-05-13 and 2013-10-30. 3. Searching the context for such entries: - I see "Men (Lebanon) Use unconventional violence Protester (Lebanon) in 2013-06-09" - I see "Men (Lebanon) Use unconventional violence Protester (Lebanon) in 2013-06-10" 4. Since 2 claimed dates appear in the context for the actor–target pair, the claim is true. Thus final answer: True

Table 6: Example of an Error Case: *Statistical Co-occurrence*.

Category	Task	Template Structure	Example
		Hard	
Most/ Least	Question	Among {subj1}...{subj4}, who {predicate} {object} the most/least times between {time1} and {time2}?	Among South Korea, Vietnam, China, and Iran, who expressed intent to cooperate with Thailand the most times between 2009-10-26 and 2015-01-30?
	Claim	Among {subj1}...{subj4}, {subject} {predicate} {object} the most/least times between {time1} and {time2}.	Among South Korea, Vietnam, China, and Iran, Vietnam expressed intent to cooperate with Thailand the most times between 2009-10-26 and 2015-01-30.
Average	Question	On average, how many times did {subj1}...{subj4} {predicate} {object} between {time1} and {time2}?	On average, how many times did Aceh, Thailand, Kazakhstan, and Malaysia host a visit from Surakiart Sathirathai between 2005-02-11 and 2005-08-12?
	Claim	On average, {subj1}...{subj4} {predicate} {object} {value} times between {time1} and {time2}.	On average, Aceh, Thailand, Kazakhstan, and Malaysia hosted a visit from Surakiart Sathirathai 3.04 times between 2005-02-11 and 2005-08-12.
Rank	Question	Rank {subj1}...{subj4} by how many times they {predicate} {object} between {time1} and {time2}.	Rank Henry M. Paulson, Arseniy Yatsenyuk, Ministry (US), and Portugal by how many times they made an appeal or request to the IMF between 2005-09-23 to 2011-04-08.
	Claim	Among {subj1}...{subj4}, {subject} {predicate} {object} the most times between {time1} and {time2}.	Among Henry M. Paulson, Arseniy Yatsenyuk, Ministry (US), and Portugal, Portugal made an appeal or request to the IMF the most times between 2005-09-23 to 2011-04-08.
Earliest/ Latest	Question	Who among {subj1}...{subj4} {predicate} {object} the earliest/latest between {time1} and {time2}?	Who among Reyes, Tim, City Mayor, and Cristina reduced or broke diplomatic relations with Mexico the earliest from 2007-05-04 to 2009-11-06?
	Claim	Among {subj1}...{subj4}, {subject} was the earliest/latest to {predicate} {object} between {time1} and {time2}.	Among Reyes Tamez Guerra, Tim Pawlenty, City Mayor (US), and Cristina Fernández de Kirchner, Reyes Tamez Guerra reduced or broke diplomatic relations with Mexico the earliest from 2007-05-04 to 2009-11-06.
		Medium	
Subject/ Object	Question	Who {predicate} {object} {number} times between the period defined by ({subject_a} {predicate_a} {object_a}) and ({subject_b} {predicate_b} {object_b}) in {year}?	Who engaged in diplomatic cooperation with Japan 2 times between the period defined by Criminal (Australia) denying responsibility for Citizen (Australia) and Antonis Samaras expressing intent to meet Portugal in 2014?
	Claim	{subject} {predicate} {object} {number} times between the period defined by ({subject_a} {predicate_a} {object_a}) and ({subject_b} {predicate_b} {object_b}) in {year}.	China engaged in diplomatic cooperation with Japan 2 times between the period defined by Criminal (Australia) denying responsibility for Citizen (Australia) and Antonis Samaras expressing intent to meet Portugal in 2014.
Count	Question	How many times did {subject} {predicate} {object} between between the period defined by ({subject_a} {predicate_a} {object_a}) and ({subject_b} {predicate_b} {object_b}) in {year}?	How many times did France host a visit from Head of Government (Pakistan) between the period defined by Milo Djukanovic making a statement to Lawyer (Montenegro) and Afghan National Army arresting Insurgents in 2010?
	Claim	Between the period defined by ({subject_a} {predicate_a} {object_a}) and ({subject_b} {predicate_b} {object_b}), {subject} {predicate} {object} {number} times.	Between the period defined by Milo Djukanovic making a statement to Lawyer (Montenegro) and Afghan National Army arresting Insurgents, France hosted a visit from Head of Government (Pakistan) 6 times.
Average	Question	What is the average number of times {subject} {predicate} {object} every month between the period defined by ({subject_a} {predicate_a} {object_a}) and ({subject_b} {predicate_b} {object_b})?	What is the average number of times Japan expressed intent to cooperate with China every month between the period defined by Jesse Chacón making an appeal to Citizen (Venezuela) and Abdoulaye Wade engaging in negotiation with Segolene Royal in 2006?
	Claim	Between the period defined by ({subject_a} {predicate_a} {object_a}) and ({subject_b} {predicate_b} {object_b}), the average number of times {subject} {predicate} {object} per month was {value}.	Between the period defined by Jesse Chacón making an appeal to Citizen (Venezuela) and Abdoulaye Wade engaging in negotiation with Segolene Royal, the average number of times Japan expressed intent to cooperate with China every month was 4.5.
Most/ Least	Question	Who {predicate} {object} the most times between the period defined by ({subject_a} {predicate_a} {object_a}) and ({subject_b} {predicate_b} {object_b})?	Which person/organization is Host a visit by Mahmoud Abbas the most times between the period defined by UN Security Council Praise or endorse Citizen (Africa) and Agriculture / Fishing / Forestry Ministry (Israel) Sign formal agreement Israeli Defense Forces?
	Claim	Between the period defined by ({subject_a} {predicate_a} {object_a}) and ({subject_b} {predicate_b} {object_b}), {subject} {predicate} {object} the most times.	During the period defined by UN Security Council Praise or endorse Citizen (Africa) and Agriculture / Fishing / Forestry Ministry (Israel) Sign formal agreement Israeli Defense Forces, Mahmoud Abbas Host a visit William E. Ward the most times.
		Simple	
Subject/ Object	Question	Who {predicate} {object} {number} times between {time1} and {time2}?	Who did Cristina Fernández de Kirchner express intent to meet or negotiate with 3 times between 2009-11-30 and 2010-05-19?
	Claim	{Subject} {predicate} {object} {number} times between {time1} and {time2}.	Cristina Fernández de Kirchner expressed intent to meet or negotiate with China 3 times between 2009-11-30 and 2010-05-19.
Count	Question	How many times did {subject} {predicate} {object} between {time1} and {time2}?	How many times did Vietnam host a visit from Ricardo Alarcón de Quesada after 2007-03-06 and before 2007-06-18?
	Claim	{Subject} {predicate} {object} {number} times between {time1} and {time2}.	Vietnam hosted a visit from Ricardo Alarcón de Quesada 2 times after 2007-03-06 and before 2007-06-18.
Average	Question	What is the average number of times {subject} {predicate} {object} every month between {time1} and {time2}?	What is the average number of times South Africa praised or endorsed China every month between 2008-02-15 and 2008-07-21?
	Claim	The average number of times {subject} {predicate} {object} every month between {time1} and {time2} was {value}.	The average number of times South Africa praised or endorsed China every month between 2008-02-15 and 2008-07-21 was 7 times.

Table 7: Comprehensive templates for Questions and Claims for temporal numerical reasoning.

Category	Task	Template Structure	Example
		Hard	
Timeline Selection	Question	Consider the reference period T1 from ({subject_a} {predicate_a} {object_a}) to ({subject_b} {predicate_b} {object_b}). A second timeline T2 ends with the event ({subject_d} {predicate_d} {object_d}). Given that T2 starts exactly when T1 begins, identify the event that **starts** T2.	Consider the reference period T1 from (Christine Lagarde Consult Representatives (France)) to (Gholamali Haddad Adel Express intent to meet or negotiate Elmar Mammadyarov). A second timeline T2 ends with the event (Health Ministry (China) Make statement China). Given that T2 starts exactly when T1 begins, identify the event that starts T2. Options: A. China Express intent to engage in diplomatic cooperation (such as policy support) Japan on 2007-08-30 B. Police (Indonesia) Arrest, detain, or charge with legal action Men (Indonesia) on 2013-05-11 C. Police (Canada) Arrest, detain, or charge with legal action Children (Canada) on 2012-03-15 D. South Korea Host a visit Islam Karimov on 2008-02-26 E. Wen Jiabao Make statement China on 2008-10-18 F. Citizen (Thailand) Express intent to meet or negotiate Thailand on 2007-10-29.
	Claim	Consider T1 from ({subject_a} {predicate_a} {object_a}) to ({subject_b} {predicate_b} {object_b}) and T2 from ({subject_c} {predicate_c} {object_c}) to ({subject_d} {predicate_d} {object_d}). The temporal relation is formally classified as {relation}.	Considering T1 from (Christine Lagarde Consult Representatives (France)) to (Gholamali Haddad Adel Express intent to meet or negotiate Elmar Mammadyarov) and T2 from (Citizen (Thailand) Express intent to meet or negotiate Thailand) to (Health Ministry (China) Make statement China), the temporal relation is formally classified as 'starts'.
Relation Classification	Question	Given the time interval T1 from ({subject_a} {predicate_a} {object_a}) to ({subject_b} {predicate_b} {object_b}) and T2 from ({subject_c} {predicate_c} {object_c}) to ({subject_d} {predicate_d} {object_d}) in {year}, what is the temporal relation between them? Options: equal, meets, met by, overlaps, overlapped by, during, contains, starts, started by, finishes, finished by.	Given the time interval T1 from (Barack Obama Make a visit Japan) to (Iran Return, release person(s) Citizen (United Kingdom)) and T2 from (Japan Host a visit Barack Obama) to (Isaias Afewerki Make an appeal or request Ethiopia) in November 2009, what is the temporal relation between them? Options: equal, meets, met by, overlaps, overlapped by, during, contains, starts, started by, finishes, finished by.
	Claim	Considering T1 from ({subject_a} {predicate_a} {object_a}) to ({subject_b} {predicate_b} {object_b}) and T2 from ({subject_c} {predicate_c} {object_c}) to ({subject_d} {predicate_d} {object_d}), the temporal relation is formally classified as '{relation}'.	Considering T1 from (Barack Obama Make a visit Japan) to (Iran Return, release person(s) Citizen (United Kingdom)) and T2 from (Japan Host a visit Barack Obama) to (Isaias Afewerki Make an appeal or request Ethiopia), the temporal relation is formally classified as 'finished by'.
		Medium	
Before & After	Question	Who {predicate} {object} {number} days before {second_subject} {second_predicate} {second_object}?	Who Consult High Ranking Military Personnel (Canada) 262 days before Ethiopia Praise or endorse Japan?
	Claim	{subject} {predicate} {object} {number} days before {second_subject} {second_predicate} {second_object}.	Benny Gantz Consult High Ranking Military Personnel (Canada) 262 days before Ethiopia Praise or endorse Japan.
In & During	Question	Which organization did {subject} {predicate} when {second_subject} {second_predicate} {second_object}?	Which organization Express intent to accept mediation South Korea when Iraq Reject Barack Obama?
	Claim	{subject} {predicate} {object} when {second_subject} {second_predicate} {second_object}.	Japan Express intent to accept mediation South Korea when Iraq Reject Barack Obama.
Time & Duration	Question	What is the time duration between the first and second time {subject} {predicate} {object} in {year}?	What is the time duration between the first and the second time Naval (Sri Lanka) Use conventional military force Combatant (Sri Lanka) in 2006?
	Claim	In {year}, the time duration between the first and second occurrence of '{subject} {predicate} {object}' was {value}.	In 2006, the time duration between the first and second occurrence of 'Naval (Sri Lanka) Use conventional military force Combatant (Sri Lanka)' was 112 days.
		Simple	
Explicit	Question	Who {predicate} {object} in {time}?	Who Praise or endorse National Transitional Council in 2011-09-16?
	Claim	{subject} {predicate} {object} in {time}.	Vietnam Praise or endorse National Transitional Council in 2011-09-16.
Relative	Question	Who {predicate} {object} {offset} days {direction} {time}?	Who Praise or endorse National Transitional Council 2 days after 2011-09-14?
	Claim	{subject} {predicate} {object} {offset} days {direction} {time}.	Vietnam Praise or endorse National Transitional Council 2 days after 2011-09-14.
Time & Duration	Question	When did {subject} {predicate} {object}?	When did Democratic Party Make an appeal or request Mwai Kibaki?
	Claim	{subject} {predicate} {object} in {time}.	Democratic Party Make an appeal or request Mwai Kibaki in 2015-11-07.

Table 8: Comprehensive templates for Questions and Claims for temporal relational reasoning.