LTLBench: Towards Benchmarks for Evaluating Temporal Reasoning in Large Language Models

Anonymous authors
Paper under double-blind review

Abstract

Temporal Reasoning (TR) is a critical ability for LLMs to understand and reason over temporal information and relationships between events. To study the TR ability in LLMs, prior works provide different ways for evaluating various aspects of TR ability. In this work, we propose an alternative perspective for evaluating TR ability by leveraging Linear Temporal Logic (LTL), and develop a pipeline to automatically synthesize challenges for assessing the TR ability of LLMs. Based on this pipeline, we construct a dataset, namely LTLBench, consisting of 2000 TR challenges, and benchmark 12 LLMs across 5 different methods. Furthermore, we conduct additional experiments to investigate the impact of increasing the number of formula operators and events on both LLM performance and the complexity of TR problems. We also perform qualitative analyses of their reasoning processes and the effects of varying the number of events and formula operators, which reveal 3 main issues in their temporal reasoning processes and the unexpected performance changes observed as problem complexity increases. We expect this work to provide valuable insights into the TR ability of LLMs[1].

1 Introduction

Temporal Reasoning (TR) is a critical reasoning ability of LLMs, encompassing the understanding, processing, and reasoning over temporal information and relationships between events, which is essential for solving problems across diverse scenarios (Shoham & Goyal, 1988; Chittaro & Montanari, 2000; Vila, 1994). Prior studies have demonstrated that although LLMs show some promise in TR, they still struggle with TR and a substantial performance gap persists between the state-of-the-art LLMs and humans on TR (Chu et al., 2023; Wang & Zhao, 2024; Beniwal et al., 2024). Nevertheless, since TR encompasses various aspects, existing investigations of the TR ability of LLMs remain incomplete. Therefore, in this work, we adopt Linear Temporal Logic to explore TR from a new perspective, focusing on formal logical reasoning over temporal information.

Linear Temporal Logic (LTL) is a formal logic, specifically a modal temporal logic, that is widely used and studied for expressing and reasoning about sequences of events over time (Kröger & Merz, 2008; Goranko & Rumberg, 2024). Although complex LTL is typically discussed and used in the context of formal and program verification, basic and moderately complex LTL tasks are, in fact, ubiquitous in daily tasks. For example, if people are out of milk, they will eventually buy it, which can be formalized in an LTL formula as $G(\text{OutOfMilk} \rightarrow F(\text{BuyMilk}))$. Likewise, if the traffic light is green, it will then turn to yellow, which can be formalized as $G(\text{Green} \rightarrow X(\text{Yellow}))$. From this perspective, LTL provides a natural and formal way to represent and operate on temporal relations between events. Therefore, we leverage LTL as a main component to construct TR problems.

To explore the TR ability of LLMs, we propose a novel approach, a TR challenges generation pipeline, to automatically synthesize TR problems. Each generated TR problem mainly consists of a context that depicts the situation of a TR problem and a hypothesis that requires LLMs to judge its validity against the given context. The core components of the pipeline involve a randomly generated directed graph that serves as

[1]Our code is open-sourced and available at `https://anonymous.4open.science/r/LTLBench-Anonymous/`.

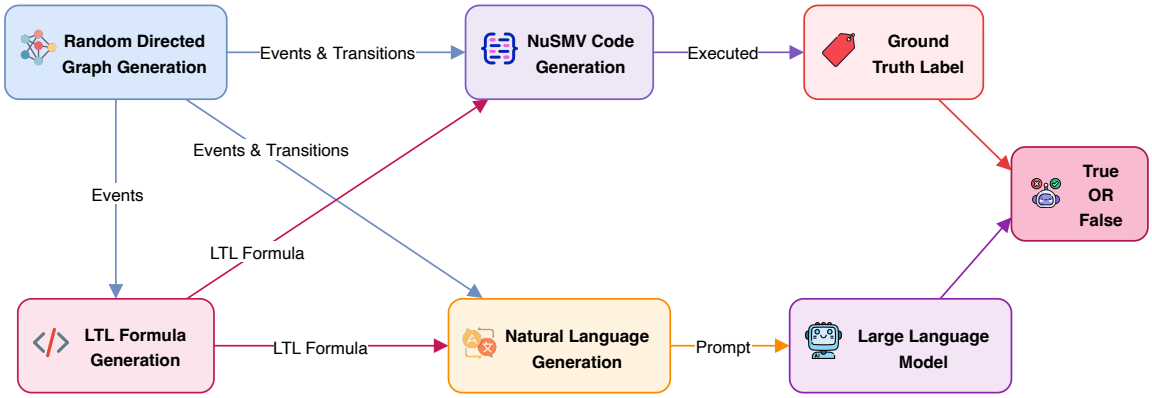


Figure 1: The overview of the TR problem generation pipeline.

the preparation for subsequent problem generations, a randomly generated LTL formula that acts as the hypothesis for the given context, and the NuSMV model checker (Cimatti et al., 2002) which executes the code of events transitions and the LTL formula to provide the ground truth label. As shown in Figure 1, during the generation process for a TR problem, we first generate a random directed graph. Then, we adopt and slightly modify the LTL formulas generation algorithm designed by Zhu (2021) to generate an LTL formula based on the events given in the graph. Subsequently, both the event information and the LTL formula are converted into NuSMV (Cimatti et al., 2002) code and executed to obtain the ground truth label. Finally, the event information and the LTL formula are translated into a TR problem in the form of natural language.

Furthermore, to conduct an intensive and comprehensive evaluation, we generate a total of 2000 TR challenges using our proposed pipeline, referred to as LTLBench, and we evaluate 12 LLMs across 5 different methods. We not only demonstrate their TR ability but also provide several qualitative insights that reveal 3 main issues they fail to reason over these challenges. In addition, we conduct additional experiments to investigate the impact of varying the number of events and operators on both the performance of LLMs and the complexity of TR problems. The key contributions of our study are summarized as follows:

1. We develop a novel TR challenges generation pipeline, which lays on Linear Temporal Logic, to evaluate the TR ability of LLMs from a new perspective, focusing on formal logical reasoning over temporal information;
2. Based on the pipeline, we construct a dataset, LTLBench, consisting of 2000 TR challenges, and evaluate 12 LLMs across 5 different methods, providing both quantitative results and qualitative insights which reveal 3 main issues of their reasoning failures;
3. We further conduct two additional experiments to demonstrate that increasing the number of formula operators and events leads to more challenges for LLMs, and offer qualitative insights to discuss the unexpected performance changes observed as problem complexity increases.

2 Related Work

2.1 TR in LLMs

Temporal Reasoning in LLMs has recently obtained substantial attention (Xiong et al., 2024; Fatemi et al., 2024; Beniwal et al., 2024; Chu et al., 2023; Hu et al., 2023; Liu et al., 2023; Vashishtha et al., 2020). Beniwal et al. (2024) point out notable deficiencies of LLMs in understanding and reasoning over temporal information and reasoning, while Xiong et al. (2024) propose TG-LLM, a framework aimed at improving LLMs performance on TR tasks. Furthermore, Liu et al. (2025) introduce a Time-R1, which employs

reinforcement learning fine-tuning to enhance temporal reasoning ability of LLMs, enabling smaller models to match or even surpass larger ones. While these works collectively highlight the progress and importance of TR ability in LLMs, the exploration of TR ability in LLMs remains ongoing, requiring further efforts to reveal and understand the boundaries.

2.2 TR Benchmarks

To discover and evaluate the TR ability of LLMs, a variety of TR datasets and benchmarks have been proposed, targeting different aspects of TR at varying levels of complexity (Fatemi et al., 2024; Wang & Zhao, 2024; Xiong et al., 2024; Beniwal et al., 2024; Qin et al., 2021; Tan et al., 2023; Virgo et al.). For example, Xiong et al. (2024) construct a TR dataset by leveraging a large temporal knowledge graph, YAGO11k (Dasgupta et al., 2018), and utilizing GPT-3.5 and rule-based Python scripts to generate TR challenges. Additionally, Fatemi et al. (2024) employ random graph generation as a foundation and preparation to form rule-based and different types of temporal facts and questions without introducing LLMs to generate TR tasks, focusing on temporal semantics and arithmetic reasoning and proposing a benchmark called Test of Time. Furthermore, Wang & Zhao (2024) introduce a TR dataset consisting of various temporal aspects such as order, arithmetic, frequency, and duration. While these approaches provide valuable perspectives for evaluating TR ability, they generally lack systematic support for representing and reasoning with temporal operators such as *eventually* and *always*, as well as their complex compositions. By contrast, our work that leverages LTL enables a natural support over those temporal operators and can explore TR ability from another perspective.

3 TR Problem Generation Pipeline

The pipeline to generate a single TR problem consists of four stages: (1) Random Directed Graph Generation, (2) LTL Formula Generation, (3) NuSMV Code Generation, and (4) Natural Language Generation. We demonstrate the overview of the process for a TR problem generation in Figure 1.

3.1 Random Directed Graph Generation

During this stage, a directed graph is randomly generated with a given number of events n where $n > 1$ to ensure the formation of transitions between events.

In this graph, each $node_i$ represents an individual $event_i$, and each $edge_j^i$ is a directed edge pointing from $node_i$ to another $node_j$, which forms the relationships and transitions between events, indicating that $event_j$ represented by $node_j$ occurs after the $event_i$ represented by $node_i$. It is important to note that each $node_i$ within the graph can have multiple outgoing edges, signifying that several subsequent events can follow $event_i$, as well as multiple incoming edges, indicating that $event_i$ can be preceded by several other events. To form a clear semantics aligned with the follow-up stages, for an event $event_i$ followed by only one event $event_j$, we say that after $event_i$, $event_j$ must happen, while for an event $event_i$ followed by more than one event $event_1, event_2, \ldots, event_n$, we say that after $event_i$, either $event_1, event_2, \ldots,$ or $event_n$ must happen.

As an example illustrated in Figure 2, given $n = 3$, three events are generated: $event_1$, $event_2$, and $event_3$. The case that $event_1$ points to $event_2$ indicates that $event_2$ must happen after $event_1$. The case that $event_1$ points to $event_3$ and $event_3$ also points to $event_1$ means that $event_1$ must happen after $event_3$ and $event_3$ must happen after $event_1$. In addition, to note, $event_1$ not only points to $event_2$ but also $event_3$, indicating that either $event_2$ or $event_3$ must happen after $event_1$. By contrast, if $event_1$ only points to $event_2$, it indicates that $event_2$ must happen after $event_1$.

The generation of the random directed graph during this stage serves as the foundation and preparation for generating the LTL formula, NuSMV code, and TR problem represented in natural language by providing the information of events and also the transitions between events.

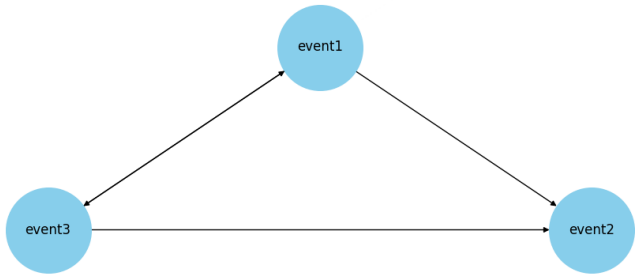


Figure 2: An example of a generated random directed graph.

```
(event1 -> (G (F event2)))
```

Listing 1: An example of a generated LTL formula.

3.2 LTL Formula Generation

Based on the events generated in the graph, we employ the algorithm designed by Zhu (2021) to generate an LTL formula with a given number of operators m where $m > 0$. The LTL operators include unary and binary operators. Unary operators, for example, include but are not limited to X which indicates that for a given event ϕ, $X\phi$ denotes that the event ϕ will occur at the next moment, and F for which $F\phi$ means that event ϕ will eventually occur at some point in the future. Binary operators include but are not limited to & representing logical AND and | representing logical OR. The given number of operators refers to the number of unary and binary LTL operators contained in an LTL formula.

An example of a generated LTL formula is also provided and shown in Listing 1. The formula means that if $event_1$ happens, it will be globally true that, at some point in the future, $event_2$ will eventually happen.

The generation of the LTL formula aims to be the preparation for generating the LTLSPEC part of the NuSMV code and also to generate the hypothesis part of the TR problem represented in natural language.

3.3 NuSMV Code Generation

Given the information of events from the graph and the LTL formula, this stage converts them into NuSMV code, which consists of two parts: (1) context generation and (2) LTLSPEC generation. The context describes the situation of the TR problem, while LTLSPEC represents a hypothesis regarding the context.

For context generation, it includes event definitions, initial event setup, and event transitions setup. Based on the generated graph, events and their transitions are converted into the context part of the NuSMV code, and an initial event is randomly selected. As shown in Listing 2, Lines 2-3 define three events, Line 5 specifies the initial event which is $event_3$, and Lines 6-11 construct the event transitions, in which, for example, Line 8 indicates that $event_2$ and $event_3$ can follow $event_1$, while $event_2$ remains to itself and after $event_2$, no other events can happen.

For LTLSPEC generation, the generated LTL formula is translated into the NuSMV code, as illustrated at Line 13 in Listing 2 which represents the LTL formula shown in Listing 1.

The resulting NuSMV code is executed by the NuSMV model cheker during the generation process to obtain the ground truth label for the TR problem.

```
MODULE main
VAR
    state : {event1, event2, event3};
ASSIGN
    init(state) := event3;
    next(state) := case
        state = event1 : {event2, event3};
        state = event2 : event2;
        state = event3 : {event1, event2};
    esac;
LTLSPEC ((state=event1) -> (G (F (state=event2))))
```

Listing 2: An example of NuSMV code.

```
1 === Context ===
2
3 Initially, event3 happened. After event1, either event2 or event3 must happen. After event2,
    ↪ no other events can happen. After event3, either event1 or event2 must happen.
4
5 === Hypothesis ===
6
7 C1: Event2 eventually happens.
8 C2: C1 always holds.
9 C3: If event1 happens, then C2 holds.
10
11 C3 is True or False?
```

Listing 3: An example of TR problem in the form of natural language.

3.4 Natural Language Generation

During this stage, the events information in the graph and the LTL formula are converted into natural language. Similar to the NuSMV code generation, this stage consists of two parts: (1) context generation and (2) hypothesis generation. The context describes the problem situation while the hypothesis specifies what LLMs are required to determine given the context.

For context generation, based on the generated graph and the initial event, event information and transitions are converted into natural language. As shown at Lines 1-4 in Listing 3, the initial event and events transitions are represented in natural language.

For hypothesis generation, the generated LTL formula is transformed into natural language as shown at Lines 5-10 in Listing 3. Additionally, Line 11 is used to prompt LLMs to determine the validity of the hypothesis.

The TR problem represented in natural language generated in this stage is the core and final product of the generation process, which will be used to evaluate the TR ability of LLMs.

4 Experiment Settings

To evaluate the TR ability of LLMs, based on the pipeline, we construct a dataset, LTLBench, consisting of 2000 problems. Each problem is generated with a fixed number of events $n = 3$ and formula operators $m = 3$. Additionally, to explore the impact of changes in the number of formula operators, we conduct evaluations on newly generated problems with a fixed number of events $n = 2$ while varying the number of formula operators $m \in \{1, 2, 3, 4, 5, 7, 9\}$. For each (n, m_i), such as $(2, 1)$ indicating that the number of events is 2 and the number of operators is 1, we generate 300 problems as a dataset for evaluation. Similarly, to examine the effects of the number of events, we fix the number of formula operators $m = 2$ and vary the number of events $n \in \{2, 3, 4, 5, 7, 9\}$. For all generated datasets, their distributions of ground truth labels are meticulously balanced, with half of the problems labeled as *True* and the other half of the problems

as *False*. Furthermore, we mainly adopted Accuracy, defined as the proportion of the correctly answered problems, as the primary evaluation metric.

For comprehensive evaluations, we select a total of 12 LLMs. Specifically, we evaluate DeepSeek-V3 (DeepSeek-AI et al., 2025) in both non-thinking mode and thinking mode, which are officially noted as *DeepSeek-Chat* model and *DeepSeek-Reasoner* model. We also include 3 OpenAI models, which are *GPT-3.5-Turbo* (Brown et al., 2020), *GPT-4o-Mini* (OpenAI et al., 2024), and *GPT-5-Mini*. In addition, we select 4 Qwen models (Yang et al., 2025; Qwen et al., 2025), namely *Qwen-Turbo*[2], *Qwen3:32B*, *Qwen3:14B*, and *Qwen2.5:72B-Instruct*. Furthermore, we also include 3 additional models, which are *Gemma3:12B-Instruct* (Team et al., 2025), *Mistral:7B-Instruct* (Jiang et al., 2023), and *Phi4:14B* (Abdin et al., 2024).

Moreover, to examine how different methods affect LLMs performance on TR challenges, we adopt 5 methods, which are *Direct Prompting*, *Zero-Shot CoT* (Kojima et al., 2022), *Few-Shot CoT* (Wei et al., 2022), *Self-Consistency* (Wang et al., 2023), and *Least-to-Most* (Zhou et al., 2022). *Direct Prompting* directly queries an LLM to generate answer in *True* or *False* for a given TR problem. *Zero-Shot CoT* prompts an LLM to reason step by step without examples, whereas *Few-Shot CoT* provides 2 examples with each consisting of a TR problem, thinking process, and final answer. *Self-Consistency* extends *Zero-Shot CoT* by sampling 3 reasoning paths and taking majority voting to select the final answer. In *Least-to-Most*, an LLM is first prompted to answer a set of breakdown smaller questions and is then prompted again with the response to previous questions to reason on the main problem to give the final answer. For *Direct Prompting*, *Zero-Shot CoT*, *Few-Shot CoT*, and *Least-to-Most* methods, we set the *temperature* to 0 and the *max completion tokens* to 2000 for all models. For *Self-Consistency*, since it requires diversity of candidates, we set the *temperature* to 0.7 and the *max completion tokens* to 2000 for all models[3]. The prompt templates for *Direct Prompting*, *Zero-Shot CoT*, and *Few-Shot CoT* are provided in Prompt Template 1, 2, and 3 in Appendix A, respectively, with *Self-Consistency* using the same prompt template as *Zero-Shot CoT*. The prompt templates for the two stages of *Least-to-Most* are shown in Prompt Template 4 and 5 in Appendix A.

5 Results and Analyses

5.1 Evaluation with LTLBench

We evaluate the selected 12 LLMs on LTLBench consisting of 2000 generated TR problems with the number of events $n = 3$ and formula operators $m = 3$. The results of all models equipped with different methods are reported in Table 1. We also demonstrate the performance of each model with different methods in a grouped horizontal bar chart as shown in Figure 3 for a better visual comparison of the results.

From the results, among all models and methods, the best one is the *GPT-5-Mini* with the *Few-Shot CoT* method, achieving an accuracy of 93.95%, whereas the worst one is the *GPT-3.5-Turbo* with the *Least-to-Most* method, whose accuracy of 51.05% is only slightly above random guessing. In addition, the average accuracies of each method across models are 65.83% for *Direct Prompting*, 67.18% for *Zero-Shot CoT*, 76.33% for *Few-Shot CoT*, 68.29% for *Self-Consistency*, and 68.21% for *Least-to-Most*, in which *Few-Shot CoT* demonstrates its significant performance improvement while the other three methods slightly outperform the *Direct Prompting*. As demonstrated in both Table 1 and Figure 3, *Few-Shot CoT* consistently and significantly outperforms other methods regardless of models. In addition, *Self-Consistency* tends to fail to further improve the performance of recent high-performing models regardless of the parameter sizes, such as *DeepSeek-Reasoner*, *DeepSeek-Chat*, *GPT-5-Mini*, *GPT-4o-Mini*, *Qwen-Turbo*, *Qwen3:32B*, *Qwen3:14B*, and *Phi4:14B*, while for earlier models such as *Mistral:7B-Instruct*, *GPT-3.5-Turbo*, and *Qwen2.5:72B-Instruct*, it can still significantly enhance their performance. We consider the reason is that after multiple iterations of LLMs, recent LLMs approach a problem in a more self-consistent way, so that the *Self-Consistency* method may fail to further improve the performance on these models. In addition, across most models, *Zero-Shot CoT* can only marginally increase the performance, which could be attributed to that most of the evaluated models have already incorporated reasoning or thinking instructions during their pretraining or instruction-tuning process, such as DeepSeek series models, OpenAI's GPT-5 series models, and also Qwen3 series

[2]The exact version of *Qwen-Turbo* is *Qwen-Turbo-2025-04-28*.

[3]For the GPT-5 series models, the *temperature* is not settable, so we do not use *temperature* for *GPT-5-Mini*.

Table 1: The accuracy (%) of LLMs equipped with different methods evaluated on LTLBench, in which DP stands for *Direct Prompting*, ZS CoT for *Zero-Shot CoT*, FS CoT for *Few-Shot CoT*, SC for *Self-Consistency*, and L2M for *Least-to-Most*, and the best accuracy across models for each method is highlighted in bold.

Model	**DP**	**ZS CoT**	**FS CoT**	**SC**	**L2M**
DeepSeek-Reasoner	71.55	69.95	81.10	73.70	72.75
Deepseek-Chat	**80.10**	**78.65**	90.40	79.40	**81.05**
GPT-5-Mini	79.70	78.45	**93.95**	**80.53**	77.75
GPT-4o-Mini	63.50	62.60	71.55	63.25	63.70
GPT-3.5-Turbo	53.25	56.55	60.40	57.10	51.05
Qwen-Turbo	66.50	67.60	78.70	68.90	69.60
Qwen2.5:72B-Instruct	59.05	68.10	76.05	67.75	69.90
Qwen3:32B	66.45	66.65	76.50	68.10	70.45
Qwen3:14B	67.45	67.45	75.15	67.80	71.25
Phi4:14B	66.65	65.60	73.50	65.90	66.15
Gemma3:12B-Instruct	60.15	67.20	78.40	68.20	67.20
Mistral:7B-Instruct	55.60	57.35	60.25	59.00	57.70

models, so that even in the case of using *Direct Prompting*, they automatically invoke the chain-of-thought to reason over the problems, as our observation of their responses.

In addition, notably, we observe that *DeepSeek-Chat* outperforms *DeepSeek-Reasoner*. Typically, we believe models equipped with enhanced and polished reasoning ability should outperform ones without it. From our perspective, this unexpected result may arise because our LTL-based TR challenges require a more complex formal logic reasoning along with temporal reasoning, in which the reason steps may be different from other challenges without formal logic injected in such as Test of Time (Fatemi et al., 2024) and may cause untypical reasoning steps. The reasoning steps that *DeepSeek-Reasoner* are trained on during pretraining and perform during inference may not align with the reasoning steps demanded by LTL-based TR problems. Thereby, although it is explicitly asked to think in a specific way to tackle LTL-based TR problems, the pretrained distribution of output tokens of reasoning steps is not suitable for this kind of problem and could even further disturb the reasoning steps to some extent, leading to even worse than equipping non-thinking or non-reasoning models with explicit task-specific CoT methods.

Moreover, to get more insights into the reasoning process of LLMs on the TR challenges, we further conduct a qualitative analysis. For each model paired with each method, we randomly select 10 problem and answer pairs, thus resulting in a total of 600 pairs. We then filter out all pairs with correct answers and finally retain 145 pairs with incorrect answers, to focus on analyzing the potential factors that lead LLMs to fail in answering problems correctly. Among these 145 pairs, we identify three major findings. The first is *Temporal Semantics and Reasoning Misalignment*, which indicates that although LLMs demonstrate an understanding of temporal semantics during reasoning steps, they may still apply these semantics incorrectly when performing actual reasoning. A representative example is observed in the results of *GPT-4o-Mini* with the *Few-Shot CoT* method, in which although it understands that after *event 1*, the *event 3* must happen, which is then followed by *event 2*, when it reasons about whether *event 2* happens at next state after *event 1*, it does not precisely capture that the exact next state is *event 3* but it reasons in a misaligned way such that since that *event 1* is followed by *event 3* and *event 3* is followed by *event 2*, *event 2* happens at next state of *event 1*, which is however incorrect[4]. The second is *Context Hypothesis Detachment*, which indicates that LLMs largely ignore the context and only leverage information from the hypothesis to perform reasoning. A representative example is observed from the results of *Qwen-Turbo* with the *Direct Prompting* method, in which since *C1* and *C2* are identical, when asked whether both *C1* and *C2* hold, it reasons in a shortcut way without leveraging context information about the transitions of events, and directly gives the answer that *C3* is true, which however is not. We observe this happens in several models such as *Qwen-*

[4]An example of *Temporal Semantics and Reasoning Misalignment* is demonstrated in Listing 4 in Appendix A.

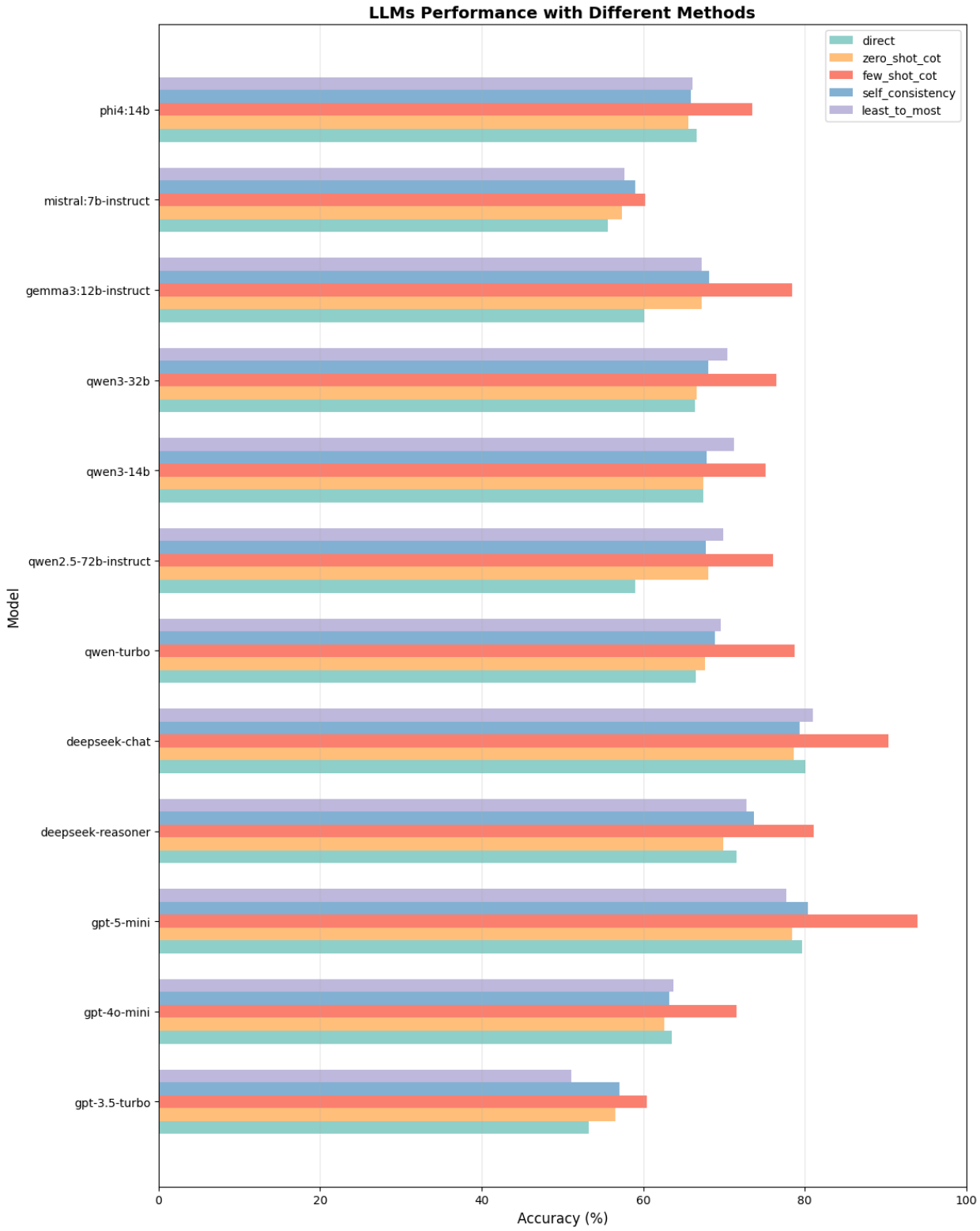


Figure 3: LLMs Performance with Different Methods.

Turbo, *GPT-4o-Mini*, *Mistral:7B-Instruct*, etc[5]. The third is *Reasoning Error Amplification*, which means

[5]An example of *Context Hypothesis Detachment* is demonstrated in Listing 5 in Appendix A.

the previous reasoning errors without retrospection for correcting lead to the later reasoning ignoring the errors and using the previous incorrect steps or conclusions to future reason, resulting in that the reasoning errors are amplified during the reasoning process. This is a normal and typical error when LLMs perform reasoning and we observe it in the reasoning steps of several models such as *GPT-3.5-Turbo*, *DeepSeek-Chat*, *Qwen3:14B*, etc., which is due to the methods equipped by LLMs lacks of ability for retrospection. This case is not only observed in this work but also reported in prior works (Zhu et al., 2025; Feng et al., 2025; Tyen et al., 2024).

5.2 Impact of Increasing m

We select 3 LLMs, namely *DeepSeek-Reasoner*, *GPT-5-Mini*, and *Qwen-Turbo*, and evaluate them on the additional constructed datasets, where the number of events is fixed to $n = 2$ while the number of formula operators m increases from 1 to 9, specifically $m \in \{1, 2, 3, 4, 5, 7, 9\}$. This aims to explore whether the performance of LLMs on TR challenges is stable as the number of formula operators increases and also whether the increase of the number of formula operators can introduce more complexity.

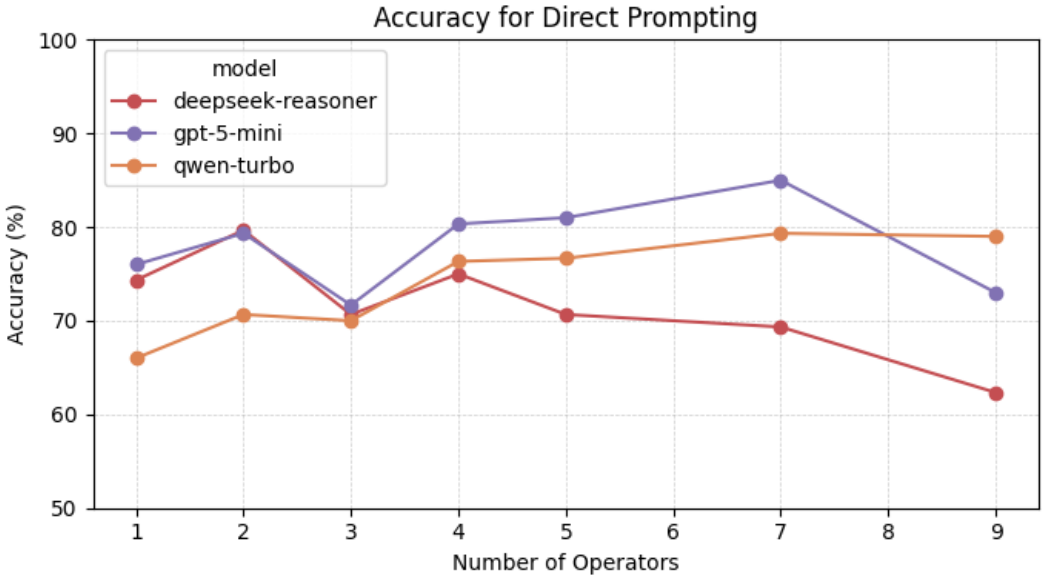
The results of the models equipped with each method are demonstrated in Figure 4. While four methods exhibit an oscillation of accuracy while $m < 7$, the *Few-Shot CoT* method maintains a stable and robust performance. Furthermore, a sudden accuracy drop is observed at $m = 3$ for the other four methods. In particular, *Direct Prompting* shows an accuracy decrease at $m = 3$ followed by an increase up to $m = 7$, indicating poorer performance on simpler problems but improved performance on harder ones, which is counter-intuitive. Thus, to further investigate the reason, we conduct a qualitative analysis. We find that when $m <= 3$, the problems are relatively simple and the LLMs do not automatically and frequently invoke explicit reasoning to arrive at answers, in which while $m < 3$, the problems could be easy for LLMs to solve even without explicit reasoning, whereas while $m = 3$, the problems become hard for the LLMs and they still fail to invoke explicit reasoning process or they put less effort in reasoning process, leading to worse performance, which is a case that especially happens frequently in *Direct Prompting*. However, while $m > 3$, the LLMs invoke reasoning more frequently and take more effort in the reasoning process, resulting in the increase of accuracy.

In addition, while $m \geq 7$, the accuracy shows an obvious decreasing trend. From $m = 7$ to $m = 9$, grouped by methods across models, accuracy decreases by 6.44% in *Direct Prompting*, 7.44% in *Zero-Shot CoT*, 3.89% in *Few-Shot CoT*, 6.22% in *Self-Consistency*, 7.11% in *Least-to-Most*, while grouped by models across methods, accuracy decreases by 6.87% in *DeepSeek-Reasoner*, 7.53% in *GPT-5-Mini*, and 4.27% in *Qwen-Turbo*. Overall, increasing the number of formula operators from 7 to 9 results in an average accuracy drop of 6.22%, with a minimum drop of 0.3% and the maximum drop of 12%. This indicates that the increase of the number of formula operators can significantly introduce more complexity to the problems and also shows that the TR ability of LLMs lacks consistency and robustness as TR problems complexity grows due to involving more formula operators.

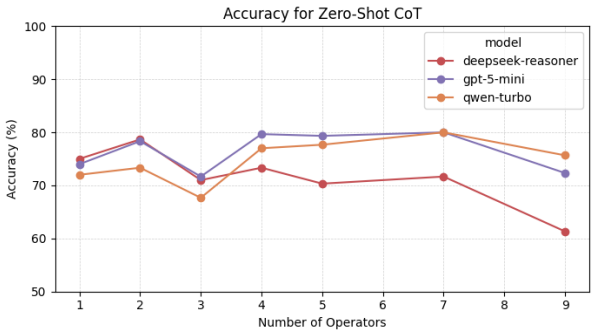
5.3 Impact of increasing n

In addition, the selected 3 models are evaluated on another additional constructed datasets, where the number of formula operators is fixed to $m = 2$ while the number of events n, where $n > 1$, increases from 2 to 9, specifically $n \in \{2, 3, 4, 5, 7, 9\}$. Similar to Section 5.2, this, however, aims to explore how the TR ability of LLMs is affected as the number of events increases and also whether the increase of the number of events can lead to more complexity.

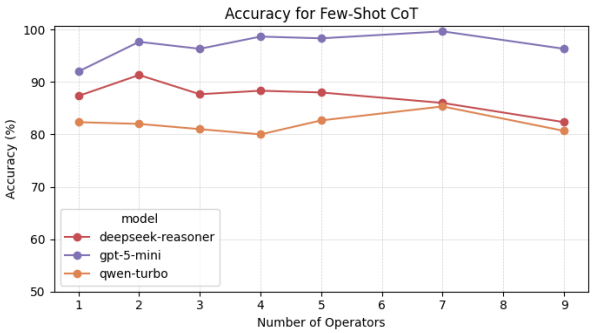
As shown in Figure 5, across all models and methods, they show a more consistent trend of accuracy decrease as the number of events increases. From $n = 2$ to $n = 9$, grouped by methods across models, accuracy drops by 16.22% in *Direct Prompting*, 18.78% in *Zero-Shot CoT*, 19.44% in *Few-Shot CoT*, 18% in *Self-Consistency*, and 17% in *Least-to-Most*, while grouped by models across methods, accuracy decreases by 19.80% in *DeepSeek-Reasoner*, 16.20% in *GPT-5-Mini*, and 17.67% in *Qwen-Turbo*. Overall, increasing the number of events from 2 to 9 results in an average accuracy drop of 17.89%, with a minimum drop of 11% and a maximum drop of 24.67%. This decline is more pronounced than that observed when increasing the number of formula operators, indicating that the increase of the number of events can considerably lead



(a) Direct Prompting

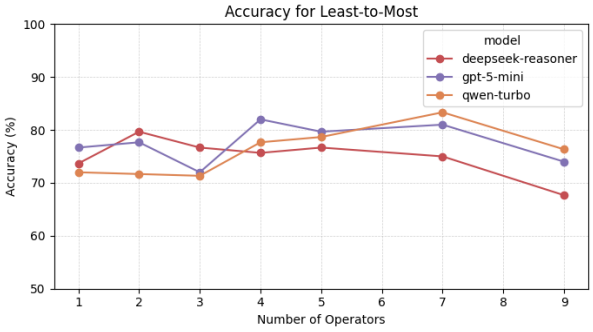


(b) Zero-Shot CoT



(c) Few-Shot CoT

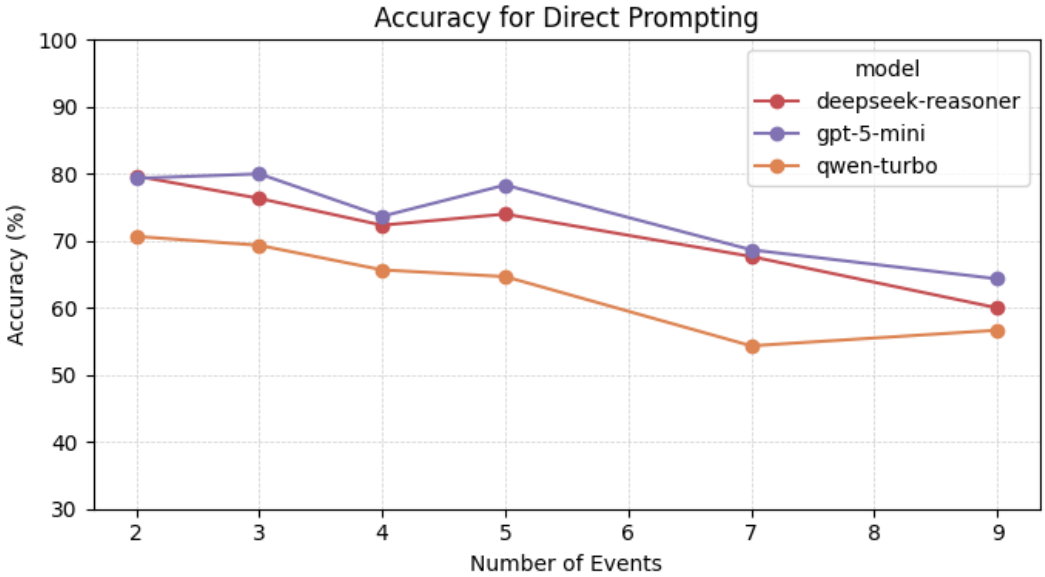
(d) Self-Consistency



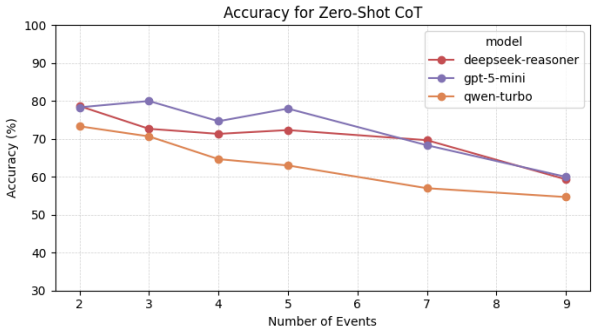
(e) Least-to-Most

Figure 4: Accuracy for *DeepSeek-Reasoner*, *GPT-5-Mini*, and *Qwen-Turbo* equipped with each method as the number of formula operators increases.

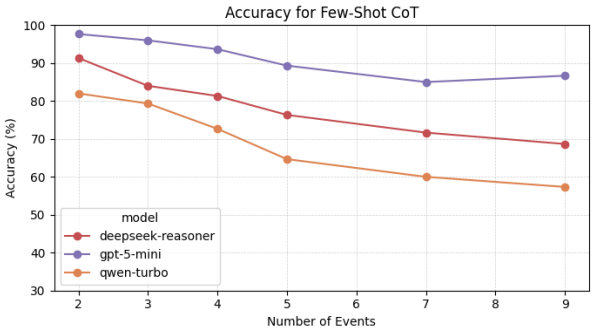
to more complexity of TR problems and that the TR ability of LLMs is unstable and not robust as the complexity of TR problems increases due to involving more events.



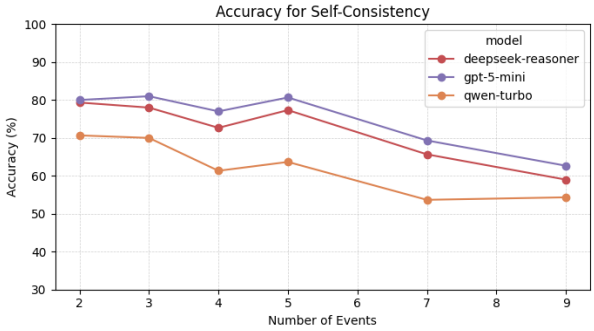
(a) Direct Prompting



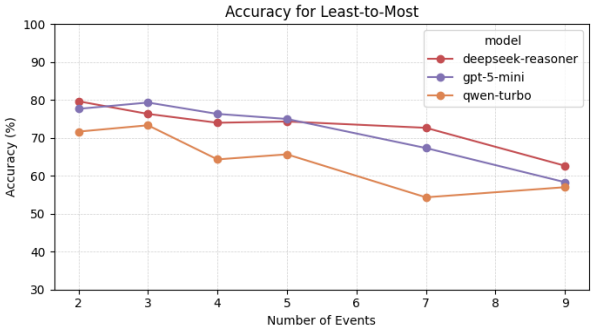
(b) Zero-Shot CoT



(c) Few-Shot CoT



(d) Self-Consistency



(e) Least-to-Most

Figure 5: Accuracy for *DeepSeek-Reasoner*, *GPT-5-Mini*, and *Qwen-Turbo* equipped with each method as the number of events increases.

6 Conclusion

In this work, we evaluate TR ability of LLMs from the perspective of formal logical reasoning over temporal information and design a pipeline for TR challenges generation, by leveraging random graph generation,

LTL formula, and the NuSMV model checker. Using this pipeline, we construct a dataset, LTLBench, consisting of 2000 TR challenges, and conduct intensive evaluations on 12 LLMs across 5 methods. The results demonstrate the limitations of LLMs in handling the TR problem. We further qualitatively analyze their reasoning processes and identify 3 main reasoning issues. In addition, with additional experiments, we demonstrate that their performance is unstable and not robust as the number of formula operators and events increases, and also show that the pipeline can synthesize TR challenges in various levels of complexity and size.

Limitations

This work mainly leverages LTL to generate TR problems, of which the expression ability is limited at linear event sequences, lacking of expression ability to handle events in branching-time. Thus, a natural step of the future direction is to utilize Computation Tree Logic (CTL) or CTL* to further construct more complicated TR problems. For example, in CTL, we can use $AE\phi$ to express that for all event sequences, ϕ eventually holds. We believe integrating the CTL and CTL* into this pipeline could further explore the TR ability boundaries of LLMs.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. Phi-4 technical report, 2024. URL https://arxiv.org/abs/2412.08905.

Himanshu Beniwal, Dishant Patel, Kowsik Nandagopan D, Hritik Ladia, Ankit Yadav, and Mayank Singh. Remember this event that year? assessing temporal information and understanding in large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 16239–16348, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.953. URL https://aclanthology.org/2024.findings-emnlp.953/.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

Luca Chittaro and Angelo Montanari. Temporal representation and reasoning in artificial intelligence: Issues and approaches. *Annals of Mathematics and Artificial Intelligence*, 28(1):47–106, 2000.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Haotian Wang, Ming Liu, and Bing Qin. Timebench: A comprehensive evaluation of temporal reasoning abilities in large language models. (arXiv:2311.17667), November 2023. URL http://arxiv.org/abs/2311.17667. arXiv:2311.17667 [cs].

A. Cimatti, E. Clarke, E. Giunchiglia, F. Giunchiglia, M. Pistore, M. Roveri, R. Sebastiani, and A. Tacchella. NuSMV Version 2: An OpenSource Tool for Symbolic Model Checking. In *Proc. International Conference on Computer-Aided Verification (CAV 2002)*, volume 2404 of *LNCS*, Copenhagen, Denmark, July 2002. Springer.

Shib Sankar Dasgupta, Swayambhu Nath Ray, and Partha Talukdar. HyTE: Hyperplane-based temporally aware knowledge graph embedding. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2001–2011, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1225. URL https://aclanthology.org/D18-1225.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. Deepseek-v3 technical report, 2025. URL https://arxiv.org/abs/2412.19437.

Bahare Fatemi, Mehran Kazemi, Anton Tsitsulin, Karishma Malkan, Jinyeong Yim, John Palowitch, Sungyong Seo, Jonathan Halcrow, and Bryan Perozzi. Test of time: A benchmark for evaluating llms on temporal reasoning, 2024. URL https://arxiv.org/abs/2406.09170.

Yiyang Feng, Yichen Wang, Shaobo Cui, Boi Faltings, Mina Lee, and Jiawei Zhou. Unraveling misinformation propagation in LLM reasoning. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 11683–11707, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.627. URL https://aclanthology.org/2025.findings-emnlp.627/.

Valentin Goranko and Antje Rumberg. Temporal Logic. In Edward N. Zalta and Uri Nodelman (eds.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2024 edition, 2024.

Xuming Hu, Junzhe Chen, Xiaochuan Li, Yufei Guo, Lijie Wen, Philip S. Yu, and Zhijiang Guo. Do large language models know about facts?, 2023. URL https://arxiv.org/abs/2310.05177.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. (arXiv:2310.06825), October 2023. URL http://arxiv.org/abs/2310.06825. arXiv:2310.06825 [cs].

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Fred Kröger and Stephan Merz. *Temporal logic and state systems*. Texts in theoretical computer science: An EATCS series. Springer, Berlin, 2008. ISBN 978-3-540-67401-6.

Jason Xinyu Liu, Ziyi Yang, Ifrah Idrees, Sam Liang, Benjamin Schornstein, Stefanie Tellex, and Ankit Shah. Grounding complex natural language commands for temporal tasks in unseen environments. (arXiv:2302.11649), October 2023. URL http://arxiv.org/abs/2302.11649. arXiv:2302.11649 [cs].

Zijia Liu, Peixuan Han, Haofei Yu, Haoru Li, and Jiaxuan You. Time-r1: Towards comprehensive temporal reasoning in llms, 2025. URL https://arxiv.org/abs/2505.13508.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla

Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

Lianhui Qin, Aditya Gupta, Shyam Upadhyay, Luheng He, Yejin Choi, and Manaal Faruqui. Timedial: Temporal commonsense reasoning in dialog. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 7066–7076, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.549. URL https://aclanthology.org/2021.acl-long.549.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Yoav Shoham and Nita Goyal. Chapter 11 - temporal reasoning in artificial intelligence. In Howard E. Shrobe and the American Association for Artificial Intelligence (eds.), *Exploring Artificial Intelligence*, pp. 419–438. Morgan Kaufmann, 1988. ISBN 978-0-934613-67-5. doi: https://doi.org/10.1016/B978-0-934613-67-5.50015-0. URL https://www.sciencedirect.com/science/article/pii/B9780934613675500150.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 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)*, pp. 14820–14835, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.828. URL https://aclanthology.org/2023.acl-long.828.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh,

Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL `https://arxiv.org/abs/2503.19786`.

Gladys Tyen, Hassan Mansoor, Victor Carbune, Peter Chen, and Tony Mak. LLMs cannot find reasoning errors, but can correct them given the error location. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13894–13908, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.826. URL `https://aclanthology.org/2024.findings-acl.826/`.

Siddharth Vashishtha, Adam Poliak, Yash Kumar Lal, Benjamin Van Durme, and Aaron Steven White. Temporal reasoning in natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4070–4078, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.363. URL `https://www.aclweb.org/anthology/2020.findings-emnlp.363`.

Lluis Vila. A survey on temporal reasoning in artificial intelligence. *Ai Communications*, 7(1):4–28, 1994.

Felix Virgo, Fei Cheng, and Sadao Kurohashi. Improving event duration question answering by leveraging existing temporal information extraction data.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL `https://openreview.net/forum?id=1PL1NIMMrw`.

Yuqing Wang and Yun Zhao. Tram: Benchmarking temporal reasoning for large language models. (arXiv:2310.00835), May 2024. URL `http://arxiv.org/abs/2310.00835`. arXiv:2310.00835 [cs].

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Siheng Xiong, Ali Payani, Ramana Kompella, and Faramarz Fekri. Large language models can learn temporal reasoning. (arXiv:2401.06853), June 2024. URL `http://arxiv.org/abs/2401.06853`. arXiv:2401.06853 [cs].

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, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng,

Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Olivier Bousquet, Quoc Le, and Ed H. Chi. Least-to-most prompting enables complex reasoning in large language models. *CoRR*, abs/2205.10625, 2022. doi: 10.48550/ARXIV.2205.10625. URL https://doi.org/10.48550/arXiv.2205.10625.

Kunlun Zhu, Zijia Liu, Bingxuan Li, Muxin Tian, Yingxuan Yang, Jiaxun Zhang, Pengrui Han, Qipeng Xie, Fuyang Cui, Weijia Zhang, Xiaoteng Ma, Xiaodong Yu, Gowtham Ramesh, Jialian Wu, Zicheng Liu, Pan Lu, James Zou, and Jiaxuan You. Where llm agents fail and how they can learn from failures, 2025. URL https://arxiv.org/abs/2509.25370.

Weijun Zhu. Big data on linear temporal logic formulas. In *2021 IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*, pp. 544–547, Chongqing, China, June 2021. IEEE. ISBN 978-1-72818-535-4. doi: 10.1109/IMCEC51613.2021.9482368. URL https://ieeexplore.ieee.org/document/9482368/.

A Appendix

Prompt Template 1: Direct Prompting, in which *question* is the placeholder for TR problem.

{question}

The answer must be either 'True' or 'False'.

Prompt Template 2: Zero-Shot CoT, in which *question* is the placeholder for TR problem.

{question}

Let's think step by step.

First, let's identify what the hypothesis is asking.
Then, let's understand and examine each condition.
Let's trace and reason through the event sequences.
Finally, let's determine if the hypothesis can be satisfied.

Please reason through this problem thoroughly before answering.
The final answer must be either 'True' or 'False'.

Prompt Template 3: Few-Shot CoT, in which *question* is the placeholder for TR problem.

=== Example 1 ===

Context:
Initially, event3 happened. After event3, either event1, or event2 must happen. After event1, either event2, or event3 must happen. After event2, event3 must happen.

Hypothesis:
C1: Event1 happens or event3 happens.
C2: Event1 happens or C1 holds.
C3: C2 eventually holds.

Let's think step by step:
1. We need to check if C3 (C2 eventually holds) is true.
2. C2 is: Event1 happens or C1 holds
3. C1 is: Event1 happens or event3 happens
4. So C2 simplifies to: Event1 happens or (Event1 happens or event3 happens)
5. This is logically equivalent to: Event1 happens or event3 happens
6. The initial state has event3 = true
7. Therefore, C2 holds at the initial state (since event3 is true)
8. Since C2 holds at the initial state, C3 (C2 eventually holds) is True

Answer: True

=== Example 2 ===

Context:
Initially, event2 happened. After event2, event3 must happen. After event3, no other events can happen. After event1, either event2, or event3 must happen.

Hypothesis:
C1: Event2 happens and event3 happens.
C2: C1 holds in the next state.
C3: C2 eventually holds.

Let's think step by step:
1. We need to check if C3 (C2 eventually holds) is true.
2. C2 is: C1 holds in the next state, where C1 is (event2 AND event3)
3. For C1 to be true, both event2 and event3 must be true simultaneously
4. In any single state, only one event can be true at a time
5. Starting from event2, the next state must be event3
6. From event3, it stays at event3 (no other events can happen)
7. At no point can event2 and event3 be true simultaneously in the same state
8. Therefore, C1 is always false
9. If C1 is always false, then C2 (C1 in the next state) is always false
10. If C2 is always false, then C3 (C2 eventually holds) is False

Answer: False

=== Now solve this problem ===

""" + {question} + """

Let's think step by step.

First, let's identify what the hypothesis is asking.
Then, let's understand and examine each condition.
Let's trace and reason through the event sequences.
Finally, let's determine if the hypothesis can be satisfied.

Please reason through this problem thoroughly before answering.
The final answer must be either 'True' or 'False'.

Prompt Template 4: Least-to-Most Question Breakdown, in which *question* is the placeholder for TR problem.

{question}

Let me break this down into smaller, manageable parts:

1. What events are involved and what are the initial conditions?
2. What are the transition rules between events?
3. What does each sub-condition (C1, C2, C3, etc.) mean?
4. How do these conditions relate to each other?

Please answer each of these questions step by step.

Prompt Template 5: Least-to-Most Question Solving, in which *question* is the placeholder for TR problem and *breakdown_response* is the response of LLMs after prompted to breakdown a given question.

{question}

Based on the following breakdown of the problem:

{breakdown_response}

Now, let's use this understanding to solve the main problem step by step:

1. Start with the simplest conditions and work up to the complex ones
2. Trace through the event sequences
3. Check if the final hypothesis can be satisfied

The final answer must be either 'True' or 'False'.

```
=== Context ===

Initially, event1 happened. After event2, either event1, or event3 must happen. After event1
    ↪ , event3 must happen. After event3, event2 must happen. After event3, event2 must
    ↪ happen.

=== Hypothesis ===

C1: Event2 happens in the next state.
C2: C1 eventually holds.
C3: Event3 happens and C2 holds.

C3 is True or False?

=== Response ===

Let's analyze the problem step by step.

### Step 1: Understanding the Context
- Initially, **event1** happens.
- After **event2**, either **event1** or **event3** must happen.
- After **event1**, **event3** must happen.
- After **event3**, **event2** must happen.
- After **event3**, **event2** must happen (this is repeated).

### Step 2: Analyzing the Hypothesis
- **C1**: Event2 happens in the next state.
- **C2**: C1 eventually holds.
- **C3**: Event3 happens and C2 holds.

### Step 3: Tracing the Event Sequences
1. **Initial State**: event1 is true.
2. From event1, the next event must be event3 (as per the context).
3. Now, we have event3 true.
4. After event3, event2 must happen (as per the context).
5. Now, we have event2 true.

### Step 4: Evaluating C1, C2, and C3
- **C1**: Event2 happens in the next state.
  - In the initial state, event1 is true, and the next state will be event3. After event3,
    ↪ event2 will happen. Therefore, C1 will hold true in the next state after event3.

- **C2**: C1 eventually holds.
  - Since C1 holds true after event3, C2 is also true because C1 eventually holds.

- **C3**: Event3 happens and C2 holds.
```

```
  - Event3 does happen (as we established in the sequence), and C2 holds true (as we established that C1 eventually holds). Therefore, C3 is true.

### Conclusion
Since C3 is true (both conditions of C3 are satisfied), the final answer is:

**Answer: True**
```

Listing 4: An example of Temporal Semantics and Reasoning Misalignment.

```
=== Context ===

Initially, event2 happened. After event3, event1 must happen. After event1, event3 must happen. After event2, event1 must happen.

=== Hypothesis ===

C1: Event2 always happens.
C2: Event2 always happens.
C3: C1 holds and C2 holds.

C3 is True or False?

=== Response ===

The answer is: **True**

### Explanation:

- The hypothesis states:
  - **C1: Event2 always happens.**
  - **C2: Event2 always happens.**
  - **C3: C1 holds and C2 holds.**

Since **C1** and **C2** are **identical statements**, both asserting that "Event2 always happens," then **C3** (which says "C1 holds and C2 holds") is logically **true** because both conditions are the same and therefore both true.

Thus, **C3 is True**.
```

Listing 5: An example of Context Hypothesis Detachment.