

Inferring Events from Time Series using Language Models

Mingtian Tan

University of Virginia
wtd3gz@virginia.edu

Mike A. Merrill

Stanford University
mikeam@cs.stanford.edu

Zack Gottesman

University of Virginia
qdw5jf@virginia.edu

Tim Althoff

University of Washington
althoff@cs.washington.edu

David Evans

University of Virginia
evans@virginia.edu

Thomas Hartvigsen

University of Virginia
hartvigsen@virginia.edu

Abstract

A common goal in analyzing time series data is to understand how events cause observed variations. We study whether Large Language Models (LLMs) can infer natural language events associated with time series data. We introduce an automated method for generating tasks that test a model’s ability to reason about events associated with time series data based on sports data, and develop a new benchmarking method. In experiments spanning 18 LLMs, we prompt LLMs to infer unobserved events given time series data and observe surprising successes, even when providing minimal context. We then show that combining distillation with Reinforcement Learning (RL) can improve the performance for small language models to approach that of large proprietary reasoning models.*

1 Introduction

Time series are time-stamped sequences of values. They are pervasive and come from many sources, like wearable devices (Garcia-Gonzalez et al., 2020), clinical records (Harutyunyan et al., 2019), and asset market prices (Wang et al., 2024b; Li et al., 2024a). Although time series analysis typically studies real-valued time series data in isolation, many time series have events that can be described in natural language associated with the numerical data. Although the causal connections between the numerical data and the natural language events are often uncertain, the events can be correlated with the numerical data and may be causally related. Figure 1 illustrates an example using estimated win probabilities for a basketball game—events favorable to Team A increase its win probability, while unfavorable events decrease it.

Prior work on reasoning about time series in conjunction with natural language has largely overlooked event descriptions (Merrill et al., 2024;

Williams et al., 2025) and focused on tasks involving just numeric time series data like trend analysis (Cai et al., 2024) or anomaly detection (Dong et al., 2025). While some studies collect sequences of news related to time series (Wang et al., 2024b; Liu et al., 2024a; Cheng and Chin, 2024), they are curated for forecasting and do not explore reasoning from the numeric data to events. Meanwhile, due to the potential inclusion of event descriptions that do not impact the time series, as well as failure to include important events, these data are not ideal as a benchmark for measuring LLMs reasoning.

To address this gap, we introduce a method to evaluate a model’s ability to reason about the events associated with time series data. The key idea is to measure a model’s ability to select the sequence of event descriptions that best correspond to the input time series data. We introduce a benchmark comprising time series data and associated natural language event descriptions. Our dataset (Section 3.3) includes 4,200 games from the NBA (basketball) and NFL (American football) sports leagues, comprising a total of 1.7 million data points and events. The real-valued data is *win probability*[†] and the task, as illustrated in Figure 4, is to determine which sequence of events is most consistent with the given win probability sequence.

To evaluate the effectiveness of our benchmark in assessing reasoning ability, we test 18 models across various factors, including the impact of available context, varying sequence lengths, and time series similarity on reasoning. We also examine the impact of replacing or removing time series and real entity names through three ablation studies. To explore the generalizability of our approach, we extend the evaluation to open-domain settings, in-

[†]We use the win probability values output by ESPN’s game analysis (<https://www.espn.com/analytics/>). As we discuss in Section 5, win probability is an effective measure of game state but is a model’s predictions about the future not a measured ground truth value.

*All resources needed to reproduce our work are available: <https://github.com/hartvigsen-group/GAMETime>

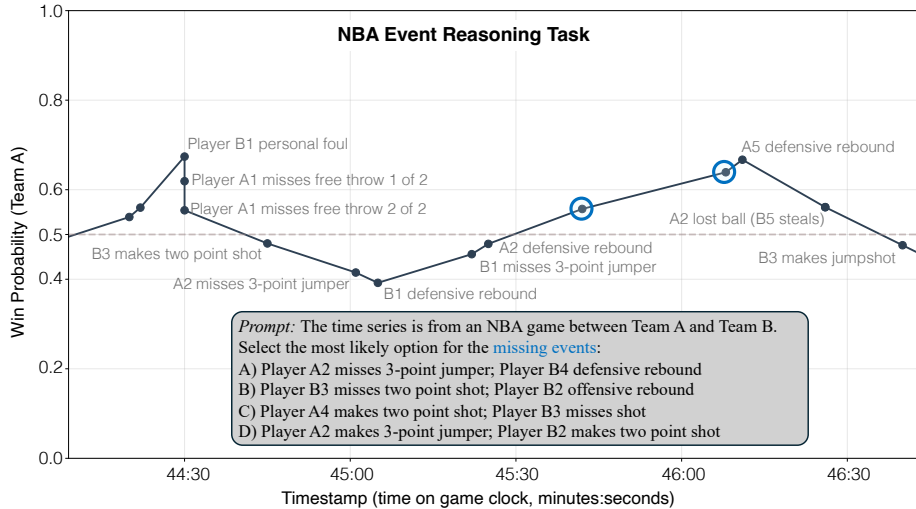


Figure 1: Illustration of time series event reasoning. The illustrative prompt provides (in text form, see details later in the paper) a time series of win probability values and optional additional context. The model is prompted to select the most likely sequence of events for some segment of the time series data. (This example is derived from an NBA game between the Dallas Mavericks (Team A) and Los Angeles Lakers (Team B), 1 November 2019.) For more details, refer to Figure 13 in Appendix D.

cluding cryptocurrency prices (Li et al., 2024a) and U.S. health data (Liu et al., 2024a).

Our findings indicate that several LLMs exhibit promising reasoning capabilities. OpenAI’s o1 achieved the highest accuracy, predicting 83% of the NBA events reasoning tasks correctly, followed by DS-R1-distill-Qwen-32B (68%) and GPT4o (41%). Through post-training with a distillation phase followed by self-improvement optimization (based on GRPO (Shao et al.)), we significantly improved the performance of the Qwen2.5 1.5B model from being the worst performing model to outperforming every model except for o1, and approaching its performance on the NBA task.

Our key contribution is a method for programmatically generating tasks that assess an LLMs’ ability to reason about events associated with a time series (Section 3.2). We demonstrate this method using tasks based on sports data and extend it to other domains (Section 4.5). We introduce an easily extensible dataset based on continuously-updating sports data with 1.7 million timesteps with values and events (Section 3.3), where changes in time series are explicitly influenced by events. In benchmarking 18 LLMs, we find promising reasoning capabilities and find clear avenues to enhance reasoning (Section 4.1 and Section 4.3) and demonstrate that post-training optimization can significantly enhance LMs ability to infer events associated with a time series (Section 4.2).

2 Related Work

Despite a growing body of work on LLMs and time series reasoning which we summarize in this section, previous benchmarks for LLMs in time series and event reasoning have not addressed the task of inferring event sequences from time series.

2.1 Time Series Reasoning with LLMs

Many studies use text to assist in time series reasoning (including forecasting) with promising results (Cao et al., 2024; Wang et al., 2025; Xie et al., 2025). These works have been applied in various domains including sociology (Cheng and Chin, 2024), energy (Wang et al., 2024b; Xu et al., 2024), and finance (Li et al., 2024a; Wang et al., 2024a). For example, Williams et al. (2025) manually curated time series forecasting data along with related text to demonstrate the value of incorporating textual information for forecasting tasks. Wang et al. (2024b) used news about energy to help LLMs predict local electricity conditions. Intrinsically, those approaches depend on LLMs’ multi-modal transfer of knowledge from natural language to time series.

However, there are also critical areas where reasoning about real-world events through time series analysis holds significant potential to enhance performance (Jin et al., 2024; Jiang et al., 2024b), compared to unimodal methods. Using LLMs for anomaly detection (Dong et al., 2025; Zhou and Yu, 2025) often involves processing time series data,

	Benchmark/Evaluation	Context	Source	Task
Multimodal	Williams et al. (2025)	Description	Manual	Forecasting
	Merrill et al. (2024)	Description	Synthetic	Reasoning & Forecasting
	Cai et al. (2024)	Question	Manual	Understanding
	Liu et al. (2024a)	News Series	Real-World	Forecasting
Unimodal	Fatemi et al. (2025)	Events & Time	Synthetic	Temporal Reasoning
	Xiong et al. (2024)	Events & Time	Synthetic	Temporal Reasoning
	Chu et al. (2024)	Events & Time	Real-World	Temporal Reasoning
	Quan and Liu (2024)	Events	Synthetic	Sequential Reasoning
	Karger et al. (2025)	Events	Real-World	Future Forecasting
	GAMETime (ours)	Events & Time	Real-World	Event Reasoning

Table 1: Time series benchmarks typically lack a focus on inferring event sequences, while event reasoning evaluations do not incorporate multimodal reasoning over numerical sequences. Our work emphasizes reasoning about events associated with time series data.

such as CPU usage rates from system monitors, and then generating an interpretable anomaly report (Liu et al., 2024b). Similarly, other domains, such as medical care (Chan et al., 2024), market analysis (Lee et al., 2024; Ye et al., 2024), and human activity analysis (Li et al., 2024b), also rely on this multi-modal reasoning capability to make actionable decisions.

Table 1 summarizes benchmarks intended to evaluate LLMs’ capability in processing time series data. Cai et al. (2024) proposed a benchmark using synthetic data to evaluate LLMs’ understanding of time series, focusing on tasks such as pattern recognition. Similarly, Merrill et al. (2024) introduced synthetic time series data and relevant textual descriptions, containing a single event (cause), to evaluate LLMs’ performance in matching time series to the scenarios that generated them (i.e., etiological reasoning). Due to the lack of paired event sequence, none of these works evaluated the LLMs’ ability to reason about events from time series.

The one exception is Liu et al. (2024a), which collects news sequences corresponding to time series dating back to 1983. However, due to the limited dataset size and potential contamination issues, it is challenging to use as a fair evaluation source, especially since the exact impact of news on time series remains unclear. To fill this gap, we propose a living benchmark with data sourced from continuously refreshed naturally-occurring data (in our case, from widely available sports data). This avoids the pitfalls associated with synthetic data, and because it can be easily refreshed avoids the contamination risks with fixed benchmarks.

2.2 Reasoning about Events

Reasoning is an ill-defined but critical capability, and reasoning benchmarks have been developed for myriad tasks including coding (Zhuo et al., 2025; Jain et al., 2024), mathematics (Cobbe et al., 2021; White et al., 2024), and finance (Xie et al., 2023; Islam et al., 2023). Other benchmarks aim to evaluate the general reasoning abilities of LLMs (Bang et al., 2023; White et al., 2024; bench authors, 2023), including BBH (Suzgun et al., 2023) and MMLU (Hendrycks et al., 2020).

Table 1 overviews benchmarks that have been proposed to evaluate LLMs’ understanding of relationships between events (Quan and Liu, 2024), as well as temporal reasoning capabilities for understanding the relationships between events and time (Xiong et al., 2024; Kenneweg et al., 2025). Karger et al. (2025) introduced a dynamically updated benchmark to evaluate event forecasting. Fatemi et al. (2025) used synthetic data to assess LLMs’ perception and reasoning between events and time. However, these benchmarks do not consider the interplay between time series and associated events, which is the focus of our work.

3 Method

This section describes our method for evaluating reasoning about events associated with a time series. Section 3.1 defines the problem setting, Section 3.2 outlines how we generate tasks from the data, and Section 3.3 introduce the dataset we use for our main experiments.

3.1 Problem Definition

A time series is a sequence of timestamped values: $\mathbf{x} = [(t_0, x_0), (t_1, x_1), \dots, (t_T, x_T)]$. The values could be real numbers or textual descriptions of events. For each sequence, the timestamps t are monotonically increasing ($i < j \implies t_i \leq t_j$), but there can be multiple (ordered) data points with the same timestamp. For our purposes, we consider time series where each data point contains a timestamp along with a real value number and a text event description, although for any particular data point either of these could be missing.

Although there may not always be a direct causal relationship between the events and measurements, we assume there is some connection between the numeric values and described events. Note that we do not assume that every timestamp where there is a numeric value includes an event description, or that every timestamp for which there is an event description has a numeric value.

Our goal is to assess an LLM’s ability to reason about time series data by measuring its ability to infer unobserved values in a time series \mathbf{x} . As illustrated in Figure 1, when the event descriptions are partially missing, the LLM is tasked to infer the missing events using the provided context and the numeric time series data for the timestamps with missing event descriptions.

3.2 Generating Event Understanding Tasks

We formulate our event reasoning evaluation as a multiple-choice question (Hendrycks et al., 2020; bench authors, 2023; Pal et al., 2022; White et al., 2024) where the model is prompted to select the event descriptions that are most likely to correspond to the provided real-valued time series data. The prompt follows this template:

System Prompt: {{sys_prompt}}
 $[t_i]$ v_i $[s_i]$ $[e_i]$
 $[t_{i+1}]$ v_{i+1} $[s_{i+1}]$
 \vdots
 $[t_{i+k-2}]$ v_{i+k-2} $[s_{i+k-2}]$
 $[t_{i+k-1}]$ v_{i+k-1} $[s_{i+k-1}]$ $[e_{i+k-1}]$
 Four options to choose: {{options}}
 Respond with this format: {{format}}

where we provide contextual task information (i.e., `sys_prompt`), along with real-valued time series of length k , $[(t_i, x_i), \dots, (t_{i+k-1}, x_{i+k-1})]$. Here, each x_i consists of three components: v_i , s_i , and e_i , representing the win probability, score, and event

in our dataset. Some information about some of the intermediate events from e_{i+1} to e_{i+k-2} are missing, and the LLM is tasked with inferring these events. For evaluation, v_i is consistently provided as a baseline for events inferring, while the remaining components are introduced separately as auxiliary conditions. Figure 12 in Appendix D provides examples of the full prompts used for the baseline setting, where only v_i is provided. To make the task tractable we frame the task as a multiple choice question with four options, one of which corresponds to the actual sequence of events, and prompt the model to select the most likely option.

To further isolate the LLM’s reasoning on time series, we replace specific named entities in our dataset with general, non-identifying descriptors. Specific team names are replaced with *Team A* or *Team B*. Actual player names are replaced with generic labels, such as *Player from Team A*, ensuring that the associations between players and their teams are preserved but revealing no other information about their identities. In evaluations from other domains, such as cryptocurrency prices (Li et al., 2024a), we replace all numerical values in news (events) sequence with symbols (e.g., α) to prevent LLMs from matching events to time series using dates or price. In open-domain settings, the impact of news on time series may exhibit a minor delay. Therefore, we provide two events occurring before t_i to better capture the full range of events that may influence the time series.

3.3 A Sports Dataset with Event Descriptions

To obtain paired data of time series and event sequences, we use data from sports. Sports data has two key advantages for our purposes: (1) the in-game events are directly correlated with in-game real-valued data; and (2) the data are continuously refreshed because new games are played every day. For the natural language events, we used play-by-play data provided by ESPN that describes key events throughout each game, such as scoring, turnovers, or fouls in basketball. ESPN also provides each team’s predicted win probability throughout the game, which we use as the real-valued time series data. These win probabilities reflect the game’s status at each time step and changes depending on the events. Since a game is a relatively closed environment, there is a clear relationship between the events and the time series: an event favoring Team A increases Team A’s win probability. This closed environment, along with

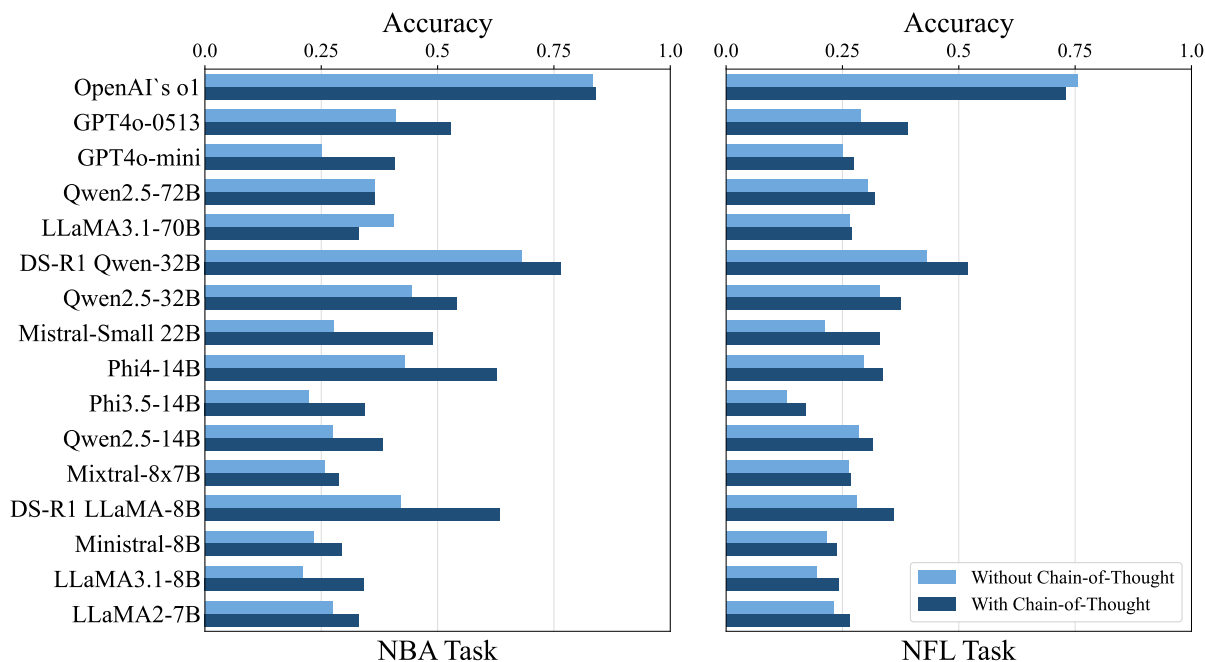


Figure 2: Prediction accuracy on NBA and NFL tasks. The reasoning-focused models such as OpenAI’s o1 and DeepSeek-R1 distilled models (DS-R1) significantly outperform others. Chain-of-Thought prompting improves reasoning capabilities for most models. The three proprietary models are first, followed by the open-weight models sorted by decreasing model size.

the continuous generation of new data to avoid contamination problems, makes sports data a good candidate for a benchmark evaluating how effectively LLMs infer events through time series.

We curate GAMETIME, a new dataset containing 4,200 time series (games) collected through Sep 2025, 3,276 from NBA basketball games and 924 from NFL American football games. Examples can be found in Appendix A.1. Each basketball game contains an average of 460 timesteps, while the football time series average 179 timesteps. The full dataset comprises 1.7 million time data points (win probabilities) paired with corresponding event descriptions.

4 Experiments

To investigate LLMs’ event reasoning capabilities under diverse conditions, we explore five research questions: **RQ1:** Can LLMs reason about events, and does Chain-of-Thought (CoT) prompting enhance this reasoning? **RQ2:** Can post-training optimization improve event reasoning? **RQ3:** What is the effect of various available contexts beyond time series? **RQ4:** Are LLMs able to distinguish underlying time series similarities? **RQ5:** How do LLMs’ event reasoning performance compare across different domains?

Figure 2 summarizes our results evaluating 18 language models (LMs), including closed-weight models such as GPT-4o (Achiam et al., 2023) and open-weights models like LLaMA3.1 (Dubey et al., 2024), and Qwen2.5 (Yang et al., 2024). Models designed for reasoning, including DeepSeek Distilled model, like DS-R1-distill-Qwen-32B (DeepSeek-AI, 2025), and OpenAI’s o1 (OpenAI, 2024) generally perform better than non-reasoning models.

4.1 Accuracy Evaluation

To evaluate LLMs on event inference, we first follow the format in Figure 11 from Appendix D. In this setting, the model is prompted to select the most likely sequence of events corresponding to a given segment of time series data, where only Team A’s win probabilities (WP_A)[‡] are provided and the negative options are sequences of the same length randomly sampled from other games. Each model is evaluated on 200 questions. To eliminate memorization effects in reasoning, we select games that occurred after the models’ training cutoff dates and replace real team and player names with generic labels such as *Player from Team A*.

Figure 2 summarizes the models’ performance

[‡]In NBA basketball there are no draws, and in NFL football draws are exceedingly rare, so the win probability for Team B is $1 - WP_A$.

Model	Post-training		NBA Task			NFL Task		
	Warmup	RL	Correct	Incorrect	Invalid	Correct	Incorrect	Invalid
Qwen2.5 (1.5B)	–	–	11	27	162	29	69	102
Qwen2.5 (1.5B)	✓	–	111	82	7	69	128	3
Qwen2.5 (1.5B)	–	✓	32	114	54	43	111	46
Qwen2.5 (1.5B)	✓	✓	151	49	0	88	112	0
GPT4o	–	–	82	118	0	58	142	0
DS-R1-32B	–	–	136	64	0	86	114	0
OpenAI’s o1	–	–	166	31	3	151	49	0

Table 2: Comparing the performance of the Qwen2.5 (1.5B) model under different post-training strategies with production models. The numbers are counts of correct, incorrect, and invalid responses from 200 test cases. When applying both warm-up (knowledge distillation) and reinforcement learning (GRPO) post-training, the 1.5B-size model achieves the competitive performance, surpassed only by OpenAI’s o1 and outperforming the distillation source DS-R1-distill-Qwen-32B (DS-R1-32B).

on the NBA and NFL task. Although the weakest models barely outperform random guessing, several models, particularly those designed for reasoning, demonstrate strong reasoning performance. GPT-4o achieves an accuracy of 41%, and DS-R1-distill-Qwen-32B reaches 68%, while o1 performs the best, with an accuracy of 83%. Similar results are observed on the NFL data, though the task appears to be more challenging. The performance of GPT-4o drops to 29%, while DS-R1-distill-Qwen-32B and o1 achieve 43% and 75.5%. Models that failed to generate valid responses, such as Qwen-7B, are not included. Appendix C shows a case study of how models perform events inferring through time series.

Chain-of-Thought prompting Next, we investigate if a longer reasoning process with Chain-of-Thought (CoT) prompting (Wei et al., 2022) improves results of LLMs on event reasoning. LLMs show an average improvement from CoT prompting of 4.5% for the basketball task and 9.6% for the NFL task. In our CoT prompt, we provide an example with a reasoning process (see Figure 14 in the Appendix). The longer reasoning process with CoT, however, also slightly increases the overall likelihood of LLMs failing to return answers in the required format by 0.6%. We acknowledge the potential for CoT strategy, but we do not further explore this due to computational constraints.

4.2 Post-training Improves Reasoning

The effectiveness of post-training has been demonstrated in math (DeepSeek-AI, 2025), coding (Hugging Face, 2025), or vision (Shen et al., 2025) tasks.

This improvement can be achieved either through training using data containing distilled reasoning processes (Muennighoff et al., 2025) or through reinforcement learning, such as GRPO (Shao et al.). To improve LMs’ performance on the event reasoning tasks, we first warm up the LM with knowledge distilled from DS-R1-distill-Qwen-32B, and subsequently apply GRPO training to the warmed-up model.

Results in Table 2 show that even a model with only 1.5B parameters can achieve competitive performance through post-training, surpassing the distilled source and ranking second only to OpenAI o1. For instance, for the NBA task, when we warmed up the model using 3, 200 correctly reasoned Q&A pairs and their reasoning processes, the warmed-up model correctly infers 111 samples and returns only 7 invalid answers (i.e., no valid result could be extracted) out of 200 test cases, compared to the base model, which answered only 11 correctly and produced 162 invalid results. After further GRPO training with 7, 500 Q&A pairs, The number of correct reasoning cases reach 151 with no invalid answers, surpassing the distilled model’s 136 and approaching the performance of OpenAI’s o1. However, employing RL alone without the warm-up phase resulted in only 32 correct responses. Appendix B.2 provide more details on the training process, including reward signals and prompt formats.

4.3 Impact of Context

In different applications, the available context that LLMs can access varies. Compared to the baseline setting, where only the real-valued time series data is provided, we also evaluate LLMs’ perfor-

	Model	Available Context				Ablations			Options
		WP Only	WP+Times	WP+Score	WP+Event	w/ Name	Remove	Replace	Reorder
NBA Task	GPT4o (0513)	41.0%	39.0%	47.5%	39.0%	55.0%	28.5%	24.0%	69.5%
	GPT4o (mini)	25.0%	24.5%	25.0%	26.0%	43.5%	21.0%	27.5%	39.0%
	Qwen2.5 (72B)	36.5%	39.0%	43.5%	39.5%	41.0%	24.5%	30.0%	66.0%
	LLama3.1 (70B)	40.5%	37.0%	50.5%	38.5%	51.0%	26.5%	26.0%	47.5%
	Qwen2.5 (32B)	44.5%	43.5%	57.5%	43.5%	50.0%	22.5%	26.0%	59.0%
	Phi4 (14B)	43.0%	35.0%	40.0%	36.0%	42.5%	25.0%	24.0%	47.0%
	Qwen2.5 (14B)	27.5%	34.5%	33.0%	32.0%	48.0%	22.0%	22.0%	44.5%
Avg. Impact of the Condition		↓ -0.6%	↑ 14.6%	↑ 0.0%	↑ 33.2%	↓ -32.2%	↓ -27.3%	↑ 46.8%	
NFL Task	GPT4o (0513)	29.0%	75.5%	43.5%	53.0%	71.0%	18.5%	22.0%	60.0%
	GPT4o (mini)	25.0%	52.0%	26.5%	35.5%	33.5%	24.5%	25.5%	42.0%
	Qwen2.5 (72B)	30.5%	69.0%	42.0%	40.5%	52.0%	25.0%	23.0%	54.0%
	LLama3.1 (70B)	26.5%	71.0%	47.5%	35.5%	65.5%	20.5%	17.0%	46.0%
	Qwen2.5 (32B)	33.0%	74.5%	43.5%	46.0%	40.5%	27.5%	27.0%	43.5%
	Phi4 (14B)	29.5%	46.5%	36.0%	38.5%	43.5%	25.0%	23.5%	28.5%
	Qwen2.5 (14B)	28.5%	55.5%	28.5%	34.5%	63.5%	25.5%	26.0%	33.0%
Avg. Impact of the Condition		↑ 120.1%	↑ 32.4%	↑ 40.4%	↑ 84.2%	↓ -17.3%	↓ -18.6%	↑ 52.8%	

Table 3: LLMs’ event reasoning accuracy (%) under various contexts and ablations. We provide each model with 200 questions for each condition. The best performing models are **bold** with **blue shading**, and second best are shaded **light blue**.

mance when different reasoning-relevant contexts are made available or modified. In addition, to evaluate the impact of time series and real names in reasoning and causal relationship between time series and events, we conduct three ablations. The results are summarized in Table 3.

Available Context Due to differences between the football and basketball data, various conditions influence differently. For example, timestamps (*TS+Times*) provide the significant improvement in reasoning for football. Similarly, when providing the score (*TS+Score*) or partial events (*TS+Event*), e_i and e_{i+k-1} , performance also improves. Note that, given computational constraints and the strong performance of reasoning models, we will primarily focus on avenues to improve base models.

Ablations Real player and team names are expected to provide cues that help models identify the correct answer. For example, through potential data contamination or directly matching team names with player names in the options. Results from *w/ Name* column in Table 3 demonstrate that real names notably improves accuracy, highlighting the necessity of removing them when evaluating reasoning (Fatemi et al., 2025). Another two ablations—removing (*Remove*) or replacing (with series from other games) (*Replace*) the time series—model performance drops to near-random levels, indicating that LLMs rely on time series for

event inferring and that a strong association exists between the time series and the events.

Options Due to the nature of possession changes in football and basketball, event sequences follow sequential constraints. To further test whether LLMs can detect logically inconsistent information to aid reasoning, we shuffle the order of ground-truth events to create negative options. Results from the *Reorder* column in Table 3 show a clear improvement, indicating that LLMs are capable of leveraging logical sequences through reasoning.

4.4 Disparity of Data

To assess how the time series similarity impacts LLMs’ reasoning, we control the distance between the time series associated with positive and negative options. We compute the distance between time series as Euclidean distance after *z-score normalization*: $D = \|norm(\mathbf{p}_{win}) - norm(\mathbf{p}'_{win})\|_{l2}$. We divided the distances into seven levels, based on the distribution of win probability differences (see Figure 7 in Appendix B.4 for details), starting from 0.4 with an increment of 0.1 per level.

We follow the setup in Section 3.2, setting the sequence length to 10 and evaluating each LM on 200 questions, keeping the ground-truth events and question time series consistent. The results are presented in Figure 3, showing a slight upward trend in LLM performance as similarity decreases for most of models. This is due to the inherent con-

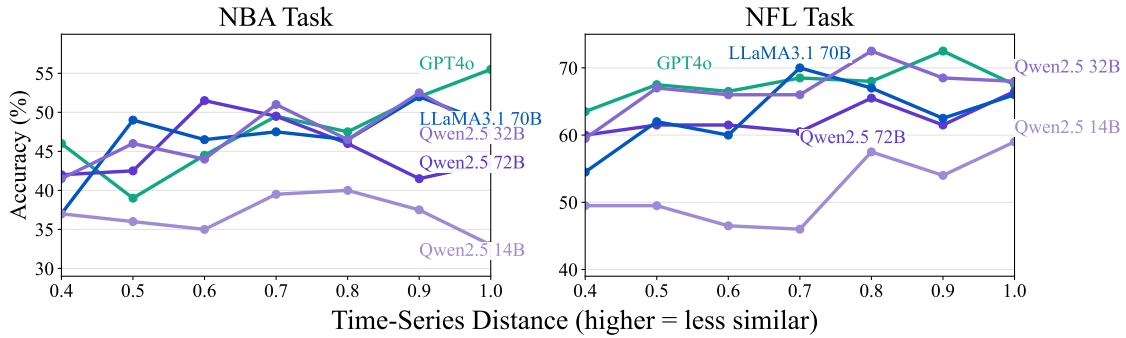


Figure 3: The performance of LLMs in distinguishing events corresponding to time series (win probabilities) with different levels of similarity. Time series *similarity* decreases as x (i.e., time series *distance*) increases.

sistency between time series and event sequences, which LLMs are able to recognize.

4.5 Other Domains

Time series data often coexist with related textual sequences. To evaluate the generalizability of our approach, we extend our evaluation to four other domains: trade (import/export, IMEX), health (influenza rates), and energy (gasoline prices) from Time-MMD (Liu et al., 2024a), as well as cryptocurrency time series from CryptoTrade (Li et al., 2024a). To limit the length of prompts, we use news titles as events for cryptocurrency. We selected the “factual” field as the events occurring at each timestamp from Time-MMD. Liu et al. (2024a) extracted these “factual” statements from news and reports to describe real-world events (Appendix B.1). Our question follows the format in Section 3.2, with an event sequence length of 10, corresponding to ten trading days for Bitcoin data or ten weeks of influenza statistics in the U.S. health dataset.

We evaluate two settings: one where events contain numerical information (i.e., *Complete*) and another where all numerical values, such as dates or real values (e.g., Bitcoin prices or trading volumes), are replaced with symbols like α (i.e., *Filtered*). Table 4 summarizes the results. Even after stripping numerical data, however, LLMs still demonstrate moderate reasoning ability. GPT-4o, for instance, consistently achieves over 50% accuracy. Additionally, open-weights models such as Qwen2.5 72B or DS-R1-distill-Qwen-32B demonstrate comparable performance to GPT-4o. Detailed results can be found in Table 5 in Appendix B.

	Filtered	GPT-4o (0513)	GPT-4o (mini)	Qwen2.5 (72B)	DS-R1 (Qwen 32B)
Crypto (Bitcoin)	✓	84% 65%	58% 40%	71% 40%	62% 39%
Trading (IMEX)	✓	91% 50%	90% 35%	90% 51%	93% 47%
Health (Influenza)	✓	62% 33%	53% 26%	77% 34%	74% 37%
Energy (Gasoline)	✓	97% 52%	95% 40%	96% 48%	98% 49%

Table 4: Accuracy of LLMs inferring events across other domains among 100 questions. Replacing numerical information in the events (Filtered rows with ✓) results in a performance decline compared to retaining the original numbers.

5 Conclusions

Data comprising time series real values paired with event sequences occur in many important domains. We introduce a dataset containing 1.7 million real-valued time series paired with events and a method for evaluating the ability of an LLM to reason about events corresponding to real-valued time series data. Our evaluation of 18 language models using this benchmark reveals that both open-weights and proprietary models exhibit promising reasoning capabilities, with reasoning models such as DS-R1-distill-Qwen-32B outperforming larger proprietary model such as GPT-4o, while OpenAI’s o1 achieves the best performance. By applying post-training optimization, we significantly improve the performance of the Qwen2.5 1.5B to surpassing every model except o1, and approaching o1’s per-

formance on the NBA task.

Limitations

Our dataset includes time series representing win probabilities in sports, which serve as an effective measurement of how events affect a team's state and have a clear relationship with events. Since it is impossible to know the true underlying probability of the game outcome, these probabilities are estimates of each team's chances to win the game produced by ESPN's proprietary model, and not the ground truth. Note that we focus on evaluating the performance of current models rather than exploring how our data can be used for reasoning model training, which we leave for future work.

Acknowledgements

This work is supported in part by funds provided by the National Science Foundation, Department of Homeland Security, and IBM through the ACTION AI Institute (Award #2229876) and by Microsoft's Accelerating Foundation Model Research program. TH was supported by CapitalOne and the University of Virginia's National Security Data & Policy Institute through the U.S. Department of Defense Contracting Activity #2024-24070100001. We thank the University of Virginia's High Performance Computing team for providing excellent computing resources.

References

- Marah Abidin, Jyoti Aneja, Harkirat Behl, et al. 2024. Phi-4 technical report. *arXiv:2412.08905*.
- Josh Achiam, Steven Adler, Sandhini Agarwal, et al. 2023. GPT-4 technical report. *arXiv:2303.08774*.
- Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. [A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity](#). In *13th International Joint Conference on Natural Language Processing and 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*.
- BIG bench authors. 2023. [Beyond the imitation game: Quantifying and extrapolating the capabilities of language models](#). *Transactions on Machine Learning Research*.
- Yifu Cai, Arjun Choudhry, Mononito Goswami, and Artur Dubrawski. 2024. TimeSeriesExam: A time series understanding exam. In *NeurIPS Workshop on Time Series in the Age of Large Models*.
- Defu Cao, Furong Jia, Sercan O Arik, Tomas Pfister, Yixiang Zheng, Wen Ye, and Yan Liu. 2024. Tempo: Prompt-based generative pre-trained transformer for time series forecasting. In *International Conference on Learning Representations (ICLR)*.
- Nimeesha Chan, Felix Parker, William Bennett, Tianyi Wu, Mung Yao Jia, James Fackler, and Kimia Ghobadi. 2024. MedTsLLM: Leveraging LLMs for multimodal medical time series analysis. In *9th Machine Learning for Healthcare Conference*.
- Junyan Cheng and Peter Chin. 2024. Sociodojo: Building lifelong analytical agents with real-world text and time series. In *International Conference on Learning Representations (ICLR)*.
- 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 *62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training Verifiers to Solve Math Word Problems. *arXiv:2110.14168*.
- DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv:2501.12948*.
- Manqing Dong, Hao Huang, and Longbing Cao. 2025. Can LLMs serve as time series anomaly detectors? In *International Conference on Learning Representations (ICLR)*.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. 2024. The Llama 3 herd of models. *arXiv:2407.21783*.
- Bahare Fatemi, Mehran Kazemi, Anton Tsitsulin, Karishma Malkan, Jinyeong Yim, John Palowitch, Sungyong Seo, Jonathan Halcrow, and Bryan Perazzi. 2025. Test of Time: A benchmark for evaluating LLMs on temporal reasoning. In *International Conference on Learning Representations (ICLR)*.
- Daniel Garcia-Gonzalez, Daniel Rivero, Enrique Fernandez-Blanco, and Miguel R. Luaces. 2020. A public domain dataset for human activity recognition using smartphones. *Sensors*, 20(8).
- Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. 2019. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):96.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv:2009.03300*.

- Hugging Face. 2025. Open R1: A fully open reproduction of DeepSeek-R1. <https://github.com/huggingface/open-r1>.
- Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. 2023. Financebench: A new benchmark for financial question answering. *arXiv:2311.11944*.
- Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2024. Livecodebench: Holistic and contamination free evaluation of large language models for code. *arXiv:2403.07974*.
- Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, et al. 2024a. Mixtral of Experts. *arXiv:2401.04088*.
- Yushan Jiang, Zijie Pan, Xikun Zhang, Sahil Garg, Anderson Schneider, Yuriy Nevmyvaka, and Dongjin Song. 2024b. Empowering time series analysis with large language models: A survey. *arXiv:2402.03182*.
- Ming Jin, Yifan Zhang, Wei Chen, Kexin Zhang, Yuxuan Liang, Bin Yang, Jindong Wang, Shirui Pan, and Qingsong Wen. 2024. Position paper: What can large language models tell us about time series analysis. In *International Conference on Machine Learning (ICML)*.
- Ezra Karger, Houtan Bastani, Chen Yueh-Han, Zachary Jacobs, Danny Halawi, Fred Zhang, and Philip E. Tetlock. 2025. **ForecastBench: A dynamic benchmark of AI forecasting capabilities**. In *International Conference on Learning Representations (ICLR)*.
- Svenja Kenneweg, Jörg Deigmöller, Philipp Cimiano, and Julian Eggert. 2025. Traveler: A benchmark for evaluating temporal reasoning across vague, implicit and explicit references. *arXiv:2505.01325*.
- Hoyoung Lee, Youngsoo Choi, and Yuhee Kwon. 2024. Quantifying qualitative insights: Leveraging LLMs to market predict. *arXiv:2411.08404*.
- Yuan Li, Bingqiao Luo, Qian Wang, Nuo Chen, Xu Liu, and Bingsheng He. 2024a. CryptoTrade: A reflective LLM-based agent to guide zero-shot cryptocurrency trading. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Zechen Li, Shohreh Deldari, Linyao Chen, Hao Xue, and Flora D Salim. 2024b. Sensorllm: Aligning large language models with motion sensors for human activity recognition. *arXiv:2410.10624*.
- Haoxin Liu, Shangqing Xu, Zhiyuan Zhao, Ling kai Kong, Harshavardhan Kamarthi, Aditya B Sasanur, Megha Sharma, Jiaming Cui, Qingsong Wen, Chao Zhang, et al. 2024a. Time-MMD: Multi-domain multimodal dataset for time series analysis. In *38th Conference on Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS)*.
- Jun Liu, Chaoyun Zhang, Jiaxu Qian, Minghua Ma, Si Qin, Chetan Bansal, Qingwei Lin, Saravan Rajmohan, and Dongmei Zhang. 2024b. Large language models can deliver accurate and interpretable time series anomaly detection. *arXiv:2405.15370*.
- Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025. Understanding r1-zero-like training: A critical perspective. In *Conference on Language Modeling (COLM)*.
- Sara Vera Marjanović, Arkil Patel, Vaibhav Adlakha, et al. 2025. DeepSeek-R1 Thoughtology: Let’s think about LLM reasoning. *arXiv:2504.07128*.
- Mike A Merrill, Mingtian Tan, Vinayak Gupta, Thomas Hartvigsen, and Tim Althoff. 2024. Language models still struggle to zero-shot reason about time series. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- OpenAI. 2024. **Learning to reason with LLMs**.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikandan Sankarasubbu. 2022. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning*.
- Yinzhu Quan and Zefang Liu. 2024. EconLogicQA: A question-answering benchmark for evaluating large language models in economic sequential reasoning. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. Wu Y.K. Li, and Daya Guo. **DeepSeekMath: Pushing the limits of mathematical reasoning in open language models**. *arXiv:2402.03300*.
- Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. 2025. VLM-R1: A stable and generalizable R1-style large vision-language model. *arXiv:2504.07615*.
- Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. 2023. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*.

- Chengsen Wang, Qi Qi, Jingyu Wang, Haifeng Sun, Zirui Zhuang, Jinming Wu, Lei Zhang, and Jianxin Liao. 2025. ChatTime: A unified multimodal time series foundation model bridging numerical and textual data. In *AAAI Conference on Artificial Intelligence*.
- Shengkun Wang, Taoran Ji, Linhan Wang, Yanshen Sun, Shang-Ching Liu, Amit Kumar, and Chang-Tien Lu. 2024a. StockTime: A time series specialized large language model architecture for stock price prediction. *arXiv:2409.08281*.
- Xinlei Wang, Maik Feng, Jing Qiu, Jinjin Gu, and Junhua Zhao. 2024b. From news to forecast: Integrating event analysis in llm-based time series forecasting with reflection. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems (NeurIPS)*.
- Colin White, Samuel Dooley, Manley Roberts, et al. 2024. LiveBench: A challenging, contamination-free LLM benchmark. *arXiv:2406.19314*.
- Andrew Robert Williams, Arjun Ashok, Étienne Marcotte, Valentina Zantedeschi, Jithendaraa Subramanian, Roland Riachi, James Requeima, Alexandre Lacoste, Irina Rish, Nicolas Chapados, and Alexandre Drouin. 2025. Context is key: A benchmark for forecasting with essential textual information. In *42nd International Conference on Machine Learning (ICML)*.
- Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. PIXIU: A large language model, instruction data and evaluation benchmark for finance. *arXiv:2306.05443*.
- Zhe Xie, Zeyan Li, Xiao He, Longlong Xu, Xidao Wen, Tieying Zhang, Jianjun Chen, Rui Shi, and Dan Pei. 2025. ChatTS: Aligning time series with llms via synthetic data for enhanced understanding and reasoning. In *Proceedings of the VLDB Endowment, 2025*.
- Siheng Xiong, Ali Payani, Ramana Kompella, and Faramarz Fekri. 2024. Large language models can learn temporal reasoning. In *62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Zhijian Xu, Yuxuan Bian, Jiayuan Zhong, Xiangyu Wen, and Qiang Xu. 2024. Beyond trend and periodicity: Guiding time series forecasting with textual cues. *arXiv:2405.13522*.
- An Yang, Baosong Yang, Beichen Zhang, et al. 2024. Qwen2.5 technical report. *arXiv:2412.15115*.
- Wen Ye, Yizhou Zhang, Wei Yang, Lumingyuan Tang, Defu Cao, Jie Cai, and Yan Liu. 2024. Beyond forecasting: Compositional time series reasoning for end-to-end task execution. *arXiv:2410.04047*.
- Rosie Zhao, Alexandru Meterez, Sham Kakade, Cengiz Pehlevan, Samy Jelassi, and Eran Malach. 2025. Echo chamber: RL post-training amplifies behaviors learned in pretraining. *arXiv:2504.07912*.
- Zihao Zhou and Rose Yu. 2025. Can LLMs understand time series anomalies? In *International Conference on Learning Representations (ICLR)*.
- Terry Yue Zhuo, Minh Chien Vu, Jenny Chim, Han Hu, Wenhao Yu, Ratnadira Widyasari, Imam Nur Bani Yusuf, Haolan Zhan, Junda He, Indraneil Paul, et al. 2025. BigCodeBench: Benchmarking code generation with diverse function calls and complex instructions. In *International Conference on Learning Representations (ICLR)*.

A Experimental Setup

A Datasets and Language models

In this section, we introduce NBA and NFL event and time series data through examples from sports datasets. Additionally, we present the models we evaluate and provide details on data from other domains.

A.1 Events and Time Series in Sports

Figure 4 illustrates the time series and event sequences for basketball and football. When an event favorable to Team A occurs, Team A’s win probability typically increases. For example, in basketball, this could be a successful score by Team A or a turnover by Team B. In football, it could include defensive plays and sacks by Team A, penalties against Team B, or offensive success by Team A. Conversely, unfavorable events lead to a decrease in win probability.

A.2 Language Models and Setups

We have run our evaluation and experiments on Nvidia A100 GPUs. The specific settings for LLMs, as well as the packages used for data processing, are provided in the repository[§]. We evaluated a total of 16 models, including open-weight models such as LLaMA3.1 (Dubey et al., 2024), proprietary models like GPT4o (Achiam et al., 2023), and reasoning-focused models such as DeepSeek-R1 (DeepSeek-AI, 2025). The full list of tested models is as follows:

- **GPT4o** (Achiam et al., 2023): We test GPT4o-0513, a high-performance variant of GPT-4 optimized for both general-purpose generation and specialized tasks, and GPT4o-mini,

[§]All information and settings needed are available: <https://github.com/hartvigsen-group/GAMETime>

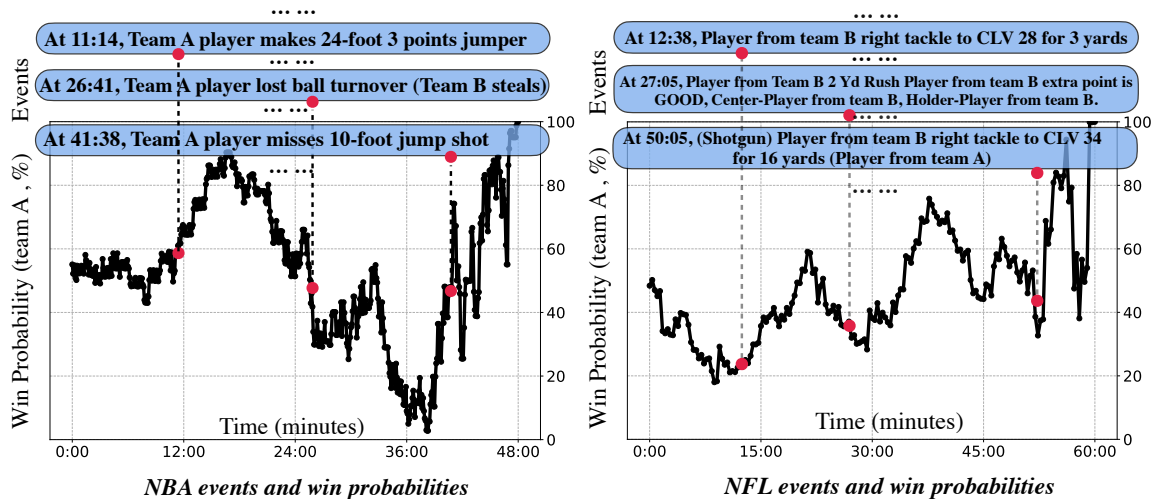


Figure 4: Examples of events and win probabilities in the NBA and NFL dataset. As the game progresses, ESPN provides descriptions of on-field events along with the corresponding win probabilities for each team at that moment. These probabilities can be considered a representation of the team’s current state.

a scaled-down version of GPT-4 designed for resource-constrained environments.

- **LLaMA** (Dubey et al., 2024): We evaluate instruction-tuned models of various parameter sizes, including LLaMA3.1-Instruct 70B, 8B, and LLaMA2-Instruct 7B.
- **Qwen2.5** (Yang et al., 2024): Our experiments included various instruction-tuned models such as Qwen2.5-Instruct 72B, 32B, 14B, and 8B.
- **Mixtral** (Jiang et al., 2024a): We test the 8x7B Mixture of Experts (MoE) model, along with Mixtral-Small 22B and Ministral-8B.
- **Phi** (Abdin et al., 2024): We included Phi-4 14B and Phi-3.5-Instruct 14B in our evaluations.
- **DeepSeek-R1** (DeepSeek-AI, 2025): Given computational constraints, we still evaluated reasoning-focused models such as DeepSeek-R1 32B and 8B. These models are distilled versions of DeepSeek-R1, using synthetic data from R1 to finetune Qwen 32B and LLaMA 8B, respectively.

A.3 Open-world Domains

To validate whether LLMs can reason about events through time series in other domains, we utilized four open-world datasets from different fields: Time-MMD (Liu et al., 2024a) (covering Trading, US Health, and Energy) and CryptoTrade (Li et al.,

2024a) (Bitcoin prices). The details are outlined as follows:

- **Trading:** Includes monthly U.S. International Trade Balance data from January 1987 to March 2024 (total length of 423 months), covering both import and export trade volumes. The corresponding text consists of keyword searches and institutional reports relevant to that month, such as "U.S. International Trade in Goods and Services".
- **U.S. Health:** Includes weekly Influenza Patients Proportion data from September 1997 to May 2024 (total length of 1 389 weeks). The corresponding text sequences are sourced from weekly keyword searches or reports from the "CDC’s ILINet system".
- **Energy:** Contains weekly Gasoline Prices (Dollars per Gallon) from April 1993 to April 2024 (total length of 1 479 weeks). The text sequences are obtained through searches or reports from institutions such as the U.S. Energy Information Administration.
- **Bitcoin:** Contains daily Bitcoin price data from January 1, 2023, to February 1, 2024 (time series length of 397), including opening and closing prices, as well as the highest and lowest prices of the day. The corresponding text sequence is derived from authoritative sources such as Bloomberg and Yahoo Finance, filtered through keyword searches to

provide five of the most relevant news articles per day. We use their headlines as event descriptions.

B Additional Experimental Results

B.1 Other Domains

Detailed results of LLMs on other domains are shown in Table 5. It can be observed that LLMs are capable of reasoning about events even in open-world domains. Moreover, when potentially confounding information in the events—such as numbers and dates—is removed (i.e., *Filtered*), LLMs still demonstrate strong reasoning performance.

B.2 Post-training Improves Inference

In the post-training phase, we primarily utilize question-answer pairs that included explicit reasoning processes, along with GRPO training, to facilitate the model’s self-improvement. The base model employed was Qwen2.5 (1.5B) (Yang et al., 2024), which demonstrates very limited initial event-inference capabilities. For example, as shown in Table 2 for the NFL dataset, it correctly infers 29 out of 200 test cases and produces 102 invalid answers; for the NBA dataset, it correctly reasons only 11 cases and yielded 162 invalid answers.

Inspired by recent work on warming up language models (Muennighoff et al., 2025; DeepSeek-AI, 2025), we apply knowledge distillation on a relatively strong-performing language model. To avoid data contamination, we selected training data exclusively from games that were different from those used in the test set. Considering the cost and computational resources, we chose DS-R1-distill-Qwen-32B as the distillation source. For the NFL task, we collected a total of 5,434 samples with an accuracy of 44.6%, and for the NBA task, we collected 4,814 samples with an accuracy of 67.5%, which is consistent with the results reported in Section 4.1. We ultimately selected all correctly reasoned samples, along with their reasoning trajectory, to warm up the Qwen (1.5B) model. The prompt structure used for the warm-up is illustrated in Figure 10. The results (Table 2) demonstrate that the warm-up phase significantly improved the model’s performance as well as its ability to return valid outputs.

Extensive research has shown that self-improvement through optimization leads to significant gains in tasks such as mathematics, coding, and visual reasoning (Shen et al., 2025; DeepSeek-

AI, 2025; Shao et al.). Building on the warmed-up model, we further applied reinforcement learning using 7,500 Q&A pairs for each task. The results in Table 2 show that, after RL optimization (e.g., GRPO (Shao et al.)), the model surpassed or matched the performance of the distilled model, even though its size was considerably smaller than that of the distillation source. The reasoning template we adopted is shown in Figure 10. Specifically, we primarily supervised two types of rewards: *format* and *correctness*, with the training reward trajectories illustrated in Figure 5. The training was conducted using the open-r1 (Hugging Face, 2025) framework and completed on 8 H200 GPUs. Detailed training hyper-parameters and settings are provided in our accompanying repository.

The essence of reinforcement learning in optimizing reasoning is strengthening reasoning trajectory based on reward signals (Liu et al., 2025; Zhao et al., 2025; Marjanović et al., 2025), which requires the language model to possess a certain level of inherent reasoning ability in the task’s domain. Therefore, we also applied GRPO training directly to the base model. Under the same data and training settings, the improvement in performance was limited; however, gains were still observed in the question-answering format, as reflected by a significant reduction in the number of invalid outputs. This further highlights the importance of warming up the model, especially in domains where the base model may have knowledge gaps.

B.3 Number of Events

To further study the effect of event quantity, we follow the setup in Section 3.2 and vary the number of events. Increasing the number of events has two potential effects. On one hand, a competent reasoner should leverage the additional information to identify logical inconsistencies. On the other hand, as the reasoning length increases, the likelihood of errors also rises. A longer reasoning process does not necessarily lead to more accurate results (Wei et al., 2022). A capable LLM should ignore any superfluous information and effectively leverage useful context to enhance its reasoning.

The results, summarized in Figure 6, reveal that for the NBA task LLMs generally perform slightly worse as the number of events increases, but for the NFL task performance improves with more events. All else being equal, having more events provides more information and should improve performance; at worst, a strong reasoning model would just ig-

LLMs → Domains ↓		GPT-4o (0513)	GPT-4o (mini)	Qwen2.5 (72B)	LLaMA3.1 (70B)	Mixtral (8x7B)	DeepSeek (R1 32B)	Qwen2.5 (32B)	Mistral (22B)	Phi4 (14B)	Qwen2.5 (14B)	DeepSeek (R1 8B)
Crypto (Bitcoin)	Complete	84%	58%	71%	49%	36%	62%	72%	28%	46%	51%	42%
	Filtered	65%	40%	40%	34%	29%	39%	39%	27%	28%	32%	25%
		↓22.6%	↓31.0%	↓43.7%	↓30.6%	↓19.4%	↓37.1%	↓45.8%	↓3.6%	↓39.1%	↓37.3%	↓40.5%
Trading (IMEX)	Complete	91%	90%	90%	85%	52%	93%	86%	54%	75%	71%	78%
	Filtered	50%	35%	51%	36%	21%	47%	45%	27%	29%	31%	22%
		↓45.1%	↓61.1%	↓43.3%	↓57.6%	↓59.6%	↓49.5%	↓47.7%	↓50.0%	↓61.3%	↓56.3%	↓71.8%
Health (Influenza)	Complete	62%	53%	77%	64%	34%	74%	60%	24%	52%	42%	48%
	Filtered	33%	26%	34%	27%	25%	37%	32%	23%	30%	33%	25%
		↓46.8%	↓50.9%	↓55.8%	↓57.8%	↓26.5%	↓50.0%	↓46.7%	↓4.2%	↓42.3%	↓21.4%	↓47.9%
Energy (Gasoline)	Complete	97%	95%	96%	84%	63%	98%	90%	57%	89%	72%	79%
	Filtered	52%	40%	48%	46%	28%	49%	45%	24%	43%	37%	29%
		↓46.4%	↓57.9%	↓50.0%	↓45.2%	↓55.6%	↓50.0%	↓50.0%	↓57.9%	↓51.7%	↓48.6%	↓63.3%

Table 5: The number of correct event reasoning (through time series) made by LLMs across other domains among testing samples ($N = 100$). Replacing numerical information in the option events—such as dates or prices—with symbols like α (*Filtered*) results in a performance decline compared to retaining the original numerical information (*Complete*). **Red** indicates the best model in this task, while **Blue** represents the second-best.

nore additional events and never perform worse. This discrepancy may stem from fundamental differences between the two sports. In a football game, because teams alternate possessions that comprise multiple correlated plays, or events, making it easier to recognize and match patterns. In basketball, each possession is typically connected to only one event, and events are more independent, and most events impact the score of the game directly. In football, each possession involves many events (at least one recorded for each down in football), but most events do not impact the score of the game. *One insight is that the amount of useful information is different across different domains.*

B.4 Time Series Similarity

We bootstrap $10k$ pairwise distances between win probabilities (i.e., p_{win}) of length 10 in our dataset and normalize them to the range $(0, 1)$. The results show that a large proportion of time series pairs fall within the $(0.4, 1)$ range, e.g., 90.6% for NFL and 91.3% for NBA data. Their distribution can be shown in Figure 7.

C Case Study: How LLMs Infer Events

To further understand how LLMs infer events from time series, we analyze their reasoning process. In this section, we summarize the types of correct and incorrect reasoning process.

C.1 How do language models reason?

As shown in Figure 15, this illustrates the reasoning process of DS-R1-distill-Qwen-32B (DeepSeek-AI, 2025) for NBA events (under a CoT prompt).

The model first interprets the trend in the time series and then matches it with potential events—If the time series exhibits an upward trend, the model aligns it with events favorable to Team A, and vice versa. After sequentially analyzing all data points and their corresponding events, LLMs synthesize their step-by-step analyses to formulate a final reasoning conclusion. High-performing models, such as GPT-4o (Achiam et al., 2023), LLaMA3.1 70B (Dubey et al., 2024), Qwen2.5 72B (Yang et al., 2024), and even smaller language model, like Phi-4 (Abdin et al., 2024) 14B, demonstrate similar reasoning trajectories with CoT Prompting. In addition, for the DS-R1-distill-Qwen-32B, we also observed numerous "aha moments" during the events reasoning process, i.e., self-reflection. For example, in the NBA task, the model reflects midway with "Wait, maybe the rebound isn't enough".

C.2 How do language models fail?

We analyzed 5,000 reasoning samples from DS-R1-distill-Qwen-32B, with the results presented in Figure 8. Both excessively short and overly long reasoning processes tend to result in higher error reasoning result. Model accuracy peaks when the reasoning spans approximately 1,400 tokens.

Too Little Reasoning We observe that the reasoning errors with short process can largely be attributed to what we term "*rushed reasoning*". Instead of carefully analyzing each event in the options, as illustrated in Figure 8, the LLM tends to make hasty generalizations and prematurely draws conclusions. An example is shown in Figure 16,

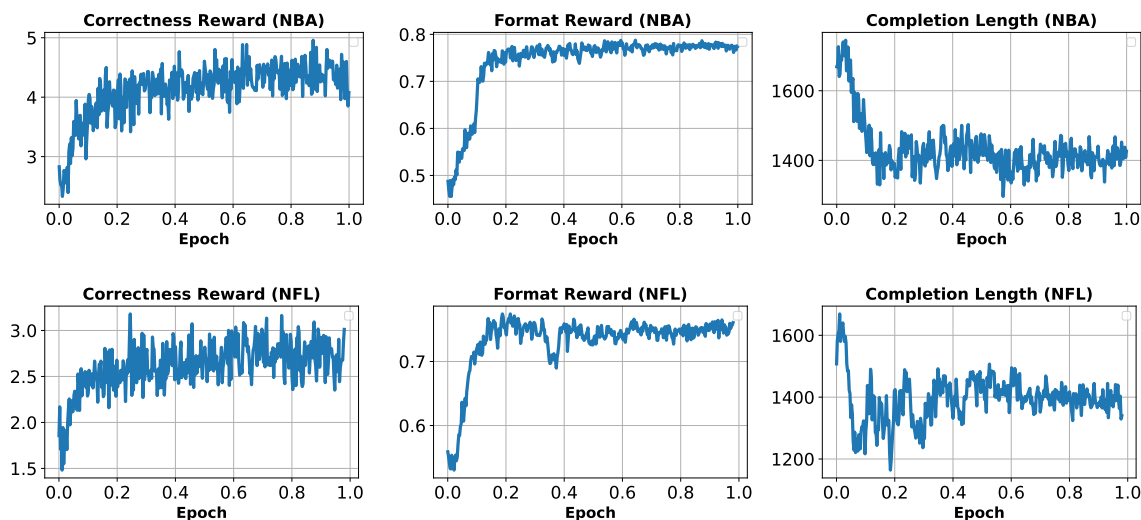


Figure 5: As the number of training steps increases, both the correctness reward (5 is maximum) and the format reward (0.8 is maximum) show clear improvements, while tokens required to complete the reasoning shows a decreasing trend.

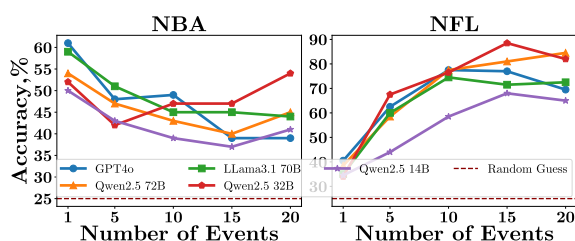


Figure 6: The reasoning performance of LLMs across event sequences of various lengths. The figure includes only models that consistently outperform the baseline.

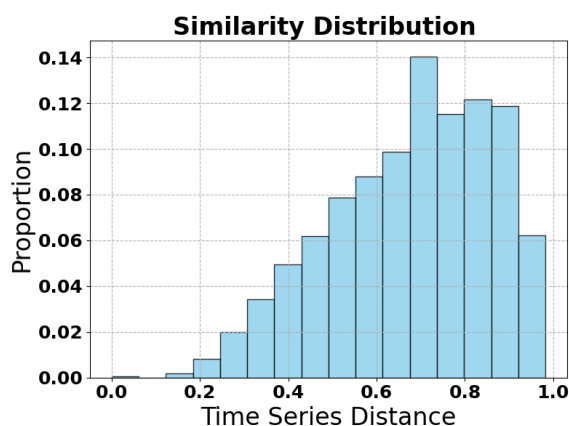


Figure 7: The similarity distribution of time series in sports data, with a time series length of 10. There is a 91% probability that the distance between two time series falls beyond 0.4.

where the LLM is able to recognize the time series pattern and attempts to reason accordingly. However, it merely provides a superficial summary of each option and arrives at a conclusion after insufficient reasoning.

Too Much Reasoning We are not the first to observe that reasoning models, particularly those in the DeepSeek series (Shao et al.; DeepSeek-AI, 2025), tend to engage in excessively long reasoning when making incorrect inferences (Liu et al., 2025; Marjanović et al., 2025). We categorize these types of errors as cases of “*overthinking*,” characterized by excessive *self-reflection* that leads to confusion and prevents the model from arriving at a correct conclusion. For instance, in Figure 8, case B shows the model repeatedly engaging in self-reflection (e.g., “Wait...”) without reaching a final answer. In this example, the model makes 18 self-corrections. In comparison, the average

number of self-reflections in the best-performing range (i.e., token usage between 1,200 to 1,500) is **7.4**, whereas in “*overthinking*” cases, where token usage exceeds 3,000, it rises to an average of **14.1**.

D Prompt Templates

Figure 11 presents the complete template for NBA and U.S Health event reasoning. For NFL data and other domains, we adopt a similar template with minor variations to accommodate domain-specific characteristics. For instance, in cryptocurrency data (Li et al., 2024a), we specify that the provided time series represents daily “Closing Prices,” while in Energy data (Liu et al., 2024a), it corresponds to

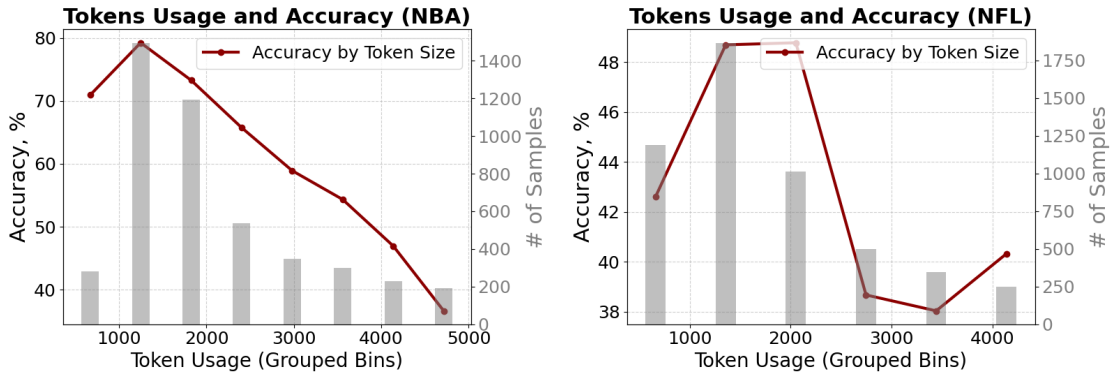


Figure 8: The relationship between token usage and reasoning accuracy. For both tasks, we sampled around 5,000 examples. We observe that DS-R1-distill-Qwen-32B achieves higher reasoning accuracy when using either fewer or more tokens, with peak accuracy occurring around 1,400 tokens.

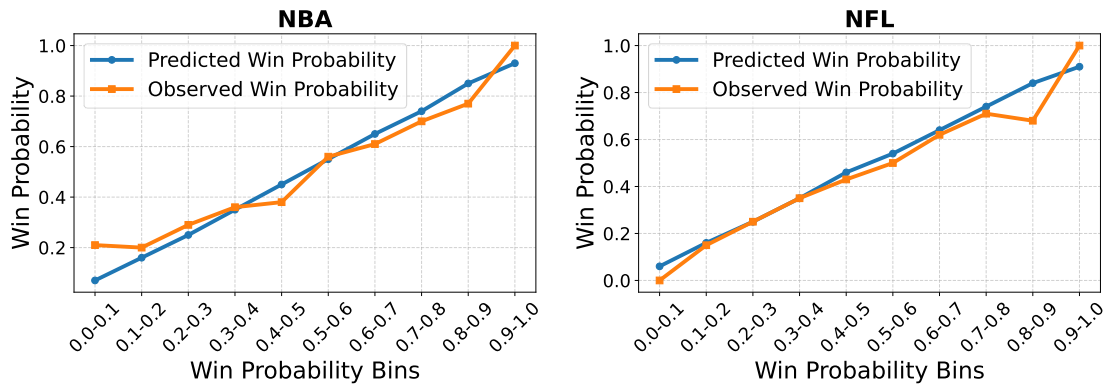


Figure 9: The Calibration of Win Probability Predictions. The results show a high degree of alignment between the model's predictions and the actual game outcomes.

the "Dollars per Gallon." (Gasoline). In addition, considering the delayed impact of real-world news, we included news events from the previous two timestamp in the options. Figure 12 and Figure 13 provide concrete examples of the baseline reasoning setup and the reasoning process with partial events provided, respectively.

Figure 14 illustrates the Chain-of-Thought (CoT) prompt for NBA event reasoning, with the format up to the "options" section remaining consistent across prompts. The CoT prompt for NFL follows a similar structure with slight modifications, such as ensuring that example events and background knowledge align with the context of American football. Note that we acknowledge that the current CoT prompt still has room for improvement, however, due to time and computational constraints, we have not conducted further explorations.

E Win Probability Calibration

To evaluate ESPN's win probability model, we performed model calibration using the predicted win

probabilities at the start of each game and the corresponding outcomes. Specifically, we compared the predicted win rates within each probability bin to the actual win rates observed in those bins, and results in Figure 9 show high degree of consistency between the predictions and the true outcomes.

F Licensing

The code from our work is released under the MIT License, while the dataset is made available under the Creative Commons Attribution-NonCommercial-ShareAlike (CC BY-NC-SA) license. This allows anyone to use, distribute, and modify the data for non-commercial purposes, provided they give proper attribution and share any derivative works under the same license terms. The sports data we use is licensed by the sports leagues to third-party providers like ESPN, which processes the data and adds its win probability model data to it. Because of the complex licensing issues with this data, we do not redistribute it, but our repository includes code for obtaining similar data.

Prompt Format of Post-training

```
< |im_start| > Respond reasoning process in the following format:  
<reasoning>  
...  
</reasoning>  
Return your answer in **X**, where **X** is your answer and can only be one of the  
selected options, such as **a**, **b**, **c**, or **d**.  
{{Question}}          <== The Reinforcement Learning Input Ends Here.  
<reasoning>  
{{Reasoning Process}}  
</reasoning>  
**{{Answer}}**< |im_end| >
```

Figure 10: The format of post-training, enclosing the reasoning process within “<reasoning>” tags and wrapped the final answer with “****X****” to maintain consistency with other evaluation formats.

Event Reasoning in Sports (Basketball)

You are an assistant for NBA basketball task. We will provide a series of consecutive timestamps, win probabilities from a basketball game, though some intermediate events will be missing. You will need to infer the likely events that occurred in the missing intervals. Below is provided timestamps, win probabilities (team A).

Step 1.	TimeStamp ₁	WP ₁
Step 2.	TimeStamp ₂	WP ₂
Step 3.	TimeStamp ₃	WP ₃
...		
Step k.	TimeStamp _k	WP _k

Please select the correct sequence of events for steps 2, ..., $k - 1$ from the four options below, Here are the potential options:{{options}}

Here is the instruction for returning reasoning results in:{{format}}

Event Reasoning in Other Domains (U.S Health)

You are an assistant for an Influenza Patients task. We will provide a series of consecutive timestamps along with the Influenza Patients Proportion. Additionally, we will present four potential event (news) sequences that occurred during that period, as well as from the previous two days. Your task is to identify and select the correct sequence of events. Below is provided date and Patients Proportion (%),

Step 1.	Date ₀	x_0
...		
Step k.	Date _k	x_k

Please select the correct sequence of events for steps 2, ..., $k - 1$ from the four options below, Here are the potential options:{{options}}

Here is the instruction for returning reasoning results in:{{format}}

Figure 11: Template used to infer events through win probabilities in sports (e.g., basketball) and open-world domains (e.g., U.S Health). For NFL data and other domains, we adopt a similar template with minor variations to accommodate domain-specific characteristics. In addition, we use the format “****X****”. Time-series timestamps (e.g., dates) can be kept or removed depending on the domain.

An example of basketball event inferring prompt (Baseline Setting).

You are an assistant for NBA basketball task. We will provide a series of consecutive win probabilities from a basketball game, though some intermediate events will be missing. You will need to infer the likely events that occurred in the missing intervals. Below is provided win probabilities (team A).

0. 17.1%
1. 16.1%
2. 19.8%
3. 16.6%
4. 15.6%
5. 18.3%
6. 15.3%
7. 10.4%
8. 10.1%
9. 10.1%
10. 10.1%
11. 10.1%

Please select the correct sequence of events for steps 1,2,3,4,5,6,7,8,9,10 from the four options below,

a.

1.Team A defensive team rebound

(... Skip)

10.Player from team B enters the game for Player from team B

b.

1.Player from team A lost ball turnover (Player from team A steals)

(... Skip)

10.Player from team B misses two point shot

c.

1.Player from team B enters the game for Player from team B

(... Skip)

10.Player from team A makes dunk

d.

1.Player from team A makes two point shot (Player from team A assists)

(... Skip)

10.Player from team A misses 23-foot three point jumper

Return your answer in the format ****X****, where X only contains the chosen option, such as ****a****, ****b****, ****c****, or ****d****.

Figure 12: This illustrates an actual prompt for basketball event reasoning, where the LLM is tasked with inferring the underlying events based solely on a sequence of win probabilities.

An example of event inferring prompt (Time Series and Events are provided).

You are an assistant for NBA basketball task. We will provide a series of consecutive win probabilities from a basketball game along with a subset of events, though some intermediate events will be missing. You will need to infer the likely events that occurred in the missing intervals.

Below is provided win probabilities (team A).

0. 71.1% Player from team B defensive 3-seconds (technical foul)

1. 75.9%

2. 80.9%

3. 81.7%

4. 81.7%

5. 79.1%

6. 76.1%

7. 79.2%

8. 73.7%

9. 83.1%

10. 88.0%

11. 85.8% Player from team B makes driving layup

Please select the correct sequence of events for steps 1,2,3,4,5,6,7,8,9,10 from the four options below,

a.

1.Player from team B enters the game for Player from team B

(... Skip)

10.Player from team B enters the game for Player from team B

b.

1.Player from team B defensive rebound

(... Skip)

10.Player from team A misses free throw 2 of 2

c.

1.Player from team A makes free throw

(... Skip)

10.Player from team A makes 26-foot three point jumper (Player from team A assists)

d.

1.Player from team B makes free throw 1 of 2

(... Skip)

10.Player from team A shooting foul

Please return your answer directly in the format ****X****, where X only contains the chosen option, such as ****a****, ****b****, ****c****, or ****d****.

Figure 13: This illustrates an actual prompt for basketball event reasoning, where the LLM is tasked with inferring the underlying events based on a sequence of win probabilities and the first and the last events during this period.

Chain-of-Thought (CoT) Prompt Template for Basketball

You are an assistant for NBA basketball task ...

...

Here are the potential options:{{options}}

When reasoning, please refer to the following guidelines:

1. Analyze Team A's 1st win probability change. For example:

- 0. 53.4%

- 1. 57.1%

- Team A's win probability increased from 53.4% to 57.1%, indicating that an event positive to Team A or negative to Team B has occurred.

2. Pause the win probability analysis for now and focus on the 1st event in options a, b, c, d, and eliminate the inconsistent options or skip this step if unsure. Please ignore very minor changes in win probabilities, such as 0.1%-0.4%. Base your decision on significant changes.

3. Similarly, turn to analyze the change in the 2nd win probability and match it with the 2nd event in the remaining options, and eliminate the inconsistent options or skip this step if unsure.

4. Repeat the above process until all incorrect options are eliminated.

Consider the following information when reasoning, it will assist you in reasoning.

1. Positive events for Team A, such as "Player from Team A scores (e.g., makes a layup or jump shot) or rebounds," or negative events for Team B, such as "Player from Team B makes a mistake (e.g., misses a shot, commits a foul, or loses the ball)."

2. Some events do not impact the win probability, such as a player entering the game or a timeout.

3. Ignore very minor changes in win probabilities, such as 0.1%-0.4%.

Return your answer in the format ****X****, where X only contains the chosen option, such as ****a****, ****b****, ****c****, or ****d****.

Figure 14: Chain-of-Thought (CoT) Template used to infer events through win probabilities in sports (e.g., basketball). For NFL data, we adopt a similar CoT Prompt with minor variations to accommodate domain-specific characteristics.

<think>

Alright, I'm trying to figure out which sequence of events corresponds to the given win probabilities for Team A. The probabilities are as follows:

0. 31.9%

1. 43.9%

... ..

10. 56.7%

11. 50.9%

I need to **analyze the changes in win probability after each event** and match them with the options provided (a, b, c, d). Let's go through each step one by one.

Starting with the first change: from 31.9% to 43.9%. That's an increase of about 12%, which is significant. This suggests a positive event for Team A or a negative event for Team B. Looking at the options, the first events are:

- a. Player A makes a 3-pointer with an assist.
- b. Player B makes a free throw 2 of 2.
- c. Player B defensive rebound.
- d. Timeout and coach's challenge.

Option a's first event is a 3-pointer made by Team A, which would increase their win probability. Option b's first event is a free throw made by Team B, which would decrease Team A's win probability, but the probability went up, so b might be inconsistent.

...

So, option a seems the most likely for the first event because scoring a 3-pointer would increase Team A's chances.

...

Wait, maybe the rebound isn't enough, ...

...

The sixth change: from 43.0% to 49.3%, an increase of about 6.3%. Positive event for Team A.

....

Putting it all together, option a seems to fit all the significant changes in win probability.

The other options don't align as well with the increases and decreases observed.

<think>

a

Figure 15: Reasoning process example from basketball data: In this case, the ground-truth answer is "a", which aligns with the reasoning of DeepSeek-R1 32B (DeepSeek-AI, 2025). Other LLMs, such as GPT-4o (Achiam et al., 2023), follow a similar reasoning pattern. To save space, some content has been omitted.

Case Study A: Failed reasoning with a Short Process

<think>

The win probabilities start at 91.6% and fluctuate a bit, but overall, Team A's chances seem to be quite high, so I'm guessing the events should reflect a game where Team A is maintaining or slightly increasing their lead.

Let me look at each option one by one.

Option a:

Team A defensive rebound. That would likely keep their ...

Option b:

Team B offensive rebound. That's bad for Team A because ...

Option c:

Substitutions happening. That might not directly affect ...

Option d:

Team A makes a three-pointer, which is good. Then Team B ...

I think Option b is better because it shows Team A making baskets and Team B missing, which ...

So I'm leaning towards option b.

</think>

b

Case Study B: Failed reasoning with a Long Process

Okay, so I need to figure out which sequence of events (a, b, c, or d) corresponds best to the given win probabilities ...

So, the initial increase from 27.5% to 29.7% suggests a positive event for Team A. Then a slight increase again ...

Starting with option a:

...

Wait, but the first step is from 27.5...

Option b:

...

Wait, but the win probability went up again,

...

Wait, maybe I'm miscounting. Let me list the events in option b: ...

...

Wait, maybe I'm miscounting the steps. Let me make sure ...

...

Wait, the user provided win probabilities from 0 to 11 ...

...

Wait, maybe the missed layup is followed by a defensive rebound, ...

...

Wait, perhaps the win probabilities are for Team A, so when Team B scores, ...

...

Wait, maybe I miscounted the events. Let me check again.

...

Wait, but the events are listed as steps 1-10, each corresponding

Given all, I think option b is the best fit despite the inconsistency at step 8, because ...

</think>

b

Figure 16: Case A illustrates an example of "rushed reasoning," where the LLM provides only short descriptions of each option before arriving at a conclusion. In contrast, Case B presents an instance of "overthinking," characterized by excessive self-reflection that leads to confusion and ultimately prevents the model from reaching a correct conclusion.