Inferring Events from Time Series using Language Models

Anonymous ACL submission

Abstract

Time series data measure how environments change over time and drive decision-making in critical domains like finance and healthcare. A common goal in analyzing time series data is to understand the underlying events that cause the observed variations. We conduct the first study of whether Large Language Models (LLMs) can infer events described with natural language from time series data. We evaluate 18 LLMs on a task to match event sequences with real-valued time series data using a new benchmark we develop using sports data. Several current LLMs demonstrate promising abilities, with OpenAI's o1 performing the best but with DS-R1-distill-Qwen-32B outperforming proprietary models such as GPT-4o. From insights derived from analyzing reasoning failures, we also find clear avenues to improve performance. By applying post-training optimizations, i.e., distillation and self-improvement, we significantly enhance the performance of the Qwen2.5 1.5B, achieving results second only to o1.*

1 Introduction

Time series data are pervasive. Examples of time series include wearable device measurements of users' actions (Anguita et al., 2013), clinical records about changes in health (Harutyunyan et al., 2019), and asset market prices (Wang et al., 2024c; Li et al., 2024a). Each of these examples represents a real-valued sequence of data points with time stamps. In addition to the real-valued data, a time series often has associated events described in natural language. Although the causal connections between the real-valued data and the natural language events is often uncertain, the events are thought to be correlated in some way with the real-valued data. Figure 1 illustrates an example from sports—events favorable to Team A increase its win probability, while unfavorable events decrease it. Benefiting from the promising potential of integrating natural language with time series analysis (Jin et al., 2024), along with the rapid advances in natural language processing, LLMs have been employed for important time series analysis tasks including forecasting (Wang et al., 2024c; Williams et al., 2024; Liu et al., 2024a; Tan et al., 2024), anomaly detection (Dong et al., 2024; Liu et al., 2024b), and time series understanding (Cai et al., 2024; Li et al., 2024a,b). The goal of analyzing time series data is often to infer events occurring in the measured environment (Liu et al., 2024b). This motivates our work to explore how LLMs infer event descriptions given context and time series data.

Prior work on reasoning about time series in conjunction with natural language has largely overlooked event descriptions (Merrill et al., 2024; Williams et al., 2024) and primarily focused on numerical sequences, such as trend analysis (Cai et al., 2024) or anomaly detection (Dong et al., 2024). Some studies collect sequences of news related to time series (Wang et al., 2024c; Liu et al., 2024a; Cheng and Chin, 2024), however they are curated for forecasting and do not explore reasoning from the real-valued 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 benchmark comprising time series data and associated natural language event descriptions where there is a clear and strong connection between the events and real-valued data. Our dataset (Section 3.3) includes 4,200 games from NBA (basketball) and NFL (American football) sports leagues, comprising a total of 1.7 million data points and events. The real-valued data is the win probability† and

*All resources needed to reproduce our work are available: https://anonymous.4open.science/r/Event_Infer-1B20/

†We use win probability values output from ESPN's game

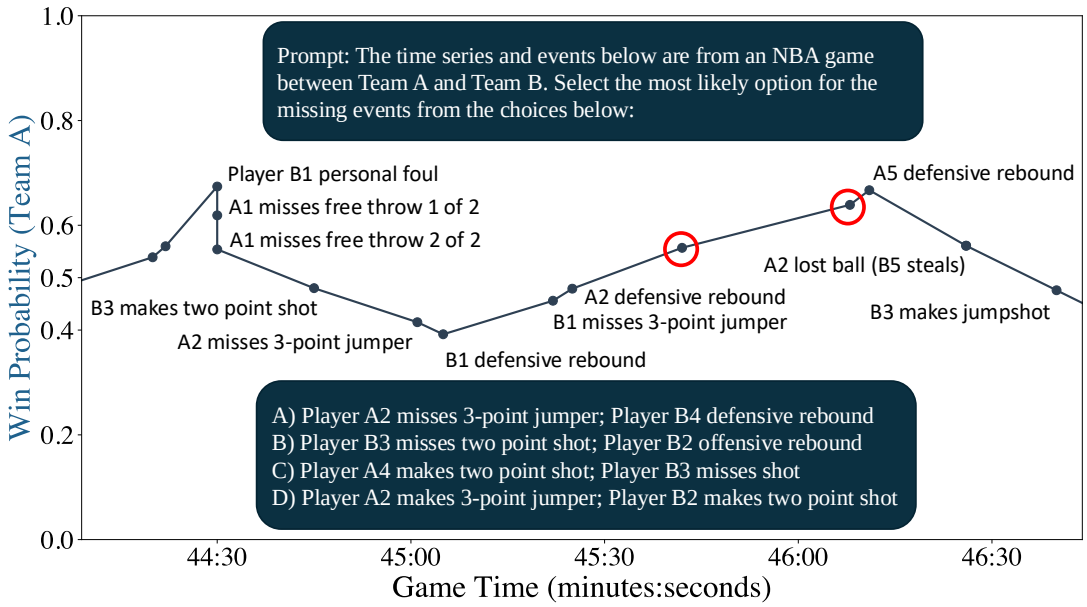

Figure 1: Illustration of time series event reasoning. The prompt provides (in text form, see details later in the paper) a time series of real-valued data (win probabilities) and corresponding natural language event descriptions. The model is prompted to select the most likely sequence of events for some segment of the time series data where no events are provided. (This example is taken from near the end of an NBA game, which is 48 minutes regulation time, between the Dallas Mavericks (Team A) and Los Angeles Lakers (Team B), 1 November 2019.)

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 LLMs' 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, including 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*. For instance, OpenAI o1 achieved the highest accuracy of 83% in NBA events reasoning, followed by DS-R1-distill-Qwen-32B with 68%, and GPT4o with 41%. However, through post-training with a distillation phase followed by self-improvement, i.e., GRPO (Shao et al., 2024), optimization, 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 contributions include introducing an evaluation approach (Section 3.2) to assess LLMs' ability to reason about event sequences through time series and extend it to multiple domains (Section 4.5). We create an easily extensible dataset 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 & Section 4.3). And we identify that post-training optimization can significantly enhance LMs on events inferring (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 used text to assist in time series reasoning (including forecasting), achieving promising results (Cao et al., 2024; Wang et al., 2024a; Xie et al., 2024). These works have made significant contributions to fields such as sociology (Cheng

analysis (https://www.espn.com/analytics/). As we discuss in Section 6, win probability is an effective measure of game state but potentially differs from ground truth.

Benchmark/ Evaluation	Properties (with Time Series)		
	Context	Source	Task
Williams et al. (2024)	Description	Manual	Forecasting
Merrill et al. (2024)	Description	Synthetic	Reason & Forecast
Cai et al. (2024)	Question	Manual	Understanding
Liu et al. (2024a)	News Series	Real-World	Forecasting
	Properties (without Time Series)		
Fatemi et al. (2024)	Event & Time	Synthetic	Temporal Reasoning
Xiong et al. (2024)	Event & Time	Synthetic	Temporal Reasoning
Chu et al. (2023)	Event & Time	Real-World	Temporal Reasoning
Quan and Liu (2024)	Event Sequence	Synthetic	Sequential Reasoning
Karger et al. (2024)	Event	Real-World	Future Forecasting
Ours	Time Series & Event Sequence	Real-World	Events 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. We propose reasoning about event sequences through time series data, incorporating corresponding timestamps.

and Chin, 2024), energy (Wang et al., 2024c; Xu et al., 2024), and finance (Li et al., 2024a; Wang et al., 2024b). For example, Williams et al. (2024) manually curated time series forecasting data along with related text to highlight the importance of incorporating textual information when using LLMs for forecasting. Wang et al. (2024c) 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., 2024; Zhou and Yu, 2024) often involves processing 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 LLMs for Events Reasoning

Reasoning is an ill-defined and broad, yet critical, capability that determines LLMs' performance across many complex tasks. Therefore, numerous reasoning benchmarks have been developed for valuable tasks, such as coding (Zhuo et al., 2024; Jain et al., 2024), mathematics (Cobbe et al., 2021; White et al., 2024), and finance (Xie et al., 2023; Islam et al., 2023). Additionally, some bench-

marks have evaluated the general reasoning abilities of LLMs (Bang et al., 2023; White et al., 2024; bench authors, 2023), including BBH (Suzgun et al., 2022) and MMLU (Hendrycks et al., 2020).

Several benchmarks, as listed in Table 1, 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). For instance, Karger et al. (2024) introduced a dynamically updated benchmark to evaluate LLMs' forecasting of future events. Fatemi et al. (2024) 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 event sequences, which is the focus of our work.

3 Benchmark

We next define the benchmark task, outline the evaluation format, and introduce the dataset details.

3.1 Problem Definition

A time series is a sequence of timestamped real values: $x = [(t_0, x_0), (t_1, x_1), ..., (t_T, x_T)]$. An event sequence is a sequence of timestamped text descriptions of events: $e = [(t_0, e_0), (t_1, e_1), ..., (t_T, e_T)]$. For each sequence, the timestamps t are monotonically increasing ($t_i \leq t_j$ if $i < j$). While the timestamps of the time series and event sequence need not be identical, there is often a one-to-one correspondence, with an event description associated with each real value. Critically, the events describe changes in the environment that result in changes in the time series values.

Given a dataset $\mathcal{D} = (\mathcal{X}, \mathcal{E})$ containing N real-valued time series and timestamp t with corresponding event sequences of length T, we are concerned with time series data represented as a pair of sequences:

$$\mathcal{X} = [(t_0, x_0), (t_1, x_1), ..., (t_{N-1}, x_{N-1})]$$

consisting of real-valued measurements, and

$$\mathcal{E} = [(s_0, e_0), (s_1, e_1), ..., (s_{T-1}, e_{T-1})]$$

comprising natural language event descriptions. Although there may not always be a direct causal relationship between the events and measurements, we assume there is some connection between the events and measurements and that the timestamps, s_j and t_i, are synchronized. Note that we do not assume that there is one event associated with each data value, or even that the timestamps of events and data values match, only that they are aligned so the ordering relationships between values in $\mathcal{X}$ and events in $\mathcal{E}$ are known.

Our goal is to interrogate an LLM's understanding of time series data by measuring its ability to infer unobserved values in $\mathcal{E}$ given $\mathcal{X}$. As illustrated in Figure 1, when the intermediate event sequence is missing, the LLM is expected to infer it using the provided real-valued time series and corresponding timestamps.

3.2 Events Reasoning Format

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 x_i
t_{i+1} x_{i+1}
...
t_{i+k-2} x_{i+k-2}
t_{i+k-1} x_{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 (e.g., $x_{i:i+k-1}$). Since time series data are typically accompanied by timestamps, the corresponding timestamps $t_{i:i+k-1}$ are provided in the prompt. The intermediate events are missing, and the LLM is tasked with inferring these events. To make the task tractable we provide *four* options, one of which corresponds to the actual sequence of events, and prompt the model to select the most likely option. Figure 12 in Appendix D gives examples of the full prompts used in our experiments.

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 Sports Dataset

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 events are directly correlated with the real-valued data; and (2) the data are continuously refreshed with new games being played every data.For the natural language events, we used play-by-play data provided by ESPN that captures key occurrences during a game, such as scoring, turnovers, or fouls in basketball. ESPN also provides each teams' predicted win probability throughout the game, which we use as the real-valued time series data. These win probabilities reflect the state of the game, as well as some knowledge about the teams, at each time step. Since a game constitutes 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 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.

Our dataset contains 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 these 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?, and **RQ5:** How does LLMs' event reasoning performance compare across different domains?

We evaluate 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). Additionally, we test models designed for reasoning, including DeepSeek Distilled model, like DS-R1-distill-Qwen-32B (DeepSeek-AI, 2025), and OpenAI's o1 (OpenAI, 2024). Our findings indicate that reasoning models perform particularly better.

4.1 Accuracy Evaluation

To evaluate LLMs on event inference, we first follow the format in Figure 12 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 on the NBA 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 overall more challenging. The performance of GPT-4o drops to 29%, while DS-R1-distill-Qwen-32B and o1 achieve 43% and 75.5%. 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-

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

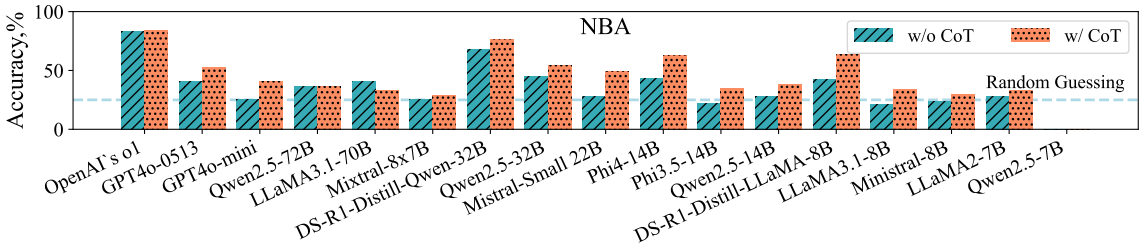

Figure 2: The performance on NBA data indicates that open-weight models, such as Qwen2.5 72B, achieve results comparable to or even surpassing proprietary models like GPT-4o. In particular, reasoning-focused models such as DS-R1-distill-Qwen-32B and OpenAI's o1 significantly outperform others. Additionally, Chain-of-Thought (CoT) prompting further enhances reasoning capabilities. Similar trends are observed in the NFL data, with details provided in Figure 5 in Appendix B.1. Note that open-weight models are presented in order of model size.

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 13 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 (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., 2024). 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' performance 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, highlight-

Language Models	Performance (NBA, # of)			Performance (NFL, # of))			Post-training	
	Correctness	Error	Invalid	Correctness	Error	Invalid	Warmup	RL
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: The performance of the Qwen2.5 (1.5B) model under different post-training strategies is evaluated using 200 test cases and compared against three other large language models. When applying both warm-up (i.e, Knowledge Distillation) and reinforcement learning (i.e., GRPO) training, the 1.5B-size model achieves the second-best performance, surpassed only by OpenAI's o1, and outperforming the distillation source DS-R1-distill-Qwen-32B (DS-R1-32B). Red indicates the best model in this task, while Blue represents the second-best.

ing the necessity of removing them when evaluating reasoning (Fatemi et al., 2024). 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 distance D between time series using the 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 8 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. We keep the ground-truth events and question time series consistent across all levels. The results are presented in Figure 3, showing a

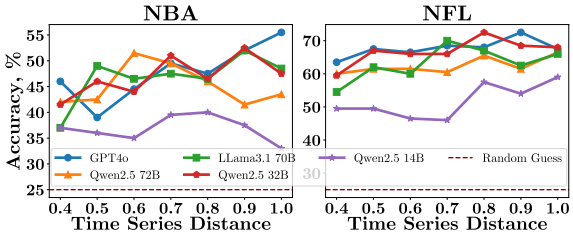

Figure 3: The performance of LLMs in distinguishing events corresponding to time series (win probabilities) with different levels of similarity. Time series *similarity* decreases as x (i.e., time series *distance*) increases.

slight upward trend in LLM performance as similarity decreases. This is due to the inherent consistency between time series and event sequences, which LLMs are able to recognize.

4.5 Other Domains

In real-world open environments, time series data usually coexist with related textual sequences, such as cryptocurrency prices alongside relevant financial news (Li et al., 2024a). 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 prevent the questions from becoming too long, 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 occurrences, more details are in Appendix A.3. Our question follows the format in Section 3.2, with an event sequence length of

Tasks	Language Models	Baseline (TS Only)	Available Context			Ablations			Options
			TS+Times	TS+Score	TS+Event	w/ Name	Remove	Replace	Reorder
Basketball Reasoning	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%**
Football Reasoning	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 conditions based on the baseline (i.e., providing only time series) and relative impacts. We provide each model with 200 questions for each condition ($N = 200$). Red highlights the best-performing model under a given condition, while Blue represents the second-best.

10, corresponding to 10 trading days for Bitcoin data or 10 weeks of influenza statistics in the U.S. health dataset.

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

Table 4: The accuracy of LLMs inferring events across other domains among 100 questions. Replacing numerical information in the events (*Filtered*) results in a performance decline compared to retaining the original numbers (*Complete*). Red indicates the best model in this task, while Blue represents the second-best.

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.1.

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 performance on the NBA task.

6 Limitations and Ethical Considerations

Our dataset includes time series representing win probabilities in sports, which serve as a 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.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, Jorge Luis Reyes-Ortiz, et al. 2013. A public domain dataset for human activity recognition using smartphones. In *Esann*, volume 3, page 3.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

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. *arXiv preprint arXiv:2410.14752*.

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 *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. *arXiv preprint arXiv:2408.07773*.

Junyan Cheng and Peter Chin. 2024. Sociodojo: Building lifelong analytical agents with real-world text and time series. In *ICLR*.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Haotian Wang, Ming Liu, and Bing Qin. 2023. Timebench: A comprehensive evaluation of temporal reasoning abilities in large language models. *arXiv preprint arXiv:2311.17667*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Manqing Dong, Hao Huang, and Longbing Cao. 2024. Can llms serve as time series anomaly detectors? *arXiv preprint arXiv:2408.03475*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Hugging Face. 2025. Open r1: A fully open reproduction of deepseek-r1.

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

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 preprint arXiv:2009.03300*.

Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. 2023. Financebench: A new benchmark for financial question answering. *arXiv preprint 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 preprint arXiv:2403.07974*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024a. Mixtral of experts. *arXiv preprint 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 preprint 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 *ICML*.

Ezra Karger, Houtan Bastani, Chen Yueh-Han, Zachary Jacobs, Danny Halawi, Fred Zhang, and Philip E Tetlock. 2024. Forecastbench: A dynamic benchmark of ai forecasting capabilities. *arXiv preprint arXiv:2409.19839*.

Svenja Kenneweg, Jörg Deigmöller, Philipp Cimiano, and Julian Eggert. 2025. Traveler: A benchmark for evaluating temporal reasoning across vague, implicit and explicit references. *arXiv preprint arXiv:2505.01325*.

Hoyoung Lee, Youngsoo Choi, and Yuhee Kwon. 2024. Quantifying qualitative insights: Leveraging llms to market predict. *arXiv preprint 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 *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1094–1106.

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 preprint arXiv:2410.10624*.

Haoxin Liu, Shangqing Xu, Zhiyuan Zhao, Lingkai 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 *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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 preprint arXiv:2405.15370*.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective, 2025. *URL https://arxiv.org/abs/2503.20783*.

Sara Vera Marjanović, Arkil Patel, Vaibhav Adlakha, Milad Aghajohari, Parishad BehnamGhader, Mehar Bhatia, Aditi Khandelwal, Austin Kraft, Benno Krojer, Xing Han Lù, et al. 2025. Deepseek-r1 thoughtology: Let's< think> about llm reasoning. *arXiv preprint 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 *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. *Preprint*, arXiv:2501.19393.

OpenAI. 2024. Learning to reason with llms, september.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Yinzhu Quan and Zefang Liu. 2024. Econlogicqa: A question-answering benchmark for evaluating large language models in economic sequential reasoning. *arXiv preprint arXiv:2405.07938*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. Wu Y.K. Li, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models.

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 preprint arXiv:2504.07615*.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

Mingtian Tan, Mike A Merrill, Vinayak Gupta, Tim Althoff, and Thomas Hartvigsen. 2024. Are language models actually useful for time series forecasting? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Chengsen Wang, Qi Qi, Jingyu Wang, Haifeng Sun, Zirui Zhuang, Jinming Wu, Lei Zhang, and Jianxin Liao. 2024a. Chattime: A unified multimodal time series foundation model bridging numerical and textual data. *AAAI*.

Shengkun Wang, Taoran Ji, Linhan Wang, Yanshen Sun, Shang-Ching Liu, Amit Kumar, and Chang-Tien Lu. 2024b. Stocktime: A time series specialized large language model architecture for stock price prediction. *arXiv preprint arXiv:2409.08281*.

Xinlei Wang, Maike Feng, Jing Qiu, Jinjin Gu, and Junhua Zhao. 2024c. From news to forecast: Integrating event analysis in llm-based time series forecasting with reflection. *arXiv preprint arXiv:2409.17515*.

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

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Siddartha Naidu, et al. 2024. Livebench: A challenging, contamination-free llm benchmark. *arXiv preprint arXiv:2406.19314*.

Andrew Robert Williams, Arjun Ashok, Étienne Marcotte, Valentina Zantedeschi, Jithendaraa Subramanian, Roland Riachi, James Requeima, Alexandre Lacoste, Irina Rish, Nicolas Chapados, et al. 2024. Context is key: A benchmark for forecasting with essential textual information. *arXiv preprint arXiv:2410.18959*.

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 preprint arXiv:2306.05443*.

Zhe Xie, Zeyan Li, Xiao He, Longlong Xu, Xidao Wen, Tieying Zhang, Jianjun Chen, Rui Shi, and Dan Pei. 2024. Chatts: Aligning time series with llms via synthetic data for enhanced understanding and reasoning. *arXiv preprint arXiv:2412.03104*.

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

Zhijian Xu, Yuxuan Bian, Jianyuan Zhong, Xiangyu Wen, and Qiang Xu. 2024. Beyond trend and periodicity: Guiding time series forecasting with textual cues. *arXiv preprint arXiv:2405.13522*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2.5 technical report. *arXiv preprint 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 preprint 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 preprint arXiv:2504.07912*.

Zihao Zhou and Rose Yu. 2024. Can llms understand time series anomalies? *arXiv preprint arXiv:2410.05440*.

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. 2024. Bigcodebench: Benchmarking code generation with diverse function calls and complex instructions. *arXiv preprint arXiv:2406.15877*.

A Appendix

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, 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.

[§]All information and settings needed are available:

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.

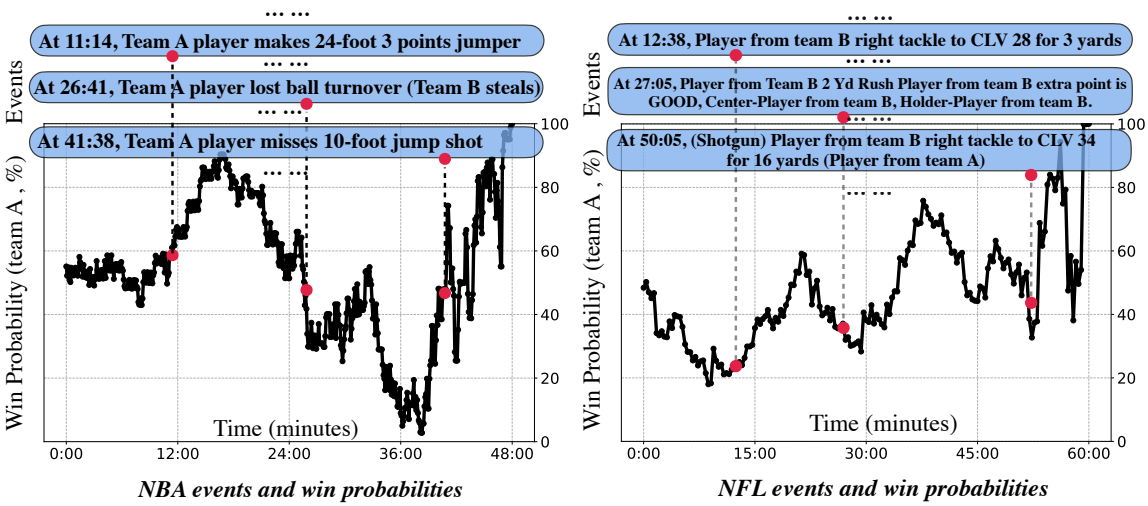

Figure 4: Examples of events and win probabilities in the NBA and NFL dataset. As the game progresses, ESPN provides descriptions of on-field events along with the corresponding win probabilities for each team at that moment. These probabilities can be considered a representation of the team's current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

B Detailed Experimental Results.

B.1 LLMs performance

The performance of language models on NFL event inference is presented in Figure 5. Overall, NFL event inference is more challenging than that of the NBA. Nevertheless, OpenAI o1 remains the best-performing model, followed by DS-R1-distill-Qwen-32B.

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 Events Inferring

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. Specifically, as shown in Table 2, in the NFL dataset, it correctly infers 29 out of 200 test cases and produces 102 invalid answers, while in 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 11. The results in 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., 2024). 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., 2024)), 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 11. Specifically, we primarily supervised two types of rewards: *format* and *correctness*, with the training reward trajectories illustrated in Figure 6. The training was conducted using the open-r1 (Face,

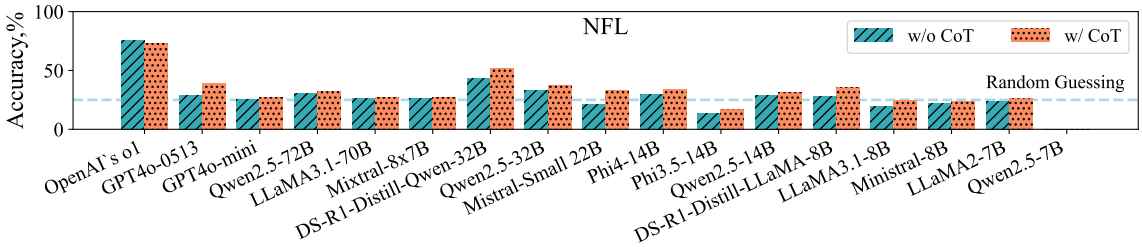

Figure 5: The performance of various language models on NFL events inferring through time series. Overall, this task is more challenging than NBA event reasoning.

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% ↓22.6%	40% ↓31.0%	40% ↓43.7%	34% ↓30.6%	29% ↓19.4%	39% ↓37.1%	39% ↓45.8%	27% ↓3.6%	28% ↓39.1%	32% ↓37.3%	25% ↓40.5%
Trading (IMEX)	Complete	91%	90%	90%	85%	52%	93%	86%	54%	75%	71%	78%
	Filtered	50% ↓45.1%	35% ↓61.1%	51% ↓43.3%	36% ↓57.6%	21% ↓59.6%	47% ↓49.5%	45% ↓47.7%	27% ↓50.0%	29% ↓61.3%	31% ↓56.3%	22% ↓71.8%
Health (Influenza)	Complete	62%	53%	77%	64%	34%	74%	60%	24%	52%	42%	48%
	Filtered	33% ↓46.8%	26% ↓50.9%	34% ↓55.8%	27% ↓57.8%	25% ↓26.5%	37% ↓50.0%	32% ↓46.7%	23% ↓4.2%	30% ↓42.3%	33% ↓21.4%	25% ↓47.9%
Energy (Gasoline)	Complete	97%	95%	96%	84%	63%	98%	90%	57%	89%	72%	79%
	Filtered	52% ↓46.4%	40% ↓57.9%	48% ↓50.0%	46% ↓45.2%	28% ↓55.6%	49% ↓50.0%	45% ↓50.0%	24% ↓57.9%	43% ↓51.7%	37% ↓48.6%	29% ↓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.

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.; 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 7, 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 ignore 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,

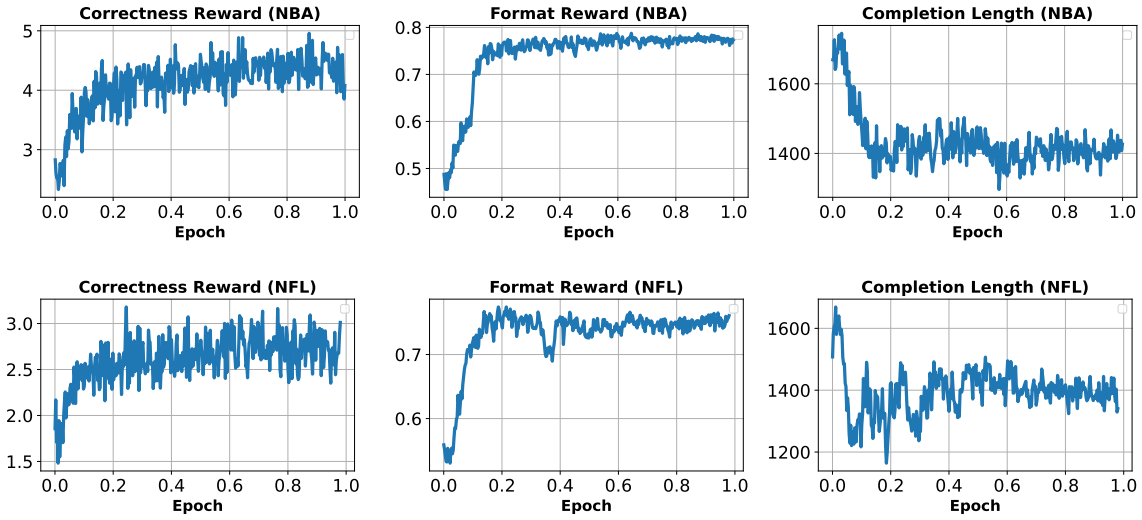

Figure 6: As the number of training steps increases, both the correctness reward (5 is maximum) and the format reward (0.8 is maximum) show clear improvements, while tokens required to complete the reasoning shows a decreasing trend.

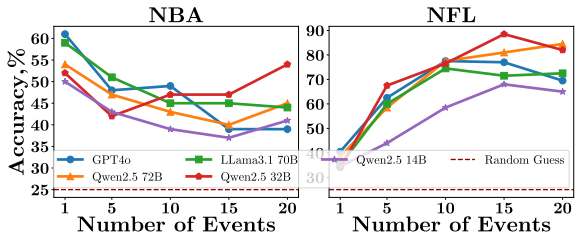

Figure 7: The reasoning performance of LLMs across event sequences of various lengths. The figure includes only models that consistently outperform the baseline.

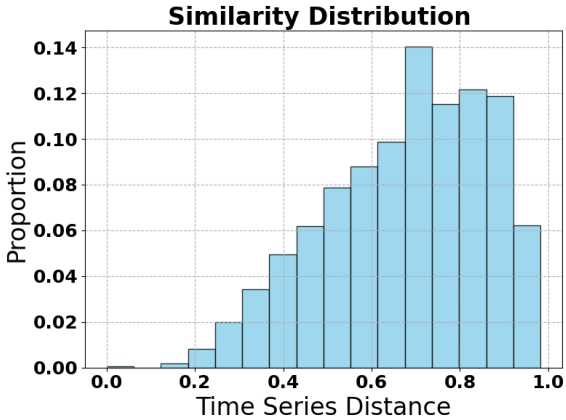

Figure 8: The similarity distribution of time series in sports data, with a time series length of 10. There is a 91% probability that the distance between two time series falls beyond 0.4.

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., $\mathbf{p}_{win}$) 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 8.

C Case Study: How Language Models 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 14, 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

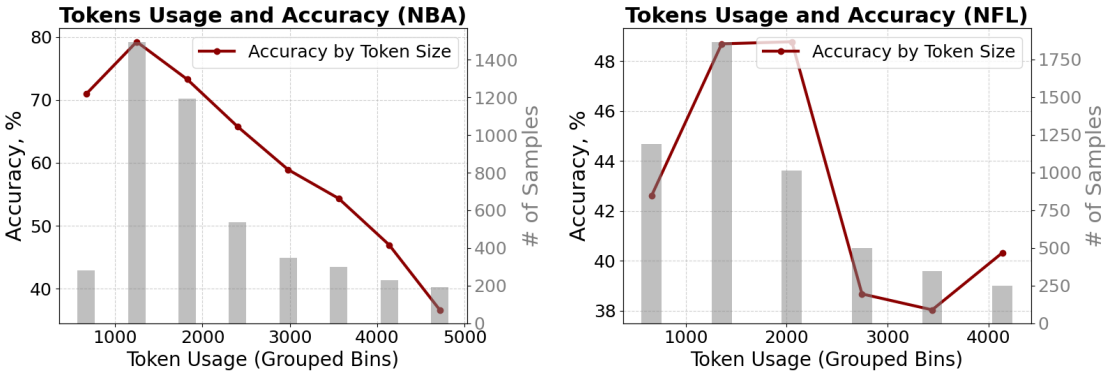

Figure 9: The relationship between token usage and reasoning accuracy. For both tasks, we sampled around 5,000 examples. We observe that DS-R1-distill-Qwen-32B achieves higher reasoning accuracy when using either fewer or more tokens, with peak accuracy occurring around 1,400 tokens.

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 9. 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 Short 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 9, the LLM tends to make hasty generalizations and prematurely draws conclusions. An example is shown in Figure 15, 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 Long Reasoning. We are not the first to observe that reasoning models, particularly those in the DeepSeek series (Shao et al., 2024; DeepSeek-AI, 2025), tend to engage in excessively long reasoning when making incorrect inferences (Liu et al.; 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 9, 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 Template

Figure 12 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 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 13 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 exam-

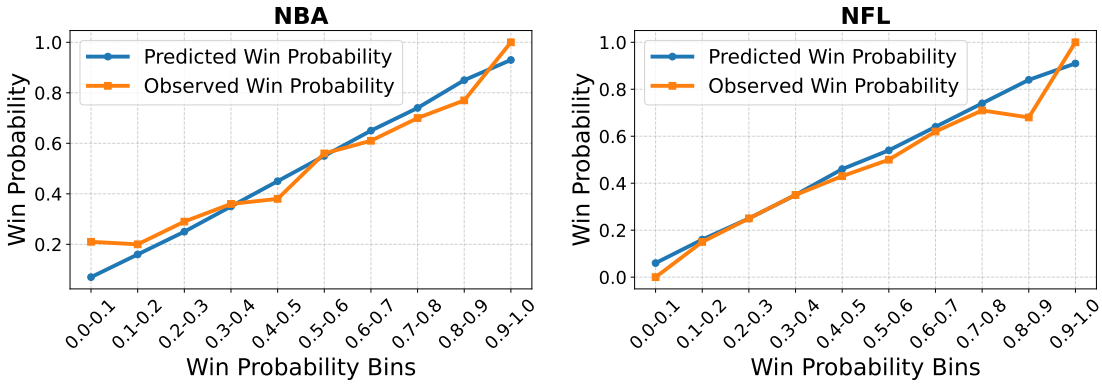

Figure 10: The Calibration of Win Probability Predictions. The results show a high degree of alignment between the model's predictions and the actual game outcomes.

ple 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 10 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.

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 11: 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 12: 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**".

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 13: 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.

Process of Event Reasoning via Time Series From DS-R1-distill-Qwen-32B

<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 14: 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 probabilitie ...
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 15: 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.