Can Language Models Infer Event Descriptions from Time Series?

Anonymous ACL submission

Abstract

Time series data measure how environments change over time and drive decision-making in critical domains like finance and healthcare. When analyzing time series, we often seek to understand the underlying events occurring in the measured environment. For example, one might ask: "What corporate announcement may have caused a sharply dropping stock price?" Events are aptly described with language, so we conduct the first study of whether Large Language Models (LLMs) can infer natural language events from time series. We curate a new benchmark featuring win probabilities collected from 4,200 basketball and American football games, featuring 1.7M timesteps with corresponding events. Building on the recent wave of using LLMs for time series, we extensively evaluate 16 LLMs, finding they demonstrate promising abilities to infer events from time series. An open-weights model, DeepSeek-R1 32B, beats proprietary models like GPT-40. Despite this impressive initial performance, we also find clear avenues to improve recent models, as we identify failures when altering the provided context, event sequence lengths, and evaluation strategy.*

1 Introduction

002

007

011

012

015

017

019

037

038

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 time series associated with natural language-described events in the measured environment, which influence trends and patterns in the time series. For instance, in Figure 1, a power outage causes a rise in temperature, which returns



Figure 1: Time series accompanied by event sequences in a cold storage environment. If an anomaly occurs (e.g., a power outage), it is reflected in the time series (in this case a rise in temperature). As the issue is resolved, the temperature gradually returns to normal.

to normal once the issue is resolved. Benefiting from the promising potential of integrating natural language with time series analysis (Jin et al., 2024), along with the rapid advancement of 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). When analyzing time series data, the goal is often to infer events occurring in the measured environment (Liu et al., 2024b). Therefore, there is a need to explore how LLMs infer event descriptions given context and time series data.

041

042

043

044

047

049

050

051

053

055

060

061

063

Prior work on reasoning about time series in conjunction with natural language has largely overlooked event sequences (Merrill et al., 2024; Williams et al., 2024). Previous evaluations of LLMs on time series instead have 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;

^{*}All resources needed to reproduce our work are available: https://anonymous.4open.science/r/reason_ events-9861/

094

100 102

105 106

108 109 110

111

112

113

114

104

Liu et al., 2024a; Cheng and Chin, 2024), however they are curated for forecasting and do not explore reasoning from time series to events. Meanwhile, due to the limited size and the potential inclusion of event descriptions that do not impact the time series, these data are not ideal as a benchmark source.

To address this gap, we introduce a benchmark that integrates time series data with associated events described in natural language. The events exhibit clear correlations with time series fluctuations, making it feasible to evaluate LLMs' event reasoning capabilities. Our benchmark includes 4,200 games, comprising a total of 1.7 million data points and events, with the capability for continuous updates (details in Section 3.3). During testing, we replace real entity names in games after the training cut-off date to mitigate potential memory effects in reasoning. We use win probability as the time series. As we discuss in Section 6, win probability is an effective measure of game state but potentially differs from ground truth.

To evaluate the effectiveness of our benchmark in assessing LLMs' reasoning under different conditions, we test 16 models across various factors, including the impact of available context, varying sequence lengths, and time series similarity on reasoning. Additionally, we examine the impact of real entity names and time series on reasoning through two ablation studies. To explore the generalizability of our approach, we extend the evaluation to open-domain settings, including cryptocurrency (Li et al., 2024a) and U.S. health data (Liu et al., 2024a). Our findings indicate that LLMs exhibit promising reasoning capabilities. For instance, the distilled DeepSeek-R1 32B achieves a 68% accuracy in basketball event reasoning (76.5%, w/CoT prompt), while GPT-40 reaches 41% accuracy (53%, w/ CoT prompt), both outperforming the baseline of random guessing (25%). Our key contributions are:

- We introduce 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 release 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 16 LLMs, we find promising reasoning capabilities and find clear avenues to enhance reasoning (Section 4.2).

2 **Related Work**

Existing benchmarks for LLMs in time series and event reasoning have yet to address the task of inferring event sequences from time series.

115

116

117

118

119

120

121

122

123

124

125

126

127

128

129

130

131

132

133

134

135

136

137

139

140

141

142

143

144

145

146

147

148

149

151

152

153

154

155

156

157

158

159

160

161

162

163

164

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 advancements have made significant contributions to fields such as sociology (Cheng 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) utilized news about energy in NSW, Australia 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 have further evaluated the ability to reason about events (causes). Liu et al.

Donohmonk/ Evolution	Properties (w/ Time Series)						
Dencimark/ Evaluation	Context	Source	Task				
Williams et al. (2024) Merrill et al. (2024) Cai et al. (2024) Liu et al. (2024a)	Description Description Question News Series	Manual Synthetic Manual Real-World	Forecasting Reason & Forecast Understanding Forecasting				
	Properties (w/o Time Series)						
Fatemi et al. (2024) Xiong et al. (2024) Chu et al. (2023) Quan and Liu (2024) Karger et al. (2024)	Event & Time Event & Time Event & Time Event Sequence Event	Synthetic Synthetic Prev. Dataset Synthetic Real-World	Temporal Reasoning Temporal Reasoning Temporal Reasoning Sequential Reasoning 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.

(2024a) collect 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

165

166

167

168

169

170

171

172

174

175

176

177

Reasoning is an ill-defined and broad, yet criti-178 cal, capability that determines LLMs' performance 179 across many complex tasks. Therefore, numerous 181 reasoning benchmarks have been developed for valuable tasks, such as coding (Zhuo et al., 2024; 182 Jain et al., 2024), mathematics (Cobbe et al., 2021; 183 White et al., 2024), and finance (Xie et al., 2023; Islam et al., 2023). Additionally, some benchmarks 185 have evaluated the general reasoning abilities of LLMs (Bang et al., 2023; White et al., 2024), in-187 cluding BBH (Suzgun et al., 2022) and MMLU 188 (Hendrycks et al., 2020). Several benchmarks have been proposed to evaluate LLMs' understanding of 190 relationships between events (Quan and Liu, 2024), 191 as well as temporal reasoning capabilities (i.e., understanding the relationships between events and 194 time) (Xiong et al., 2024; Chu et al., 2023). For instance, Karger et al. (2024) introduced a dynami-195 cally updated benchmark to evaluate LLMs' fore-196 casting of future events. Fatemi et al. (2024) used synthetic data to assess LLMs' perception and rea-198

soning 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. 199

200

201

202

203

204

206

207

209

210

211

212

213

214

215

216

217

218

219

221

222

223

224

225

3 Benchmark

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

3.1 Problem Definition

Α time series is a length-Tsequence timestamped of 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: $e = [(t_0, e_0), (t_1, e_1), ..., (e_T, e_T)].$ For each, the timestamps t are monotonically increasing $(t_i \leq t_j \text{ if } i < j)$. While the timestamps of the time series and event sequence need not be identical, they often are. 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 realvalued 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: one consisting of real-valued measurements

$$\mathcal{X} = \{(t_i, x_i)\}_{i=0}^N$$

and the other comprising natural language event descriptions

$$\mathcal{E} = \{(s_j, e_j)\}_{j=0}^T$$
 227

where the event e and the value x have a causal relationship, with s_i and t_i synchronized. Note that 229



Figure 2: Given time series $x_{1:N}$ and the event sequence $e_{1:T}$ that possibly generated them, where $t_{1:N}$ and $s_{1:T}$ are the timestamps of $x_{1:N}$ and $e_{1:T}$, respectively. We provide the time series data from j - k to j + k, along with the timestamps, to evaluate LLMs on inferring missing events in between. A variation of this framework supports inferring measurements, such as forecasting.

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 relationships between values in \mathcal{X} and events in \mathcal{E} are known.

Our goal is to interrogate LLMs' understanding of time series data by measuring their ability to infer likely unobserved values in \mathcal{E} given \mathcal{X} . As shown in Figure 2, when the intermediate event e_j (corresponding to s_j) is missing, the LLM is expected to infer it using the provided time series from x_{j-k} to x_{j+k} and corresponding timestamp from t_{j-k} to t_{j+k} . When k > 1, the LLM is required to infer multiple intermediate events.

3.2 Events Reasoning Format

We formulate our event reasoning evaluation as a multiple-choice question, as detailed below,

The <i>Role</i> and <i>Task</i> Prompts: {{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}					
Here are four potential options:{{options}}						
Here is the instruction for returning reasoning						
results in:{{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, we also provide the corresponding timestamps $t_{i:i+k-1}$ during reasoning. The intermediate events are missing, and the LLM is tasked with inferring these events. To make the task of matching time series to ground truth more tractable, we provide *four* options, one of which corresponds to the actual sequence of events. The final results are required to be presented in a fixed format. Figure 8 in Appendix B gives examples of the full prompts used in our experiments.

257

258

259

260

261

262

263

264

265

266

267

269

270

271

272

273

274

275

276

277

278

279

282

284

290

291

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 with "Team A" or "Team B." Player names are replaced with generic labels, such as "Player from Team A" or "Player from Team B," ensuring that the associations between players and their teams are preserved. In evaluations from other domains, such as cryptocurrency (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 Time Series and Events Sequence Dataset

To obtain paired data of time series and event sequences, we utilize data from sports, a domain where the effects of events on the time series are well-defined,. ESPN's[†] "play-by-play" captures key occurrences during a game, such as scoring, turnovers, or fouls in basketball. As the game progresses, these events form a sequence. Each event impacts the course of the game, and ESPN also provides the teams' win probabilities throughout the game as a representation of their status. Since a game constitutes a relatively closed environment, there is a clear relationship between the events and

4

246

247

248

256

[†]The source of the sports data can be accessed via https: //www.espn.com/

the time series: an event favoring Team A will raise Team A's win probability, and vice versa for Team B. This closed environment makes sports data a good candidate for a benchmark evaluating how effectively LLMs reason about events through time series.

Our dataset includes 4,200 time series (games) collected up to January 9, 2025, with the capability for continuous updates. The dataset contains 3,276 time series from basketball and 924 from football. On average, each basketball time series contains 460 timesteps, while football time series average 179 timestep. In total, the dataset consists of 1.7 million time series data points (win probabilities) paired with corresponding in-game events.

4 Experiments

293

294

301

305

307

310

311

313

314

315

317

319

321

322

323

324

327

330

331

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

We evaluate 16 language models, including closed-weight models such as GPT-40 (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 reasoning-oriented models, such as DeepSeek-R1 32B and 8B (DeepSeek-AI et al., 2025). Our findings indicate that LLMs, such as GPT-40, LLama3.1 70B, and particularly the reasoningoriented DeepSeek-R1 32B, demonstrate promising reasoning capabilities.

4.1 LLMs Performance and Impact of CoT

In our event reasoning baseline, we follow the format in Figure 8 from Appendix B, but provide only Team A's win probabilities and set the event sequence length to 10 (k = 12), with negative event sequences 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 more generic expressions, such as "*Player from Team A.*" Results, shown in Figure 3, indicate that reasoning about football events is more challenging, as GPT-40 achieves only 29% accuracy, compared to 41% in basketball. Among the tested LLMs, the distilled DeepSeek-R1 32B performs the best, achieving 43% accuracy in football event reasoning and 68% in basketball. In addition, other open-weights models, such as LLama3.1 70B and Qwen2.5 72B, or even Phi-4 14B, reaches performance levels comparable to GPT-40.

341

342

343

344

345

346

347

348

350

351

352

353

354

355

357

358

359

360

361

363

364

365

366

367

369

370

371

372

373

374

375

376

377

378

379

380

381

382

383

385

386

387

388

390

391

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. In short, the answer is "yes"-LLMs show an average reasoning performance boost from CoT prompting of 4.5% for basketball events and 9.6%for football events. The longer reasoning process with CoT, however, also increases the likelihood of LLMs failing to return answers in required format. For example, GPT-40 achieves an 11% improvement in accuracy but experiences a slight increase of 2.5% in failure to follow instructions. In contrast, some models instruction following benefits from CoT prompting. For Phi-3.5 14B-Instruct, the failure rate decreases by 39%, which may be attributed to the instruction-tuning process. In addition, adding CoT prompting improves DeepSeek-R1 32B accuracy by 9% while maintaining a 0% failure rate, demonstrating stability in instruction following. The overall instruction following failure rate increased by 0.6% after adding CoT prompting. Details related to the CoT prompt are in Figure 9. We acknowledge the potential for additional CoT prompt optimization, but we do not further explore this due to computational constraints.

4.2 Impact of Available Context in Reasoning

In different applications, the available context that LLMs can access varies. In addition to the baseline setting, where only the time series is provided, we also evaluate LLMs' performance when different reasoning-relevant contexts are made available or modified. For example, in basketball-related tasks, if the score is provided at each step, a proficient reasoner should use the score information to eliminate inconsistent event options, thereby improving inference accuracy. To evaluate the role of time series in the reasoning process, we conduct a simple ablation study. The impact of different conditions is summarized in Table 2. We observe that when time series is removed (i.e., *w/o TS*), all reason-



Figure 3: The performance of various language models on events reasoning indicates that open-weights models, such as Qwen2.5 72B (Yang et al., 2024), achieve comparable or even superior results to proprietary models like GPT-40 (API version 2024-05-13) (Achiam et al., 2023), particularly reasoning-focused models such as DeepSeek-R1 32B (DeepSeek-AI et al., 2025), which significantly outperforms others. Additionally, Chain-of-Though (CoT) prompting further enhances reasoning performance.

ing performances drop to near-random guessing levels (i.e., accuracy is around 25%), *indicating that LLMs are leveraging time series for event reasoning*. In contrast, providing LLMs with team names and real entities in events (e.g., team and player names) (*w/ Name*) notably improves reasoning, highlighting the necessity of removing them when evaluating reasoning (Fatemi et al., 2024).

396

400

401

402

403

404

405

406

407

408

409

410

411

412

413

414

415 416

417

418

419

420

Due to differences between the football and basketball data, various conditions influence reasoning differently. For example, timestamps provide the most significant improvement in reasoning about football events. Similarly, when providing the score (TS+Score) or partial events, e_i and e_{i+k-1} (TS+Event), reasoning performance also improves. Since reasoning-oriented models have demonstrated relatively strong reasoning capabilities, especially given computational constraints, we focus on evaluating other LLMs that outperform random guessing in the remainder of this evaluation as well as the following two.

To test whether LLMs can detect logically inconsistent information to aid reasoning, we shuffle the order of ground-truth events to create negative options (i.e., "*Reorder*"). Due to the nature of possession changes in football and basketball, event sequences follow sequential constraints. When logical inconsistencies are introduced in the events, LLMs effectively leverage it to enhance reasoning.

4.3 Impact of Quantity of Events

Increasing the number of events in the reasoning process introduces two potential effects. On one hand, a competent reasoner should leverage additional information to identify logical inconsistencies, such as contradictions between events and corresponding win probability changes. 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). Ideally, a capable LLM should focus on beneficial information and remain unaffected by the impact of reasoning length. 421

422

423

424

425

426

427

428

429

430

431

432

433

434

435

436

437

438

439

440

441

442

443

444

445

446

447

448

To study the effect of event quantity, we follow the setup in Section 3.2 and design questions with an increasing number of events (i.e., k - 2). The results, summarized in Figure 4, reveal an interesting pattern: the impact of event quantity varies across different datasets. When reasoning about basketball events, LLMs generally perform slightly worse as the number of events increases. For football events, however, LLM performance improves. 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, possessions are typically connected to only

Tasks	Language	Baseline	Ava	ailable Con	text	Abla	Options	
	Models	(TS Only)	TS+Time	me TS+Score TS+Event		w/ Name	w/o TS	Reorder
Basketball Reasoning	GPT-4o(0513)	41.0%	39.0%	47.5%	39.0%	55.0%	28.5%	69.5%
	GPT-4o(mini)	25.0%	24.5%	25.0%	26.0%	43.5%	21.0%	39.0%
	Qwen2.5(72B)	36.5%	39.0%	43.5%	39.5%	41.0%	24.5%	66.0%
	LLama3.1(70B)	40.5%	37.0%	50.5%	38.5%	51.0%	26.5%	47.5%
	Qwen2.5(32B)	44.5%	43.5%	57.5%	43.5%	50.0%	22.5%	59.0%
	Phi4(14B)	43.0%	35.0%	40.0%	36.0%	42.5%	25.0%	47.0%
	Qwen2.5(14B)	27.5%	34.5%	33.0%	32.0%	48.0%	22.0%	44.5%
	Avg. Impact of t	↓ -0.6%	\uparrow 14.6%	↑ 0.0%	↑ 33.2 %	↓-32.2%	↑ 46.8%	
-	GPT-4o(0513)	29.0%	75.5%	43.5%	53.0%	71.0%	18.5%	60.0%
ning	GPT-4o(mini)	25.0%	52.0%	26.5%	35.5%	33.5%	24.5%	42.0%
easo	Qwen2.5(72B)	30.5%	69.0%	42.0%	40.5%	52.0%	25.0%	54.0%
Football Re	LLama3.1(70B)	26.5%	71.0%	47.5%	35.5%	65.5%	20.5%	46.0%
	Qwen2.5(32B)	33.0%	74.5%	43.5%	46.0%	40.5%	27.5%	43.5%
	Phi4(14B)	29.5%	46.5%	36.0%	38.5%	43.5%	25.0%	28.5%
	Qwen2.5(14B)	28.5%	55.5%	28.5%	34.5%	63.5%	25.5%	33.0%
	Avg. Impact of t	he Condition	↑ 120.1%	↑ 32.4 %	↑ 40.4%	↑ 84.2%	↓ -17.3%	↑ 52.8%

Table 2: LLMs' event reasoning accuracy (%) under various conditions based on the baseline (i.e., providing only time series). 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.



Figure 4: The reasoning performance of LLMs across event sequences of various lengths. The figure includes only models that consistently outperform the baseline.

one event, and events are more independent. One insight is that LLMs exhibit different efficacy in utilizing information in reasoning across different domains.

4.4 Impact of Time Series Similarity

449

450

451

452

453

454

455

456

457

458

459

460

461

462

Time series serve as an alternate modality for event sequences. Consequently, whentwo time series are similar, distinguishing the two corresponding event sequences based on the time series alone becomes challenging. 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:

$$\mathbf{D} = ||norm(\mathbf{P}_{win}) - norm(\mathbf{P}_{win}')||_{l2}$$

463

465

466

467

468

469

470

471

472

473

474

475

476

477

478

479

480

481

482

483

484

485

486

We bootstrap the pairwise distances between win probabilities (i.e., \mathbf{P}_{win}) of length 10 in our dataset and normalized 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 (refer to Figure 7 in Appendix A.4 for details). Therefore, we divided the distances into seven levels, 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. To ensure a fair comparison, we keep the ground-truth events and question time series consistent across all levels. The results are presented in Figure 5, showing a slight upward trend in LLM performance as similarity increases. This is due to the inherent consistency between time series and event sequences, which LLMs are able to recognize.

4.5 Evaluating Applicability Across Domains

In real-world open environments, time series data usually coexist with related textual sequences, such

$\begin{array}{c} LLMs \rightarrow \\ Domains \downarrow \end{array}$		GPT-40 (0513)	GPT-40 (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_US (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 3: 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.



Figure 5: 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.

488

489 490

491

492

493

494

495

496

497

498

499

503

504

507

as the proportion of infectious diseases alongside U.S. government influenza reports (Liu et al., 2024a) or cryptocurrency (e.g., Bitcoin) prices alongside relevant news from financial media (Li et al., 2024a). To evaluate the generalizability of our approach, we extend our evaluation to four other domains: Trading (Import/Export, IMEX), U.S. Health (Influenza Rates), and Energy (Gasoline Prices) from Time-MMD (Liu et al., 2024a), as well as cryptocurrency time series from Crypto-Trade (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. Similarly, our question follows the format in subsection 3.2, with an event sequence length of 10, corresponding to 10 trading days for Bitcoin data or 10 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*). Since numerical information provides crucial context for reasoning, removing it degrades performance. Even after stripping numerical data, however, LLMs still demonstrate moderate reasoning ability. GPT-40, for instance, consistently achieves over 50% accuracy. Additionally, open-weights models such as Qwen2.5 72B and reasoning-oriented models like DeepSeek-R1 32B demonstrate comparable performance to GPT-40. Full results are shown in Table 3. 508

509

510

511

512

513

514

515

516

517

518

519

520

521

522

523

524

525

526

527

528

529

530

531

532

533

534

535

536

537

5 Conclusions

LLMs have been developed for time series analysis, and data comprising time series paired with event sequences are prevalent in real-world scenarios. We introduce a dataset containing 1.7 million real-valued time series paired with events and a corresponding evaluation approach. We then evaluate 16 language models using this dataset and extend our method to multiple domains, including cryptocurrency and U.S. health. Our findings reveal that both open-weights and proprietary models exhibit promising reasoning capabilities, with reasoning-focused models such as DeepSeek-R1 32B outperforming larger proprietary model such as GPT-40.

Limitations and Ethical Considerations 6

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 win probability 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.

> We release all code and data necessary to replicate our complete experiments at https://anonymous.4open.science/r/ reason_events-9861/. As we await approval from the data provider, however, we may not be able to release the final curated dataset. In that case, we will provide the tools necessary to replicate our data collection process.

References

538

540

541

542

544

547

549

550

553

554

558

559

560

561

563

568

569

570

571

572

573

574

580

582

585

586

588

- 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 Reves-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.
- 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, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, and ... 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.
- 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.

635

636

637

638

639

640

641

642

643

644

645

589

590

748

749

750

751

752

700

701

702

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.

647

663

674

675

677

679

687

690

691

- 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*.
- 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 Ilm-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 Thirtyeight 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*.
- Mike A Merrill, Mingtian Tan, Vinayak Gupta, Thomas Hartvigsen, and Tim Althoff. 2024. Language models still struggle to zero-shot reason about time series. *arXiv preprint arXiv:2404.11757*.
- 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*.
- 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 Ilm benchmark. *arXiv preprint arXiv:2406.19314*.

- 753 754 756
- 765 770 772 773 774 775 776
- 778
- 787 788 789
- 794
- 795

768

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

arXiv:2410.18959.

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.

Andrew Robert Williams, Arjun Ashok, Étienne Mar-

cotte, Valentina Zantedeschi, Jithendaraa Subrama-

nian, Roland Riachi, James Requeima, Alexan-

dre Lacoste, Irina Rish, Nicolas Chapados, et al. 2024. Context is key: A benchmark for forecasting with essential textual information. arXiv preprint

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.

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 syn-

thetic data for enhanced understanding and reasoning.

arXiv preprint

arXiv preprint arXiv:2306.05443.

arXiv preprint arXiv:2412.03104.

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.

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.

890

842

843

844

845

A Appendix

796

797

801

802

803

807

810

811

812

813

814

815

816

817

819

820

821

826

827

830

835

837

841

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 6 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 openweight models such as LLaMA3.1 (Dubey et al., 2024), proprietary models like GPT40 (Achiam et al., 2023), and reasoning-focused models such as DeepSeek-R1 (DeepSeek-AI et al., 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 et al., 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.

^{*}All information and settings needed are available:https: //anonymous.4open.science/r/reason_events-9861/



Figure 6: 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.



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.

A.4 Time Series Similarity

891

892

894

900

901

902

904

We bootstrapped 10k pairwise distances between win probabilities of length 10 for both NBA and NFL data. After normalizing the distances to the range (0, 1), their distribution is shown in Figure 4, where 91% of the distances fall beyond 0.4.

B Prompt Template

Figure 8 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 9 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.

905

906

907

908

909

910

911

912

913

914

915

916

917

918

919

920

921

922

923

924

925

926

927

928

929

930

931

932

933

934

935

936

C Language Models Reasoning Process

To further understand how LLMs infer events from time series, we analyzed their output. As shown in Figure 10, this illustrates the reasoning process of DeepSeek-R1 32B (DeepSeek-AI et al., 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-40 (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.

D Licensing

940

The code from our work is released under the 941 MIT License, while the dataset is made avail-942 able under the Creative Commons Attribution-943 NonCommercial-ShareAlike (CC BY-NC-SA) li-944 cense. This allows anyone to use, distribute, and 945 modify the data for non-commercial purposes, pro-946 vided they give proper attribution and share any 947 derivative works under the same license terms. 948

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. TimeStamp₁ WP_1 Step 2. WP_2 TimeStamp₂ Step 3. TimeStamp₃ WP_3 ... WP_k

Step k. TimeStamp_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. Date₁ x_1 Step 2. Date₂ x_2 Step 3. Date₃ x_3 ••• 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 8: 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 9: 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 DeepSeek-R1 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.

••••

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 10: 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 et al., 2025). Other LLMs, such as GPT-40 (Achiam et al., 2023), follow a similar reasoning pattern. To save space, some content has been omitted.