Can Slow-Thinking LLMs Reason Over Time? Empirical Studies in Time Series Forecasting

Mingyue Cheng
State Key Laboratory of Cognitive Intelligence, University of Science and Technology of China
Hefei, Anhui Province, China
mycheng@ustc.edu.cn

Jiahao Wang
State Key Laboratory of Cognitive Intelligence, University of Science and Technology of China
Hefei, Anhui Province, China
jiahao.wang@mail.ustc.edu.cn

Daoyu Wang
State Key Laboratory of Cognitive Intelligence, University of Science and Technology of China
Hefei, Anhui Province, China
wdy030428@mail.ustc.edu.cn

Xiaoyu Tao
State Key Laboratory of Cognitive Intelligence, University of Science and Technology of China
Hefei, Anhui Province, China
txytiny@mail.ustc.edu.cn

Qi Liu*
State Key Laboratory of Cognitive Intelligence, University of Science and Technology of China
Hefei, Anhui Province, China
qiliuql@ustc.edu.cn

Enhong Chen
State Key Laboratory of Cognitive Intelligence, University of Science and Technology of China
Hefei, Anhui Province, China
cheneh@ustc.edu.cn

Abstract

Time series forecasting (TSF) traditionally relies on fast-thinking paradigms that map historical observations directly to future sequences of continuous values. While effective, such approaches often frame forecasting as a pattern-matching problem and tend to overlook explicit reasoning over temporal dynamics and contextual factors, which are critical for modeling long-range dependencies and non-stationary behaviors in real-world scenarios. Recent slow-thinking large language models (LLMs), such as OpenAI o1 and DeepSeek-R1, demonstrate strong inference-time multi-step reasoning abilities. This raises a fundamental question: can slow-thinking LLMs reason over temporal dynamics to support accurate TSF, even without task-specific training? To investigate this question, we present TimeReasoner, a systematic empirical study that reformulates TSF as a conditional reasoning process performed entirely at inference time. TimeReasoner integrates hybrid instructions consisting of task directives, timestamps, sequential values, and optional contextual features, and induces multi-step temporal reasoning in pretrained slow-thinking LLMs through chain-of-thought prompting and rollout-based reasoning strategies. Extensive experiments across diverse TSF benchmarks show that slow-thinking LLMs consistently outperform prior baselines or achieve competitive training-free forecasting performance. Beyond accuracy, we analyze how different inference-time reasoning strategies influence forecasting behaviors, highlighting both the potential and limitations of slow-thinking paradigms for TSF.[1]

CCS Concepts

• **Mathematics of computing → Time series analysis**.

Keywords

Forecasting, Long Chain-of-thought, Time Series Reasoning

ACM Reference Format:
Mingyue Cheng, Jiahao Wang, Daoyu Wang, Xiaoyu Tao, Qi Liu, and Enhong Chen. 2026. Can Slow-Thinking LLMs Reason Over Time? Empirical Studies in Time Series Forecasting. In *Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining (WSDM '26), February 22–26, 2026, Boise, ID, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3773966.3777931

1 Introduction

Time series forecasting (TSF) is a fundamental data mining task with wide-ranging applications in domains such as finance [56], energy [14], and healthcare [29] where accurate forecasts are critical for strategic planning and effective decision-making under uncertainty [3, 47]. The goal of TSF is to predict future values of target variables by leveraging historical observations together with contextual features, including temporal covariates and static attributes [22, 59, 72].

Over the past decades, numerous approaches have been proposed for time series forecasting (TSF) [10, 41], which can be broadly categorized into traditional statistical modeling approaches [35] and data-driven approaches. Traditional statistical models, such as ARIMA [71] and Holt–Winters [6], rely on strong assumptions of linearity, stationarity, and seasonality. While offering clear explainability through explicit model structures, they often struggle to capture complex, nonlinear, or non-stationary temporal patterns. In contrast, data-driven approaches, particularly deep learning–based models, capture intricate temporal patterns such as trends and seasonality through nonlinear architectures, achieving strong performance across standard benchmarks [2, 77] and demonstrating further improvements in various TSF scenarios [23, 37]. With recent advances in large language models (LLMs), TSF has also been explored from multimodal understanding and generative modeling perspectives, leveraging the ability of LLMs to incorporate heterogeneous contextual features [25]. Subsequent studies extend this line of work through large-scale LLM integration and generation-oriented architectures [75]. Despite these advances, most existing

*Qi Liu is the corresponding author.
[1] https://github.com/realwangjiahao/TimeReasoner

WSDM '26, Boise, ID, USA.

ACM ISBN 979-8-4007-2292-9/2026/02
https://doi.org/10.1145/3773966.3777931

LLM-based TSF approaches still treat forecasting as a direct sequence mapping problem, producing predictions in a single inference pass without explicit intermediate reasoning. Consequently, the potential of LLMs for step-wise temporal reasoning—an ability that may be critical for long-horizon and context-aware forecasting—remains largely under-explored [16, 52].

Building on this observation, we further abstract existing TSF methods under a unified perspective and note that most of them largely follow a fast-thinking paradigm [24]. Specifically, forecasting is typically performed through primarily single-step inference, mapping historical inputs directly to future targets with limited explicit intermediate reasoning [13]. Under this paradigm, models often capture short-term temporal dependencies effectively, while systematically integrating heterogeneous contextual information in a step-by-step manner remains challenging [28]. In contrast, slow-thinking LLMs, such as OpenAI o1 [46] and DeepSeek-R1 [19], demonstrate strong multi-step reasoning abilities in domains such as mathematics, typically by generating explicit intermediate reasoning steps before producing final answers [44]. Such reasoning processes may offer advantages for modeling long-range dependencies and contextual regime shifts in TSF. However, it remains underexplored whether these reasoning capabilities can be effectively leveraged for time series forecasting, particularly in training-free settings without task-specific fine-tuning.

These questions, centered on the potential of slow-thinking LLMs for TSF, motivate a re-examination of how forecasting problems are formulated and solved. In this work, we investigate whether slow-thinking LLMs can function as forecasters through explicit, inference-time step-by-step reasoning. Specifically, we reformulate TSF as a training-free conditional reasoning task, in which predictions are guided by historical observations together with various contextual features. Accordingly, we propose TimeReasoner, a inference-time reasoning framework that leverages LLMs' semantic reasoning capabilities to produce both interpretable forecasts and structured reasoning traces. We further analyze how different reasoning strategies, prompt construction, and modeling choices influence forecasting performance, and investigate the following research questions:

RQ1: How does TimeReasoner perform compared to state-of-the-art baselines? We evaluate the forecasting performance of TimeReasoner across multiple benchmarks to establish its overall effectiveness relative to existing methods.

RQ2: Which key settings in hybrid instruction and reasoning strategies are essential for accurate forecasting? Through ablation on TimeReasoner's key configurations—including hybrid prompt components (timestamps, contextual features, raw series) and alternative reasoning strategies—we identify the factors that most influence predictive accuracy.

RQ3: How trustworthy is TimeReasoner under real-world input imperfections? We simulate incomplete and noisy inputs by introducing missing values and perturbations, evaluating the model's ability to maintain accuracy and reasoning quality under such conditions. We further quantify its capability to estimate and communicate forecasting uncertainty.

RQ4: How does TimeReasoner perform under different TSF task settings? We evaluate the model's forecasting accuracy and reasoning trace across varying look-back windows and prediction lengths to understand its adaptability to different temporal settings.

RQ5: How do inference-time hyperparameters affect forecasting quality? We analyze the sensitivity of TimeReasoner to key hyperparameters such as temperature and base LLM selection, examining their impact on prediction accuracy.

RQ6: Do reasoning traces from TimeReasoner provide explainable insights into its predictions? We examine intermediate reasoning steps and representative case studies to assess their consistency with observed input patterns, and potential to enhance the explaination of the model's forecasts.

2 Related Work

In this section, we briefly introduce the related work, including time series forecasting and LLM-based reasoning.

2.1 Time Series Forecasting

Time series forecasting has been extensively studied over the past decades [10]. Early statistical methods, such as ARIMA [3, 71], relied on auto-regressive and moving-average formulations with strong theoretical guarantees, but often assumed ideal data properties that rarely hold in practice [21]. With the growth of data and computing power, deep learning-based sequence-to-sequence models have emerged [50]. Recurrent neural networks were among the first to capture temporal dependencies [63], but suffered from limited receptive fields and error accumulation in recursive inference [32]. Subsequent advances introduced architectures for long-range dependency modeling [79], including self-attention and convolutional designs [34, 38]. Recent work has also integrated classical techniques—such as trend-seasonal decomposition and time-frequency transforms [42, 81]—into neural networks [62, 76, 80]. Notably, even simple linear models enhanced with decomposition can be competitive [68], and slice-based approaches achieve strong performance in long-term forecasting [9, 45, 78].

In recent years, large language models (LLMs) have attracted considerable attention for their ability to understand and generate human-like text [31, 54, 67]. Beyond their success in natural language processing [55, 75], LLMs have shown strong potential in time series analysis [27]. Their application to time series generally follows two paradigms: *fine-tuning* and *prompt-based zero-shot learning* [25]. In fine-tuning, a pre-trained LLM is further adapted to time series tasks by training on task-specific data [4, 5], enabling it to capture domain patterns and improve performance [26, 39], albeit with substantial labeled data and computational costs [10]. Prompt-based zero-shot methods instead guide predictions via well-designed prompts without task-specific training [18, 36], offering greater flexibility and efficiency [53] but sometimes underperforming fine-tuned models in specialized scenarios [59]. Both paradigms underscore the rising interest in LLMs for time series, while revealing challenges in optimizing their effectiveness for such tasks.

2.2 LLM-based Reasoning

LLM-based reasoning refers to leveraging large language models' ability to perform multi-step, structured thinking for complex tasks. Within this paradigm, *inference-time reasoning* [70] has emerged as an effective strategy to enhance a pre-trained model's performance

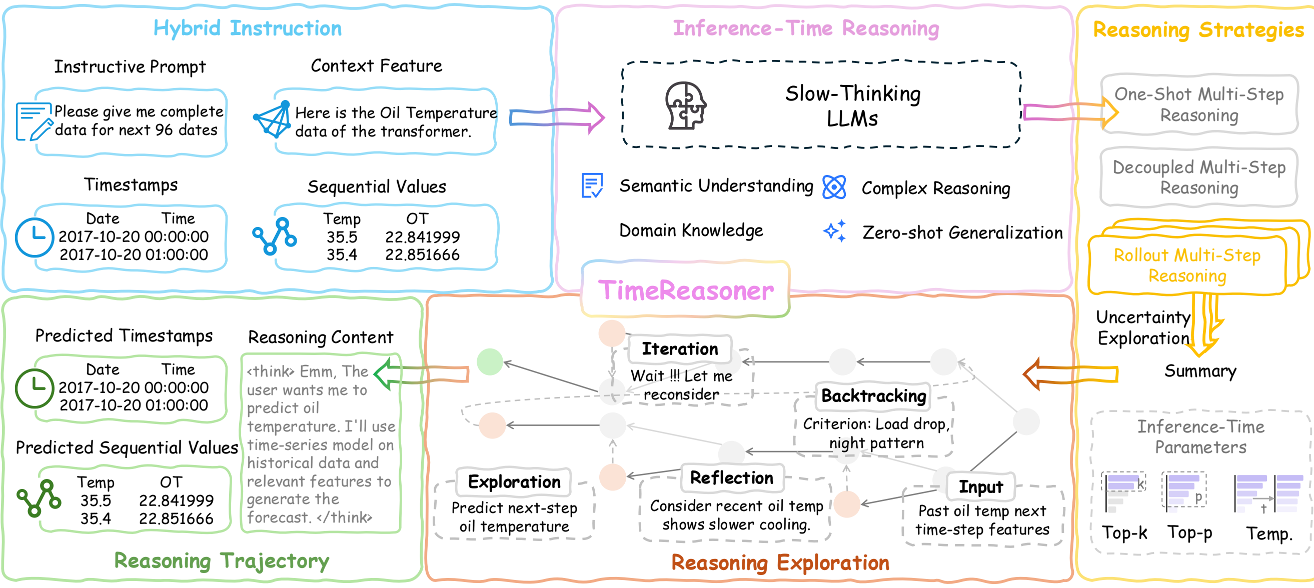


Figure 1: Overall framework of TimeReasoner for training-free reasoning-based time series forecasting.

on challenging reasoning problems by allocating additional computational resources during inference [51, 66]. These resources may include extended processing time, increased computational steps, or advanced strategies such as multi-path sampling and iterative refinement [33]. In contrast to training-time scaling—which improves model capability through more parameters, data, or computation during training [48]—inference-time reasoning exploits extra computation at inference to activate deeper reasoning potential [73] without modifying model weights [8]. Recent LLMs demonstrate the effectiveness of this approach: DeepSeek-R1 [19] applies iterative refinement to boost coding and reasoning performance [43]; Gemini-2.5-pro [57] is expected to excel on sophisticated tasks with extended processing or multi-step protocols; and OpenAI's GPT-o1 [20] achieves significant gains on challenging reasoning benchmarks using chain-of-thought prompting [61]. These examples underscore the growing role of inference-time reasoning in trading inference-time computation for higher accuracy [44].

Despite the transformative progress slow-thinking reasoning has brought to numerous complex reasoning tasks, whether such powerful reasoning models can be effectively leveraged to improve time series forecasting has yet to be explored.

3 Problem Definition

Given a multivariate time series $\mathcal{Y} = \{\mathbf{y}_t\}_{t=1}^{T}$, where $\mathbf{y}_t \in \mathbb{R}^d$ denotes the d-dimensional target variables at time step t, the goal of multivariate time series forecasting (TSF) is to predict future values $\{\mathbf{y}_{t+1}, \ldots, \mathbf{y}_{t+H}\}$ over a fixed predict window H. At each forecasting step t, the model is provided with a lookback window of length L, i.e., $\{\mathbf{y}_{t-L}, \ldots, \mathbf{y}_{t-1}\}$, summarizing recent historical observations. In addition, the model may access a context vector $\mathcal{Z} \in \mathbb{R}^k$ containing supplementary information such as temporal covariates, static attributes, or other relevant features to aid forecasting.

4 The Proposed TimeReasoner

In this section, we present the overall architecture of TimeReasoner, detailing how it leverages LLMs to perform inference-time reasoning for time series forecasting.

4.1 Overview of TimeReasoner

Figure 1 illustrates the overall architecture of TimeReasoner, a training-free, inference-time reasoning framework for time series forecasting using slow-thinking LLMs. The framework starts with Hybrid Instruction, where instructive prompts are integrated with contextual features (e.g., timestamps, sequential values) to construct structured inputs. These inputs are processed in the Inference-Time Reasoning stage, where slow-thinking LLMs leverage semantic understanding, domain knowledge, and multi-step reasoning capabilities to generate forecasts without task-specific fine-tuning. To further improve forecasting accuracy, reasoning strategies including one-shot multi-step, decoupled multi-step, and rollout multi-step reasoning are employed, along with uncertainty exploration and inference-time parameter control. Finally, Reasoning Exploration records the reasoning trajectories and supports iterative refinement, backtracking, reflection, and explanation, thereby producing accurate predictions accompanied by interpretable reasoning traces.

4.2 Hybrid Instruction

To support time series reasoning, we design multimodal prompts including hybrid instructions that combine:

- *Instructive prompts.* TimeReasoner utilizes an instructive prompt to explicitly specify the forecasting objective, such as the number of future time steps to be predicted. This directive helps the LLM focus on the target prediction range and aligns its generative reasoning with the user's intent.

- *Contextual features.* TimeReasoner is provided with natural language descriptions that offer contextual information like domain-level knowledge and channel-specific semantics. These help align numerical data with real-world meaning, enhancing explainability and task relevance [64].
- *Raw time series in original space.* TimeReasoner takes time series in their original scale as input, without normalization or standardization. Preserving the raw magnitudes allows the model to reason over absolute values and real-world fluctuations that might otherwise be distorted.
- *Timestamps.* TimeReasoner incorporates original timestamps into the prompt, enabling it to recognize both absolute temporal positions (such as specific hours or dates) and relative patterns (such as periodicity), which are crucial for temporal reasoning [58, 69].

4.3 Inference-Time Reasoning

Based on the constructed multimodal prompts, TimeReasoner performs inference-time reasoning by leveraging the semantic understanding and complex reasoning capabilities of slow-thinking LLMs. Specifically, the model first interprets the forecasting objective and contextual descriptions to establish a high-level semantic grounding of the task, including the prediction horizon, domain constraints, and plausible temporal behaviors. It then reasons over the raw time series values and timestamps jointly, identifying salient temporal patterns such as trends, periodicity, regime shifts, and potential anomalies directly in the original value space. During generation, the LLM incrementally integrates these observations, reflecting on recent dynamics and historical regularities to form a coherent forecasting trajectory. This inference-time reasoning process allows TimeReasoner to adaptively combine numerical evidence with domain semantics, enabling flexible and interpretable time series forecasting without task-specific parameter tuning.

4.4 Reasoning Strategies

We explore three distinct reasoning paradigms to evaluate the forecasting capability of slow-thinking LLMs.

- *One-Shot Reasoning.* TimeReasoner performs a single, comprehensive reasoning pass on the given input. During the generation, it processes all necessary information and internally executes the multi-step logic required. The model then outputs the complete set of predicted results, without any subsequent iteration.
- *Decoupled Reasoning.* Unlike one-shot reasoning, the reasoning process is decoupled: the model generates a partial thought, reflects on it, and then continues, forming a generate–reflect–generate loop within a single inference, enabling multi-step reasoning.
- *Rollout Reasoning.* In rollout reasoning, TimeReasoner generates predictions incrementally, producing one part of the result in each step. This process involves several iterative rounds, where each output serves as context for the next, enabling TimeReasoner to gradually build toward a comprehensive final result.

4.5 Reasoning Exploration and Trajectory

Beyond directly generating numerical forecasts, TimeReasoner leverages the slow-thinking capability of LLMs to perform iterative reasoning over time series dynamics at inference time. Given historical observations and timestamps, the model incrementally analyzes recent temporal behaviors, such as local trends, periodic patterns, and potential regime changes, and reflects on how these dynamics may evolve over the predict window. During this process, the LLM may revisit earlier hypotheses, backtrack from inconsistent assumptions, and refine its interpretation of the time series when new evidence (e.g., abrupt fluctuations or weakened seasonality) is identified. This iterative reasoning and reflection enable the model to form a coherent forecasting rationale rather than relying on a single-pass mapping from past to future values.

To obtain a stable and reliable prediction, we execute the slow-thinking forecasting procedure multiple times under the same input and aggregate the generated results; specifically, we perform three independent generations and take the mean of the resulting forecast sequences as the final output. This aggregation strategy mitigates variability introduced by stochastic inference and provides a more robust estimate of future time series values, while preserving the explainability afforded by explicit reasoning traces.

Table 1: Statistical information of experimental datasets.

Dataset	Domain	Length	Features	Frequency
ETTh1&ETTh2	Electricity	17,420	7	1 hour
ETTm1&ETTm2	Electricity	69,680	7	15 mins
AQWan&AQShunyi	Environment	35,064	11	1 hour
Exchange	Economy	7,588	8	1 day
NASDAQ	Economy	1,244	5	1 day
Wind	Energy	48,673	7	15 mins
VitalDB	Healthcare	2,855	2	3s

5 Experimental Settings

In this section, we describe the experimental settings for evaluating TimeReasoner, covering datasets, baselines, implementation details.

5.1 Datasets

To evaluate the generalization of our approach across diverse scenarios, we deliberately select datasets from multiple domains rather than a few popular benchmarks. The datasets include: ETT (electricity load, 2016–2018) [79]; Exchange (foreign exchange rates, 1990–2016) [32]; Wind (wind measurements, 2020–2021) [32]; AQ (air quality, four years) [74]; NASDAQ (stock market prices, volumes, and high–low values) [15]; and VitalDB (physiological signals sampled every 3 seconds) [11]. These datasets span energy, finance, climate, environmental monitoring, and healthcare, offering heterogeneous temporal patterns, sampling rates, and contextual features. Such diversity supports robust reasoning-based forecasting evaluation beyond narrow benchmark settings and reduces bias toward any single data distribution. Following standard practice, each dataset is chronologically split into training, validation, and testing sets. Table 1 provides detailed descriptions.

5.2 Baselines

To ensure a thorough and fair evaluation, we compare TimeReasoner against a diverse set of strong baselines covering three major categories: (i) Deep learning-based models, which represent state-of-the-art architectures for time series forecasting, including DLinear [68], PatchTST [45], iTransformer [40], and TimeXer [60]; (ii) LLM-based approaches, which leverage large language models for time series analysis, including GPT4TS [82], Time-LLM [26], and LLMTime [18]; (iii) Time series foundation models, which aim to provide general-purpose forecasting capability across domains, including MOIRAI [65], MOMENT [17], and Chronos [1].

Table 2: Performance comparison of TimeReasoner and baseline models with best values in bold and second-best underlined. MSE ↓ and MAE ↓ are used as the evaluation metrics. Overall, TimeReasoner achieves superior performance across datasets.

Methods	Metric	ETTh1	ETTh2	ETTm1	ETTm2	Exchange	AQWan	AQShunyi	Wind	NASDAQ	VitalDB
DLinear	MSE	7.7656	10.3584	14.0352	7.8604	0.0014	21090.0523	21077.8247	1617.3343	**0.0007**	68.2254
	MAE	1.4981	2.0567	1.7591	1.8981	0.0253	51.1701	50.5644	18.3691	0.0215	6.7148
PatchTST	MSE	9.2430	10.9489	16.3988	5.6375	0.0009	12528.5571	16747.3794	2013.9736	**0.0007**	50.7402
	MAE	1.6500	1.9930	1.9294	**1.3834**	0.0198	39.8232	42.0165	20.1837	**0.0211**	**5.6983**
iTransformer	MSE	7.5195	9.9716	**12.7432**	5.7392	0.0010	13577.9277	18147.8075	1600.0376	0.0008	79.2864
	MAE	1.4983	1.9010	**1.6374**	1.4422	0.0202	39.8069	42.7643	17.9244	0.0235	7.3322
TimeXer	MSE	8.4878	11.4123	14.0473	**5.5801**	0.0009	14509.0913	16505.3908	1668.6824	**0.0007**	65.8567
	MAE	1.5430	2.0606	1.7534	1.4238	0.0193	41.2693	40.9784	18.2386	0.023	6.1253
GPT4TS	MSE	6.9454	9.7156	15.9028	5.6327	0.0009	13543.6153	16821.9321	1786.3085	0.0010	65.3715
	MAE	1.4127	1.8891	1.9364	1.4743	0.0200	39.7527	41.8708	18.2427	0.0253	6.7207
Time-LLM	MSE	6.8152	9.9876	15.8078	5.6787	0.0010	13405.9845	16678.0982	1773.9651	0.0011	65.9761
	MAE	1.4115	1.8880	1.9350	1.4733	0.0199	39.7481	41.8598	18.2365	0.0252	6.7151
LLMTime	MSE	10.6261	11.2011	14.6394	6.9001	0.0026	29141.5143	30266.3043	3979.7996	0.0021	99.5761
	MAE	1.6987	1.5876	1.9543	1.5109	0.0345	65.6789	60.3456	30.1234	0.0234	8.9054
Chronos	MSE	10.6328	15.2152	22.7706	9.6567	0.0103	13858.3941	16601.0443	2874.5295	0.0043	**50.2799**
	MAE	1.5128	2.1104	2.2046	1.7979	0.0324	42.7446	45.1168	23.1097	0.0341	5.7185
Moirai	MSE	10.5981	11.9302	38.5828	10.4213	**0.0008**	14864.1762	19450.7758	2279.2915	0.0009	180.8878
	MAE	1.5690	1.9536	2.7939	1.7779	0.0182	42.0587	44.1069	20.0049	0.0245	10.6444
MOMENT	MSE	18.3100	12.1693	26.2155	8.1986	0.0012	14687.9019	15596.0312	1726.4521	0.0006	73.3861
	MAE	2.3847	2.1208	2.8222	1.7389	0.0223	43.5303	43.4243	20.0301	0.0201	6.8837
TimeReasoner	MSE	**5.4469**	**8.6020**	14.2055	6.4384	0.0009	**11305.6345**	**12874.9412**	**1556.5316**	0.0008	79.4890
	MAE	**1.1984**	**1.6811**	1.6984	1.4209	**0.0168**	**36.1614**	**37.0193**	**17.5034**	0.0215	6.9735

5.3 Implementation Details

The TimeReasoner is model-independent; in this work, we adopt DeepSeek-R1[2] for its open-source availability and ability to expose complete reasoning traces. By default, all experiments are conducted under the one-shot reasoning paradigm, with inference performed via the official API[3] using temperature=0.6, top-p=0.7, and max-tokens=8,192. Foundation model baselines follow the same zero-shot inference protocol, while deep learning baselines are trained using the Adam optimizer [30] in PyTorch [49] on a single NVIDIA GeForce RTX 4090D GPU, with official implementations and recommended parameters. The lookback window is set to $L = 36$ for NASDAQ, $L = 300$ for VitalDB, and $L = 96$ for other datasets; the predict window is $H = 36$, $H = 200$, and $H = 96$ respectively. We adopt a channel-independent evaluation. Following standard TSF practice, we report mean squared error (MSE) to measure squared deviations and mean absolute error (MAE) to measure absolute deviations between predictions and ground truth.

6 Experimental Results

In this section, we conduct a comprehensive set of experiments to empirically evaluate the performance of TimeReasoner and validate the potential of slow-thinking for time series forecasting.

[2]https://huggingface.co/deepseek-ai/DeepSeek-R1
[3]https://platform.deepseek.com

6.1 Forecasting Performance Comparison

To address RQ1, we evaluate TimeReasoner's forecasting performance across multiple benchmarks against state-of-the-art baselines. Table 2 summarizes the forecasting results of TimeReasoner, demonstrating its competitive performance over other baseline models. Results on ten datasets show that TimeReasoner achieves competitive forecasting results on most datasets, especially on ETTh1 and AQShunyi. We observe that TimeReasoner particularly excels on datasets characterized by complex temporal dynamics. On datasets like AQWan, AQShunyi and Wind, TimeReasoner consistently surpasses other methods by effectively capturing underlying dependencies. These findings underscore TimeReasoner's outstanding capability in tackling difficult forecasting tasks and its generalization to complex time series. We also observe that the model achieves performance comparable to strong baselines on other datasets. This suggests that the proposed approach does not introduce performance degradation outside scenarios involving complex temporal dynamics.

6.2 Analysis of Key Settings in TimeReasoner

To address RQ2, we conduct ablations on key components of TimeReasoner, including hybrid prompt constructions and reasoning strategies, to assess their impact on accuracy.

6.2.1 Impact of Hybrid Prompt. We study how prompt-design choices affect TimeReasoner's forecasting performance by ablating

Table 3: Ablation study on timestamps, context information and normalization strategies.

Key Settings	Models	ETTh1		Wind		NASDAQ	
		MSE	MAE	MSE	MAE	MSE	MAE
Components	TimeReasoner	**5.4469**	**1.1984**	**1556.5316**	**17.5034**	**0.0008**	**0.0215**
Timestamps	*w/o* timestamps	25.3058	2.4705	3278.3356	29.7851	0.0021	0.0476
	w/ forward shifting	7.9222	1.3044	1570.5614	17.8587	0.0009	0.0218
	w/ backward shifting	6.4596	1.2640	2195.4411	19.3174	0.0010	0.0218
Context Information	*w/o* context information	7.8781	1.4325	2239.5041	19.4609	0.0009	0.0238
Normalization	*w/* Z-score Normalization	8.7691	1.4601	2076.8085	18.8674	0.0011	0.0236
	w/ RevIN	122.0912	11.7891	6641.7854	67.1268	0.0025	0.0671

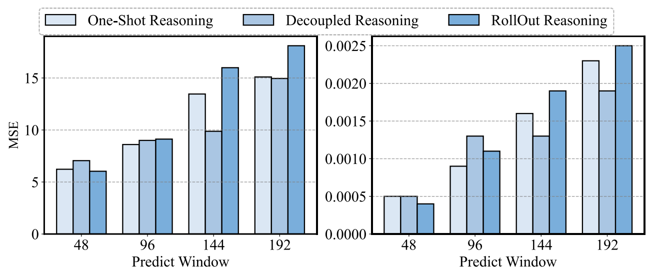


Figure 2: Comparison of TimeReasoner's forecasting capabilities under the three reasoning strategies with experiments conducted on ETTh2 (left) and Exchange (right) datasets.

three components: timestamps, contextual information, and normalization, as shown in Table 3. For timestamps, we consider two perturbations: removal, which discards all timestamp annotations, and shifting, which applies a uniform forward/backward offset while preserving relative intervals. Shifting results in only marginal degradation, suggesting robustness to changes in calendar-time anchoring, whereas removal causes a pronounced drop, highlighting the importance of temporal cues.

For contextual information, removing domain-specific descriptors degrades performance on datasets with strong seasonal or location patterns (e.g., Wind) but has little impact on highly volatile series (e.g., NASDAQ), indicating that the benefit of context is dataset-dependent. For normalization, we compare Z-score and RevIN with using raw time series and find that unnormalized inputs perform best. This suggests that standard normalization, while often helpful for deep forecasting models, can attenuate salient structures (e.g., trends and abrupt changes) that are useful for slow-thinking reasoning, and that preserving the original scale may be more compatible with our prompting setup.

6.2.2 Impact of Various Reasoning Strategies. We hypothesize that Rollout Reasoning may be more effective for short-horizon forecasting, as multi-round prediction can limit error propagation when the horizon is small, while Decoupled Reasoning may be better suited for long-horizon forecasting by enabling more thorough refinement within a single inference pass. We fix the lookback window to $L = 96$ and evaluate four predict windows $H \in \{48, 96, 144, 192\}$.

Table 4: Experimental results on ETTh1 and ETTm1 datasets showing model stability under different input settings.

Variants	ETTh1		ETTm1	
	MSE	MAE	MSE	MAE
Full	**5.4469**	**1.1984**	**14.2055**	**1.6984**
No-Imp	9.4523	1.4781	24.4894	2.2308
None-Imp	6.6059	1.2498	14.5843	1.7418
Lin-Imp	<u>5.7765</u>	<u>1.2218</u>	<u>14.4251</u>	<u>1.7154</u>

Figure 2 shows that One-Shot Reasoning remains competitive across all horizons. Rollout Reasoning performs best at short horizons ($H = 48, 96$), but its error grows more rapidly as the horizon increases ($H = 144, 192$), consistent with error accumulation from iterative rollouts. In contrast, Decoupled Reasoning becomes the best-performing method at longer horizons, suggesting that its within-pass refinement yields more stable predictions when long-range dependencies dominate.

6.3 Trustworthiness Evaluation

To address RQ3, we assess TimeReasoner's robustness by simulating missing and noisy inputs, and further evaluate its uncertainty estimation capability.

6.3.1 Stability Evaluation w.r.t. Missing Input. We evaluate the stability of TimeReasoner under missing values, a common challenge in real-world time series (e.g., due to sensor failures [7]). While traditional models typically rely on imputation, slow-thinking LLM-based approaches can process raw, incomplete data directly. We consider three variants: (1) No Imputation (No-Imp), where missing entries are removed; (2) None Placeholder Imputation (None-Imp), where missing values are replaced with a `None` token; and (3) Linear Interpolation Imputation (Lin-Imp) [12]. As shown in Table 4, TimeReasoner performs best with fully observed data, demonstrating its ability to capture complete temporal dependencies. It remains robust under linear interpolation, while using `None` placeholders causes moderate degradation due to increased uncertainty. In contrast, dropping missing entries leads to a sharp decline, as it disrupts the temporal structure. These results suggest that preserving temporal continuity is more beneficial than deleting missing entries even when using placeholders.

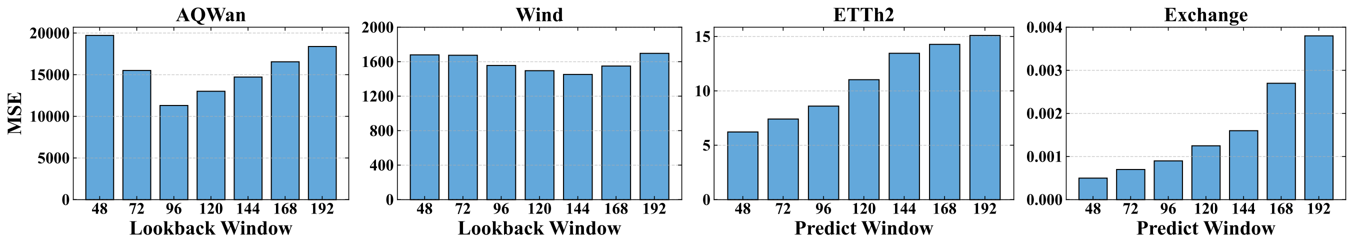


Figure 3: Relationship between TimeReasoner's forecasting performance and lookback window and predict window.

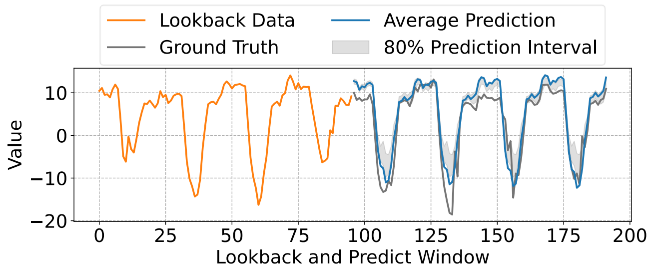


Figure 4: The 80% confidence interval of prediction and average prediction results during 50 independent generations given the same input on ETTh1 dataset.

6.3.2 Evaluation of Forecasting Uncertainty. In this section, we aim to evaluate TimeReasoner's capability in forecasting uncertainty quantification, with the goal of examining how well it captures the inherent reasoning uncertainty of large language models during multi-step inference and assessing both the accuracy and reliability of its probabilistic outputs.

As shown in Figure 4, predictions from the same LLM given identical input can vary notably, with the 80% prediction interval spanning a wide range across the forecast horizon. This highlights the inherent uncertainty in LLM-based forecasting and underscores the need to account for it when interpreting outputs. To quantify the reasoning uncertainty in slow-thinking forecasting, we compute the standard deviation of multiple predictions at each forecast step. Figure 5 shows per-step uncertainty curves across four datasets, each line representing the deviation trajectory over the forecast window. Uncertainty generally increases with forecast length, especially on ETTh1 and ETTh2, suggesting accumulation over longer generations. Sharp spikes—most notably on ETTh1 and ETTm1 reflect moments of high uncertainty. These patterns underscore the stochastic nature of LLM-based forecasting and the need for fine-grained uncertainty estimation.

6.4 Evaluation Under Different TSF Settings

To address RQ4, we vary the look-back window and prediction length to assess TimeReasoner's adaptability across different TSF task configurations.

6.4.1 Performance Comparison w.r.t Varying Lookback Window. To examine how the lookback window influences TimeReasoner's performance—given that longer histories often benefit traditional time series forecasting—we fix the predict window at $H = 96$ and vary $L \in \{48, 72, 96, 120, 144, 168, 192\}$. As shown in Figure 3, performance follows a non-monotonic trend: increasing L initially improves accuracy by providing richer patterns and context, but beyond an optimal length it degrades due to irrelevant or noisy information. This suggests that while TimeReasoner benefits from historical context, excessively long inputs can dilute salient signals, making it essential to balance context richness and relevance.

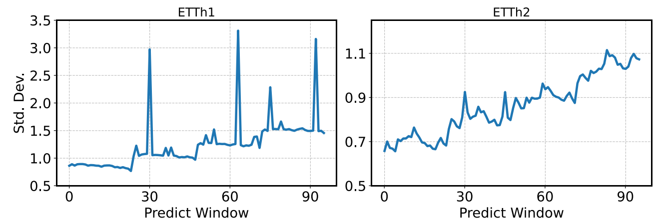


Figure 5: Standard deviation of predictions at each forecast step across ETTh1 and ETTh2 datasets.

6.4.2 Performance Comparison w.r.t Varying Predict Window. In previous approaches, forecasting errors generally exhibit a positive correlation with the predict window length H since longer horizons introduce greater error propagation. Whether this principle still holds for TimeReasoner, which relies on step-by-step slow reasoning rather than learned representations, remains unclear. To investigate this, we fix the lookback window L = 96 and gradually increase the predict window $H \in \{48, 72, 96, 120, 144, 168, 192\}$. As shown in Figure 3, MSE increases with a longer predict window. This suggests that the previous principle is applicable to TimeReasoner. This performance degradation suggests that TimeReasoner's reasoning becomes more difficult since it extrapolates further from observed data, limiting its effectiveness for long-term forecasting.

6.5 Hyperparameter Sensitivity Analysis

To address RQ5, we analyze the effects of inference-time hyperparameters, such as temperature and base LLM selection, on forecasting performance and stability.

6.5.1 Performance Comparison under Varying Temperatures. Temperature controls the randomness of token sampling during LLM decoding, thereby affecting output determinism and variability. To quantify its impact on TimeReasoner, we vary the decoding temperature while keeping the prompt construction and all other settings (including the lookback window and predict window) fixed, and evaluate forecasting performance. As shown in Figure 6, MSE exhibits a non-monotonic pattern: starting from $\tau = 0$, performance improves and achieves the lowest MSE at $\tau = 0.6$, but degrades as τ increases further. Lower temperatures yield more stable yet less flexible generations that may miss subtle temporal variations, whereas higher temperatures introduce excessive randomness, increasing variance and reducing reliability. These results suggest that an intermediate temperature provides a better trade-off between stability and controlled diversity in generation.

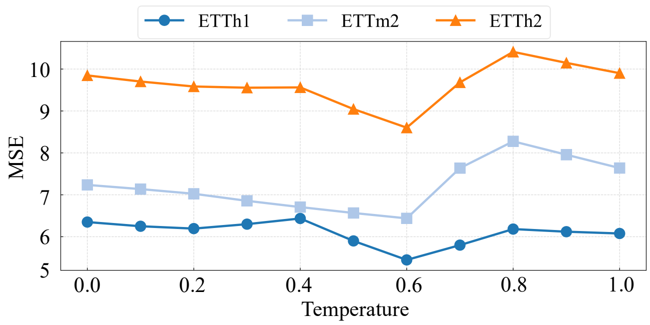


Figure 6: Sensitivity analysis of TimeReasoner to the temperature parameter across ETTh1, ETTh2, and ETTm2 datasets.

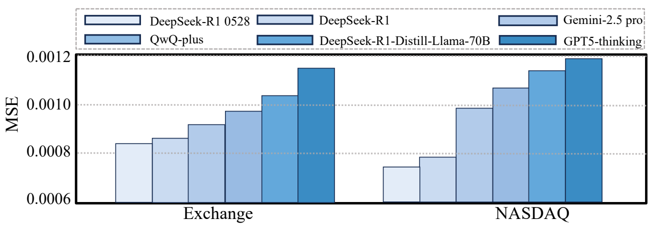


Figure 7: Comparative performance of TimeReasoner with different base LLMs on Exchange and NASDAQ datasets.

6.5.2 Performance Comparison across Base LLMs. We compare TimeReasoner across six backbone reasoning LLMs: DeepSeek-R1-0528, DeepSeek-R1, Gemini-2.5 Pro, QwQ-plus, DeepSeek-R1-Distill-Llama-70B, and GPT-5-Thinking. As shown in Figure 7, DeepSeek-R1-0528 delivers the best forecasting performance on both Exchange and NASDAQ, followed closely by DeepSeek-R1, while the distilled variant performs worst, highlighting the performance gap introduced by distillation. The improvement from DeepSeek-R1 to DeepSeek-R1-0528 further indicates that TimeReasoner benefits from stronger backbone reasoning capabilities.

Interestingly, GPT-5-Thinking underperforms in this setting. Qualitative inspection suggests that it sometimes relies on simple code-style extrapolation heuristics (e.g., linear or mean extrapolation) instead of closely tracking distributional cues in the look-back window, which can reduce responsiveness to regime shifts and degrade accuracy. Overall, these results suggest that the *style* of reasoning—data-attentive inference versus heuristic shortcutting—can be as important as model scale, and that observation-first prompting and data-focused exemplars may be a promising direction for improving robustness.

6.6 Visualization Analysis

To address RQ6, we analyze intermediate reasoning traces and a representative forecasting case to assess whether the output is aligned with the input data.

6.6.1 Reasoning Trace Analysis. Building on TimeReasoner's zero-shot forecasting performance, we examine its model-generated CoT-style reasoning traces. We observe a common three-stage structure: (1) Data inspection and pattern identification, where the model summarizes salient temporal characteristics (e.g., periodicity/seasonality, trend, and level shifts); (2) Strategy selection and justification, where it evaluates candidate forecasting heuristics (from seasonal baselines to trend extrapolation) and discusses their trade-offs; and (3) Self-reflection and contextual checks, where it critiques the chosen strategy, noting strengths (e.g., capturing seasonality) and limitations (e.g., lacking explicit statistical mechanisms). Overall, these traces suggest a diagnostic–select–verify workflow prior to producing the final forecast.

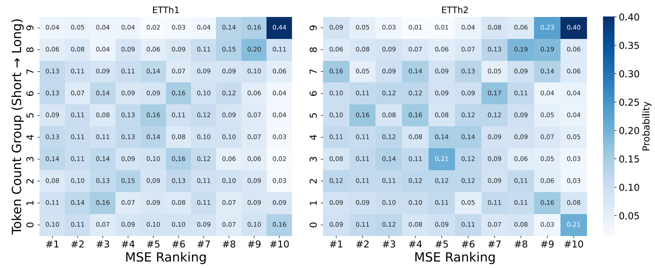


Figure 8: Heatmap of MSE rankings across CoT length groups on ETTh1 and ETTh2. The x-axis shows MSE ranks (#1–#10), and the y-axis shows CoT token deciles from shortest (bottom) to longest (top).

We further investigate the influence of the generated Chain-of-Thought (CoT) length on forecasting accuracy. As shown in Figure 8, all predictions are grouped into deciles according to their CoT token counts, with shorter CoTs placed in the lower deciles and longer ones in the upper deciles. A distinct pattern emerges: longer CoT generations are disproportionately concentrated in the worst MSE rankings, whereas shorter CoTs exhibit a more diverse distribution that occasionally includes top-ranked predictions. This observation implies that while moderate reasoning depth can aid pattern recognition and improve temporal alignment, excessively long reasoning chains may introduce irrelevant details, overfit transient patterns, or drift away from salient signals in the input. Such overextension not only inflates computational cost but can also amplify stochastic variability in token-by-token generation, ultimately degrading predictive performance. These findings highlight the importance of calibrating reasoning length, balancing depth with focus to achieve reliable and efficient LLM-based forecasting.

6.6.2 Forecasting Case Analysis. Figure 9 presents a representative example of TimeReasoner's inference-time reasoning, illustrating its deliberative inference behavior for time-series forecasting with lookback window $L = 96$ and predict window $H = 96$. The model first summarizes salient temporal patterns (e.g., trend and seasonality) and then evaluates candidate forecasting strategies. It subsequently inspects the most recent observations, noting potential anomalies and regime shifts. Based on these cues, TimeReasoner forms a forecasting plan by referencing comparable historical periods and enforcing continuity across the predicted sequence. The visualization on the right shows that the predicted window preserves both seasonal and trend components observed in the look-back window, suggesting that the model can produce interpretable, step-by-step forecasts.

6.6.3 Failure Case Analysis. We observe that not all forecasts generated by TimeReasoner are successful, with some exhibiting systematic and recurring errors. To better understand these limitations, we conduct a detailed manual inspection of poorly performing cases and categorize them into four representative failure patterns, as

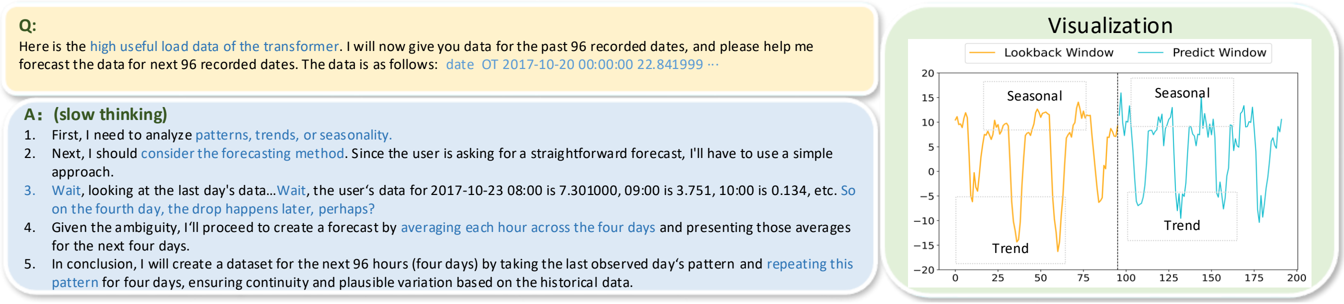
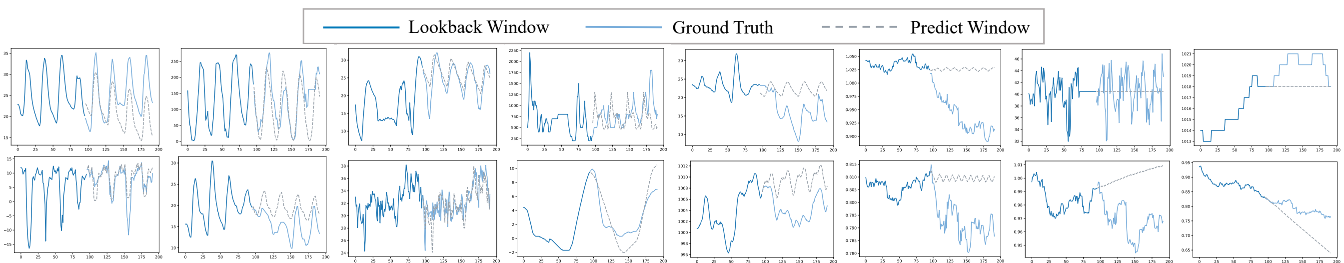


Figure 9: A case demonstrating how TimeReasoner performs slow-thinking capability for 96-96 time series reasoning.



(a) Peak Clipping (b) Phase-Shift Error (c) Copy-Paste Repeat (d) Constant Collapse

Figure 10: Four representative failure modes of TimeReasoner, including (a) Peak Clipping, (b) Phase-Shift Error, (c) Copy-Paste Repeat, and (d) Constant Collapse.

shown in Figure 10. (a) Peak clipping refers to systematic underestimation or overestimation at peaks and troughs, which flattens critical turning points and leads to the loss of important extreme-value information. (b) Phase-shift error describes cases where predictions capture the correct overall shape of the ground truth but are misaligned in time, causing decisions to be triggered too early or too late. (c) Copy-paste repeat reflects the tendency to directly replicate segments from the lookback window without adapting to new conditions, resulting in stale and inaccurate forecasts. (d) Constant collapse denotes degeneration into nearly constant outputs, disregarding inherent variability in the data.

These patterns reveal distinct weaknesses in amplitude capture, temporal alignment, adaptability, and output diversity. Such errors can severely undermine forecast reliability—misaligned peaks may cause decision failures in time-critical applications, excessive repetition can overlook regime changes or unexpected events, and constant predictions fail to reflect dynamic trends, leading to missed opportunities or inadequate responses. Beyond lowering pointwise accuracy, these issues also reduce the explainability and trustworthiness of the model's reasoning process. Addressing them will require enhancing sensitivity to extremes, improving phase anchoring to the correct temporal context, and discouraging trivial or degenerate outputs, thereby ensuring that the model remains both accurate and responsive across diverse forecasting scenarios.

7 Conclusion and Discussions

This work explored whether large language models (LLMs) with slow-thinking capabilities can reason over temporal patterns for time series forecasting. We reformulated forecasting as a conditional reasoning task, encoding historical time series and contextual features into structured prompts to guide trajectory generation. To realize this paradigm, we proposed TimeReasoner, a prompting-based framework that integrates hybrid inputs, including time series, timestamps, and semantic descriptors, further supporting multiple inference-time reasoning strategies (one-shot, decoupled, and roll-out). Experiments across diverse TSF benchmarks demonstrate that slow-thinking LLMs achieve strong zero-shot forecasting performance, while their intermediate reasoning traces reveal coherent patterns of temporal dynamics. This study represents an initial step toward reasoning-centric forecasting systems.

Discussions. Our results demonstrate the potential of slow-thinking approaches for time series forecasting, showing that even a relatively simple framework like TimeReasoner can benefit from the evolving ecosystem of reasoning-oriented LLMs, such as DeepSeek-R1, and is expected to gain further improvements as more advanced models become available. However, challenges remain: the reasoning trajectories generated by LLMs still exhibit uncertainty, calling for future work on principled uncertainty quantification; and the existence of erroneous predictions in some cases highlights the need for more robust reasoning and error-correction mechanisms. Addressing these limitations will be essential for realizing reliable and generalizable reasoning-based forecasting systems.

8 Acknowledgments

This research was supported by grants from the National Natural Science Foundation of China (No. 62502486, 62337001), the grants of Provincial Natural Science Foundation of Anhui Province (No. 2408085QF193), USTC Research Funds of the Double First-Class Initiative (No. YD2150002501), the Fundamental Research Funds for the Central Universities of China (No. WK2150110032).

Ethical Considerations

All experiments in this study are conducted on publicly available, fully anonymized benchmark datasets, which contain no personal or sensitive information. The research does not involve human subjects, protected attributes, or any data that could lead to privacy violations. No foreseeable negative societal impacts are identified, such as unfair treatment, malicious surveillance, or deliberate misuse by adversarial actors. The proposed TimeReasoner framework is designed for general-purpose time series forecasting research and does not target sensitive or safety-critical application domains. The work follows established ethical guidelines for data usage, reproducibility, and responsible scientific research, and all experimental settings are documented to ensure transparency and replicability.

References

[1] Abdul Fatir Ansari, Lorenzo Stella, Ali Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. [n. d.]. Chronos: Learning the Language of Time Series. *Transactions on Machine Learning Research* ([n. d.]).

[2] Konstantinos Benidis, Syama Sundar Rangapuram, Valentin Flunkert, Yuyang Wang, Danielle Maddix, Caner Turkmen, Jan Gasthaus, Michael Bohlke-Schneider, David Salinas, Lorenzo Stella, et al. 2022. Deep learning for time series forecasting: Tutorial and literature survey. *Comput. Surveys* 55, 6 (2022), 1–36.

[3] George EP Box and Gwilym M Jenkins. 1968. Some recent advances in forecasting and control. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 17, 2 (1968), 91–109.

[4] Ching Chang, Wen-Chih Peng, and Tien-Fu Chen. 2023. Llm4ts: Two-stage fine-tuning for time-series forecasting with pre-trained llms. *CoRR* (2023).

[5] Ching Chang, Wei-Yao Wang, Wen-Chih Peng, Tien-Fu Chen, and Sagar Samtani. 2024. Align and Fine-Tune: Enhancing LLMs for Time-Series Forecasting. In *NeurIPS Workshop on Time Series in the Age of Large Models*. https://openreview.net/forum?id=AaRCmJieG4

[6] Chris Chatfield. 1978. The Holt-winters forecasting procedure. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 27, 3 (1978), 264–279.

[7] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. 2018. Recurrent neural networks for multivariate time series with missing values. *Scientific reports* 8, 1 (2018), 6085.

[8] Yanxi Chen, Xuchen Pan, Yaliang Li, Bolin Ding, and Jingren Zhou. 2024. A simple and provable scaling law for the test-time compute of large language models. *arXiv preprint arXiv:2411.19477* (2024).

[9] Mingyue Cheng, Yiheng Chen, Qi Liu, Zhiding Liu, Yucong Luo, and Enhong Chen. 2025. InstrucTime: Advancing Time Series Classification with Multimodal Language Modeling. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining* (Hannover, Germany) *(WSDM '25)*. Association for Computing Machinery, New York, NY, USA, 792–800. doi:10.1145/3701551.3703499

[10] Mingyue Cheng, Zhiding Liu, Xiaoyu Tao, Qi Liu, Jintao Zhang, Tingyue Pan, Shilong Zhang, Panjing He, Xiaohan Zhang, Daoyu Wang, et al. 2025. A Comprehensive Survey of Time Series Forecasting: Concepts, Challenges, and Future Directions. *Authorea Preprints* (2025).

[11] Mingyue Cheng, Jintao Zhang, Zhiding Liu, and Chunli Liu. 2025. A hybrid multi-factor network with dynamic sequence modeling for early warning of intraoperative hypotension. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence* (Montreal, Canada) *(IJCAI '25)*. Article 548, 9 pages. doi:10.24963/ijcai.2025/548

[12] Gregory C Chow and An-loh Lin. 1971. Best linear unbiased interpolation, distribution, and extrapolation of time series by related series. *The review of Economics and Statistics* (1971), 372–375.

[13] Winnie Chow, Lauren Gardiner, Haraldur T Hallgrímsson, Maxwell A Xu, and Shirley You Ren. 2024. Towards time series reasoning with llms. *arXiv preprint arXiv:2409.11376* (2024).

[14] Chirag Deb, Fan Zhang, Junjing Yang, Siew Eang Lee, and Kwok Wei Shah. 2017. A review on time series forecasting techniques for building energy consumption. *Renewable and Sustainable Energy Reviews* 74 (2017), 902–924.

[15] Fuli Feng, Xiangnan He, Xiang Wang, Cheng Luo, Yiqun Liu, and Tat-Seng Chua. 2019. Temporal relational ranking for stock prediction. *ACM Transactions on Information Systems (TOIS)* 37, 2 (2019), 1–30.

[16] Yair Gat, Nitay Calderon, Amir Feder, Alexander Chapanin, Amit Sharma, and Roi Reichart. 2023. Faithful explanations of black-box nlp models using llm-generated counterfactuals. *arXiv preprint arXiv:2310.00603* (2023).

[17] Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. 2024. Moment: A family of open time-series foundation models. *arXiv preprint arXiv:2402.03885* (2024).

[18] Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew G Wilson. 2023. Large language models are zero-shot time series forecasters. *Advances in Neural Information Processing Systems* 36 (2023), 19622–19635.

[19] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948* (2025).

[20] Joe Hasei, Ryuichi Nakahara, Koichi Takeuchi, Aki Yoshida, Takuto Itano, Tomohiro Fujiwara, Eiji Nakata, Toshiyuki Kunisada, and Toshifumi Ozaki. 2025. Comparative analysis of a standard (GPT-4o) and reasoning-enhanced (o1 pro) large language model on complex clinical questions from the Japanese orthopaedic board examination. *Journal of Orthopaedic Science* (2025).

[21] Hansika Hewamalage, Christoph Bergmeir, and Kasun Bandara. 2021. Recurrent neural networks for time series forecasting: Current status and future directions. *International Journal of Forecasting* 37, 1 (2021), 388–427.

[22] Shushan Hu, Edward Corry, Edward Curry, William JN Turner, and James O'Donnell. 2016. Building performance optimisation: A hybrid architecture for the integration of contextual information and time-series data. *Automation in Construction* 70 (2016), 51–61.

[23] Romain Ilbert, Malik Tiomoko, Cosme Louart, Ambroise Odonnat, Vasilii Feofanov, Themis Palpanas, and Ievgen Redko. 2024. Analysing multi-task regression via random matrix theory with application to time series forecasting. *Advances in Neural Information Processing Systems* 37 (2024), 115021–115057.

[24] Kartikeya Jha, Nishita Sinha, Shriniwas S. Arkatkar, and Ashoke K. Sarkar. 2016. A comparative study on application of time series analysis for traffic forecasting in India: prospects and limitations. *Current Science* 110, 3 (2016), 373–385. http://www.jstor.org/stable/24906782

[25] Yushan Jiang, Zijie Pan, Xikun Zhang, Sahil Garg, Anderson Schneider, Yuriy Nevmyvaka, and Dongjin Song. 2024. Empowering time series analysis with large language models: a survey. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence* (Jeju, Korea) *(IJCAI '24)*. Article 895, 9 pages. doi:10.24963/ijcai.2024/895

[26] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, et al. 2023. Time-llm: Time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2310.01728* (2023).

[27] Ming Jin, Yifan Zhang, Wei Chen, Kexin Zhang, Yuxuan Liang, Bin Yang, Jindong Wang, Shirui Pan, and Qingsong Wen. 2024. Position: what can large language models tell us about time series analysis. In *Proceedings of the 41st International Conference on Machine Learning* (Vienna, Austria) *(ICML'24)*. JMLR.org, Article 895, 17 pages.

[28] Tomi Kauppinen and Eero Hyvönen. 2007. Modeling and reasoning about changes in ontology time series. In *Ontologies: A handbook of principles, concepts and applications in information systems*. Springer, 319–338.

[29] Shruti Kaushik, Abhinav Choudhury, Pankaj Kumar Sheron, Nataraj Dasgupta, Sayee Natarajan, Larry A Pickett, and Varun Dutt. 2020. AI in healthcare: time-series forecasting using statistical, neural, and ensemble architectures. *Frontiers in big data* 3 (2020), 4.

[30] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[31] Pranjal Kumar. 2024. Large language models (LLMs): survey, technical frameworks, and future challenges. *Artificial Intelligence Review* 57, 10 (2024), 260.

[32] Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. 2018. Modeling Long- and Short-Term Temporal Patterns with Deep Neural Networks. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (Ann Arbor, MI, USA) *(SIGIR '18)*. Association for Computing Machinery, New York, NY, USA, 95–104. doi:10.1145/3209978.3210006

[33] Hanchung Lee. 2024. Reasoning Series, Part 2: Reasoning and Inference Time Scaling. https://leehanchung.github.io. (10 2024). https://leehanchung.github.io/blogs/2024/10/21/reasoning-inference-scaling/

[34] Shiyang Li, Xiaoyong Jin, Yao Xuan, Xiyou Zhou, Wenhu Chen, Yu-Xiang Wang, and Xifeng Yan. 2019. *Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting*. Curran Associates Inc., Red Hook, NY, USA.

[35] Bryan Lim and Stefan Zohren. 2021. Time-series forecasting with deep learning: a survey. *Philosophical Transactions of the Royal Society A* 379, 2194 (2021), 20200209.

[36] Haoxin Liu, Zhiyuan Zhao, Jindong Wang, Harshavardhan Kamarthi, and B. Aditya Prakash. 2024. LSTPrompt: Large Language Models as Zero-Shot Time Series Forecasters by Long-Short-Term Prompting. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 7832–7840. doi:10.18653/v1/2024.findings-acl.466

[37] Jingwei Liu, Ling Yang, Hongyan Li, and Shenda Hong. 2024. Retrieval-augmented diffusion models for time series forecasting. In *Proceedings of the 38th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '24)*. Curran Associates Inc., Red Hook, NY, USA, Article 91, 21 pages.

[38] Minhao Liu, Ailing Zeng, Muxi Chen, Zhijian Xu, Qiuxia Lai, Lingna Ma, and Qiang Xu. 2022. Scinet: Time series modeling and forecasting with sample convolution and interaction. *Advances in Neural Information Processing Systems* 35 (2022), 5816–5828.

[39] Peiyuan Liu, Hang Guo, Tao Dai, Naiqi Li, Jigang Bao, Xudong Ren, Yong Jiang, and Shu-Tao Xia. 2025. CALF: aligning LLMs for time series forecasting via cross-modal fine-tuning. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence (AAAI'25/IAAI'25/EAAI'25)*. AAAI Press, Article 2109, 9 pages. doi:10.1609/aaai.v39i18.34082

[40] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. 2023. itransformer: Inverted transformers are effective for time series forecasting. *arXiv preprint arXiv:2310.06625* (2023).

[41] Zhiding Liu, Mingyue Cheng, Zhi Li, Zhenya Huang, Qi Liu, Yanhu Xie, and Enhong Chen. 2023. Adaptive Normalization for Non-stationary Time Series Forecasting: A Temporal Slice Perspective. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 14273–14292. https://proceedings.neurips.cc/paper_files/paper/2023/file/2e19dab94882bc95ed094c4399cfda02-Paper-Conference.pdf

[42] Zhiding Liu, Jiqian Yang, Mingyue Cheng, Yucong Luo, and Zhi Li. 2024. Generative Pretrained Hierarchical Transformer for Time Series Forecasting. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Barcelona, Spain) *(KDD '24)*. Association for Computing Machinery, New York, NY, USA, 2003–2013. doi:10.1145/3637528.3671855

[43] Sarah Mercer, Samuel Spillard, and Daniel P Martin. 2025. Brief analysis of DeepSeek R1 and its implications for Generative AI. *SuperIntelligence-Robotics-Safety & Alignment* 2, 1 (2025).

[44] 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. *arXiv preprint arXiv:2501.19393* (2025).

[45] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. 2023. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=Jbdc0vTOcol

[46] OpenAI. 2024. Learning to Reason with LLMs. https://openai.com/index/learning-to-reason-with-llms/

[47] Tingyue Pan, Mingyue Cheng, Shilong Zhang, Zhiding Liu, Xiaoyu Tao, Yucong Luo, Jintao Zhang, and Qi Liu. 2025. OneCast: Structured Decomposition and Modular Generation for Cross-Domain Time Series Forecasting. *arXiv preprint arXiv:2510.24028* (2025).

[48] Shubham Parashar, Blake Olson, Sambhav Khurana, Eric Li, Hongyi Ling, James Caverlee, and Shuiwang Ji. 2025. Inference-time computations for llm reasoning and planning: A benchmark and insights. *arXiv preprint arXiv:2502.12521* (2025).

[49] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. *PyTorch: an imperative style, high-performance deep learning library*. Curran Associates Inc., Red Hook, NY, USA.

[50] David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. 2020. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International journal of forecasting* 36, 3 (2020), 1181–1191.

[51] Nishad Singhi, Hritik Bansal, Arian Hosseini, Aditya Grover, Kai-Wei Chang, Marcus Rohrbach, and Anna Rohrbach. 2025. When To Solve, When To Verify: Compute-Optimal Problem Solving and Generative Verification for LLM Reasoning. *arXiv preprint arXiv:2504.01005* (2025).

[52] Mingtian Tan, Mike Merrill, Vinayak Gupta, Tim Althoff, and Tom Hartvigsen. 2024. Are language models actually useful for time series forecasting? *Advances in Neural Information Processing Systems* 37 (2024), 60162–60191.

[53] Hua Tang, Chong Zhang, Mingyu Jin, Qinkai Yu, Zhenting Wang, Xiaobo Jin, Yongfeng Zhang, and Mengnan Du. 2025. Time series forecasting with llms: Understanding and enhancing model capabilities. *ACM SIGKDD Explorations Newsletter* 26, 2 (2025), 109–118.

[54] Xiaoyu Tao, Tingyue Pan, Mingyue Cheng, and Yucong Luo. 2024. Hierarchical multimodal llms with semantic space alignment for enhanced time series classification. *arXiv preprint arXiv:2410.18686* (2024).

[55] Xiaoyu Tao, Shilong Zhang, Mingyue Cheng, Daoyu Wang, Tingyue Pan, Bokai Pan, Changqing Zhang, and Shijin Wang. 2025. From values to tokens: An llm-driven framework for context-aware time series forecasting via symbolic discretization. *arXiv preprint arXiv:2508.09191* (2025).

[56] Francis EH Tay and Lijuan Cao. 2001. Application of support vector machines in financial time series forecasting. *omega* 29, 4 (2001), 309–317.

[57] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).

[58] Chengsen Wang, Qi Qi, Jingyu Wang, Haifeng Sun, Zirui Zhuang, Jinming Wu, and Jianxin Liao. 2024. Rethinking the power of timestamps for robust time series forecasting: a global-local fusion perspective. In *Proceedings of the 38th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '24)*. Curran Associates Inc., Red Hook, NY, USA, Article 700, 27 pages.

[59] Jiahao Wang, Mingyue Cheng, Qingyang Mao, Yitong Zhou, Daoyu Wang, Qi Liu, Feiyang Xu, and Xin Li. 2025. TableTime: Reformulating Time Series Classification as Training-Free Table Understanding with Large Language Models. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management* (Seoul, Republic of Korea) *(CIKM '25)*. Association for Computing Machinery, New York, NY, USA, 3009–3019. doi:10.1145/3746252.3761056

[60] Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Guo Qin, Haoran Zhang, Yong Liu, Yunzhong Qiu, Jianmin Wang, and Mingsheng Long. 2024. TimeXer: empowering transformers for time series forecasting with exogenous variables. In *Proceedings of the 38th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '24)*. Curran Associates Inc., Red Hook, NY, USA, Article 15, 30 pages.

[61] 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 (2022), 24824–24837.

[62] Qingsong Wen, Zhe Zhang, Yan Li, and Liang Sun. 2020. Fast RobustSTL: Efficient and Robust Seasonal-Trend Decomposition for Time Series with Complex Patterns. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Virtual Event, CA, USA) *(KDD '20)*. Association for Computing Machinery, New York, NY, USA, 2203–2213. doi:10.1145/3394486.3403271

[63] Ruofeng Wen, Kari Torkkola, Balakrishnan Narayanaswamy, and Dhruv Madeka. 2017. A multi-horizon quantile recurrent forecaster. *arXiv preprint arXiv:1711.11053* (2017).

[64] 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* (2024).

[65] Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. 2024. Unified training of universal time series forecasting transformers. In *Proceedings of the 41st International Conference on Machine Learning* (Vienna, Austria) *(ICML'24)*. JMLR.org, Article 2178, 25 pages.

[66] Wenkai Yang, Shuming Ma, Yankai Lin, and Furu Wei. 2025. Towards thinking-optimal scaling of test-time compute for llm reasoning. *arXiv preprint arXiv:2502.18080* (2025).

[67] Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. 2024. A survey on large language model (LLM) security and privacy: The Good, The Bad, and The Ugly. *High-Confidence Computing* 4, 2 (2024), 100211. doi:10.1016/j.hcc.2024.100211

[68] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. 2023. Are transformers effective for time series forecasting?. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence (AAAI'23/IAAI'23/EAAI'23)*. AAAI Press, Article 1248, 8 pages. doi:10.1609/aaai.v37i9.26317

[69] Chaolv Zeng, Yuan Tian, Guanjie Zheng, and Yunjun Gao. 2024. How Much Can Time-related Features Enhance Time Series Forecasting? *arXiv preprint arXiv:2412.01557* (2024).

[70] Zhiyuan Zeng, Qinyuan Cheng, Zhangyue Yin, Yunhua Zhou, and Xipeng Qiu. 2025. Revisiting the Test-Time Scaling of o1-like Models: Do they Truly Possess Test-Time Scaling Capabilities?. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 4651–4665. doi:10.18653/v1/2025.acl-long.232

[71] G Peter Zhang. 2003. Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing* 50 (2003), 159–175.

[72] Jintao Zhang, Mingyue Cheng, Xiaoyu Tao, Zhiding Liu, and Daoyu Wang. 2025. Conditional denoising meets polynomial modeling: a flexible decoupled framework for time series forecasting. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence* (Montreal, Canada) *(IJCAI '25)*. Article 778, 9 pages. doi:10.24963/ijcai.2025/778

[73] Qiyuan Zhang, Fuyuan Lyu, Zexu Sun, Lei Wang, Weixu Zhang, Zhihan Guo, Yufei Wang, Irwin King, Xue Liu, and Chen Ma. 2025. What, How, Where, and How Well? A Survey on Test-Time Scaling in Large Language Models. *arXiv preprint arXiv:2503.24235* (2025).

[74] Shuyi Zhang, Bin Guo, Anlan Dong, Jing He, Ziping Xu, and Song Xi Chen. 2017. Cautionary tales on air-quality improvement in Beijing. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 473, 2205 (2017),

20170457.

[75] Xiyuan Zhang, Ranak Roy Chowdhury, Rajesh K. Gupta, and Jingbo Shang. 2024. Large language models for time series: a survey. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence* (Jeju, Korea) *(IJCAI '24)*. Article 921, 9 pages. doi:10.24963/ijcai.2024/921

[76] Xiaohan Zhang, Tian Gao, Mingyue Cheng, Bokai Pan, Ze Guo, Yaguo Liu, and Xiaoyu Tao. 2025. AlphaCast: A Human Wisdom-LLM Intelligence Co-Reasoning Framework for Interactive Time Series Forecasting. *arXiv preprint arXiv:2511.08947* (2025).

[77] Xi Nicole Zhang, Yuan Pu, Yuki Kawamura, Andrew Loza, Yoshua Bengio, Dennis Shung, and Alexander Tong. 2024. Trajectory flow matching with applications to clinical time series modelling. *Advances in Neural Information Processing Systems* 37 (2024), 107198–107224.

[78] Yunhao Zhang and Junchi Yan. 2023. Crossformer: Transformer Utilizing Cross-Dimension Dependency for Multivariate Time Series Forecasting. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=vSVLM2j9eie

[79] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. 2021. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 11106–11115.

[80] Tian Zhou, Ziqing Ma, Qingsong Wen, Liang Sun, Tao Yao, Wotao Yin, Rong Jin, et al. 2022. Film: Frequency improved legendre memory model for long-term time series forecasting. *Advances in neural information processing systems* 35 (2022), 12677–12690.

[81] Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. 2022. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International conference on machine learning*. PMLR, 27268–27286.

[82] Tian Zhou, Peisong Niu, Liang Sun, Rong Jin, et al. 2023. One fits all: Power general time series analysis by pretrained lm. *Advances in neural information processing systems* 36 (2023), 43322–43355.