

Exploring Reasoning Approaches for Zero-Shot Time Series Forecasting: A Benchmark and Insights

Abstract

Reasoning ability is crucial for solving challenging tasks. With the advancement of foundation models, especially large language models (LLMs), a wide range of reasoning strategies has been proposed, including test-time enhancements, such as Chain-of-Thought, and post-training optimizations, as used in DeepSeek-R1. While these reasoning strategies have demonstrated effectiveness across various challenging language or vision tasks, their applicability and impact on time-series forecasting (TSF), particularly the challenging zero-shot TSF, remain largely unexplored. In particular, it is unclear whether zero-shot TSF benefits from reasoning and, if so, what types of reasoning strategies are most effective.

To bridge this gap, we propose REC4TS, the first benchmark that systematically evaluates the effectiveness of popular reasoning strategies when applied to zero-shot TSF tasks. REC4TS conducts comprehensive evaluations across datasets spanning eight domains, covering both unimodal and multimodal with short-term and long-term forecasting tasks. More importantly, REC4TS provides key insights: (1) Self-consistency emerges as the most effective plug-in reasoning strategy; (2) Hybrid reasoning is more suitable for TSF than pure system 1 or 2; (3) Group-relative policy optimization emerges as a more suitable approach for incentivizing reasoning ability during post-training; (4) Multimodal TSF benefits more from reasoning strategies compared to unimodal TSF. Beyond these insights, REC4TS establishes two pioneering starting blocks to support future zero-shot TSF reasoning research: (1) A novel dataset, TIME-THINKING, containing forecasting samples annotated with reasoning trajectories from multiple advanced LLMs, and (2) A new and simple test-time scaling-law validated on foundational TSF models enabled by self-consistency reasoning strategy. All data and code are publicly accessible at: <https://anonymous.4open.science/r/REC4TS/>

1 Introduction

Reasoning capabilities are fundamental to solving challenging tasks and have been the focus of research for decades [13, 71]. Recent advancements in foundation models, particularly the rise of large language models (LLMs), have brought

reasoning strategy research into the era of artificial intelligence [67, 64, 47]. By integrating designed reasoning strategies with powerful foundation models, these models have demonstrated remarkable success in tackling complex real-world challenges, including math, planning, and engineering [3, 55, 56, 24]. Despite these advancements in extensive domains, the application of reasoning strategies in zero-shot time-series forecasting (TSF) remains largely unexplored. While recent time-series foundation models have initially demonstrated the feasibility of zero-shot TSF [9, 15, 2, 54], they primarily rely on the memory of large-scale time series data from the pre-training phase rather than employing explicit reasoning for zero-shot TSF.

However, reasoning is not only natural but also essential for zero-shot TSF. On the one hand, time-series data inherently contains rich causal relationships¹, making it a natural domain for reasoning-based approaches [42, 50]. For instance, in epidemiological forecasting, the outbreak of a pandemic naturally reasons a subsequent increase in infections—an essential target for flu prediction [41]. On the other hand, and more critically, reasoning plays an important role in zero-shot TSF. Unlike conventional supervised TSF, which “memorizes” underlying correlation structures from training data, zero-shot TSF—designed as the primary objective for most time-series foundation models—depends on reasoning capabilities to “interpret” the pattern correlations within limited historical data, which have never been “memorized,” in order to infer future values. Nevertheless, existing time-series foundation models have yet to fully exploit them by incorporating effective reasoning strategies.

Therefore, a natural twofold research question (RQ) for reasoning strategies on zero-shot TSF arises: “**RQ1: Can zero-shot TSF benefit from enhanced reasoning ability?**” While reasoning is an integral part of time-series analysis, such as used in feature selection and statistical analysis [57, 7], its effectiveness has yet to be systematically validated in zero-shot TSF with foundation models. Furthermore, if reasoning is indeed beneficial (as

¹Here, by “causal relationships” we refer to predictive causal structures such as Granger causality [17]. Check the research motivation of using LLMs in App. B

demonstrated later in our work), a subsequent question emerges: **“RQ2: What kind of reasoning strategies does zero-shot TSF need?”**

Existing reasoning strategies are often categorized into three cognitive systems: Reasoning enhanced System 1, which enables fast and direct thinking, System 2, which involves deep and analytical reasoning [26] built-in the foundation model, and Hybrid Reasoning, which explicitly controls reasoning effort. These three approaches differ distinctly in their implementation. Reasoning enhanced System 1 typically occurs during the inference stage, such as Chain-of-Thought (CoT) reasoning, possibly being further enhanced through self-consistency and self-correction [67, 64, 47, 31]. In contrast, System 2 reasoning asks the model itself to promote deeper, structured thinking by incorporating incentives and rewards during post-training [46, 19]. Furthermore, hybrid reasoning adjusts the mixture of System 1 and System 2 by setting the reasoning effort hyperparameter. Given these fundamental differences, addressing **RQ2** requires an in-depth investigation into their effectiveness for time-series forecasting. However, such studies remain unexplored, and a systematic benchmark to assess the impact of different reasoning strategies in TSF is still lacking.

To bridge this gap and address the proposed twofold research questions, we introduce REC4TS, a benchmark designed to systematically evaluate the effectiveness of various popular reasoning strategies in zero-shot TSF. To the best of our knowledge, this is the first benchmark study to investigate these questions. **Our main contributions are summarized as follows:**

- **New Research Direction:** We introduce a novel research direction focused on reasoning strategies for zero-shot TSF. This research aims to understand when, what, and how reasoning strategies impact zero-shot TSF tasks.
- **Comprehensive Benchmark:** We systematically evaluate the effectiveness of existing reasoning strategies for zero-shot TSF, conducting about 2,500 experiments. Our analysis covers both reasoning-enhanced System 1, System 2 and hybrid reasoning strategies across nine foundation models.
- **In-depth Insights:** Our benchmark provides in-depth insights, revealing that self-consistency is the current most effective reasoning-enhanced System 1 strategy, group-

relative policy optimization is the current most suitable System 2 approach, and hybrid reasoning with low reasoning effort can balance performance and computational cost. Furthermore, we show that reasoning strategies are generally more effective in multimodal zero-shot TSF than unimodal zero-shot TSF.

- **Open-Source Toolkits:** We release three key toolkits, including the evaluation suite, scaling law and reasoning data to comprehensively initiate future research on reasoning models for TSF. Specifically, we fully release REC4TS as an easy-to-use evaluation suite, including the code, documents and experiment logs. We validate a new and simple test-time scaling law with foundation time-series models. Additionally, we introduce TIME-THINKING, the first TSF dataset with reasoning trajectories.

We provide related work in App. A, clarify the research motivation and position in App. B, and present the limitations in App. D.

2 REC4TS: A Suite for Evaluating Reasoning Strategies for Zero-Shot TSF

REC4TS consists of four core modules: Datasets, Reasoning Strategies, Models, and Evaluations. We introduce these modules one by one in this section. More details are provided in Section E.

2.1 Dataset module

The dataset module includes datasets from eight different domains, all containing both numerical time series and aligned textual context series, providing unified data support for downstream time series forecasting. As detailed in Tab. 7, these verified datasets [36, 33] cover key domains such as Agriculture, Climate, Economy, Energy, Health, Security, Employment, and Traffic, with weekly and monthly frequencies. The textual context series consists of keyword-based web summaries, aligned by date with the numerical series. As shown in Fig. 5, these datasets exhibit diverse characteristics, enabling comprehensive evaluation.

2.2 Reasoning Strategies Module

As shown in Fig. 1, REC4TS systematically includes three mainstream approaches for reasoning, following existing works [47, 48, 72]: (1) Direct System 1 Reasoning – directly using generative models such as GPT-4o for reasoning. (2) Test-Time-Enhanced System 1 Reasoning – incorpo-

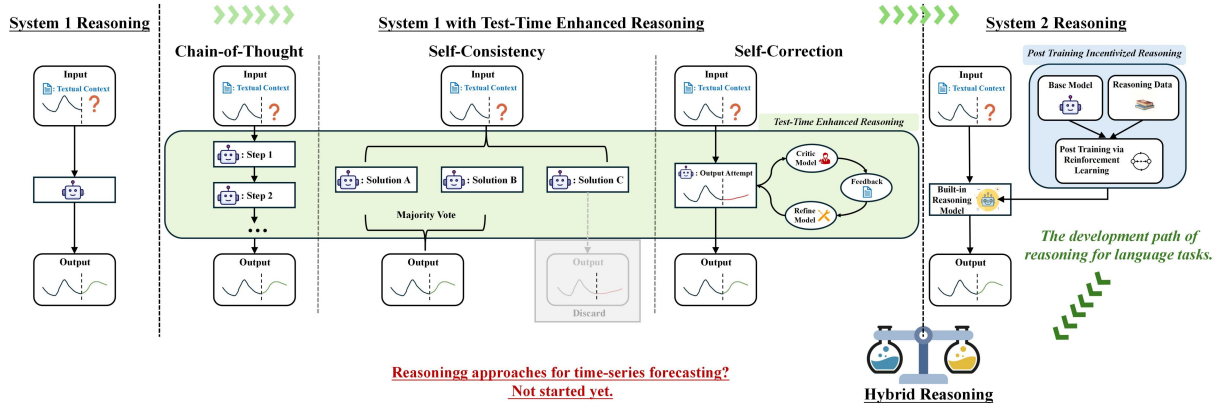


Figure 1: The reasoning strategies included in the proposed REC4TS benchmark. REC4TS systematically includes three mainstream approaches: the direct System 1, i.e., directly using generative models such as GPT-4o for reasoning; the test-time-enhanced System 1, including Chain-of-Thought, Self-Consistency, and Self-Correction; the post-training-empowered System 2, which enables built-in reasoning capabilities through reinforcement learning, such as DeepSeek-R1 [19]; the hybrid reasoning models with adjustable reasoning effort, such as Claude-3.7-Sonnet.

rating techniques such as Chain-of-Thought [67], Self-Consistency² [64], and Self-Correction [40]. These approaches improve reasoning beyond System 1’s intuitive responses by performing additional computations during inference without modifying the model’s pre-trained weights. (3) Post-Training-Enabled System 2 Reasoning – such as DeepSeek-R1 [19]. Unlike test-time-enhanced System 1, System 2 reasoning is typically achieved through reinforcement learning and reasoning data during the post-training phase and has built-in reasoning capabilities. (4) Hybrid reasoning systems, which are able to explicitly adjust reasoning effort.

Specifically, Chain-of-Thought (CoT) prompts the model to break down complex problems into a series of logical steps before providing a final answer. The Self-Consistency method further enhances reasoning diversity by generating multiple reasoning paths in parallel and selecting the most consistent result. In contrast, the Self-Correction approach iteratively refines the model’s output through feedback, aiming to improve overall accuracy and reliability. For the System 2 strategies, also known as large reasoning models [72], REC4TS includes the closed-source o1-mini from OpenAI and Gemini-2.0-flash-thinking from Google, as well as the open-source DeepSeek-R1 from DeepSeek. Compared to o1-mini, which employs Proximal Policy Optimization (PPO) [51] by training two models simultaneously—a policy model for generating responses and a critic model for evaluating them—DeepSeek-R1 adopts Group Relative Policy Optimization (GRPO) [53], which

²See the conceptual difference between self-consistency, ensembling, and uncertainty-based sampling in App. C.

eliminates the need for a separate critic model by ranking multiple responses at once.

2.3 Models Module

REC4TS covers three series of foundational models, including both closed- and open-source models. Each series provides **pure System 1 and System 2 versions**. Note that since reasoning strategies for foundational time-series models have not yet been studied and are difficult to implement directly, reusing foundational language models for zero-shot TSF—which have been widely validated by existing works [74, 18, 37, 25, 6]—is currently the best choice. Inspired by recent research [62, 22, 23], we reformulate numerical time series into a tabular format, i.e., "timestamp : numerical value", to enable LLMs as powerful time-series analysts. Specifically, REC4TS includes OpenAI’s GPT-4o and o1-mini, Google’s Gemini-2.0-Flash and Gemini-2.0-flash-thinking, and DeepSeek’s DeepSeek-V3 and DeepSeek-R1 as the corresponding System 1 and System 2 pairs, respectively.

Furthermore, REC4TS includes **state-of-the-art hybrid reasoning models** beyond pure System 1 and 2 models, including OpenAI’s o4-mini, Google’s Gemini-2.5-Pro, and Anthropic’s Claude-3.7-Sonnet. These models allow control over the extended thinking cost by adjusting reasoning effort level. Specifically, we define three levels of reasoning effort: low, medium, and high. Specifically, for o4-mini, we explicitly set the three reasoning effort levels. For the other two models, we approximate the levels using 1024, 2048, and 4096 reasoning tokens, respectively. For these models, REC4TS uniformly adopts GPT-4o

as the reference.

2.4 Evaluation Module

We comprehensively consider the following four common settings: unimodal short-term, unimodal long-term, multi-modal short-term, and multi-modal long-term. In multi-modal TSF, both numerical series and aligned textual context series are used as inputs, whereas unimodal TSF uses only numerical series. The forecasting period for long-term TSF is the next six months whereas the short-term TSF is the next three months. We follow most existing TSF works [70, 69, 45] by setting the lookback window length to 96 by default. We use Mean Squared Error (MSE) as the evaluation metric. To avoid data contamination—meaning the evaluation dataset may have been seen during the foundation model’s pretraining—we use horizon windows after October 2023, which is the knowledge cutoff date of selected foundation models.

Implementation Details. For reproducibility, we summarize all key implementation details in Tab. 4, including (i) time-series textualization and prompt templates, (ii) reasoning-strategy setting, and (iii) model versions and decoding hyperparameters. Full prompt templates and reasoning-strategy implementations are in App. E.2, and model versions plus decoding configurations are in App. E.4.

3 Experimental Results and Insights

Based on the constructed REC4TS suite, we conduct experiments to evaluate reasoning strategies for zero-shot TSF across eight datasets and four settings. We repeat each experiment three times, reporting the average MSE and standard deviation. We then discuss the two research questions raised

in Section 1 one by one, following the structure of Answer – Evidence – Analysis. We detail the experimental results comparing pure System 1 and System 2 models in Tab. 1, Tab. 8, and Tab. 2, corresponding to OpenAI’s, Google’s, and DeepSeek’s pure System 1 and 2 models, respectively. We visualize the average win rate of different reasoning strategies relative to direct System 1 inference, where 50% indicates a tie, in Fig. 3.

Similarly, we present the experimental results comparing pure System 1 and hybrid reasoning models in App Tab. 9, Tab. 10, and Tab. 11, corresponding to OpenAI’s, Google’s, and Anthropic’s hybrid reasoning models, respectively. We visualize how the average win rate relative to System 1, where 50% means a tie, in Fig. 3, and GPT-4o changes with different levels of reasoning effort in Fig. 2. We then discuss the two RQs raised in Section 1 in the following.

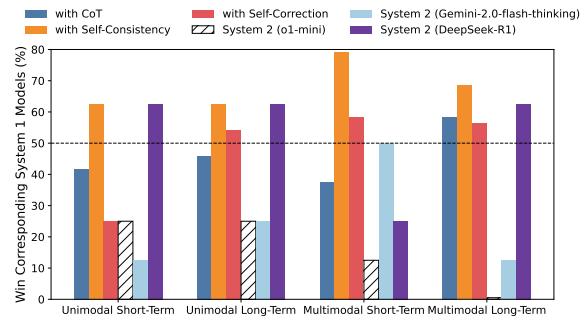


Figure 3: The average win rate of reasoning strategies compared to corresponding direct System 1. We observe the consistent and significant effectiveness of self-consistency, as well as the unique effectiveness of DeepSeek-R1 among System 2 strategies.

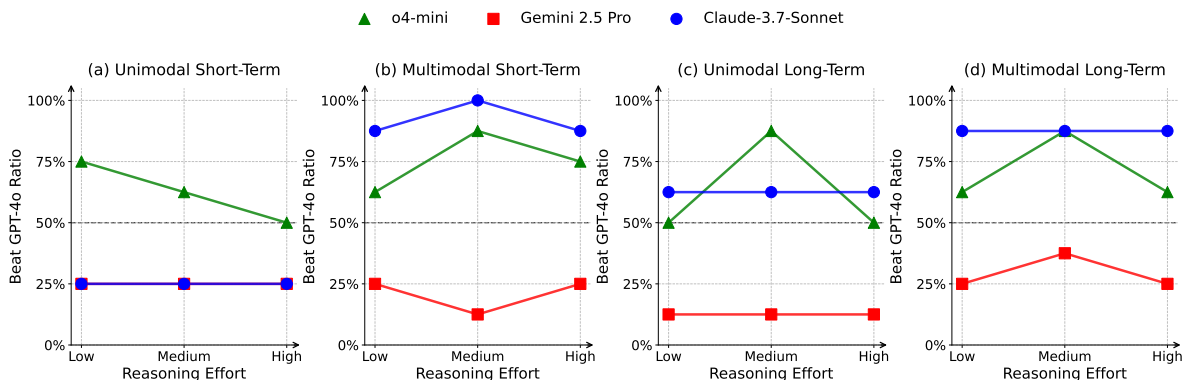


Figure 2: The average win rate of state-of-the-art hybrid reasoning models compared to GPT-4o under different levels of reasoning effort. We note that the increase of reasoning effort, i.e., extended thinking tokens, cannot improve TSF performance, but increases computational cost. Overall, hybrid reasoning under lower reasoning effort is the most cost-effective choice. See App. F for the systematic analysis and case study of higher reasoning effort’s failure.

Table 1: Results with OpenAI’s System 1 (GPT-4o) and 2 (o1-mini) Models. We report the mean MSE and standard deviation over three repeated experiments. Reasoning strategies that outperform the direct System 1 are highlighted in green, while those that perform worse or have similar performance (due to higher computational cost) are marked in red. In "Win System 1," we present the probability of each reasoning strategy outperforming System 1 across datasets. We observe that **only the self-consistency strategy is consistently effective**, while **the System 2 strategy is consistently ineffective**.

(a) Results of Unimodal Short-term TSF Settings. We use numerical series only to forecast the next three months.

| Dataset | System 1 | System 1 with Test-time Reasoning Enhancement | | | System 2 |
|--------------|-------------|---|-----------------------|----------------------|-------------|
| | GPT-4o | with CoT | with Self-Consistency | with Self-Correction | o1-mini |
| Agriculture | 0.021±0.011 | 0.909±1.275 | 0.021±0.003 | 0.025±0.007 | 0.069±0.013 |
| Climate | 1.599±0.500 | 1.704±0.164 | 1.517±0.263 | 1.998±0.677 | 1.412±0.159 |
| Economy | 0.631±0.135 | 0.638±0.410 | 0.450±0.171 | 1.018±0.184 | 0.583±0.001 |
| Energy | 0.363±0.110 | 0.258±0.029 | 0.167±0.242 | 0.396±0.086 | 0.930±0.747 |
| Flu | 0.568±0.425 | 0.592±0.291 | 0.481±0.288 | 0.663±0.078 | 1.441±1.234 |
| Security | 0.093±0.029 | 0.259±0.001 | 0.084±0.028 | 0.165±0.070 | 0.225±0.048 |
| Employment | 0.010±0.004 | 0.006±0.002 | 0.012±0.001 | 0.013±0.003 | 0.021±0.003 |
| Traffic | 0.385±0.471 | 0.113±0.063 | 0.047±0.009 | 0.053±0.009 | 0.566±0.731 |
| Win System 1 | NA | 3/8 | 5/8 | 1/8 | 2/8 |

(b) Results of Multimodal Short-term TSF Settings. We use numerical series with textual context series to forecast the next three months.

| Dataset | System 1 | System 1 with Test-time Reasoning Enhancement | | | System 2 |
|--------------|-------------|---|-----------------------|----------------------|-------------|
| | GPT-4o | with CoT | with Self-Consistency | with Self-Correction | o1-mini |
| Agriculture | 0.018±0.015 | 0.018±0.011 | 0.013±0.008 | 0.018±0.006 | 0.045±0.056 |
| Climate | 1.716±0.580 | 1.920±0.505 | 1.712±0.191 | 2.042±0.609 | 1.603±0.496 |
| Economy | 0.569±0.162 | 0.940±0.445 | 0.291±0.127 | 0.503±0.071 | 0.583±0.001 |
| Energy | 0.541±0.457 | 0.316±0.125 | 0.187±0.090 | 0.225±0.080 | 0.718±0.786 |
| Flu | 0.548±0.164 | 1.071±0.643 | 0.288±0.071 | 1.261±1.164 | 0.983±1.177 |
| Security | 0.076±0.052 | 0.110±0.087 | 0.146±0.025 | 0.151±0.035 | 0.244±0.020 |
| Employment | 0.020±0.006 | 0.020±0.003 | 0.019±0.003 | 0.021±0.004 | 0.028±0.008 |
| Traffic | 0.551±0.396 | 1.577±1.421 | 0.030±0.010 | 0.347±0.349 | 0.911±0.594 |
| Win System 1 | NA | 2/8 | 7/8 | 4/8 | 1/8 |

(c) Results of Unimodal Long-term TSF Settings. We use numerical series only to forecast the next six months.

| Dataset | System 1 | System 1 with Test-time Reasoning Enhancement | | | System 2 |
|--------------|-------------|---|-----------------------|----------------------|-------------|
| | GPT-4o | with CoT | with Self-Consistency | with Self-Correction | o1-mini |
| Agriculture | 0.093±0.057 | 0.920±1.134 | 0.057±0.011 | 0.068±0.018 | 0.293±0.089 |
| Climate | 0.754±0.051 | 1.199±0.132 | 0.811±0.081 | 0.877±0.041 | 0.708±0.058 |
| Economy | 0.463±0.146 | 1.040±0.482 | 0.620±0.116 | 0.748±0.069 | 0.359±0.001 |
| Energy | 0.197±0.038 | 0.746±0.500 | 0.177±0.062 | 0.296±0.153 | 0.926±0.771 |
| Flu | 0.219±0.053 | 0.967±0.412 | 0.230±0.077 | 0.639±0.479 | 0.862±0.597 |
| Security | 0.183±0.044 | 0.162±0.038 | 0.135±0.011 | 0.165±0.017 | 0.211±0.075 |
| Employment | 0.011±0.006 | 0.013±0.002 | 0.009±0.003 | 0.013±0.004 | 0.053±0.015 |
| Traffic | 0.066±0.046 | 0.218±0.158 | 0.046±0.016 | 0.036±0.008 | 0.091±0.042 |
| Win System 1 | NA | 1/8 | 5/8 | 3/8 | 2/8 |

(d) Results of Multimodal Long-term TSF Settings. We use numerical series with textual context series to forecast the next six months.

| Dataset | System 1 | System 1 with Test-time Reasoning Enhancement | | | System 2 |
|--------------|-------------|---|-----------------------|----------------------|-------------|
| | GPT-4o | with CoT | with Self-Consistency | with Self-Correction | o1-mini |
| Agriculture | 0.110±0.065 | 0.097±0.044 | 0.063±0.009 | 0.051±0.042 | 0.210±0.022 |
| Climate | 1.365±0.479 | 0.995±0.109 | 1.065±0.014 | 0.912±0.004 | 1.549±0.566 |
| Economy | 0.487±0.237 | 1.027±0.321 | 0.500±0.184 | 0.543±0.074 | 0.827±0.662 |
| Energy | 0.365±0.185 | 0.254±0.122 | 33.743±23.911 | 0.293±0.026 | 0.707±0.499 |
| Flu | 0.291±0.065 | 0.369±0.058 | 0.445±0.210 | 0.529±0.365 | 1.070±0.284 |
| Security | 0.196±0.056 | 0.188±0.027 | 0.140±0.028 | 0.116±0.041 | 0.207±0.001 |
| Employment | 0.015±0.002 | 0.021±0.007 | 0.021±0.002 | 0.106±0.115 | 0.031±0.003 |
| Traffic | 0.207±0.205 | 0.341±0.402 | 0.045±0.013 | 0.377±0.504 | 1.482±1.788 |
| Win System 1 | NA | 4/8 | 4/8 | 4/8 | 0/8 |

Table 2: Results with DeepSeek’s System 1 (DeepSeek-V3) and 2 (DeepSeek-R1) Models. We report the mean MSE and standard deviation over three repeated experiments. Reasoning strategies that outperform the direct System 1 are highlighted in **green**, while those that perform worse or have similar performance (due to higher computational cost) are marked in **red**. In "Win System 1," we present the probability of each reasoning strategy outperforming System 1 across datasets. We observe that **the self-consistency still consistently works**. We find that **DeepSeek-R1 is the only System 2 model that is effective** for TSF, which we attribute to its Group Relative Policy Optimization approach aligning well with the TSF task.

(a) Results of Unimodal Short-term TSF Settings. We use numerical series only to forecast the next three months.

| Dataset | System 1 | System 1 with Test-time Reasoning Enhancement | | | System 2 |
|--------------|-------------|---|-----------------------|----------------------|-------------|
| | DeepSeek-V3 | with CoT | with Self-Consistency | with Self-Correction | DeepSeek-R1 |
| Agriculture | 0.038±0.032 | 0.019±0.001 | 0.046±0.015 | 0.013±0.003 | 0.016±0.010 |
| Climate | 1.216±0.202 | 2.650±0.905 | 1.207±0.197 | 1.246±0.081 | 1.541±0.397 |
| Economy | 0.406±0.218 | 0.433±0.031 | 0.284±0.227 | 0.441±0.161 | 0.583±0.001 |
| Energy | 0.736±0.752 | 0.212±0.022 | 0.187±0.011 | 0.182±0.063 | 0.189±0.021 |
| Flu | 1.464±1.031 | 1.650±0.236 | 0.980±0.445 | 1.682±0.292 | 1.298±1.330 |
| Security | 0.283±0.140 | 0.218±0.093 | 0.185±0.052 | 0.116±0.012 | 0.247±0.017 |
| Employment | 0.036±0.019 | 0.020±0.006 | 0.035±0.019 | 0.018±0.007 | 0.012±0.005 |
| Traffic | 0.066±0.031 | 0.201±0.001 | 0.109±0.028 | 0.107±0.067 | 0.113±0.073 |
| Win System 1 | NA | 4/8 | 6/8 | 4/8 | 5/8 |

(b) Results of Multimodal Short-term TSF Settings. We use numerical series with textual context series to forecast the next three months.

| Dataset | System 1 | System 1 with Test-time Reasoning Enhancement | | | System 2 |
|--------------|-------------|---|-----------------------|----------------------|-------------|
| | DeepSeek-V3 | with CoT | with Self-Consistency | with Self-Correction | DeepSeek-R1 |
| Agriculture | 0.032±0.012 | 0.027±0.006 | 0.023±0.001 | 0.042±0.025 | 2.712±0.001 |
| Climate | 1.428±0.432 | 1.857±0.431 | 1.371±0.001 | 1.411±0.258 | 2.235±0.850 |
| Economy | 0.427±0.174 | 0.598±0.069 | 0.306±0.005 | 0.369±0.128 | 0.615±0.101 |
| Energy | 0.253±0.089 | 0.486±0.318 | 0.197±0.001 | 0.505±0.339 | 0.731±0.777 |
| Flu | 1.073±0.447 | 1.564±0.982 | 0.362±0.161 | 0.441±0.173 | 1.329±1.306 |
| Security | 0.186±0.001 | 0.206±0.010 | 0.187±0.001 | 0.130±0.018 | 0.161±0.051 |
| Employment | 0.016±0.001 | 0.022±0.003 | 0.016±0.001 | 0.016±0.001 | 0.114±0.139 |
| Traffic | 0.201±0.001 | 0.201±0.001 | 0.201±0.001 | 0.114±0.063 | 0.153±0.069 |
| Win System 1 | NA | 1/8 | 5/8 | 5/8 | 2/8 |

(c) Results of Unimodal Long-term TSF Settings. We use numerical series only to forecast the next six months.

| Dataset | System 1 | System 1 with Test-time Reasoning Enhancement | | | System 2 |
|--------------|-------------|---|-----------------------|----------------------|-------------|
| | DeepSeek-V3 | with CoT | with Self-Consistency | with Self-Correction | DeepSeek-R1 |
| Agriculture | 0.216±0.049 | 0.102±0.034 | 0.103±0.014 | 0.121±0.065 | 0.091±0.019 |
| Climate | 0.902±0.001 | 1.383±0.227 | 0.786±0.153 | 0.913±0.078 | 0.662±0.051 |
| Economy | 0.613±0.776 | 0.540±0.386 | 0.393±0.113 | 0.948±0.589 | 0.359±0.001 |
| Energy | 0.603±0.359 | 0.575±0.452 | 0.923±0.265 | 0.332±0.150 | 1.396±0.001 |
| Flu | 0.841±0.215 | 0.658±0.227 | 0.538±0.021 | 0.939±0.328 | 0.972±0.533 |
| Security | 0.275±0.060 | 0.245±0.039 | 0.280±0.004 | 0.186±0.033 | 0.168±0.028 |
| Employment | 0.051±0.013 | 0.021±0.002 | 0.039±0.006 | 0.023±0.003 | 0.021±0.001 |
| Traffic | 0.414±0.001 | 0.209±0.145 | 94.305±66.620 | 0.306±0.153 | 0.158±0.181 |
| Win System 1 | NA | 7/8 | 5/8 | 5/8 | 6/8 |

(d) Results of Multimodal Long-term TSF Settings. We use numerical series with textual context series to forecast the next six months.

| Dataset | System 1 | System 1 with Test-time Reasoning Enhancement | | | System 2 |
|--------------|-------------|---|-----------------------|----------------------|-------------|
| | DeepSeek-V3 | with CoT | with Self-Consistency | with Self-Correction | DeepSeek-R1 |
| Agriculture | 0.088±0.058 | 0.063±0.022 | 0.136±0.080 | 0.119±0.078 | 0.019±0.010 |
| Climate | 0.897±0.001 | 2.193±0.330 | 0.897±0.001 | 0.939±0.074 | 1.849±0.570 |
| Economy | 0.629±0.147 | 0.558±0.282 | 0.486±0.074 | 0.623±0.218 | 0.806±0.354 |
| Energy | 0.995±0.139 | 1.286±0.568 | 0.809±0.241 | 0.493±0.112 | 0.746±0.459 |
| Flu | 2.624±2.400 | 0.974±0.446 | 0.644±0.488 | 1.135±0.643 | 1.560±0.957 |
| Security | 0.179±0.002 | 0.250±0.024 | 0.156±0.027 | 0.274±0.071 | 0.134±0.055 |
| Employment | 0.034±0.001 | 0.029±0.008 | 0.034±0.001 | 0.030±0.005 | 0.105±0.115 |
| Traffic | 0.414±0.001 | 0.414±0.001 | 0.414±0.001 | 0.192±0.157 | 0.152±0.185 |
| Win System 1 | NA | 4/8 | 4/8 | 5/8 | 5/8 |

3.1 RQ1: Can TSF Benefit from Reasoning?

Overall Answer: TSF can benefit from enhanced reasoning ability

We observe that in all four TSF scenarios, at least two reasoning strategies are effective, by outperforming the corresponding System 1 models in over 50% of cases; at least one reasoning strategy is significant, by surpassing the corresponding System 1 model in over 60% of cases. See App. H for an empirical recipe regarding when reasoning tends to improve forecasts and when it tends to fail.

From Short Term vs. Long Term Perspective: Long-term TSF benefits more consistently.

We observe that long-term TSF, in both unimodal and multimodal settings, consistently benefits from all three System 1-based reasoning strategies across datasets and methods. Specifically, the CoT, Self-Consistency, and Self-Correction strategies outperform System 1 models in 52.08%, 58.33%, and 54.17% cases, respectively. In contrast, short-term TSF only consistently benefits from the Self-Consistency strategy. This aligns with TSF, where long-term forecasting requires more consideration of temporal and event influences, while short-term forecasting is more similar to the input.

From Uni-modal vs. Multi-modal Perspective: Multi-modal TSF benefits more significantly.

We observe that multimodal TSF, in both long-term and short-term settings, benefits more significantly from reasoning enhancement. Specifically, the Self-Consistency and Self-Correction strategies outperform System 1 models in 66.67% and 56.25% cases, respectively. In contrast, unimodal TSF only significantly benefits from the Self-Consistency strategy. This aligns with the intuition that multimodal TSF requires more reasoning to use textual context. See App. G for the probabilistic metric analysis and case study to illustrate how zero-shot TSF benefits from textual context.

3.2 RQ2: Need What Reasoning Strategies?

Overall Answer: Self-consistency and hybrid reasoning are usually effective.

We observe that the self-consistency strategy outperforms the System 1 model at a rate of 60% to 80%. Self consistency selects the most coherent reasoning path from various options, which follows the inherent logic of TSF: to consider multiple possible future scenarios and choose the most likely one. Meanwhile, the most advanced hybrid reasoning models, o4-mini and Claude-3.7-Sonnet, require only low reasoning effort settings to consistently outperform pure System 1 models. This validates again that TSF tasks require a combination of fast reaction and rigorous reasoning.

From System 1 vs System 2 Perspective: Reasoning Enhanced System 1 win

System 1 with test-time reasoning enhancement achieves an average effectiveness of 66.67%, which is much higher than the 33.33% of System 2. This suggests that pure System 2 reasoning may not be the correct answer for TSF. In contrast, reasoning-enhanced System 1 is more suitable, as it combines quick responses with slow thinking in line with TSF, which also combines superficial pattern recognition, especially periodicity and trends [8, 37], and deep reasoning, especially event influence [36]. See App. F for failure analysis and case study.

From System 2 Perspective: DeepSeek-R1 is the only effective pure reasoning model.

We observe that DeepSeek-R1 is the only effective pure reasoning model, while the other two, o1-mini and Gemini-2.0-Flash-Thinking, are ineffective. DeepSeek-R1 shows significant improvements in three out of four settings, surpassing the System 1 model (DeepSeek-V3) in 60% cases. We believe this is due to DeepSeek-R1's unique RL solution, i.e., GRPO [19], which focuses solely on outcomes rather than on labeled reasoning paths. Clearly, for TSF, relying on precise reasoning paths to forecast uncertain future numerical series is also not rational.

4 Open-Source Toolkits: Evaluation Suite, Test-Time Scaling Law, and Datasets

We provide three key open-source toolkits to support future research on foundational TSF reasoning models. Specifically, these include our REC4TS as an easy-to-use evaluation suite with experiment logs, a newly verified test-time scaling law on foundation time-series models for zero-shot TSF, and the first TSF dataset with reasoning annotations distilled from six advanced foundation models.

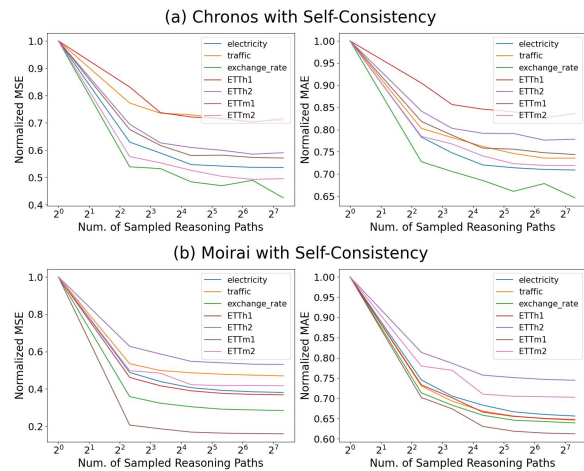


Figure 4: Verified test-time scaling law on foundation time-series models (Chronos and Moirai) inspired by our insights. MSE and MAE are normalized based on one sampled path’s performances. The performances continuously improve as the number of sampled reasoning paths in the self-consistency reasoning strategy increases. We further validate the statistical robustness of the scaling-law trend in App. I.5.

4.1 Released Evaluation Suite.

We fully release REC4TS as an easy-to-use evaluation suite to facilitate future research, including the code, documents, and experiment logs. The released code supports batch experiments, unified selection of reasoning strategy with LLMs, and automated saving and extraction of experimental results. We provide the hyperparameters and model outputs of over 1500 experiments.

4.2 Test-Time Scaling Law Verification.

We further generalize our empirical insights to foundation time-series models. Since the implementation of the GRPO-empowered reasoning model remains in the exploratory stage, we only focus on the self-consistency reasoning strategy. Specifically, we treat the sampling number of probabilistic foundation time-series models, including Moirai [68] and Chronos [2], as the number of sampled reasoning paths in self-consistency, and

we use the median as the most consistent reasoning path. We use multiple well-adopted unimodal time-series datasets [69], and more setups are in Section I. As shown in Figure 4, we clearly demonstrate a new and simple test-time scaling law for TSF. We observe that with an increase in the number of sampled reasoning paths at test time, the performance of both foundation TSF models improves consistently and significantly, reducing the MSE error by at least 20% and up to 50%. We also observe that the model performance gradually converges at about 32 sampled reasoning paths. Our verified scaling law provides promising evidence and improvement room for reasoning-empowered TSF.

4.3 Reasoning-Annotated TSF Datasets

Inspired by recent research in post-training large reasoning models [78, 66, 43], we realized that another obstacle for foundation TSF reasoning models research, aside from evaluation suites and scaling laws, is the lack of a reasoning-annotated dataset. To this end, we curate the first reasoning-annotated TSF dataset, named TIME-THINKING, which pairs TSF task queries and answers with reasoning traces. Specifically, we adopt six advanced and diverse foundation models, including GPT-4o, o1-mini, Gemini-2.0-flash, Gemini-2.0-flash-thinking, DeepSeek-V3, and DeepSeek-R1, and record both the visible final output and the intermediate reasoning chain (if available) for TSF tasks. We set an appropriate temperature for each model and repeat the sampling 15 times, covering all datasets and settings in REC4TS. Finally, we select around two thousand valid samples from a total of about 3K samples. We detail quality control for data selection in App. J including the TSF-aware LLM-as-a-judge rubric [30], the error-based sanity filter, and the 3-run majority-vote protocol. See demos in App. K.

5 Conclusion

In this work, we introduce REC4TS, the first comprehensive benchmark to evaluate the effectiveness of existing reasoning strategies in zero-shot TSF tasks. Through REC4TS’s evaluations, we address two key questions: whether (RQ1) reasoning benefits zero-shot TSF, and what (RQ2) reasoning strategy is most effective. Our findings demonstrate that zero-shot TSF can indeed benefit from reasoning, and self-consistency reasoning provides significant advantage for both foundation language models and foundation time-series models.

References

- [1] Taha Aksu, Gerald Woo, Juncheng Liu, Xu Liu, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. 2024. Gift-eval: A benchmark for general time series forecasting model evaluation. *arXiv preprint arXiv:2410.10393*.
- [2] Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, and 1 others. 2024. Chronos: Learning the language of time series.
- [3] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.
- [4] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.
- [5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- [6] Defu Cao, Furong Jia, Sercan O Arik, Tomas Pfister, Yixiang Zheng, Wen Ye, and Yan Liu. Tempo: Prompt-based generative pre-trained transformer for time series forecasting. In *The Twelfth International Conference on Learning Representations*.
- [7] Yonghong Chen, Govindan Rangarajan, Jianfeng Feng, and Mingzhou Ding. 2004. Analyzing multiple nonlinear time series with extended granger causality. *Physics letters A*, 324(1):26–35.
- [8] Robert B Cleveland, William S Cleveland, Jean E McRae, Irma Terpenning, and 1 others. 1990. Stl: A seasonal-trend decomposition. *J. off. Stat.*, 6(1):3–73.
- [9] Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. 2024. A decoder-only foundation model for time-series forecasting. In *Forty-first International Conference on Machine Learning*.
- [10] Wenjie Du, Jun Wang, Linglong Qian, Yiyuan Yang, Zina Ibrahim, Fanxing Liu, Zepu Wang, Haoxin Liu, Zhiyuan Zhao, Yingjie Zhou, and 1 others. 2024. Tsi-bench: Benchmarking time series imputation. *arXiv preprint arXiv:2406.12747*.
- [11] Michael Eichler and Vanessa Didelez. 2012. Causal reasoning in graphical time series models. *arXiv preprint arXiv:1206.5246*.
- [12] Vijay Ekambaram, Arindam Jati, Pankaj Dayama, Sumanta Mukherjee, Nam Nguyen, Wesley M Gifford, Chandra Reddy, and Jayant Kalagnanam. 2025. Tiny time mixers (ttms): Fast pre-trained models for enhanced zero/few-shot forecasting of multivariate time series. *Advances in Neural Information Processing Systems*, 37:74147–74181.
- [13] Friedrich Gebhardt, Angi Voß, Wolfgang Gräther, and Barbara Schmidt-Belz. 1997. *Reasoning with complex cases*. Springer.
- [14] Ethan Goh, Robert Gallo, Jason Hom, Eric Strong, Yingjie Weng, Hannah Kerman, Joséphine A Cool, Zahir Kanjee, Andrew S Parsons, Neera Ahuja, and 1 others. 2024. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA Network Open*, 7(10):e2440969–e2440969.
- [15] 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*.
- [16] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819.
- [17] Clive WJ Granger. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438.
- [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*, 36:19622–19635.
- [19] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shitong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- [20] Taicheng Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh Chawla, Olaf Wiest, Xiangliang Zhang, and 1 others. 2023. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. *Advances in Neural Information Processing Systems*, 36:59662–59688.
- [21] Shibo Hao, Yi Gu, Haotian Luo, Tianyang Liu, Xiyuan Shao, Xinyuan Wang, Shuhua Xie, Haodi Ma, Adithya Samavedhi, Qiyue Gao, and 1 others. 2024. Llm reasoners: New evaluation, library, and analysis of step-by-step reasoning with large language models. *arXiv preprint arXiv:2404.05221*.
- [22] Shi Bin Hoo, Samuel Müller, David Salinas, and Frank Hutter. 2025. The tabular foundation model tabpfn outperforms specialized time series forecasting models based on simple features. *arXiv preprint arXiv:2501.02945*.
- [23] Yuxiao Hu, Qian Li, Jinyue Yan, Dongxiao Zhang, and Yuntian Chen. 2025. [Context-alignment: Activating and enhancing LLMs capabilities in time](#)

| | | | |
|-----|--|---|-----|
| 515 | series. In <i>The Thirteenth International Conference on Learning Representations</i> . | [35] Haoxin Liu, Chenghao Liu, and B Aditya Prakash. 2024. A picture is worth a thousand numbers: Enabling llms reason about time series via visualization. <i>arXiv preprint arXiv:2411.06018</i> . | 569 |
| 516 | | | 570 |
| 517 | [24] Haolin Jin, Linghan Huang, Haipeng Cai, Jun Yan, Bo Li, and Huaming Chen. 2024. From llms to llm-based agents for software engineering: A survey of current, challenges and future. <i>arXiv preprint arXiv:2408.02479</i> . | [36] Haoxin Liu, Shangqing Xu, Zhiyuan Zhao, Lingkai Kong, Harshavardhan Prabhakar Kamarthi, Aditya Sasanur, Megha Sharma, Jiaming Cui, Qingsong Wen, Chao Zhang, and 1 others. 2025. Time-mmd: Multi-domain multimodal dataset for time series analysis. <i>Advances in Neural Information Processing Systems</i> , 37:77888–77933. | 571 |
| 518 | | | 572 |
| 519 | | | 573 |
| 520 | | | 574 |
| 521 | | | 575 |
| 522 | [25] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and 1 others. Time-llm: Time series forecasting by reprogramming large language models. In <i>The Twelfth International Conference on Learning Representations</i> . | [37] 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 <i>Findings of the Association for Computational Linguistics ACL 2024</i> , pages 7832–7840. | 576 |
| 523 | | | 577 |
| 524 | | | 578 |
| 525 | | | 579 |
| 526 | | | 580 |
| 527 | | | 581 |
| 528 | [26] Daniel Kahneman. 2011. <i>Thinking, fast and slow</i> . macmillan. | [38] Xu Liu, Juncheng Liu, Gerald Woo, Taha Aksu, Yuxuan Liang, Roger Zimmermann, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. 2024. Moirai-moe: Empowering time series foundation models with sparse mixture of experts. <i>arXiv preprint arXiv:2410.10469</i> . | 582 |
| 529 | | | 583 |
| 530 | | | 584 |
| 531 | [27] Harshavardhan Kamarthi and B Aditya Prakash. 2023. Large pre-trained time series models for cross-domain time series analysis tasks. <i>arXiv preprint arXiv:2311.11413</i> . | [39] 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. <i>arXiv preprint arXiv:2310.06625</i> . | 585 |
| 532 | | | 586 |
| 533 | | | 587 |
| 534 | [28] Aayush Karan and Yilun Du. 2025. Reasoning with sampling: Your base model is smarter than you think. <i>arXiv preprint arXiv:2510.14901</i> . | [40] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhunoye, Yiming Yang, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. <i>Advances in Neural Information Processing Systems</i> , 36:46534–46594. | 588 |
| 535 | | | 589 |
| 536 | | | 590 |
| 537 | [29] Tomi Kauppinen and Eero Hyvönen. 2007. Modeling and reasoning about changes in ontology time series. In <i>Ontologies: A handbook of principles, concepts and applications in information systems</i> , pages 319–338. Springer. | [41] Sarabeth M Mathis, Alexander E Webber, Tomás M León, Erin L Murray, Monica Sun, Lauren A White, Logan C Brooks, Alden Green, Addison J Hu, Roni Rosenfeld, and 1 others. 2024. Evaluation of flusight influenza forecasting in the 2021–22 and 2022–23 seasons with a new target laboratory-confirmed influenza hospitalizations. <i>Nature communications</i> , 15(1):6289. | 591 |
| 538 | | | 592 |
| 539 | | | 593 |
| 540 | | | 594 |
| 541 | | | 595 |
| 542 | [30] Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. In <i>Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing</i> , pages 4334–4353. | [42] Raha Moraffah, Paras Sheth, Mansooreh Karami, Anchit Bhattacharya, Qianru Wang, Anique Tahir, Adrienne Raglin, and Huan Liu. 2021. Causal inference for time series analysis: Problems, methods and evaluation. <i>Knowledge and Information Systems</i> , 63:3041–3085. | 596 |
| 543 | | | 597 |
| 544 | | | 598 |
| 545 | | | 599 |
| 546 | | | 600 |
| 547 | | | 601 |
| 548 | | | 602 |
| 549 | | | 603 |
| 550 | [31] Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, and 1 others. 2024. Training language models to self-correct via reinforcement learning. <i>arXiv preprint arXiv:2409.12917</i> . | [43] 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. <i>arXiv preprint arXiv:2501.19393</i> . | 604 |
| 551 | | | 605 |
| 552 | | | 606 |
| 553 | | | 607 |
| 554 | | | 608 |
| 555 | | | 609 |
| 556 | [32] Minae Kwon, Sang Michael Xie, Kalesha Bullard, and Dorsa Sadigh. 2023. Reward design with language models. <i>arXiv preprint arXiv:2303.00001</i> . | [44] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. 2023. A time series is worth 64 words: Long-term forecasting with transformers. In <i>International Conference on Learning Representations</i> . | 610 |
| 556 | | | 611 |
| 557 | | | 612 |
| 558 | | | 613 |
| 559 | [33] Minhua Lin, Zhengzhang Chen, Yanchi Liu, Xujiang Zhao, Zongyu Wu, Junxiang Wang, Xiang Zhang, Suhang Wang, and Haifeng Chen. 2024. Decoding time series with llms: A multi-agent framework for cross-domain annotation. <i>arXiv preprint arXiv:2410.17462</i> . | | 614 |
| 560 | | | 615 |
| 561 | | | 616 |
| 562 | | | 617 |
| 563 | | | 618 |
| 564 | | | 619 |
| 565 | [34] Zicheng Lin, Zhibin Gou, Tian Liang, Ruilin Luo, Haowei Liu, and Yujiu Yang. 2024. Criticbench: Benchmarking llms for critique-correct reasoning. <i>arXiv preprint arXiv:2402.14809</i> . | | 620 |
| 566 | | | 621 |
| 567 | | | 622 |
| 568 | | | 623 |

| | | |
|-----|---|-----|
| 627 | [45] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In <i>The Eleventh International Conference on Learning Representations</i> . | 683 |
| 628 | | 684 |
| 629 | | 685 |
| 630 | | 686 |
| 631 | | 687 |
| 632 | [46] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. <i>Advances in neural information processing systems</i> , 35:27730–27744. | 688 |
| 633 | | 689 |
| 634 | | 690 |
| 635 | | 691 |
| 636 | | 692 |
| 637 | | 693 |
| 638 | | 694 |
| 639 | [47] Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. 2023. Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies. <i>arXiv preprint arXiv:2308.03188</i> . | 695 |
| 640 | | 696 |
| 641 | | 697 |
| 642 | | 698 |
| 643 | | 699 |
| 644 | [48] Aske Plaat, Annie Wong, Suzan Verberne, Joost Broekens, Niki van Stein, and Thomas Back. 2024. Reasoning with large language models, a survey. <i>arXiv preprint arXiv:2407.11511</i> . | 700 |
| 645 | | 701 |
| 646 | | 702 |
| 647 | | 703 |
| 648 | [49] Alexander Rodríguez, Harshavardhan Kamarthi, Pulak Agarwal, Javen Ho, Mira Patel, Suchet Sapre, and B Aditya Prakash. 2024. Machine learning for data-centric epidemic forecasting. <i>Nature Machine Intelligence</i> , 6(10):1122–1131. | 704 |
| 649 | | 705 |
| 650 | | 706 |
| 651 | | 707 |
| 652 | | 708 |
| 653 | [50] Jakob Runge, Andreas Gerhardus, Gherardo Varando, Veronika Eyring, and Gustau Camps-Valls. 2023. Causal inference for time series. <i>Nature Reviews Earth & Environment</i> , 4(7):487–505. | 709 |
| 654 | | 710 |
| 655 | | 711 |
| 656 | | 712 |
| 657 | [51] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. <i>arXiv preprint arXiv:1707.06347</i> . | 713 |
| 658 | | 714 |
| 659 | | 715 |
| 660 | | 716 |
| 661 | [52] Omer Berat Sezer, Mehmet Ugur Gudelek, and Ahmet Murat Ozbayoglu. 2020. Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. <i>Applied soft computing</i> , 90:106181. | 717 |
| 662 | | 718 |
| 663 | | 719 |
| 664 | | 720 |
| 665 | | 721 |
| 666 | [53] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024. Deepseek-math: Pushing the limits of mathematical reasoning in open language models. <i>arXiv preprint arXiv:2402.03300</i> . | 722 |
| 667 | | 723 |
| 668 | | 724 |
| 669 | | 725 |
| 670 | | 726 |
| 671 | | 727 |
| 672 | [54] Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. 2024. Time-moe: Billion-scale time series foundation models with mixture of experts. <i>arXiv preprint arXiv:2409.16040</i> . | 728 |
| 673 | | 729 |
| 674 | | 730 |
| 675 | | 731 |
| 676 | | 732 |
| 677 | [55] Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. 2023. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In <i>Proceedings of the IEEE/CVF international conference on computer vision</i> , pages 2998–3009. | 733 |
| 678 | | 734 |
| 679 | | 735 |
| 680 | | 736 |
| 681 | | 737 |
| 682 | | 738 |
| | [56] Haotian Sun, Yuchen Zhuang, Lingkai Kong, Bo Dai, and Chao Zhang. 2023. Adaplaner: Adaptive planning from feedback with language models. <i>Advances in neural information processing systems</i> , 36:58202–58245. | 739 |
| | | 740 |
| | [57] Youqiang Sun, Jiuyong Li, Jixue Liu, Christopher Chow, Bingyu Sun, and Rujing Wang. 2015. Using causal discovery for feature selection in multivariate numerical time series. <i>Machine Learning</i> , 101:377–395. | 741 |
| | | 742 |
| | [58] Anika Tabassum, Supriya Chinthavali, Varisara Tansakul, and B Aditya Prakash. 2021. Actionable insights in urban multivariate time-series. In <i>Proceedings of the 30th ACM International Conference on Information & Knowledge Management</i> , pages 1774–1783. | 743 |
| | | 744 |
| | [59] Mingtian Tan, Mike Merrill, Vinayak Gupta, Tim Althoff, and Tom Hartvigsen. 2024. Are language models actually useful for time series forecasting? <i>Advances in Neural Information Processing Systems</i> , 37:60162–60191. | 745 |
| | | 746 |
| | [60] Amir Taubenfeld, Tom Sheffer, Eran Ofek, Amir Feder, Ariel Goldstein, Zorik Gekhman, and Gal Yona. 2025. Confidence improves self-consistency in llms. <i>arXiv preprint arXiv:2502.06233</i> . | 747 |
| | | 748 |
| | [61] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. <i>Advances in neural information processing systems</i> , 30. | 749 |
| | | 750 |
| | [62] Jiahao Wang, Mingyue Cheng, Qingyang Mao, Qi Liu, Feiyang Xu, Xin Li, and Enhong Chen. 2024. Tabletime: Reformulating time series classification as zero-shot table understanding via large language models. <i>arXiv preprint arXiv:2411.15737</i> . | 751 |
| | | 752 |
| | [63] Xinglin Wang, Shaoxiong Feng, Yiwei Li, Peiwen Yuan, Yueqi Zhang, Chuyi Tan, Boyuan Pan, Yao Hu, and Kan Li. 2025. Make every penny count: Difficulty-adaptive self-consistency for cost-efficient reasoning. In <i>Findings of the Association for Computational Linguistics: NAACL 2025</i> , pages 6904–6917. | 753 |
| | | 754 |
| | [64] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. <i>arXiv preprint arXiv:2203.11171</i> . | 755 |
| | | 756 |
| | [65] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In <i>The Eleventh International Conference on Learning Representations</i> . | 757 |
| | | 758 |
| | [66] Yuxuan Wang, Haixu Wu, Jiayang Dong, Yong Liu, Mingsheng Long, and Jianmin Wang. 2024. Deep time series models: A comprehensive survey and benchmark. <i>arXiv preprint arXiv:2407.13278</i> . | 759 |
| | | 760 |

- 740 [67] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny
741 Zhou, and 1 others. 2022. Chain-of-thought prompt-
742 ing elicits reasoning in large language models. *Ad-
743 vances in neural information processing systems*,
744 35:24824–24837.
- 746 [68] Gerald Woo, Chenghao Liu, Akshat Kumar, Caim-
747 ing Xiong, Silvio Savarese, and Doyen Sahoo. 2024.
748 [Unified training of universal time series forecasting
749 transformers](#). *Preprint*, arXiv:2402.02592.
- 750 [69] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jian-
751 min Wang, and Mingsheng Long. 2022. Timesnet:
752 Temporal 2d-variation modeling for general time se-
753 ries analysis. In *The Eleventh International Confer-
754 ence on Learning Representations*.
- 755 [70] Haixu Wu, Jiehui Xu, Jianmin Wang, and Ming-
756 sheng Long. 2021. Autoformer: Decomposition
757 transformers with auto-correlation for long-term se-
758 ries forecasting. *Advances in neural information
759 processing systems*, 34:22419–22430.
- 760 [71] Sascha Wüstenberg, Samuel Greiff, and Joachim
761 Funke. 2012. Complex problem solving—more than
762 reasoning? *Intelligence*, 40(1):1–14.
- 763 [72] Fengli Xu, Qian Yue Hao, Zefang Zong, Jingwei
764 Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan,
765 Jiahui Gong, Tianjian Ouyang, Fanjin Meng, and 1
766 others. 2025. Towards large reasoning models: A
767 survey of reinforced reasoning with large language
768 models. *arXiv preprint arXiv:2501.09686*.
- 769 [73] Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen,
770 Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao,
771 and Tianyi Zhou. 2024. A survey on knowledge
772 distillation of large language models. *arXiv preprint
773 arXiv:2402.13116*.
- 774 [74] Hao Xue and Flora D Salim. 2023. Promptcast: A
775 new prompt-based learning paradigm for time series
776 forecasting. *IEEE Transactions on Knowledge and
777 Data Engineering*, 36(11):6851–6864.
- 778 [75] Yibo Yan, Shen Wang, Jiahao Huo, Jingheng Ye,
779 Zhendong Chu, Xuming Hu, Philip S Yu, Carla
780 Gomes, Bart Selman, and Qingsong Wen. 2025. Po-
781 sition: Multimodal large language models can signif-
782 icantly advance scientific reasoning. *arXiv preprint
783 arXiv:2502.02871*.
- 784 [76] Daoan Zhang, Junming Yang, Hanjia Lyu, Zijian
785 Jin, Yuan Yao, Mingkai Chen, and Jiebo Luo. 2025.
786 A benchmark and chain-of-thought prompting strat-
787 egy for large multimodal models with multiple im-
788 age inputs. In *International Conference on Pattern
789 Recognition*, pages 226–241. Springer.
- 790 [77] Siru Zhong, Weilin Ruan, Ming Jin, Huan Li, Qing-
791 song Wen, and Yuxuan Liang. Time-vlm: Exploring
792 multimodal vision-language models for augmented
793 time series forecasting. In *Forty-second Interna-
794 tional Conference on Machine Learning*.
- [78] Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan
Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat,
Ping Yu, Lili Yu, and 1 others. 2023. Lima: Less is
more for alignment. *Advances in Neural Information
Processing Systems*, 36:55006–55021.
- [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 *Proceed-
ings of the AAAI conference on artificial intelligence*,
volume 35, pages 11106–11115.
- [80] Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang,
Liang Sun, and Rong Jin. 2022. Fedformer: Fre-
quency enhanced decomposed transformer for long-
term series forecasting. In *International conference
on machine learning*, pages 27268–27286. PMLR.
- [81] Zhehua Zhou, Jiayang Song, Kunpeng Yao, Zhan
Shu, and Lei Ma. 2024. Isr-llm: Iterative self-refined
large language model for long-horizon sequential task
planning. In *2024 IEEE International Conference on
Robotics and Automation (ICRA)*, pages 2081–2088.
IEEE.

Appendix

A Related Work

A.1 (Zero-Shot) Time-Series Forecasting.

Time-series forecasting (TSF) is a fundamental task in time-series analysis with broad real-world applications, including economics, urban computing, and epidemiology [52, 58, 49]. The introduction of transformer-based architectures [61] has significantly advanced TSF performance, setting new benchmarks in forecasting accuracy [79, 80, 44, 39]. Building upon these advancements, recent research has explored foundation TSF models, which achieve competitive zero-shot forecasting performance comparable to supervised TSF models [9, 2, 15, 54, 12, 27]. Trained on billions of data samples, these foundation models can generate accurate forecasts across diverse time series without requiring additional post-training or fine-tuning.

A.2 Foundation Model Reasoning.

With the rise of foundation models, reasoning has become a critical research direction, enabling models to move beyond surface-level pattern recognition toward structured decision-making [67, 5]. Reasoning in foundation models primarily follows two paradigms: reasoning-enhanced System 1, which improves inference during test time for fast and intuitive thinking, such as Chain-of-Thought reasoning [67, 64, 47, 31], and System 2, which enables deep, analytical reasoning as an inherent capability of the model. System 2 reasoning is typically developed through post-training techniques such as reinforcement learning from human feedback (RLHF), reward shaping, knowledge distillation, and the more recent group-relative policy optimization (GRPO) in DeepSeek-R1 [4, 32, 16, 73, 19], which allows foundation models to process more structured, deeper, and analytical thinking. Leveraging these advanced reasoning strategies, foundation models have demonstrated strong capabilities in handling complex tasks, including mathematical problem-solving, automated planning, and applications in engineering and science [3, 81, 56, 55, 20, 14].

Most recently, hybrid reasoning models are released to let users adjust the depth of their “thinking” to match task demands. OpenAI’s o4-mini exposes three effort tiers—low, medium, and high—that control the number of internal reasoning passes and tool invocations, so developers can

choose fast, cost-efficient responses or deeper analytical processing for harder problems. Anthropic’s Claude 3.7 Sonnet offers an “extended thinking” mode where users set a custom thinking budget via token limits or a simple API toggle, trading off response speed against the detail level of its chain-of-thought explanations. Google DeepMind’s Gemini 2.5 Pro and its Flash variant embed a fully hybrid pipeline that can be switched on or off on demand. In the Flash preview, a single switch toggles internal reasoning, while in Pro the reasoning pipeline can be dynamically throttled to hit specific latency or cost targets without external ensembling. This adjustable reasoning effort applies across text, code, image, and video tasks, giving practitioners precise control over speed, expense, and output fidelity.

A.3 Time-Series Reasoning.

Reasoning over time-series data is an important topic in time-series analysis [29]. Conventional time-series reasoning primarily focuses on causal analysis, including causal discovery, feature selection, and graph-based analysis [11, 57, 50, 7]. However, these conventional methods focus only on numerical modalities, making validation challenging and often limiting their application to synthetic or simplified scenarios. Recent advancements have broadened time-series reasoning beyond numerical data, incorporating multimodal perspectives. Notably, emerging research has explored reasoning over time-series data using LLMs [37, 75] and vision-based approaches [35], expanding its scope and applicability. While these studies show promising potential for time series reasoning with advanced foundation models, they do not explicitly establish a connection between reasoning strategies and TSF tasks.

Notably, reasoning techniques in language tasks have formed clear development paths, i.e., from System 1 to reasoning-enhanced System 1 to System 2 to hybrid reasoning, offering significant benefits (e.g., performance, generalization, and interpretability). However, these techniques remain largely unexplored for time-series forecasting tasks, which this work aims to explore.

Importantly, existing studies and benchmarks focus on either (zero-shot) TSF [1, 66, 10] or reasoning strategies [34, 21, 76], no existing work systematically evaluates how different reasoning strategies impact zero-shot TSF. Our study bridges this gap and introduces a multimodal reasoning enhancement setup beyond conventional unimodal

| | | | |
|-----|---|--|------|
| 917 | approaches. This multimodal setup incorporates | at an earlier stage. Many TSFM design choices are | 963 |
| 918 | textual context alongside numerical data, provid- | inspired by the LLM literature, such as decoder- | 964 |
| 919 | ing more comprehensive information for real-world | only architectures [9] and mixture-of-experts de- | 965 |
| 920 | forecasting, and setting a higher standardization for | signs [38]. Given the current focus on test-time | 966 |
| 921 | reasoning ability in zero-shot TSF tasks. | reasoning in the LLM community, it is natural to | 967 |
| | | ask whether similar test-time reasoning strategies | 968 |
| 922 | B Clarification of the research motivation | help zero-shot TSF, and what failure modes may | 969 |
| | | arise. | 970 |
| 923 | B.1 The value of studying zero-shot TSF | | 971 |
| 924 | Zero-shot TSF is a widely adopted evaluation goal | Natural support for multimodal inputs and in- | 972 |
| 925 | in the time-series foundation model community, | terpretable outputs. LLMs can process both nat- | 973 |
| 926 | with recent benchmarks such as GIFT-Eval [1] ³ . | ural language and textified numerical series, which | 974 |
| 927 | From an application perspective, many organiza- | allows them to combine aligned textual context | 975 |
| 928 | tions maintain large collections of heterogeneous | (such as news, reports, and logs) with time series | 976 |
| 929 | time series (including long-tail or newly created | and to output both numerical forecasts and textual | 977 |
| 930 | metrics); training or tuning a separate model for | explanations. In contrast, existing TSF models of- | 978 |
| 931 | each series can be costly, while an “out-of-the-box” | ten require additional components to support such | 979 |
| 932 | zero-shot forecast can be useful for cold-start sce- | multimodal integration and explanation. | 980 |
| 933 | narios, rapid prototyping, and prioritizing which | | 981 |
| 934 | series deserve further investment. We emphasize | Positioning. We position this work as a system- | 982 |
| 935 | that zero-shot performance is not meant to replace | atic evaluation of test-time reasoning strategies for | 983 |
| 936 | fine-tuning. Instead, it is a base capability that | uni-modal and multi-modal zero-shot TSF using | 984 |
| 937 | (i) directly supports the above practical scenarios | widely available LLMs, aiming to provide guid- | 985 |
| 938 | and (ii) provides a stronger starting point for subse- | ance for future time-series foundation model de- | 986 |
| 939 | quent adaptation. Recent foundation TSF models | sign, rather than to argue that LLMs dominate de- | 987 |
| 940 | also report strong out-of-distribution generaliza- | icated time-series architectures. | 988 |
| 941 | tion in several settings [2, 38], which suggests that | | |
| 942 | broad pretraining can be competitive under certain | C Self-Consistency vs. Ensembling and | 989 |
| 943 | shifts. | Uncertainty Sampling | 990 |
| 944 | B.2 Why use LLMs? | | 991 |
| 945 | The effectiveness of using LLMs for TSF remains | What we mean by self-consistency. In this work, | 992 |
| 946 | a debated question [59, 77]. Our goal is not to | we follow the classic self-consistency recipe for | 993 |
| 947 | prove that “LLMs are necessarily better than ded- | LLMs [65]: for the same model and the same | 994 |
| 948 | icated time-series models”, but to systematically | prompt, we use stochastic decoding to sample | 995 |
| 949 | answer two research questions given off-the-shelf | K outputs, discard intermediate reasoning text, | 996 |
| 950 | zero-shot LLMs stated in Section 1: RQ1 (whether | and aggregate only the final answers (for exam- | 997 |
| 951 | zero-shot TSF benefits from enhanced reasoning | ple, via majority vote). Our use of the term “self- | 998 |
| 952 | ability) and RQ2 (what kinds of reasoning strate- | consistency” refers to <i>same-model</i> , <i>same-input</i> , | 999 |
| 953 | gies are needed). | <i>multi-path sampling with a stability-based selec-</i> | 1000 |
| 954 | | <i>tion rule</i> , rather than to training-time ensembling. | 1001 |
| 955 | Reusing LLM abilities as an exploration tool. | | 1002 |
| 956 | We use LLMs because current time-series founda- | How this relates to probabilistic TSF sampling. | 1003 |
| 957 | tion models are still at an early stage: they generally | For probabilistic foundation TSF models (for ex- | 1004 |
| 958 | lack both reasoning ability and multimodal capabil- | ample, Chronos and Moirai), sampling multiple | 1005 |
| 959 | ity. LLMs have received substantial research and | predictive trajectories is standard. Our contribu- | 1006 |
| 960 | computational investment and show strong general | tion is not a new sampling algorithm; instead, we | 1007 |
| 961 | abilities related to planning and structured reason- | place TSF sampling under the same <i>path-based</i> lens as | 1008 |
| 962 | ing, while time-series foundation models are still | LLM self-consistency by treating sampled trajec- | 1009 |
| | | tories as candidate paths. When we need a single | 1010 |

³<https://huggingface.co/spaces/Salesforce/GIFT-Eval>

Table 3: Self-consistency (same model, multi-path) versus classical ensembling and uncertainty-based sampling.

| Method family | Source of diversity | Aggregation target | Typical goal / interpretation |
|------------------------------|---|---|--|
| Classical ensembling | Multiple models (different parameters, seeds, or training data) | Average / vote over model outputs | Reduce variance and improve stability by averaging over parameter uncertainty |
| Uncertainty-based sampling | Multiple samples from a predictive distribution (same model) | Use sample spread (variance/entropy/intervals), or take expectation | Estimate predictive uncertainty; the distribution itself is the object of interest |
| Self-consistency (this work) | Multiple stochastic decoding paths (LLMs) or sampled trajectories (TSF) from the same model | Select a stable answer/path (majority vote for discrete answers; median for numeric trajectories) | Search for a stable solution among candidate paths; we use aggregation as a test-time selection rule rather than as uncertainty estimation |

point forecast for accuracy evaluation, we use the *median* across sampled paths as a simple stability-oriented choice: it selects a central tendency that is robust to occasional extreme sampled trajectories, and it is not used as an uncertainty estimator.

Conceptual distinction. Table 3 further contrasts self-consistency with classical ensembling and uncertainty-based sampling along the axes (source of diversity, aggregation target, and intended use).

What is novel in this paper (and what is not). We do not claim a new self-consistency or sampling algorithm. The methodological novelty is a unified test-time *path/budget* perspective that aligns (i) LLM reasoning paths and (ii) TSF predictive paths along the same axis of “number of sampled paths / test-time budget”. This unified view lets us (a) describe LLM self-consistency and TSF path sampling using the same abstraction, and (b) validate a simple test-time scaling-law pattern for TSF models (Figure 4) under comparable budget sweeps. We include a short discussion of recent self-consistency developments in the LLM literature [65, 28, 63, 60] to clarify that our emphasis is on evaluation and unification rather than on algorithmic novelty.

D Limitations and Future Work

Since reasoning strategies for foundational time-series models have not yet been studied and are difficult to implement directly, we have to reuse foundational language models to explore effective TSF reasoning strategies. We envision that our benchmark and insights offer promising potential for future research on understanding and designing effective reasoning strategies for zero-shot TSF.

First, our results show that only limited existing System 2 approaches benefit zero-shot TSF. This can be interpreted as a distinction in reasoning logic between time-series analysis or zero-shot TSF, and other types of reasoning that are inherently learned by most System 2 models. This finding encourages future studies to design reasoning strategies more tailored to time-series data and zero-shot TSF tasks. In this work, we adopt a fixed history window of 96 time steps, following common conventions in time-series evaluation [70, 69, 45]. One future work is to study how history length interacts with reasoning strategies by stratifying series into short/medium/long-history regimes and comparing trends within each group.

Second, in validating the effectiveness of reasoning in zero-shot TSF, our open-source toolkit—comprising both REC4TS and TIME-THINKING—provides guidelines for future research. Specifically, we validate the scaling law of zero-shot TSF by incorporating reasoning within REC4TS, paving the way for future work on large-scale foundation TSF models with reasoning capabilities.

Additionally, we provide TIME-THINKING, the first TSF dataset with reasoning annotations, enabling deeper investigation and understanding of reasoning strategies in zero-shot TSF. A natural next step of TIME-THINKING is to add a human TSF expert audit, but it requires a well-defined screening rubric (to avoid rewarding fluent yet statistically naive reasoning) and multi-annotator cross-validation with agreement checks, which we leave as future work.

Another extension of this work is to study few-shot forecasting. Such a study requires a carefully controlled protocol to standardize how demonstrations are selected for each dataset and horizon, how

1085 example forecasts and reasoning traces are format-
 1086 ted, how to choose the adaptation solutions, and
 1087 how the supervision budget is controlled across in-
 1088 context demonstrations versus lightweight adapta-
 1089 tion. We view our REC4TS and TIME-THINKING
 1090 as a convenient starting point for this direction.

1091 E More Details of REC4TS

1092 E.1 Details of Datasets

1093 The datasets utilized in the REC4TS benchmark
 1094 are primarily sourced from the **Time-MMD** col-
 1095 lection [36]. Time-MMD is a multi-domain, mul-
 1096 timodal time series dataset designed to facilitate
 1097 comprehensive time series analysis by integrating
 1098 numerical series with aligned textual context. For
 1099 the purposes of REC4TS, we leverage its structure
 1100 which provides numerical time series data across
 1101 eight diverse domains along with corresponding
 1102 textual information intended to enrich the context
 1103 for forecasting tasks.

1104 The aligned textual context series in REC4TS
 1105 are specifically derived from the “Web Search Re-
 1106 sults” component of Time-MMD [36]. The textual
 1107 data is constructed by first designing 2-3 highly re-
 1108 levant keywords for each numerical target variable.
 1109 Subsequently, the Google API is used to perform
 1110 weekly web searches for these keywords, collecting
 1111 the top results (typically titles and content snippets)
 1112 from 1980 to the present. This raw textual informa-
 1113 tion then undergoes a crucial preprocessing stage
 1114 using Large Language Models (LLMs, specifically
 1115 Llama3-70B as per the Time-MMD paper). This
 1116 LLM-based preprocessing includes filtering for rel-
 1117 evance to the numerical series, disentangling fac-
 1118 tual statements from predictions or opinions, and
 1119 summarizing the content to enhance usability and
 1120 ensure that the textual data provides concise, rele-
 1121 vant context aligned with the numerical time series.
 1122 The final textual series used in REC4TS thus con-
 1123 sists of these LLM-processed, date-aligned web
 1124 summaries, aimed at providing exogenous contex-
 1125 tual signals for the zero-shot time series forecasting
 1126 tasks. We provide details of datasets in Table 7 and
 1127 visualization in Figure 5.

E.2 Details of Reasoning Strategies Implementation 1128 1129

E.2.1 Implementation of reusing LLMs for unimodal forecasting 1130 1131

Prompts used for unimodal time-series forecasting
 1132 are detailed in Figure 8 1133

E.2.2 Implementation of reusing LLMs for multimodal forecasting 1134 1135

Prompts used for multimodal time-series forecast-
 1136 ing are detailed in Figure 7 1137

E.2.3 Implementation of chain-of-thought reasoning 1138 1139

Prompts used for chain-of-thought reasoning are
 1140 detailed in Figure 9 1141

E.2.4 Implementation of self-consistency reasoning 1142 1143

We set the default number of sampled reasoning
 1144 paths in self-consistency to 3, using the median
 1145 as the most consistent reasoning path. Prompts
 1146 used for self-consistency reasoning are detailed in
 1147 Figure 10 1148

E.2.5 Implementation of self-correction reasoning 1149 1150

We set the default number of iterations for self-
 1151 correction to 2. Prompts used for self-correction
 1152 reasoning are detailed in Figure 11 1153

E.3 Details of LLMs 1154

E.3.1 OpenAI Model Series 1155

We select GPT-4o and o1-mini as representatives
 1156 of System 1 and System2 models from OpenAI.
 1157 Released on Mar 13 2024, GPT-4o is an autoregres-
 1158 sive multimodal model which integrates all inputs
 1159 and outputs into a unified neural network, offering
 1160 faster and more cost-effective performance than its
 1161 predecessors.⁴ In comparison, o1-mini is trained
 1162 with large-scale reinforcement learning to reason,
 1163 and it shows better reasoning performance over
 1164 GPT-4o.⁵ 1165

E.3.2 Google Model Series 1166

Taken from the latest Gemini models, Gemini 2.0
 1167 Flash and Gemini 2.0 Flash Thinking compose a
 1168 pair of System 1 and 2 models. Based on the non-
 1169 thinking version, Gemini 2.0 Flash Thinking is an
 1170

⁴<https://openai.com/index/gpt-4o-system-card/>

⁵<https://openai.com/index/openai-o1-mini-advancing-cost-efficient-reasoning/>

Table 4: Implementation Details at a Glance (pointers to prompts, hyperparameters, and settings).

| Item | Where | Key setting / note |
|---|--------------------------|--|
| Unimodal TSF prompt | Figure 8 | Time-series textualization template |
| Multimodal TSF prompt | Figure 7 | Numerical series + aligned textual context |
| CoT prompting | Figure 9 | Step-by-step analysis before prediction |
| Self-consistency | Figure 10 | Default $K=3$ sampled paths; median selection |
| Self-correction | Figure 11 | Default 2 refinement iterations |
| LLM decoding hyperparameters | Figure 6 | Temperature, top-p, penalties; OpenRouter configs |
| Hybrid reasoning effort (low/medium/high) | Main text, Models Module | Reasoning-token budgets / effort tiers |
| Scaling-law (TSFM) setup | Section I, Table 12 | Context=512, pred_len=96; #paths 1, 5, 10, . . . , 160 |

experimental model designed to enhance reasoning capabilities by generating its own "thinking process" during responses. This self-articulation allows the model to break down complex tasks into smaller steps, leading to more robust and accurate reasoning outcomes.⁶

E.3.3 Deepseek Model Series

DeepSeek-V3 is an open source 671B Mixture-of-Experts (MoE) language model. It employs Multi-head Latent Attention and DeepSeekMoE architectures for efficient inference and cost-effective training. DeepSeek-R1 surpasses V3 in reasoning tasks by incorporating cold-start data before reinforcement learning (RL), enhancing readability and performance. This approach distills reasoning patterns from larger models into smaller ones, resulting in superior reasoning capabilities compared to those developed through RL post training on smaller models.^{7 8}

E.4 Details of Evaluations

We use the API from popular OpenRouter⁹ platform to perform LLM inference. We strictly follow the recommended parameters for each LLM, i.e., median values from users on OpenRouter. We detail the LLM model versions and hyperparameter settings in Figure 6.

F Failure Analysis of High-Effort Reasoning

System analysis. To better understand why "thinking more" can lead to worse forecasts (Figure 2), we analyze a filtered subset of multimodal examples where the high-effort run (i.e., substantially longer reasoning traces) produces a clearly

worse forecast than the corresponding low-effort run. We manually examine 50 such failures and attribute each case to a primary failure mode based on the reasoning trace and the final numeric forecast.

Taxonomy. We group the observed failures into six types: (1) wrong mapping between context and numbers, (2) generic boilerplate reasoning, (3) overfitting to historical noise, (4) insufficient information / unclear attribution, (5) hallucination / fabricated events, and (6) reasoning contradicts the final forecast. Table 5 summarizes the operational definitions used for attribution, together with the frequency at a coarser group level.

Summary finding. Across the 50 examined cases, 35/50 failures fall into (1) and (2), 11/50 fall into (3) and (4), and only 4/50 fall into (5) and (6). This pattern suggests that the degradation is mainly driven by mis-grounded or shallow reasoning (for example, mapping qualitative context to the wrong direction or magnitude, or using template-like narratives), rather than pure hallucination. We note that these statistics are computed on a filtered failure subset (high-effort worse than low-effort), and should not be interpreted as the overall distribution across all examples.

F.1 Case studies

We provide representative cases to illustrate how the main failure modes appear in practice. Each case highlights (i) the contextual cue used in the trace, (ii) the high-effort reasoning symptom, and (iii) how it pushes the numeric forecast away from a reasonable continuation of the series.

Case 1: Wrong mapping between context and

⁶<https://ai.google.dev/gemini-api/docs/thinking>

⁷<https://github.com/deepseek-ai/DeepSeek-V3>

⁸<https://github.com/deepseek-ai/DeepSeek-R1>

⁹<https://openrouter.ai/>

Table 5: Taxonomy of failure modes for high-effort reasoning traces, with operational definitions and frequencies in a filtered subset of 50 failures. We report stable frequencies at the group level (pairs of types) because some cases exhibit multiple symptoms, and fine-grained tie-breaking can be subjective.

| Failure mode | Operational definition (trace → forecast symptom) | Count |
|--|--|-------|
| (1)–(2) Mis-grounded or shallow reasoning (group) | The trace mentions plausible factors, but either (i) maps them to numeric trends with incorrect timing/direction/magnitude, or (ii) uses template-like narratives that weakly condition on the given series/context. | 35/50 |
| (1) Wrong mapping between context and numbers | Cites real context, but maps it to the wrong numeric trend (timing, direction, or magnitude), often turning qualitative statements into monotonic growth/decline without checking the series pattern. | – |
| (2) Generic boilerplate | Uses generic, reusable reasoning (for example, seasonality stories or standard factor lists) and outputs a “typical” shape that could fit many series, with little use of the provided inputs. | – |
| (3)–(4) Noise sensitivity or unclear attribution (group) | The forecast is dominated by short-term anomalies or the trace is too vague/truncated to support a reliable attribution, even though the output is clearly worse than low-effort. | 11/50 |
| (3) Overfitting to historical noise | Overreacts to a few spikes/dips and extrapolates them forward, ignoring broader trend or seasonality; treats event-driven outliers as persistent signals. | – |
| (4) Insufficient information / unclear | The trace is too generic, internally incomplete, or truncated; key details needed to decide between (1)–(3) are missing, so we mark the case as unclear. | – |
| (5)–(6) Hallucination or inconsistency (group) | The trace introduces unsupported events/policies/numbers, or the reasoning conclusion does not match the final numeric forecast. | 4/50 |
| (5) Hallucination / fabricated events | Mentions events, policies, or specific claims that do not appear in the provided inputs and uses them to justify a trend change. | – |
| (6) Reasoning contradicts the final forecast | Text says “decrease” or “stabilize” but the forecast rises (or vice versa), indicating a mismatch between the verbal decision and the numeric output. | – |

numbers (retail broiler composite price). The high-effort trace cites a downward revision and easing cost pressure, which would normally support a dip or slower growth. However, the final forecast still shows only a brief dip followed by a slow, monotonic upward drift. This is a wrong context-to-number mapping: the model repeats relevant text but does not translate it into the correct direction or timing in the numeric path.

Case 2: Generic boilerplate (influenza patient proportion). The high-effort trace describes a textbook flu season pattern (rise in autumn, peak in winter, decline in spring) and lists standard factors (vaccines, variants, surveillance). The final forecast is a symmetric bell-shaped curve that would fit many years, but it does not reflect year-specific signals (for example, delayed peaks or weaker seasons). This is boilerplate reasoning: the trace is long, but it weakly conditions on the actual series and provided context.

Case 3: Overfitting to historical noise (disaster and emergency grants, 3-month horizon). The high-effort trace repeatedly focuses on a few unusually large peaks in recent months and treats them as strong evidence that spending will remain very high. The final forecast extrapolates these peaks forward instead of treating them as event-driven outliers relative to a lower baseline. This is noise overfitting: the model fits short-term anomalies and ignores the broader pattern.

G Analysis of the Benefits of Multimodal Reasoning

Quantitative evidence (CRPS). Beyond point-forecast MSE, we report Continuous Ranked Probability Score (CRPS) in Table 6, a proper scoring rule that evaluates the full predictive distribution (lower is better). Each setting is repeated 5 times, and we report the average relative CRPS change of multimodal versus unimodal forecasts. Across

multiple model families and horizons, multimodal inputs often reduce CRPS, which is consistent with the hypothesis that textual context provides additional signals that can improve distribution-level forecast quality, rather than only changing a single point estimate.

Qualitative evidence (case study: active text-to-number integration). We provide a representative TIME-THINKING example (US gasoline prices, 3-month horizon, DeepSeek;) to illustrate that the reasoning does not merely become longer in the multimodal setting, but actively uses text to adjust the numeric path. **Unimodal** (MSE= 0.225): the model only observes past gasoline prices and produces a generic seasonal narrative (for example, mild winter demand effects), yielding a mildly decreasing forecast that underestimates the sharp late-2023 drop. **Multimodal** (MSE= 0.024): with aligned news passages, the reasoning explicitly cites short-horizon signals (for example, rapid declines in crude oil futures and downward short-term price guidance) and maps them to a larger and faster near-term decrease; the forecast changes from a mild downward slope to a sharper drop that stabilizes near ~ 3.2 USD, substantially improving accuracy. This example suggests that the multimodal gain is not only due to higher task complexity, but also because reasoning can leverage textual context to make targeted corrections to a naive extrapolation.

Table 6: Average relative CRPS change (%) of multimodal vs. unimodal forecasts over 5 repeated experiments (negative means lower CRPS, i.e., better performances).

| Model family | Horizon | Avg multimodal CRPS change (%) |
|--------------|---------|--------------------------------|
| DeepSeek | 3m | -14.74% |
| GPT | 3m | -0.26% |
| Gemini | 3m | -44.73% |
| DeepSeek | 6m | -0.31% |
| GPT | 6m | -2.07% |
| Gemini | 6m | -15.56% |

H When Does Reasoning Help and When Does It Hurt? An Empirical Recipe

This appendix complements the aggregate results in Section 3.1 by making the “help vs. hurt” behavior of reasoning explicit.

H.1 Global “Help vs. Hurt” Analysis

Instance-level labeling. For each time-series instance, we assign one of two labels: **help** if the forecast error (MAE) with reasoning is lower than that of the corresponding no-reasoning baseline, and **hurt** otherwise. This labeling is performed under matched settings (same model family, same dataset, same horizon, and same evaluation protocol), so that the label isolates the effect of adding reasoning.

High-level trends. Across datasets and horizons, we observe a clear pattern consistent with our main findings in the paper.

- In **multimodal** and **6-month** settings where the future is strongly affected by exogenous events (for example, energy prices, disaster budgets, travel volume, and drought-related series), reasoning is more often helpful and yields larger average MAE reductions.
- In contrast, in **unimodal** settings or **3-month** horizons with relatively simple dynamics, the help/hurt ratio is closer to balanced, and some datasets show small negative net effects.

This suggests that in easier settings, additional reasoning can introduce extra variability without adding reliable signal, while in harder multimodal long-horizon settings, reasoning more often provides useful adjustments beyond naive extrapolation.

H.2 When Reasoning Helps: How Text Actively Adjusts Forecasts

We next provide representative TIME-THINKING examples to illustrate how multimodal reasoning can actively use textual context to improve numerical forecasts. Across these examples, a common positive pattern is that the reasoning leverages external signals in the text (reports, outlooks, policy guidance, or calendar cues) to adjust the *direction*, *slope*, or *seasonal shape* of a baseline extrapolation, rather than merely producing a post-hoc narrative for the same numbers.

Authority/policy signals correcting trend direction (USDA/EIA/CDC-style guidance). In a retail broiler composite price series (3-month hori-

1368 zon), the unimodal baseline follows a naive contin- 1418
1369 uation of recent movements and yields a large error, 1419
1370 while the multimodal reasoning run substantially 1420
1371 reduces MAE (approximately a 79% reduction as 1421
1372 measured in our audited example). The reasoning 1422
1373 cites report-style guidance that revises near-term 1423
1374 expectations downward and uses this to pull the
1375 forecast away from trend extension. In this case,
1376 the text provides a concrete directional cue (down-
1377 ward revision) that aligns with the realized future
1378 and directly corrects the numeric trajectory.

1380 **Distinguishing one-off shocks from new regimes**
1381 **(energy and disaster budgets).** For US gasoline 1424
1382 prices (6-month horizon), the unimodal baseline 1425
1383 underestimates how much prices fall after an ex- 1426
1384 treme peak year. The multimodal reasoning run 1427
1385 uses outlook-style context to identify the peak year 1428
1386 as exceptional and anchors the subsequent months 1429
1387 at a lower, more stable level with mild seasonality, 1430
1388 leading to a large MAE reduction (roughly 80% in 1431
1389 the examined example). A similar pattern appears 1432
1390 in disaster grants and emergency budgets: contex- 1433
1391 tual text can explain that large spikes are driven 1434
1392 by specific events (for example, major storms or 1435
1393 bills), so future spending should not repeatedly
1394 match those extremes. Here, reasoning helps by
1395 “denoising” event-driven outliers and preventing
1396 over-extrapolation.

1398 **Realigning seasonality with calendar and be-**
1399 **havior cues (travel, unemployment, influenza).**
1400 For travel volume (6-month horizon), multimodal
1401 reasoning can tie month-to-month changes to con-
1402 crete calendar events (for example, major holidays
1403 and seasonal travel behavior), producing a “high-
1404 low-high” pattern that better matches the realized
1405 seasonal shape than a smooth extrapolation. For
1406 unemployment or influenza-related series, contex-
1407 tual discussion of holiday hiring and post-holiday
1408 adjustments (or surveillance-related signals) can
1409 introduce small but meaningful seasonal bumps
1410 on top of a stable macro trend. These examples
1411 illustrate that text can provide time-aligned cues
1412 for seasonality and event timing that are not always
1413 identifiable from a short history alone.

1414 **H.3 When Reasoning Hurts: Common** 1415 **Failure Patterns**

1416 To understand harmful cases, we follow a tar- 1465
1417 geted audit protocol: we filter multimodal instances 1466

where the reasoning token budget is high but the 1418
forecast is worse than the corresponding low-effort 1419
run, and then manually inspect 50 such reasoning 1420
traces. We group the observed failures into six 1421
types. 1422
1423

Summary of audited counts. Among the 50 ex- 1424
amined failures, 35/50 fall into (i) **wrong mapping** 1425
between context and numbers and (ii) **generic** 1426
boilerplate reasoning. An additional 11/50 cor- 1427
respond to (iii) **overfitting to historical noise** or 1428
(iv) **insufficient information / unclear attribution.** 1429
Only 4/50 involve (v) **hallucination / fabricated** 1430
events or (vi) **reasoning contradicts the final fore-** 1431
cast. Overall, the degradation is therefore mainly 1432
driven by mis-grounded or shallow reasoning rather 1433
than pure hallucination. 1434
1435

Failure mode 1: wrong mapping between con- 1436
text and numbers. The model cites real informa- 1437
tion from the text but projects it into the forecast 1438
with incorrect timing, direction, or magnitude. For 1439
example, it may read that a report revises expect- 1440
ations downward or that risk is easing, yet still 1441
predicts a renewed upward drift; or it treats high- 1442
level phrases such as “increasing climate risk” as 1443
evidence that a monthly budget should rise almost 1444
linearly, ignoring cycles and caps. In these cases, 1445
parts of the reasoning may be consistent with the 1446
provided context, but the implied numerical map- 1447
ping conflicts with the observed historical dynam- 1448
ics or the forecast horizon. 1449
1450

Failure mode 2: generic boilerplate reasoning. 1451
The reasoning follows a template story (for exam- 1452
ple, “typical flu season”, “standard holiday effect”, 1453
or a generic macro supply-demand narrative) that 1454
could fit many series, with only weak grounding in 1455
the current instance. As a result, peaks, amplitudes, 1456
or turning points become misaligned because the 1457
model relies more on the template than on instance- 1458
specific signals from the data and context. 1459
1460

Failure mode 3: overfitting to historical noise. 1461
The model locks onto a few recent spikes or dips 1462
and assumes they will persist, instead of treating 1463
them as one-off events. This failure is the nega- 1464
tive counterpart of the successful “denoising” pat- 1465
tern: rather than using context to discount outliers, 1466

1467 the reasoning amplifies them and extrapolates their
1468 magnitude forward.
1469

1470 **Failure mode 4: insufficient information / un-**
1471 **clear attribution.** Some traces are too generic or
1472 too truncated to reliably attribute a single failure
1473 mode. The reasoning may contain vague phrases
1474 about “mixed signals” or “seasonal fluctuations”
1475 without concrete grounding, making it hard to de-
1476 termine whether the error stems from misreading
1477 context, under-using the series, or an unrelated
1478 modeling limitation.
1479

1480 **Failure modes 5–6: hallucination and self-**
1481 **contradiction (rare in our audit).** Hallucinated
1482 policies or fabricated events do occur, but are re-
1483 latively rare in the audited set. A small number of
1484 cases also show a mismatch between the textual
1485 conclusion and the numerical forecast, indicating
1486 that the final numbers do not faithfully follow the
1487 stated reasoning.

1488 I Details of Test-Time Scaling Law 1489 Experiments

1490 I.1 Data Setup

1491 We adopted the TimesNet datasets as shown in
1492 Table 12 for our experiments in this section. All
1493 datasets listed are open source datasets.

1494 I.2 Model Setup

1495 We utilized two contemporary popular zero-
1496 shot TSF that supports distributional forecast-
1497 ing, Chronos [2] and Moirai [68], to carry out
1498 our investigation about relationship between self-
1499 consistency and model performance. Here we take
1500 the number of sampled paths as a reflection of self-
1501 consistency. We chose the “base” version in experi-
1502 ments for both models.

1503 I.3 Hyperparameters

1504 The context for forecasting is set to 512 and the pre-
1505 diction length is set to 96, representing a short-term
1506 forecasting for 3 months. The specific numbers of
1507 sampled paths are set to 1 as a control group, then
1508 5, 10 and doubling itself to 160 for scaling up.

1509 I.4 Numerical Results

1510 In table 13,14,15,16 we include original validation
1511 results for our test-time scaling law, besides nor-
1512 malized ones mentioned above.

I.5 Statistical Robustness of the Test-Time Scaling Trend (Figure 4)

Figure 4 suggests that the main gains from multi-
path sampling concentrate in the 1→16/32 region
and then flatten. To quantify run-to-run variability
and assess whether the observed plateau is larger
than noise, we repeat each configuration five times
with independent randomness, resulting in 2 TSF
foundation models ×7 datasets ×8 path sizes ×5
repetitions (>35k inferences). For each repetition,
we compute the dataset-level metric and normalize
it by the 1-path result of the same dataset, then
average across the 7 datasets; we finally report
the mean, standard deviation, and 95% confidence
interval (CI) across the five repetitions.

The repeated results confirm that the dominant
improvement occurs in the 1→16/32 region, while
the marginal gains beyond 32 paths are much
smaller. For example, increasing paths from 32
to 128 reduces the cross-dataset normalized MAE
by 0.010 for both models (Moirai: 0.679→0.669;
Chronos: 0.764→0.754), whereas the reduction
from 1 to 32 is about 0.321 (Moirai) and 0.236
(Chronos). At the same time, the run-to-run vari-
ability is small (Table 17), enabling a direct signifi-
cance check for incremental gains.

These tests indicate that the curve flattening is
not driven by noise: the improvements from 32
to 64 remain detectable, while the effect size is
already much smaller than that in the 1→16/32
region; further increasing to 128 yields a very small
gain, and for Chronos the 64→128 difference is not
statistically significant under this test. We therefore
describe the trend as a practical “leveling off” after
16–32 paths: additional samples can still help, but
with strongly diminishing returns under near-linear
compute growth.

J TIME-THINKING Dataset: Annotation Quality Assurance

Goal and scope. A textually coherent argument
can still be statistically naive for time-series fore-
casting (TSF). Therefore, the quality control of
TIME-THINKING is designed to go beyond “log-
ical coherence” and explicitly encourage TSF-
grounded reasoning and numeric consistency. Start-
ing from approximately 3,000 initially collected
candidates, we select around 2,000 reasoning tra-
jectories through a multi-stage automated screening
pipeline.

1563

Judge model and TSF-aware rubric. We adopt an LLM-as-a-judge approach based on **Prometheus-2** [30]. Instead of only rating textual coherence, our judging prompt (Figure 13) instructs the judge to: (i) ground the reasoning in the provided time series (trend, seasonality, and possible structural breaks) and the multimodal context when present, (ii) penalize generic or task-irrelevant narratives, and (iii) verify that the final numerical forecast is consistent with the reasoning text (rather than contradicting it). Table 19 summarizes the TSF-aware checks used in our rubric.

Multi-stage screening pipeline. Beyond judge scoring, we add two additional safeguards to reduce noisy or misleading trajectories: an error-based sanity filter and repeated judging with majority voting. Table 20 summarizes the full pipeline.

Limitations of the current quality control. We emphasize that this pipeline is fully automated and does not include human TSF expert auditing. While the TSF-aware rubric, numeric consistency checks, error-based filtering, and majority voting mitigate common failure modes, future work can further strengthen the quality assurance with carefully designed human audit protocols (see our limitation statement in Section D).

K Demos of Reasoning-Annotated TSF Datasets

We use GPT-4o with the agriculture and climate dataset as an example to demonstrate the reasoning annotated TSF dataset we built. We organize the data in a Question and Answering format.

K.1 Agriculture

K.1.1 Unimodal TSF

Demo is provided in Figure 14.

K.1.2 Multimodal TSF

Demo is provided in Figure 15.

K.2 Climate

K.2.1 Unimodal TSF

Demo is provided in Figure 16.

K.2.2 Multimodal TSF

Demo is provided in Figure 17.

Table 7: Overview of dataset in REC4TS, covering key variables across nine domains with weekly, or monthly frequencies

| Domain | Target | Frequency | Timestamps | Timespan |
|-------------|-------------------------------|-----------|------------|----------------|
| Agriculture | Retail Broiler Composite | Monthly | 496496 | 1983 - 2024/04 |
| Climate | Drought Level | Monthly | 496496 | 1983 - 2024/04 |
| Economy | International Trade Balance | Monthly | 423423 | 1989 - 2024/04 |
| Energy | Gasoline Prices | Weekly | 14791479 | 1996 - 2024/04 |
| Health | Influenza Patients Proportion | Weekly | 13891389 | 1997 - 2024/04 |
| Security | Disaster and Emergency Grants | Monthly | 297297 | 1999 - 2024/04 |
| Employment | Unemployment Rate | Monthly | 900900 | 1950 - 2024/04 |
| Traffic | Travel Volume | Monthly | 531531 | 1980 - 2024/04 |

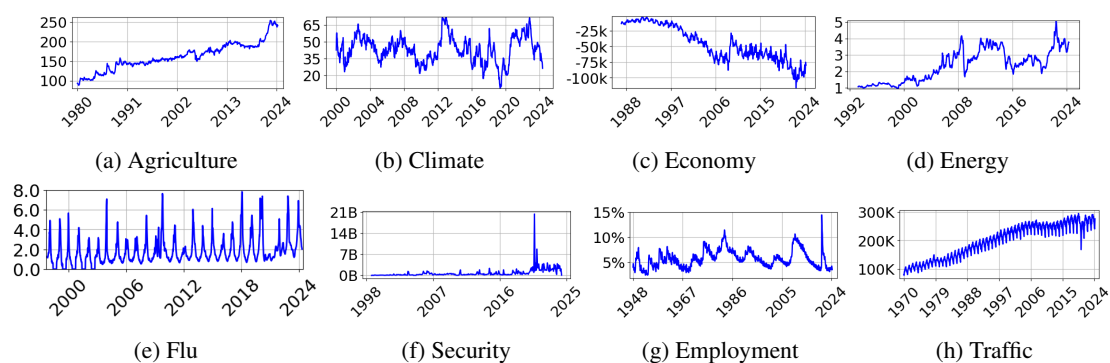


Figure 5: Visualization of time-series data.

```

model_params = {
  "openai/o1-mini-2024-09-12": {
    "top_p": 1,
    "temperature": 0.7,
    "frequency_penalty": 0,
    "presence_penalty": 0,
    "repetition_penalty": 1,
    "top_k": 0
  },
  "google/gemini-2.0-flash-thinking-exp:free": {
    "top_p": 1,
    "temperature": 0.8,
    "repetition_penalty": 1
  },
  "deepseek/deepseek-r1": {
    "top_p": 1,
    "temperature": 0.7,
    "frequency_penalty": 0,
    "presence_penalty": 0,
    "repetition_penalty": 1,
    "top_k": 0
  },
  "openai/gpt-4o-2024-05-13": {
    "top_p": 1,
    "temperature": 0.9,
    "frequency_penalty": 0,
    "presence_penalty": 0,
    "repetition_penalty": 1,
    "top_k": 0
  },
  "google/gemini-2.0-flash-exp:free": {
    "top_p": 1,
    "temperature": 0.7,
    "frequency_penalty": 0,
    "presence_penalty": 0,
    "repetition_penalty": 1,
    "top_k": 0
  },
  "deepseek/deepseek-chat": {
    "top_p": 1,
    "temperature": 0.9,
    "frequency_penalty": 0,
    "presence_penalty": 0,
    "repetition_penalty": 1,
    "top_k": 0
  }
}

```

Figure 6: Hyperparameter settings for different models.

Table 8: Results with Google’s System 1 (Gemini-2.0-flash) and 2 (Gemini-2.0-flash-thinking) Models. We report the mean MSE and standard deviation over three repeated experiments. Reasoning strategies that outperform the direct System 1 are highlighted in green, while those that perform worse or have similar performance (due to higher computational cost) are marked in red. In "Win System 1," we present the probability of each reasoning strategy outperforming System 1 across datasets. We again observe that **only the self-consistency strategy consistently works**, while **the System 2 strategy consistently fails**.

(a) Results of Unimodal Short-term TSF Settings. We use numerical series only to forecast the next three months.

| Dataset | System 1 | System 1 with Test-time Reasoning Enhancement | | | System 2 |
|--------------|------------------|---|-----------------------|----------------------|---------------------------|
| | Gemini-2.0-flash | with CoT | with Self-Consistency | with Self-Correction | Gemini-2.0-flash-thinking |
| Agriculture | 0.011±0.001 | 0.010±0.004 | 0.009±0.004 | 0.012±0.008 | 0.017±0.004 |
| Climate | 1.234±0.239 | 1.800±0.326 | 1.749±0.791 | 1.703±0.280 | 2.416±0.112 |
| Economy | 0.113±0.007 | 0.272±0.256 | 0.229±0.145 | 0.121±0.026 | 0.172±0.049 |
| Energy | 0.172±0.038 | 0.181±0.048 | 0.132±0.047 | 0.235±0.060 | 0.327±0.054 |
| Flu | 0.809±0.353 | 0.641±0.224 | 0.402±0.197 | 1.854±1.271 | 2.068±1.076 |
| Security | 0.170±0.054 | 0.252±0.104 | 0.380±0.323 | 0.191±0.095 | 0.259±0.001 |
| Employment | 0.002±0.001 | 0.005±0.003 | 0.004±0.004 | 0.004±0.002 | 0.311±0.001 |
| Traffic | 0.347±0.415 | 0.097±0.060 | 0.016±0.006 | 0.034±0.014 | 0.201±0.001 |
| Win System 1 | NA | 3/8 | 4/8 | 1/8 | 1/8 |

(b) Results of Multimodal Short-term TSF Settings. We use numerical series with textual context series to forecast the next three months.

| Dataset | System 1 | System 1 with Test-time Reasoning Enhancement | | | System 2 |
|--------------|------------------|---|-----------------------|----------------------|---------------------------|
| | Gemini-2.0-flash | with CoT | with Self-Consistency | with Self-Correction | Gemini-2.0-flash-thinking |
| Agriculture | 0.010±0.003 | 0.006±0.001 | 0.009±0.002 | 0.011±0.004 | 0.008±0.002 |
| Climate | 2.115±0.660 | 1.725±0.227 | 1.980±0.760 | 1.529±0.290 | 2.106±0.294 |
| Economy | 0.376±0.085 | 0.326±0.067 | 0.373±0.079 | 0.283±0.083 | 0.509±0.109 |
| Energy | 0.143±0.069 | 0.117±0.015 | 0.143±0.027 | 0.091±0.065 | 0.218±0.106 |
| Flu | 0.594±0.219 | 0.607±0.294 | 0.332±0.102 | 1.422±0.542 | 3.171±0.001 |
| Security | 0.558±0.604 | 0.145±0.050 | 0.172±0.119 | 0.141±0.065 | 0.259±0.001 |
| Employment | 0.013±0.002 | 0.015±0.002 | 0.011±0.002 | 0.011±0.003 | 0.311±0.001 |
| Traffic | 0.322±0.196 | 0.046±0.017 | 0.163±0.106 | 0.425±0.235 | 0.201±0.001 |
| Win System 1 | NA | 6/8 | 7/8 | 5/8 | 4/8 |

(c) Results of Unimodal Long-term TSF Settings. We use numerical series only to forecast the next six months.

| Dataset | System 1 | System 1 with Test-time Reasoning Enhancement | | | System 2 |
|--------------|------------------|---|-----------------------|----------------------|---------------------------|
| | Gemini-2.0-flash | with CoT | with Self-Consistency | with Self-Correction | Gemini-2.0-flash-thinking |
| Agriculture | 0.032±0.007 | 0.036±0.011 | 0.035±0.007 | 0.077±0.026 | 0.093±0.018 |
| Climate | 1.476±0.651 | 0.964±0.321 | 0.674±0.092 | 0.908±0.153 | 1.240±0.705 |
| Economy | 0.092±0.038 | 0.216±0.142 | 0.078±0.013 | 0.066±0.003 | 0.244±0.035 |
| Energy | 0.303±0.044 | 0.130±0.021 | 0.241±0.060 | 0.489±0.134 | 0.241±0.148 |
| Flu | 1.190±1.171 | 1.049±0.447 | 0.596±0.128 | 1.095±0.546 | 1.920±0.001 |
| Security | 0.196±0.052 | 0.533±0.493 | 0.955±0.389 | 0.154±0.031 | 0.207±0.001 |
| Employment | 0.011±0.001 | 0.019±0.007 | 0.009±0.002 | 0.013±0.005 | 0.268±0.001 |
| Traffic | 0.068±0.063 | 0.215±0.079 | 0.074±0.048 | 0.050±0.013 | 0.414±0.001 |
| Win System 1 | NA | 3/8 | 5/8 | 5/8 | 2/8 |

(d) Results of Multimodal Long-term TSF Settings. We use numerical series with textual context series to forecast the next six months.

| Dataset | System 1 | System 1 with Test-time Reasoning Enhancement | | | System 2 |
|--------------|------------------|---|-----------------------|----------------------|---------------------------|
| | Gemini-2.0-flash | with CoT | with Self-Consistency | with Self-Correction | Gemini-2.0-flash-thinking |
| Agriculture | 0.052±0.026 | 0.034±0.009 | 0.034±0.006 | 0.024±0.007 | 0.096±0.032 |
| Climate | 1.644±0.398 | 1.452±0.461 | 1.318±0.079 | 1.292±0.401 | 1.006±0.327 |
| Economy | 0.092±0.010 | 0.234±0.049 | 0.134±0.044 | 10.357±14.475 | 1.093±0.806 |
| Energy | 0.138±0.106 | 0.208±0.116 | 0.159±0.077 | 0.384±0.074 | 0.713±0.513 |
| Flu | 0.659±0.173 | 0.557±0.164 | 0.477±0.006 | 0.785±0.064 | 1.920±0.001 |
| Security | 0.123±0.062 | 0.109±0.020 | 0.142±0.043 | 0.151±0.053 | 0.207±0.001 |
| Employment | 0.029±0.004 | 0.022±0.003 | 0.026±0.003 | 0.026±0.002 | 0.268±0.001 |
| Traffic | 0.085±0.068 | 0.037±0.027 | 0.020±0.007 | 0.058±0.010 | 0.414±0.001 |
| Win System 1 | NA | 6/8 | 5/8 | 4/8 | 1/8 |

Table 9: Performance Comparison of Different Reasoning budgets for o4-mini. We report mean MSE and standard deviation across datasets. Results that outperform GPT-4o are highlighted in green, while those performing worse are marked in red. In "Beat GPT-4o" row, we present the fraction of datasets where each reasoning budget outperforms GPT-4o.

(a) Results of Unimodal Short-term TSF Settings. We use numerical series only to forecast the next three months.

| Dataset | GPT-4o | o4-mini | | |
|-------------|-------------|----------------------|-------------------------|-----------------------|
| | | low reasoning effort | medium reasoning effort | high reasoning effort |
| Agriculture | 0.021±0.011 | 0.021±0.013 | 0.044±0.013 | 0.026±0.006 |
| Climate | 1.599±0.500 | 0.960±0.081 | 1.448±0.058 | 1.643±0.109 |
| Economy | 0.631±0.135 | 0.271±0.074 | 0.182±0.065 | 0.165±0.061 |
| Energy | 0.363±0.110 | 0.262±0.131 | 0.168±0.105 | 0.091±0.014 |
| Flu | 0.568±0.425 | 0.508±0.246 | 1.843±1.143 | 2.011±0.987 |
| Security | 0.093±0.029 | 0.216±0.061 | 0.218±0.032 | 0.146±0.022 |
| Employment | 0.010±0.004 | 0.012±0.005 | 0.009±0.005 | 0.005±0.003 |
| Traffic | 0.385±0.471 | 0.069±0.037 | 0.025±0.008 | 0.024±0.010 |
| Beat GPT-4o | NA | 6/8 | 5/8 | 4/8 |

(b) Results of Multimodal Short-term TSF Settings. We use numerical series with textual context series to forecast the next three months.

| Dataset | GPT-4o | o4-mini | | |
|-------------|-------------|----------------------|-------------------------|-----------------------|
| | | low reasoning effort | medium reasoning effort | high reasoning effort |
| Agriculture | 0.018±0.015 | 0.018±0.009 | 0.014±0.008 | 0.026±0.022 |
| Climate | 1.716±0.580 | 1.026±0.397 | 1.434±0.529 | 1.660±0.179 |
| Economy | 0.569±0.162 | 0.430±0.217 | 0.212±0.118 | 0.183±0.087 |
| Energy | 0.541±0.457 | 0.181±0.086 | 0.305±0.213 | 0.241±0.247 |
| Flu | 0.548±0.164 | 1.053±0.156 | 0.951±0.175 | 0.371±0.142 |
| Security | 0.076±0.052 | 0.109±0.024 | 0.065±0.032 | 0.143±0.027 |
| Employment | 0.020±0.006 | 0.015±0.005 | 0.016±0.000 | 0.008±0.005 |
| Traffic | 0.551±0.396 | 0.168±0.047 | 0.195±0.221 | 0.023±0.011 |
| Beat GPT-4o | NA | 5/8 | 7/8 | 6/8 |

(c) Results of Unimodal Long-term TSF Settings. We use numerical series only to forecast the next six months.

| Dataset | GPT-4o | o4-mini | | |
|-------------|-------------|----------------------|-------------------------|-----------------------|
| | | low reasoning effort | medium reasoning effort | high reasoning effort |
| Agriculture | 0.093±0.057 | 0.026±0.014 | 0.035±0.007 | 0.045±0.029 |
| Climate | 0.754±0.051 | 0.477±0.077 | 0.722±0.103 | 0.694±0.057 |
| Economy | 0.463±0.146 | 0.284±0.053 | 0.106±0.059 | 0.272±0.071 |
| Energy | 0.197±0.038 | 0.392±0.151 | 0.167±0.089 | 0.259±0.194 |
| Flu | 0.219±0.053 | 0.696±0.317 | 1.657±1.006 | 2.379±0.999 |
| Security | 0.183±0.044 | 0.151±0.043 | 0.147±0.076 | 0.185±0.125 |
| Employment | 0.011±0.006 | 0.018±0.004 | 0.009±0.003 | 0.021±0.003 |
| Traffic | 0.066±0.046 | 0.209±0.162 | 0.045±0.009 | 0.038±0.009 |
| Beat GPT-4o | NA | 4/8 | 7/8 | 4/8 |

(d) Results of Multimodal Long-term TSF Settings. We use numerical series with textual context series to forecast the next six months.

| Dataset | GPT-4o | o4-mini | | |
|-------------|-------------|----------------------|-------------------------|-----------------------|
| | | low reasoning effort | medium reasoning effort | high reasoning effort |
| Agriculture | 0.110±0.065 | 0.023±0.011 | 0.034±0.004 | 0.875±1.165 |
| Climate | 1.365±0.479 | 0.608±0.058 | 0.988±0.522 | 0.850±0.128 |
| Economy | 0.487±0.237 | 0.264±0.089 | 0.201±0.124 | 0.283±0.176 |
| Energy | 0.365±0.185 | 0.492±0.104 | 0.168±0.092 | 0.088±0.035 |
| Flu | 0.291±0.065 | 0.625±0.290 | 0.916±0.153 | 0.712±0.413 |
| Security | 0.196±0.056 | 0.187±0.029 | 0.164±0.042 | 0.175±0.075 |
| Employment | 0.015±0.002 | 0.021±0.003 | 0.013±0.007 | 0.021±0.012 |
| Traffic | 0.207±0.205 | 0.057±0.048 | 0.130±0.138 | 0.025±0.006 |
| Beat GPT-4o | NA | 5/8 | 7/8 | 5/8 |

Table 10: Performance Comparison of Different Token budgets for Gemini 2.5 Pro. We report mean MSE and standard deviation across datasets. Results that outperform GPT-4o are highlighted in green, while those performing worse are marked in red. In "Beat GPT-4o" row, we present the fraction of datasets where each reasoning budget outperforms GPT-4o.

(a) Results of Unimodal Short-term TSF Settings. We use numerical series only to forecast the next three months.

| Dataset | GPT-4o | Gemini 2.5 Pro | | |
|-------------|-------------|----------------------|----------------------|----------------------|
| | | 1K extended thinking | 2K extended thinking | 4k extended thinking |
| Agriculture | 0.021±0.011 | 0.916±1.270 | 2.712±0.000 | 2.712±0.000 |
| Climate | 1.599±0.500 | 2.099±0.000 | 2.099±0.000 | 1.951±0.210 |
| Economy | 0.631±0.135 | 0.439±0.204 | 0.583±0.000 | 0.557±0.037 |
| Energy | 0.363±0.110 | 1.239±0.835 | 1.829±0.000 | 1.288±0.764 |
| Flu | 0.568±0.425 | 3.171±0.000 | 3.171±0.000 | 3.171±0.000 |
| Security | 0.093±0.029 | 0.232±0.037 | 0.195±0.091 | 0.188±0.100 |
| Employment | 0.010±0.004 | 0.208±0.146 | 0.208±0.146 | 0.311±0.000 |
| Traffic | 0.385±0.471 | 0.141±0.085 | 0.201±0.000 | 0.162±0.056 |
| Beat GPT-4o | NA | 2/8 | 2/8 | 2/8 |

(b) Results of Multimodal Short-term TSF Settings. We use numerical series with textual context series to forecast the next three months.

| Dataset | GPT-4o | Gemini 2.5 Pro | | |
|-------------|-------------|----------------------|----------------------|----------------------|
| | | 1K extended thinking | 2K extended thinking | 4k extended thinking |
| Agriculture | 0.018±0.015 | 0.910±1.274 | 2.712±0.000 | 2.712±0.000 |
| Climate | 1.716±0.580 | 3.220±0.343 | 2.132±0.046 | 2.615±0.410 |
| Economy | 0.569±0.162 | 0.507±0.107 | 0.583±0.000 | 0.265±0.103 |
| Energy | 0.541±0.457 | 1.829±0.000 | 0.641±0.840 | 0.656±0.830 |
| Flu | 0.548±0.164 | 3.171±0.000 | 2.254±1.296 | 2.338±1.178 |
| Security | 0.076±0.052 | 0.200±0.083 | 0.192±0.095 | 0.187±0.102 |
| Employment | 0.020±0.006 | 0.311±0.000 | 0.105±0.146 | 0.210±0.143 |
| Traffic | 0.551±0.396 | 0.161±0.058 | 0.201±0.000 | 0.201±0.000 |
| Beat GPT-4o | NA | 2/8 | 1/8 | 2/8 |

(c) Results of Unimodal Long-term TSF Settings. We use numerical series only to forecast the next six months.

| Dataset | GPT-4o | Gemini 2.5 Pro | | |
|-------------|-------------|----------------------|----------------------|----------------------|
| | | 1K extended thinking | 2K extended thinking | 4k extended thinking |
| Agriculture | 0.093±0.057 | 0.856±1.179 | 2.523±0.000 | 0.859±1.177 |
| Climate | 0.754±0.051 | 1.277±0.000 | 1.277±0.000 | 1.604±0.463 |
| Economy | 0.463±0.146 | 0.359±0.000 | 0.359±0.000 | 0.359±0.000 |
| Energy | 0.197±0.038 | 1.396±0.000 | 1.396±0.000 | 0.968±0.604 |
| Flu | 0.219±0.053 | 1.920±0.000 | 1.920±0.000 | 1.920±0.000 |
| Security | 0.183±0.044 | 0.207±0.000 | 0.207±0.000 | 0.207±0.000 |
| Employment | 0.011±0.006 | 0.181±0.123 | 0.268±0.000 | 0.184±0.119 |
| Traffic | 0.066±0.046 | 0.414±0.000 | 0.284±0.184 | 0.284±0.184 |
| Beat GPT-4o | NA | 1/8 | 1/8 | 1/8 |

(d) Results of Multimodal Long-term TSF Settings. We use numerical series with textual context series to forecast the next six months.

| Dataset | GPT-4o | Gemini 2.5 Pro | | |
|-------------|-------------|----------------------|----------------------|----------------------|
| | | 1K extended thinking | 2K extended thinking | 4k extended thinking |
| Agriculture | 0.110±0.065 | 0.853±1.181 | 0.864±1.173 | 0.849±1.184 |
| Climate | 1.365±0.479 | 1.277±0.000 | 1.349±0.102 | 1.277±0.000 |
| Economy | 0.487±0.237 | 0.359±0.000 | 0.301±0.082 | 0.320±0.054 |
| Energy | 0.365±0.185 | 0.982±0.586 | 0.961±0.615 | 0.948±0.633 |
| Flu | 0.291±0.065 | 1.766±0.218 | 1.920±0.000 | 1.920±0.000 |
| Security | 0.196±0.056 | 0.207±0.000 | 0.144±0.045 | 0.207±0.000 |
| Employment | 0.015±0.002 | 0.094±0.123 | 0.268±0.000 | 0.182±0.121 |
| Traffic | 0.207±0.205 | 0.414±0.000 | 0.286±0.181 | 0.414±0.000 |
| Beat GPT-4o | NA | 2/8 | 3/8 | 2/8 |

Table 11: Performance Comparison of Different Token budgets for Claude-3.7-Sonnet. We report mean MSE and standard deviation across datasets. Results that outperform GPT-4o are highlighted in green, while those performing worse are marked in red. In "Beat GPT-4o" row, we present the fraction of datasets where each reasoning budget outperforms GPT-4o.

(a) Results of Unimodal Short-term TSF Settings. We use numerical series only to forecast the next three months.

| Dataset | GPT-4o | Claude-3.7-Sonnet 1K extended thinking | Claude-3.7-Sonnet 2K extended thinking | Claude-3.7-Sonnet 4K extended thinking |
|-------------|-------------|---|---|---|
| Agriculture | 0.021±0.011 | 0.916±1.270 | 2.712±0.000 | 2.712±0.000 |
| Climate | 1.599±0.500 | 2.099±0.000 | 2.099±0.000 | 1.951±0.210 |
| Economy | 0.631±0.135 | 0.439±0.204 | 0.583±0.000 | 0.557±0.037 |
| Energy | 0.363±0.110 | 1.239±0.835 | 1.829±0.000 | 1.288±0.764 |
| Flu | 0.568±0.425 | 3.171±0.000 | 3.171±0.000 | 3.171±0.000 |
| Security | 0.093±0.029 | 0.232±0.037 | 0.195±0.091 | 0.188±0.100 |
| Employment | 0.010±0.004 | 0.208±0.146 | 0.208±0.146 | 0.311±0.000 |
| Traffic | 0.385±0.471 | 0.141±0.085 | 0.201±0.000 | 0.162±0.056 |
| Beat GPT-4o | NA | 2/8 | 2/8 | 2/8 |

(b) Results of Multimodal Short-term TSF Settings. We use numerical series with textual context series to forecast the next three months.

| Dataset | GPT-4o | Claude-3.7-Sonnet 1K extended thinking | Claude-3.7-Sonnet 2K extended thinking | Claude-3.7-Sonnet 4K extended thinking |
|-------------|-------------|---|---|---|
| Agriculture | 0.018±0.015 | 0.009±0.005 | 0.006±0.001 | 0.008±0.004 |
| Climate | 1.716±0.580 | 1.304±0.265 | 1.376±0.032 | 1.437±0.082 |
| Economy | 0.569±0.162 | 0.145±0.102 | 0.167±0.043 | 0.308±0.044 |
| Energy | 0.541±0.457 | 0.131±0.047 | 0.101±0.045 | 0.171±0.071 |
| Flu | 0.548±0.164 | 0.545±0.217 | 0.473±0.258 | 0.777±0.223 |
| Security | 0.076±0.052 | 0.103±0.022 | 0.061±0.039 | 0.039±0.009 |
| Employment | 0.020±0.006 | 0.011±0.002 | 0.010±0.006 | 0.014±0.004 |
| Traffic | 0.551±0.396 | 0.044±0.004 | 0.043±0.003 | 0.059±0.019 |
| Beat GPT-4o | NA | 7/8 | 8/8 | 7/8 |

(c) Results of Unimodal Long-term TSF Settings. We use numerical series only to forecast the next six months.

| Dataset | GPT-4o | Claude-3.7-Sonnet 1K extended thinking | Claude-3.7-Sonnet 2K extended thinking | Claude-3.7-Sonnet 4K extended thinking |
|-------------|-------------|---|---|---|
| Agriculture | 0.093±0.057 | 0.124±0.037 | 0.081±0.029 | 0.046±0.014 |
| Climate | 0.754±0.051 | 0.699±0.091 | 0.679±0.015 | 0.684±0.019 |
| Economy | 0.463±0.146 | 0.053±0.004 | 0.186±0.063 | 0.057±0.029 |
| Energy | 0.197±0.038 | 0.169±0.055 | 0.209±0.022 | 0.255±0.079 |
| Flu | 0.219±0.053 | 1.157±0.574 | 1.137±0.140 | 0.782±0.467 |
| Security | 0.183±0.044 | 0.133±0.053 | 0.094±0.008 | 0.089±0.004 |
| Employment | 0.011±0.006 | 0.014±0.003 | 0.010±0.002 | 0.012±0.002 |
| Traffic | 0.066±0.046 | 0.021±0.010 | 0.161±0.179 | 0.026±0.013 |
| Beat GPT-4o | NA | 5/8 | 5/8 | 5/8 |

(d) Results of Multimodal Long-term TSF Settings. We use numerical series with textual context series to forecast the next six months.

| Dataset | GPT-4o | Claude-3.7-Sonnet 1K extended thinking | Claude-3.7-Sonnet 2K extended thinking | Claude-3.7-Sonnet 4K extended thinking |
|-------------|-------------|---|---|---|
| Agriculture | 0.110±0.065 | 0.027±0.009 | 0.035±0.011 | 0.057±0.029 |
| Climate | 1.365±0.479 | 0.819±0.048 | 0.859±0.096 | 0.783±0.067 |
| Economy | 0.487±0.237 | 0.097±0.032 | 0.084±0.028 | 0.069±0.028 |
| Energy | 0.365±0.185 | 0.097±0.061 | 0.091±0.049 | 0.141±0.042 |
| Flu | 0.291±0.065 | 0.724±0.143 | 0.566±0.248 | 0.729±0.394 |
| Security | 0.196±0.056 | 0.117±0.007 | 0.091±0.013 | 0.089±0.016 |
| Employment | 0.015±0.002 | 0.012±0.001 | 0.007±0.002 | 0.008±0.001 |
| Traffic | 0.207±0.205 | 0.041±0.013 | 0.039±0.025 | 0.033±0.006 |
| Beat GPT-4o | NA | 7/8 | 7/8 | 7/8 |

```

base_prompt = f"""As an expert in {target_name}, predict the
    trends from {future_dates[0]} to {future_dates[-1]}
    based on both historical data and contextual information.

Historical data (chronological order):
{historical_str}

Contextual information:
{context_text}

Please enclose your final predictions between [PRED_START] and [
    PRED_END] markers exactly like this:
[PRED_START]
2024-01-01: 123.45
2024-02-01: 124.56
[PRED_END]"""

```

Figure 7: Prompt used for multimodal time-series forecasting.

```

base_prompt = f"""As an expert in {target_name}, predict the
    trends from {future_dates[0]} to {future_dates[-1]} based on
    the historical data.

Historical data (chronological order):
{historical_str}

Please enclose your final predictions between [PRED_START] and [
    PRED_END] markers exactly like this:
[PRED_START]
2024-01-01: 123.45
2024-02-01: 124.56
[PRED_END]"""

```

Figure 8: Prompt used for unimodal time-series forecasting.

```

prompt = base_prompt + "\n\nLet's approach this step by step:\n1
    . Analyze historical trends\n2. Identify patterns\n3. Make
    predictions"

```

Figure 9: Prompt used for chain-of-thought reasoning.

```

elif method == "self_consistency":
    predictions = []
    for _ in range(3):
        pred = get_single_prediction(base_prompt, future,
                                     target_name, api_key, model)
        predictions.append(pred)
    return average_predictions(predictions)

```

Figure 10: Prompt used for self-consistency reasoning strategy.

```

elif method == "self_correction":
    current_pred = get_single_prediction(base_prompt, future,
                                       target_name, api_key, model)
    for _ in range(2):
        correction_prompt = f"{base_prompt}\n\nPrevious
                             prediction:\n{format_prediction(current_pred)}\n\n
                             Please review and improve the prediction."
        current_pred = get_single_prediction(correction_prompt,
                                             future, target_name, api_key, model)
    return current_pred

```

Figure 11: Prompt used for self-correction reasoning strategy.

Table 12: Dataset for Test-Time Scaling Law Experiments

| Name | Domain | Frequency | Timestamps |
|---------------|---------|-----------|------------|
| Electricity | Energy | Hourly | 26305 |
| Traffic | Traffic | Hourly | 17545 |
| Exchange Rate | Economy | Daily | 7589 |
| ETTh1 | Energy | Hourly | 17421 |
| ETTh2 | Energy | Hourly | 17421 |
| ETTh1 | Energy | 15 Min | 69681 |
| ETTh2 | Energy | 15 Min | 69681 |

Table 13: Numerical Results of Chronos (MSE)

| Num Path \ Dataset | Electricity | Traffic | Exchange Rate | ETTh1 | ETTh2 | ETTm1 | ETTm2 |
|--------------------|-------------|---------|---------------|-------|-------|-------|-------|
| 1 | 101164.19 | 2.17e-5 | 2.12e-3 | 13.95 | 50.00 | 9.62 | 37.42 |
| 5 | 63658.19 | 1.68e-5 | 1.14e-3 | 11.59 | 34.73 | 6.50 | 21.57 |
| 10 | 59627.12 | 1.60e-5 | 1.13e-3 | 10.29 | 31.32 | 5.94 | 20.71 |
| 20 | 55347.09 | 1.58e-5 | 1.02e-3 | 10.06 | 30.47 | 5.58 | 19.68 |
| 40 | 54779.91 | 1.55e-5 | 1.00e-3 | 9.98 | 29.96 | 5.60 | 18.86 |
| 80 | 54301.96 | 1.53e-5 | 1.00e-3 | 9.80 | 29.25 | 5.52 | 18.44 |
| 160 | 54233.27 | 1.53e-5 | 0.90e-3 | 9.98 | 29.52 | 5.50 | 18.54 |

Table 14: Numerical Results of Chronos (MAE)

| Num Path \ Dataset | Electricity | Traffic | Exchange Rate | ETTh1 | ETTh2 | ETTm1 | ETTm2 |
|--------------------|-------------|---------|---------------|-------|-------|-------|-------|
| 1 | 225.51 | 2.77e-3 | 0.033 | 2.77 | 5.12 | 2.23 | 4.33 |
| 5 | 176.63 | 2.22e-3 | 0.024 | 2.51 | 4.31 | 1.83 | 3.40 |
| 10 | 168.59 | 2.17e-3 | 0.023 | 2.38 | 4.11 | 1.76 | 3.32 |
| 20 | 162.45 | 2.11e-3 | 0.022 | 2.35 | 4.05 | 1.69 | 3.20 |
| 40 | 161.03 | 2.07e-3 | 0.022 | 2.33 | 4.05 | 1.69 | 3.13 |
| 80 | 160.15 | 2.03e-3 | 0.022 | 2.29 | 3.97 | 1.67 | 3.11 |
| 160 | 159.89 | 2.03e-3 | 0.022 | 2.32 | 3.98 | 1.66 | 3.11 |

Table 15: Numerical Results of Moirai (MSE)

| Num Path \ Dataset | Electricity | Traffic | Exchange Rate | ETTh1 | ETTh2 | ETTm1 | ETTm2 |
|--------------------|-------------|---------|---------------|-------|-------|-------|-------|
| 1 | 181007.83 | 3.23e-5 | 3.80e-3 | 31.55 | 68.30 | 40.91 | 65.14 |
| 5 | 88376.75 | 1.73e-5 | 1.36e-3 | 14.57 | 42.92 | 8.45 | 32.44 |
| 10 | 79590.41 | 1.61e-5 | 1.23e-3 | 13.15 | 40.11 | 7.63 | 31.56 |
| 20 | 73638.04 | 1.57e-5 | 1.16e-3 | 12.31 | 37.42 | 6.91 | 27.54 |
| 40 | 71017.52 | 1.55e-5 | 1.11e-3 | 11.88 | 36.90 | 6.69 | 27.26 |
| 80 | 69676.72 | 1.53e-5 | 1.09e-3 | 11.70 | 36.39 | 6.58 | 27.23 |
| 160 | 69018.68 | 1.52e-5 | 1.08e-3 | 11.63 | 36.27 | 6.53 | 27.16 |

Table 16: Numerical Results of Moirai (MAE)

| Num Path \ Dataset | Electricity | Traffic | Exchange Rate | ETTh1 | ETTh2 | ETTm1 | ETTm2 |
|--------------------|-------------|---------|---------------|-------|-------|-------|-------|
| 1 | 277.88 | 3.21e-3 | 0.036 | 3.78 | 5.95 | 2.97 | 5.45 |
| 5 | 207.14 | 2.34e-3 | 0.026 | 2.77 | 4.84 | 2.09 | 4.25 |
| 10 | 196.03 | 2.23e-3 | 0.025 | 2.65 | 4.68 | 2.01 | 4.19 |
| 20 | 189.85 | 2.15e-3 | 0.024 | 2.52 | 4.50 | 1.88 | 3.87 |
| 40 | 185.28 | 2.11e-3 | 0.023 | 2.48 | 4.44 | 1.84 | 3.84 |
| 80 | 183.44 | 2.09e-3 | 0.023 | 2.46 | 4.44 | 1.83 | 3.84 |
| 160 | 182.38 | 2.07e-3 | 0.023 | 2.45 | 4.43 | 1.82 | 3.83 |

Table 17: Repeated results for Figure 4 (normalized MAE; lower is better). We report cross-dataset mean/std over five repetitions, and 95% CI computed as $t_{0.975,4} \cdot \text{std}/\sqrt{5}$ (with $t_{0.975,4}=2.776$).

| # paths | Moirai | | | Chronos | | |
|---------|--------|-------|---------------|---------|-------|---------------|
| | mean | std | 95% CI | mean | std | 95% CI |
| 1 | 1.000 | 0.006 | ± 0.00745 | 1.000 | 0.023 | ± 0.02855 |
| 4 | 0.792 | 0.002 | ± 0.00248 | 0.876 | 0.016 | ± 0.01987 |
| 8 | 0.721 | 0.002 | ± 0.00248 | 0.805 | 0.010 | ± 0.01242 |
| 16 | 0.693 | 0.002 | ± 0.00248 | 0.779 | 0.009 | ± 0.01117 |
| 32 | 0.679 | 0.001 | ± 0.00124 | 0.764 | 0.005 | ± 0.00621 |
| 64 | 0.672 | 0.001 | ± 0.00124 | 0.756 | 0.004 | ± 0.00497 |
| 128 | 0.669 | 0.001 | ± 0.00124 | 0.754 | 0.003 | ± 0.00372 |

Table 18: Significance tests for incremental gains in the plateau region (normalized MAE). We use a two-sided Welch’s t -test on the five independent repetitions for each configuration.

| Comparison | Δ mean (Moirai) | p (Moirai) | Δ mean (Chronos) | p (Chronos) |
|------------|------------------------|--------------|-------------------------|---------------|
| 16 vs 32 | 0.014 | $< 10^{-4}$ | 0.015 | 0.016 |
| 32 vs 64 | 0.007 | $< 10^{-4}$ | 0.008 | 0.025 |
| 64 vs 128 | 0.003 | 0.0015 | 0.002 | 0.399 |

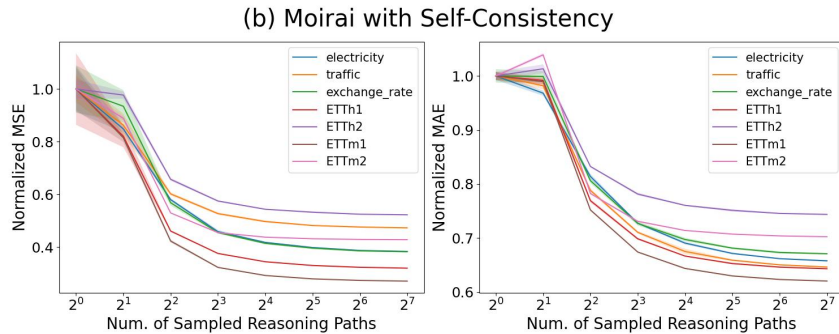


Figure 12: Moirai with multi-path sampling: mean normalized MSE/MAE across five repetitions with error bands (95% CI). Results are averaged over 7 datasets after normalizing each dataset by its 1-path score.

Table 19: TSF-aware rubric used by the LLM judge for screening reasoning trajectories in TIME-THINKING.

| Rubric dimension | Operational check (what the judge is instructed to prefer) | Main failure it prevents |
|--|---|---|
| TSF-grounded reasoning | Reasoning explicitly references observable patterns in the provided history (trend, seasonality, level shifts, and potential structural breaks) and does not ignore dominant series characteristics. | Statistically naive narratives that do not match the series |
| Context-to-number grounding (multimodal) | When textual context is present, the reasoning must connect context to the forecast through plausible mechanisms and avoid unsupported context-to-number mapping (for example, turning general statements into monotonic growth). | Mis-grounded mapping between context and numeric trends |
| Anti-boilerplate preference | Generic templates and task-irrelevant factor lists are penalized; the judge prefers reasoning that is specific to the given series and context. | Textually fluent but non-informative explanations |
| Reasoning-forecast consistency | The final numeric forecast should follow the stated reasoning (for example, if the reasoning concludes “stabilize” or “decline”, the forecast should not show a clear increase). | Verbal-numeric contradictions |

Table 20: Quality control pipeline for TIME-THINKING. The design goal is to reduce statistically naive yet fluent reasoning by combining TSF-aware judging, numeric sanity checks, and repeated decisions.

| Stage | What we do | Why it helps |
|--------------------------------------|---|--|
| (1) Candidate generation | Collect reasoning trajectories and the corresponding final numeric forecasts from base generators. | Provides a diverse pool of candidate trajectories. |
| (2) TSF-aware judging | Use Prometheus-2 with the TSF-aware rubric in Figure 13 to score trajectory quality (not limited to coherence). | Encourages grounding in the series and checks reasoning-forecast consistency. |
| (3) Repeated judging + majority vote | Run the same judging procedure three times and use majority voting to decide keep/discard. | Reduces sensitivity to a single stochastic judge run. |
| (4) Error-based sanity filter | Discard candidates whose prediction error is above a preset threshold. | Filters out trajectories that look plausible in text but produce clearly poor forecasts. |
| (5) Final selection | Keep trajectories that pass the voting decision and the error filter. | Improves the overall reliability of the released dataset. |

Prometheus 2 Judging Prompt for Reasoning Trajectory Quality

You are an expert evaluator, Prometheus 2. Your task is to assess the quality of a reasoning trajectory generated by a Large Language Model for a given Time-Series Forecasting (TSF) problem

****INPUTS:****

1. ****TSF Task Query:****

{tsf_task_query_placeholder}

(This includes the target variable, historical data format/snippet, context if multimodal, and the future period to be forecasted.)

2. ****Generated Reasoning Trajectory:****

{generated_reasoning_trajectory_placeholder}

3. ****Generated Final Forecast (Numerical Output):****

{generated_final_forecast_placeholder}

****EVALUATION CRITERIA & RUBRIC:****

Please evaluate the "Generated Reasoning Trajectory" based on the following criteria. Provide a holistic quality score from 1 (Very Poor) to 5 (Excellent) and a brief justification.

* ****1. Logical Coherence and Soundness (Weight: 30%):****

- * Are the reasoning steps logical and internally consistent?
- * Does the reasoning flow make sense in the context of TSF?
- * Are there any obvious logical fallacies or contradictions?

* ****2. Relevance to the Task (Weight: 25%):****

- * Is the entire reasoning process focused on solving the given TSF query?
- * Does it appropriately consider key aspects of the time series (e.g., trend, seasonality, context if provided)?
- * Are there irrelevant digressions or generic, non-specific statements?

* ****3. Completeness and Clarity (Weight: 20%):****

- * Is the reasoning sufficiently detailed to understand the model's thought process?
- * Is the language clear, precise, and unambiguous?
- * Are there any overly terse, vague, or missing steps that hinder understanding?

* ****4. Absence of Hallucinations/Nonsense (Weight: 15%):****

- * Is the reasoning grounded in the provided TSF data/context?
- * Are there any claims made that are clearly not supported by the input or are factually incorrect in a general sense relevant to TSF?
- * Are there any nonsensical calculations or steps?

* ****5. Faithfulness & Format (Minor, Weight: 10%):****

- * (If multimodal) Is the reasoning based on textual context a faithful interpretation?
- * Is the "Generated Final Forecast" presented in a reasonable and parsable format as might be expected from the TSF query? (Focus on major format issues, not minor numerical precision unless it stems from a reasoning flaw).

****OUTPUT FORMAT:****

Please provide your evaluation in the following format:

****Overall Quality Score (1-5):**** [Your Score]

****Justification:**** [Your brief textual justification (2-4 sentences) explaining your score, highlighting key strengths and weaknesses based on the criteria above.]

Figure 13: Conceptual prompt structure provided to Prometheus 2 for evaluating the quality of reasoning trajectories in the TIME-THINKING dataset. Placeholders like {tsf_task_query_placeholder} would be filled with the actual content for each sample.

Question

"As an expert in the US Retail Broiler Composite, predict the trends from 2023-11-30 to 2024-01-31 based on the historical data.

Historical data (chronological order):

2015-11-30: 196.99; 2015-12-31: 193.15; 2016-01-31: 192.66;
2016-02-29: 187.42; ...
2023-08-31: 252.59; 2023-09-30: 246.16; 2023-10-31: 247.64

Please enclose your final predictions between [PRED_START] and [PRED_END] markers exactly like this:

[PRED_START]

2024-01-01: 123.45

2024-02-01: 124.56

[PRED_END]"

Answer

"To make a prediction for the US Retail Broiler Composite prices for November 30, 2023, to January 31, 2024, let's first analyze the historical trends from previous years in these months.

Here's a summary of the data for November, December, and January over the years:

- Nov 2015: 196.99
- Dec 2015: 193.15
- Jan 2016: 192.66
- ...
- Nov 2022: 252.10
- Dec 2022: 247.03
- Jan 2023: 242.33

There appears to be a pattern where November values are not substantially different from December, and changes into January can vary. Recent years indicate a trend of increasing prices into December followed by a modest decrease in January.

For projection, considering recent trends showing an upward movement in prices, we'll apply a similar rate of change to the most current data points.

Given that the last available price is for October 2023 at 247.64, predicting a reasonable increase in November, a slight upward or stable trend in December, and then a small drop or stabilization in January:

[PRED_START]

2023-11-30: 249.75

2023-12-31: 250.85

2024-01-31: 248.50

[PRED_END]"

Figure 14: Reasoning annotated example of unimodal agriculture.

```

## Question ##
"As an expert in the US Retail Broiler Composite, predict the
  trends from 2023-11-30 to 2024-01-31
based on both historical data and contextual information.

Historical data (chronological order):
2015-11-30: 196.99; 2015-12-31: 193.15; 2016-01-31: 192.66; ...
2023-09-30: 246.16; 2023-10-31: 247.64

Contextual information:
2021-12-30: The United States broiler market has seen a 2%
  decrease in broiler-type chicks placed, ...
2023-10-26: The annual broiler per-capita consumption is
  reported in the Today's Beef Consumer Fall 2023 Update.

Please enclose your final predictions between [PRED_START] and [
  PRED_END] markers exactly like this:
[PRED_START]
2024-01-01: 123.45
2024-02-01: 124.56
[PRED_END]"

## Answer ##
"Based on the historical data and contextual information
provided, here is my prediction for the US Retail Broiler
Composite from 2023-11-30 to 2024-01-31:\n\n1. **Historical
Data Analysis**:\n  - Historically, prices have exhibited
seasonal patterns with a tendency to decrease towards the end
of the year and then either stabilize or slightly increase
early in the year.\n  - November to January has shown some
degree of variability, with occasional drops in prices in
December followed by modest increases or stability in January
.\n\n2. **Contextual Information**:\n  - The US broiler
market faces various influencing factors like international
demand, supply chain disruptions, labor shortages, and prices
of feed like corn.\n  - The ongoing issues such as price-
fixing charges among major poultry producers and bird flu
have potential impacts on both supply and prices.\n  - The
US market is connected to global trends, including export
demands which are showing strong signals for growth in 2024.\n\n3. **Market Trends**:\n  - From recent reports, there is
an expectation of a higher supply of chicken meat but with
weaker domestic demand, potentially leading to slight price
reductions.\n  - However, the international market remains
robust, and overall inflation pressures could balance out
some of the downward trends.\n\n4. **Conclusion**:\n  -
Given the current market conditions, we can expect a slight
decrease in prices towards the end of 2023 compared to the
middle of the year. However, prices might stabilize or
slightly increase in January 2024 as the market adjusts to
new supply-demand dynamics and international pressures.\n\n[
PRED_START]\n2023-11-30: 243.80\n2023-12-31: 241.50\n2024
-01-31: 243.10\n[PRED_END]"

```

Figure 15: Reasoning annotated example of multimodal agriculture.

```

## Question ##
"As an expert in the US Drought Level, predict the trends from
  2023-11-30 to 2024-01-31
based on the historical data.

Historical data (chronological order):
2015-11-30: 3.35; 2015-12-31: 4.05; 2016-01-31: 2.04; ...
2023-09-30: 2.12; 2023-10-31: 2.17

Please enclose your final predictions between [PRED_START] and [
  PRED_END] markers exactly like this:
[PRED_START]
2024-01-01: 123.45
2024-02-01: 124.56
[PRED_END]"

## Answer ##
"To predict the US Drought Level trends from November 30, 2023,
  to January 31, 2024,
we need to analyze historical seasonal patterns and recent
  trends.

### Seasonal Trends and Analysis:
1. **Late Fall to Early Winter (Nov to Jan):**
  - Historically, November to January sees diverse behavior but
    often a decrease from late fall to early winter.
  - Novembers have shown values like 1.75, 1.62, 2.71, 1.94,
    1.94, 1.30, 2.44.
  - December and January values usually see some fluctuations,
    but generally, there's an inclination towards
    stabilizations or mild decreases.

2. **Recent Yearly Trends:**
  - 2022 and 2023 show end-of-year levels at:
    - 2022: 1.68 (Oct) to 2.44 (Nov to Dec) to 2.94 (Jan)
    - 2023: 2.17 (Oct latest data point)
  - We notice a slight rise in early winter periods, following
    potential weather patterns.

3. **Data Synthesis:**
  - Given slight upticks as we move into winter and the fact
    that the October 2023 value is 2.17, we can predict a
    modest rise by end of November.
  - Typical increases from November to January stay within a
    moderate range, respecting the patterns seen in similar
    past periods.

### Predictions:
Considering the above factors, here are the expected values:

[PRED_START]
2023-11-30: 2.35
2023-12-31: 2.55
2024-01-31: 2.75
[PRED_END]

This prediction reflects historical patterns, recent trends, and
  typical seasonal behaviors."

```

Figure 16: Reasoning annotated example of unimodal TSF for US Drought Level.

Question

"As an expert in the US Drought Level, predict the trends from 2023-11-30 to 2024-01-31 based on both historical data and contextual information.

Historical data (chronological order):

2015-11-30: 3.35; 2015-12-31: 4.05; 2016-01-31: 2.04; ...
2023-09-30: 2.12; 2023-10-31: 2.17

Contextual information:

2021-12-30: Historical temperature, precipitation, and Palmer drought data from 1895 to present for climate divisions, states, and regions in the contiguous U.S. are available.

2023-10-26: The contiguous U.S. had its 18th-warmest October in the 129-year record.

Please enclose your final predictions between [PRED_START] and [PRED_END] markers exactly like this:

[PRED_START]

2024-01-01: 123.45

2024-02-01: 124.56

[PRED_END]"

Answer

"Based on the historical data and contextual information provided, I will make predictions for the US Drought Level for the period from 2023-11-30 to 2024-01-31. I will take into account the trends in recent years and the seasonal patterns observed in the historical data, as well as the latest contextual information.

- Historical data suggests seasonal variations, with drought levels typically higher in the winter months.
- Recent contextual data indicate an expected wetter-than-normal winter for the southern contiguous U.S., which may alleviate drought conditions in those areas while potentially leading to more pronounced drought in others.

Considering these factors, here are the predicted drought levels:

[PRED_START]

2023-11-30: 2.25

2023-12-31: 2.55

2024-01-31: 2.35

[PRED_END]"

Figure 17: Reasoning annotated example of multimodal TSF for US Drought Level.