

# TimeOmni-1: Incentivizing Complex Reasoning with Time Series in Large Language Models

Tong Guan<sup>1,2</sup>, Zijie Meng<sup>2</sup>, Dianqi Li, Shiyu Wang, Chao-Han Huck Yang<sup>3</sup>  
Qingsong Wen<sup>4</sup>, Zuozhu Liu<sup>2</sup>, Sabato Marco Siniscalchi<sup>5,6</sup>, Ming Jin<sup>1\*</sup>, Shirui Pan<sup>1\*</sup>

<sup>1</sup> Griffith University <sup>2</sup> Zhejiang University <sup>3</sup> NVIDIA  
<sup>4</sup> Squirrel Ai Learning <sup>5</sup> University of Palermo <sup>6</sup> Norwegian University of Science and Technology

**Presenter: Tong Guan    \*Correspondence: Ming Jin, Shirui Pan**



**Paper**



**Model**



**Demo**



# Outline

---

## Part 1

### Introduction

- Background
- Motivation
- Contribution

## Part 2

### Methodology

- Two design principles for genuine time series reasoning
- Reasoning-critical task formulation: three core capabilities and four atomic tasks
- TSR-Suite: data construction pipeline
- TimeOmni-1: model overview

## Part 3

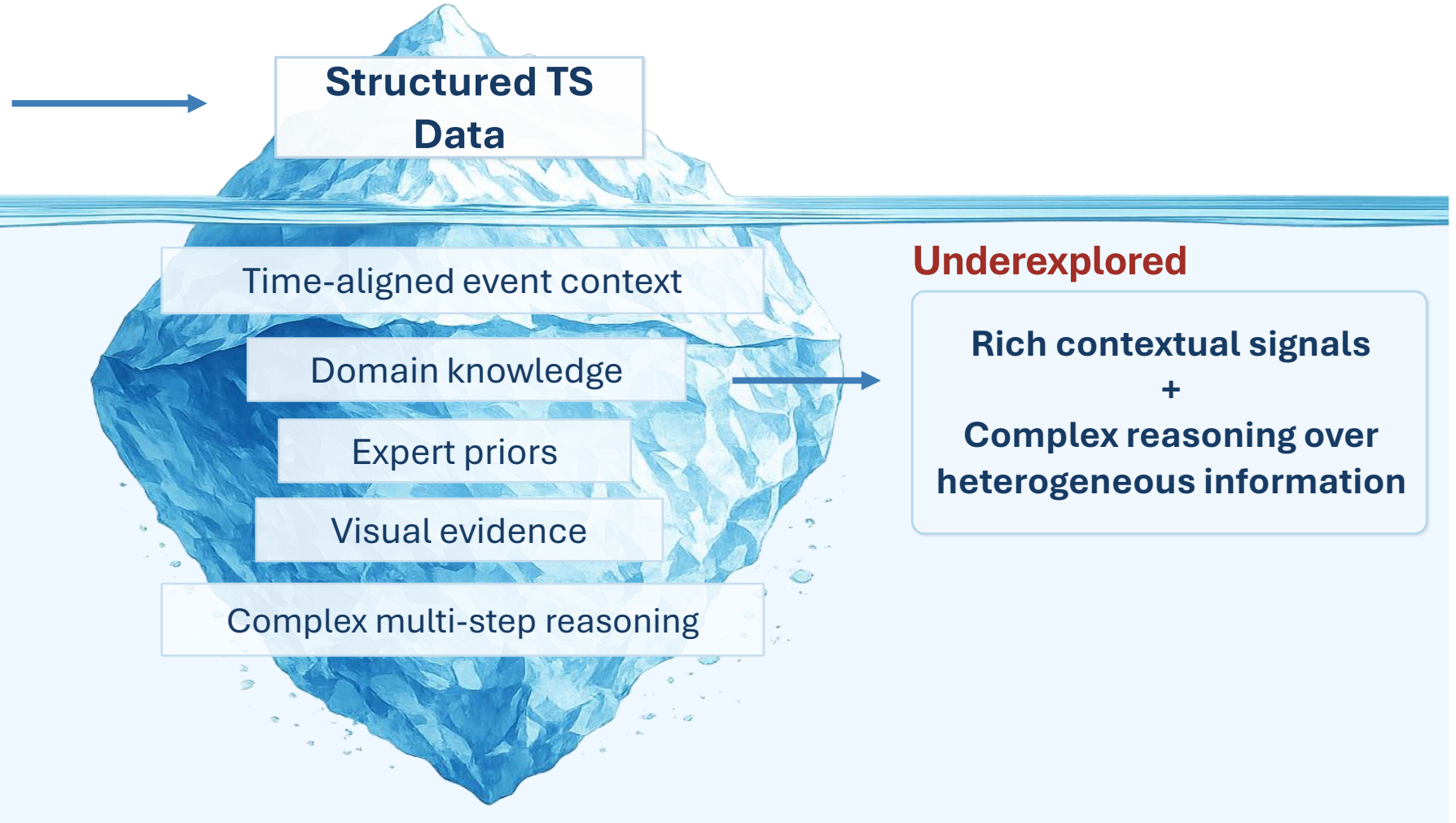
### Experiments

- Main results
- Further analysis
- Four key findings
- Takeaways

# Background

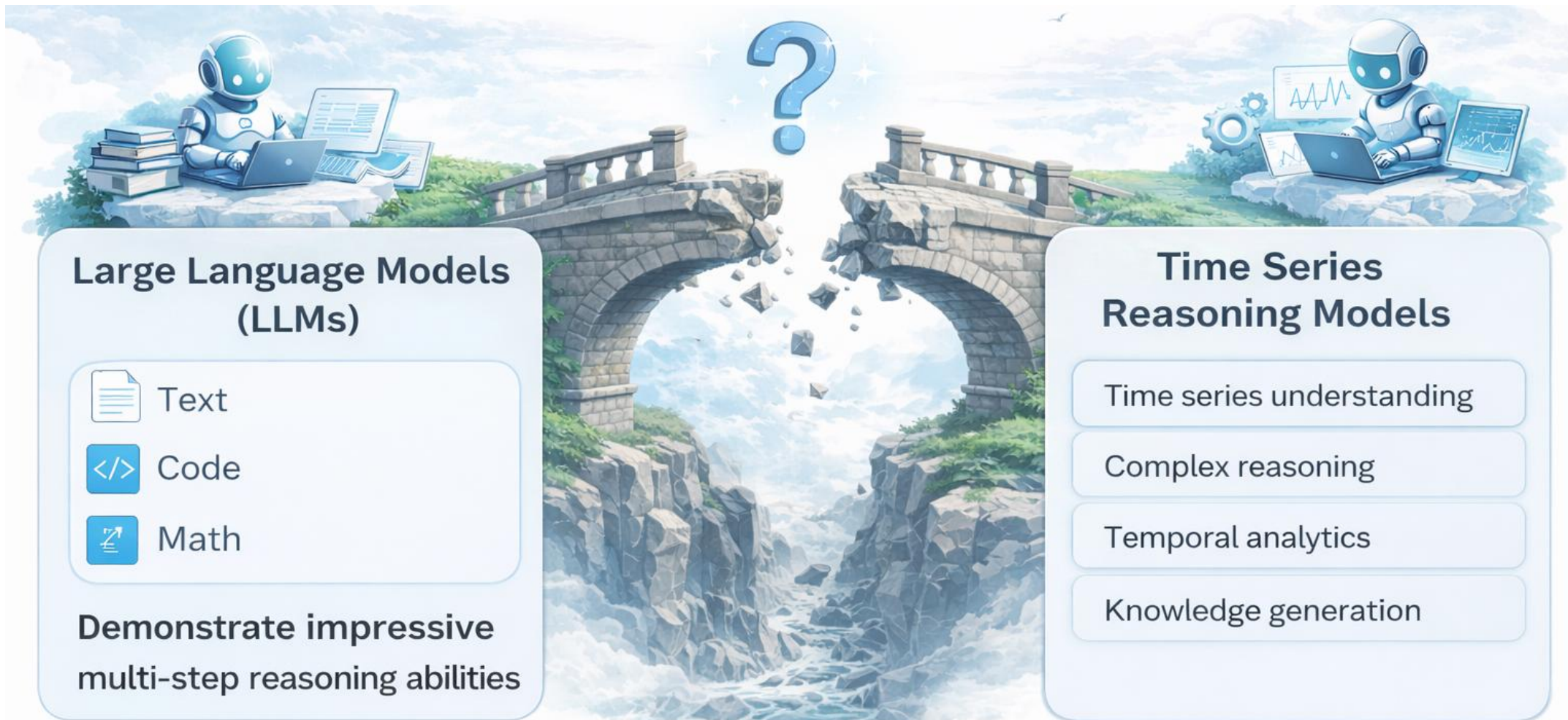
- Basic pattern analytics over structured time series (TS) data is not enough for general-purpose time series AI.

**Current time series analysis mainly focuses on**



# Background

- Despite the impressive reasoning abilities of LLMs, complex time series reasoning is still largely underexplored.



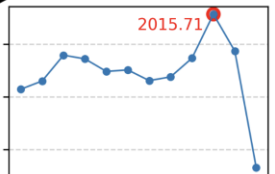
# Motivation

- What hinders the development of time series reasoning models (TSRMs)?

## Limitation 1:

The scarcity of high-quality data to support general-purpose reasoning over time series.

The data point 2015.71 is significantly higher compared to others in its immediate neighborhood in the series. True or False?



True.

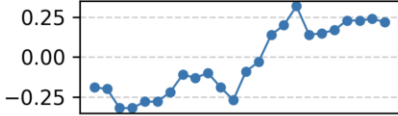
*Overthinking*

```
<think>The question is about whether the point 2015.71 is significantly higher compared to its immediate neighbors in the provided series. The series is \[2001.46, 2003.03, 2007.93, 2007.25, 2004.84, 2005.15, 2003.1, 2003.81, 2007.36, 2015.71, 2006.7, 1986.59]. The value of interest 2015.71 is at index 9 (0-based). Its immediate neighbors are 2007.36 (previous) and 2006.7 (next). Comparing: 2015.71 - 2007.36 = 8.35, and 2015.71 - 2006.7 = 9.01. Both are relatively large differences compared to typical fluctuations in the series (mostly 2000-2008 range). So yes, 2015.71 is clearly higher than its immediate neighbors. Hence the answer is True.</think> <answer>True</answer>
```

Overly simple time series questions do not require reasoning, and forcing reasoning may trigger overthinking

Analyze the series for volatility. Which of the following best describes the volatility of this segment? a) Highly volatility b) Moderately volatility c) Low volatility

Ambiguous Task Design:  
Lacking Clear Volatility Thresholds



<think>Early part stays around -0.3 to -0.1, later part shifts to positive values up to +0.32. There is a trend, but no abrupt extreme spikes. Standard deviation would be moderate, probably ~0.2.</think>  
<answer>b)Mod.</answer>

<think>The values move smoothly from slightly negative toward positive without sharp jumps; the swings are modest and fairly contained within ±0.3. Best characterized as low volatility.</think>  
<answer>c)Low</answer>

Models Are Guessing the Boundary of Volatility

Insufficient context leads to ambiguous

## Limitation 2:

The lack of a validated and feasible pathway for effective time series reasoning across tasks.

# Contribution

---

## 1. New Datasets and Testbed TSR-Suite

The first comprehensive suite for time series reasoning, covering **4 atomic tasks** across **3 core capabilities: perception, extrapolation, and decision-making**. TSR-Suite contains **23,605 QA** pairs from 10 domains, including **2,339 human-guided** hierarchical annotations for reasoning traces.

## 2. New Models TimeOmni-1

A unified model for time series reasoning, trained in two stages: **temporal-prior** SFT first, **task-grounded** RL second.

## 3. Comprehensive Evaluation and Key Insights

**Strong performance** on both in-distribution (ID) and out-of-distribution (OOD) settings, with evidence that **joint training across reasoning tasks yields mutual gains**.

# Outline

---

## Part 1

### Introduction

- Background
- Motivation
- Contribution

## Part 2

### Methodology

- Two design principles for genuine time series reasoning
- Reasoning-critical task formulation: three core capabilities and four atomic tasks
- TSR-Suite: data construction pipeline
- TimeOmni-1: model overview

## Part 3

### Experiments

- Main results
- Further analysis
- Four key findings
- Takeaways

# Two Principles for Genuine Time Series Reasoning

- What Makes a Time Series Task Truly Reasoning Critical?

## Principle 1

### QA pairs must reward reasoning

A task is reasoning critical only if reasoning models (RM) substantially outperform non-reasoning models (NRM). Otherwise, explicit reasoning does not meaningfully help.

#### Desired signal

$$\bar{S}(M_{RM}) \gg \bar{S}(M_{NRM})$$

## Principle 2

### QA pairs must ensure context sufficiency

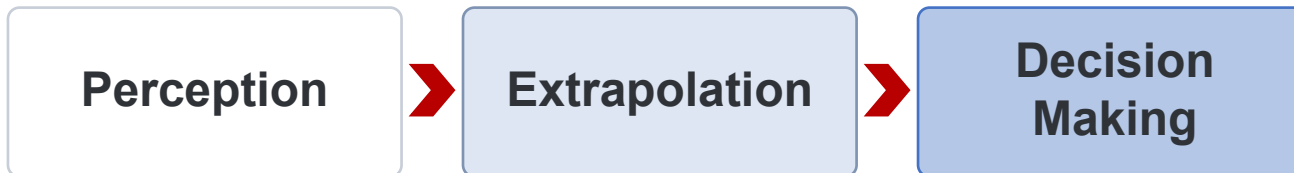
Reasoning depends on both the time series input  $X$  and auxiliary context  $C$ . Let  $K$  denote the number of ambiguous options. If either is underspecified, even an ideal reasoner is forced to guess.

#### Desired signal

$$\lim_{RC \rightarrow \infty} P(\text{correct} \mid X, C) \begin{cases} \approx \frac{1}{K}, & \text{if } X \text{ or } C \text{ is underspecified,} \\ \gg \frac{1}{K}, & \text{if } X \text{ and } C \text{ are sufficient.} \end{cases}$$

# Reasoning-Critical Time Series Tasks Formulation

- Three Core Capabilities of Time Series Reasoning



- Four Atomic Tasks

## 1 Scenario Understanding

Link one series to its most plausible real-world scenario or event.

## 2 Causal Discovery

Compare multiple series and infer directed causal relations.

## 3 Event-Aware Forecasting

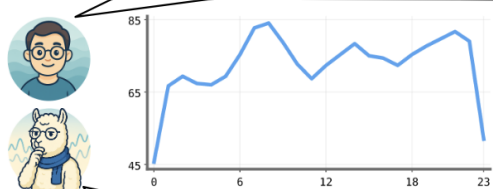
Forecast future trajectories under explicit event perturbations.

## 4 Decision Making

Use perception + forecasting to pick actions that maximize utility.

### Task 1: Scenario Understanding

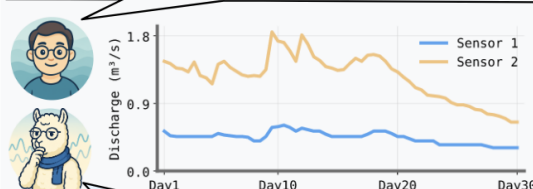
Please identify the scenario that most likely created the given time series.



- 1 Series length check.
  - 2 Magnitude & unit sanity.
  - 5 Final elimination & plausibility.
- A. Monthly humidity over two years; spikes July-November.

### Task 2: Causal Discovery

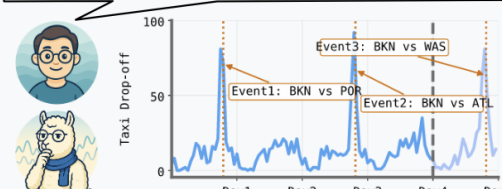
Does the discharge at one hydrometric station causally influence the other?



- 1 Trend Consistency: Closely aligned trends.
  - 2 Key Fluctuation Alignment: Major peaks and dips align.
  - 3 Direction of Causality: Mean discharges: EVBS < YEV6.
- B. Yes. Series1 → Series2.

### Task 3: Event-Aware Forecasting

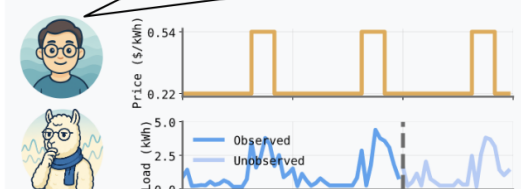
Based on the events and series, please predict for next 12 hours.



- 1 Identify the baseline patterns and trend.
  - 2 Estimate the incremental impact of any events.
  - 3 Combine the baseline and event effects.
- [2, 20, 21, 23, 19, 18, 16, 22, 31, 45, 63, 83]

### Task 4: Decision Making

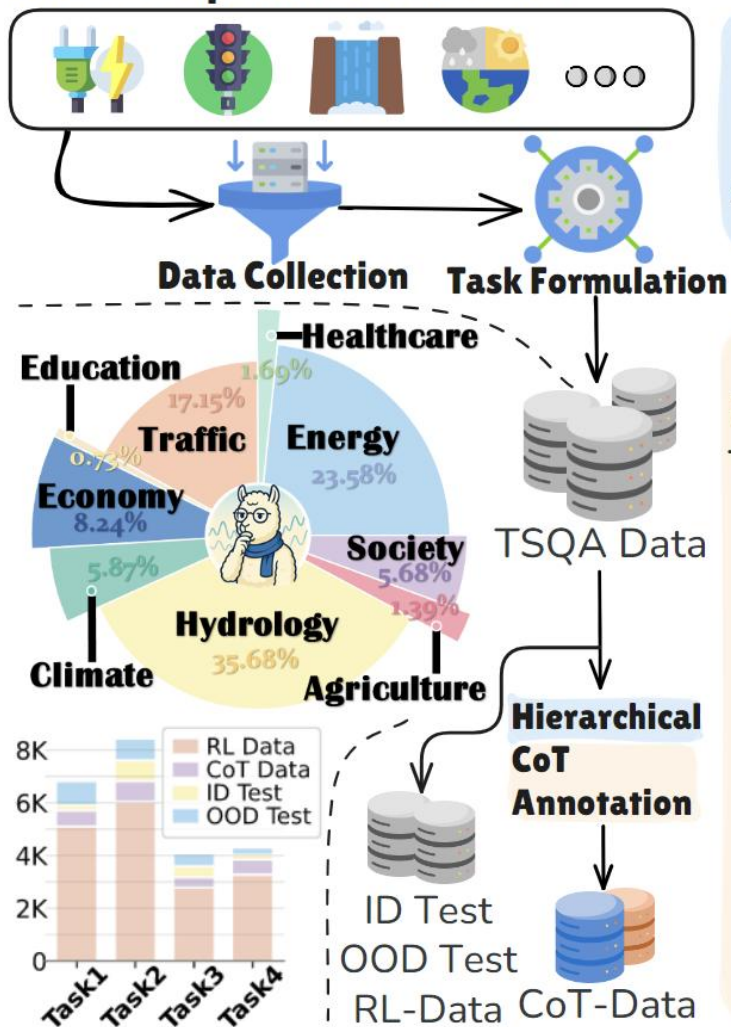
Using the observed Load series, plan tomorrow's battery charge-discharge.



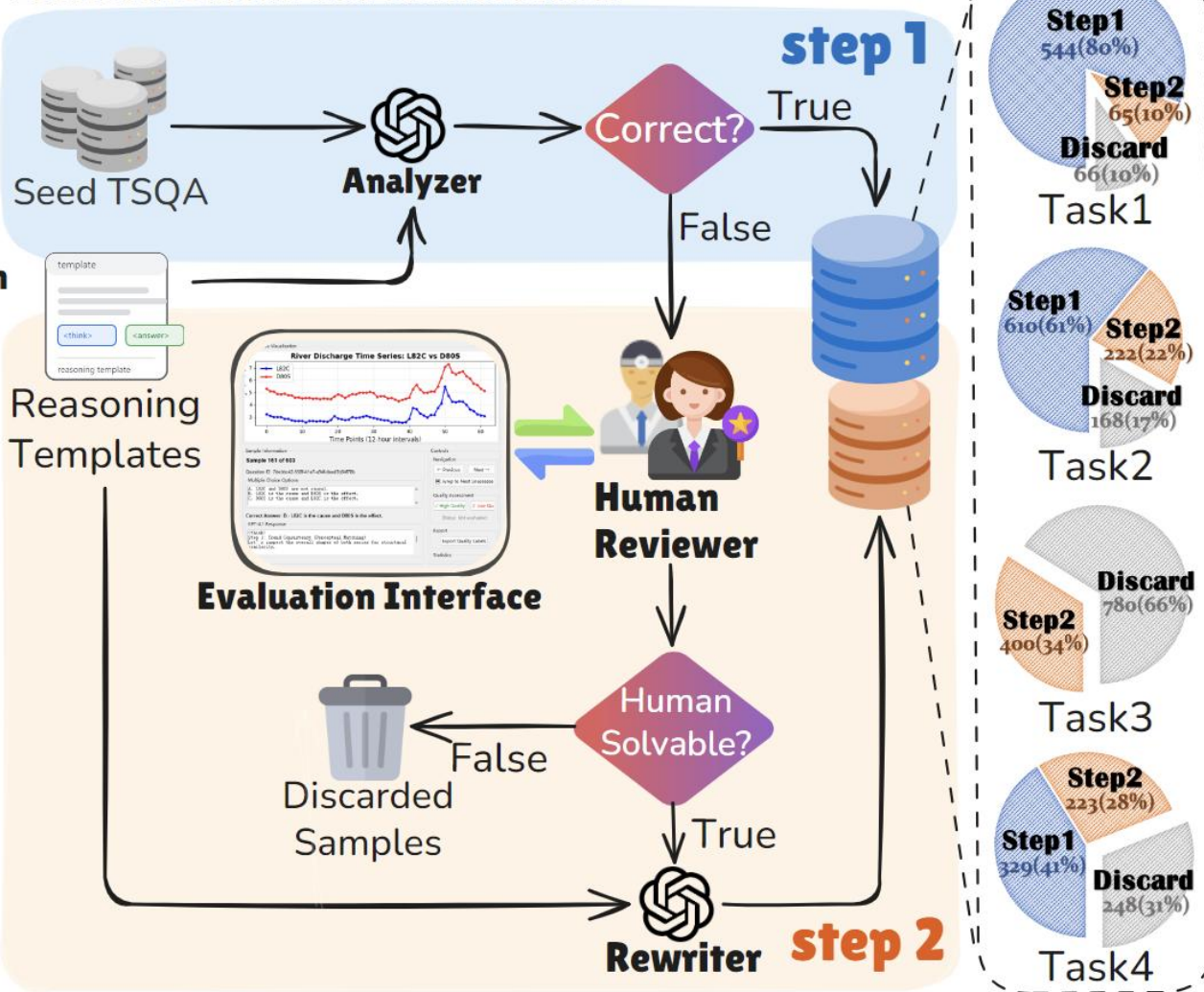
- 1 Forecast the next 24-hour load.
  - 2 Principles for evaluating strategies.
  - 3 Cost calculation and strategy comparison.
- C. Hour 1 charging, hour 8 charging, hours 18-19 discharging.

# TSR-Suite: Data Construction Pipeline

## A. Data Pipeline of TSR-Suite

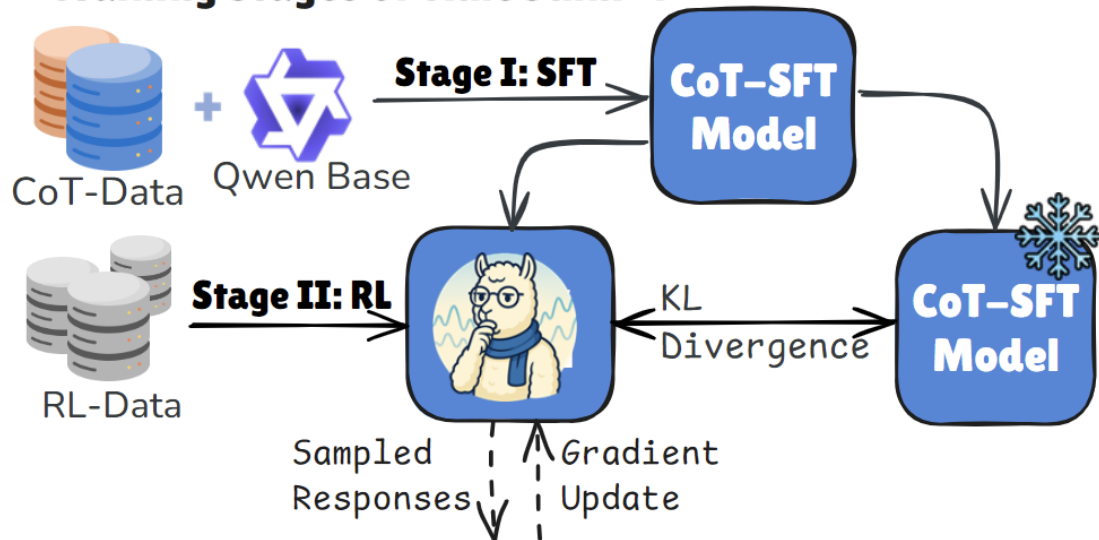


## B. Hierarchical CoT Annotation



# TimeOmni-1: A Unified Model for Time Series Reasoning

## Training Stages of TimeOmni-1



**Stage 1**  
Temporal Prior Supervised Fine-Tuning



**Stage 2**  
Task-Grounded Reinforcement Learning

<b>Reward Evaluation</b>	<b>Format Reward</b>	<b>Task1,2,4</b>	1	0	$\mathcal{R}_{\text{discrete}} = \begin{cases} 1, & \hat{y} = y \\ 0, & \text{otherwise} \end{cases}$
	$\mathcal{R}_{\text{format}}$	<b>Task3</b>	(0, 1.1]	0	$\mathcal{R}_{\text{count}} = 0.1$
	1				
	0				$\mathcal{R}_{\text{sequence}} = \begin{cases} 0, & \text{if } \text{len}(\hat{y}) \neq \text{len}(y) \\ e^{-\alpha \cdot \text{MAE}} + \mathcal{R}_{\text{count}}, & \text{if } \text{len}(\hat{y}) = \text{len}(y) \end{cases}$

# Outline

---

## Part 1

### Introduction

- Background
- Motivation
- Contribution

## Part 2

### Methodology

- Two design principles for genuine time series reasoning
- Reasoning-critical task formulation: three core capabilities and four atomic tasks
- TSR-Suite: data construction pipeline
- TimeOmni-1: model overview

## Part 3

### Experiments

- Main results
- Further analysis
- Four key findings
- Takeaways

# Main Results

Table 1: Model comparison on four tasks, reporting ACC for Tasks 1, 2, and 4, MAE for Task 3, together with SR under ID and OOD testbeds. All metrics are computed only on valid responses. “–” denotes SR below 10%, where results are omitted for lack of statistical significance and therefore we choose not to report them. **Red**: the best, **Blue**: the 2nd best.

Method	Base LLMs	Scenario Understanding				Causality Discovery				Event-aware Forecasting				Decision Making			
		ID		OOD		ID		OOD		ID		OOD		ID		OOD	
		ACC	SR%	ACC	SR%	ACC	SR%	ACC	SR%	MAE	SR%	MAE	SR%	ACC	SR%	ACC	SR%
<b>Proprietary Models</b>																	
GPT-4.1-2025-04-14		<a href="#">85.5</a>	100.0	<b>87.8</b>	100.0	28.7	99.9	<a href="#">35.9</a>	100.0	<b>13.79</b>	97.4	170.78	76.1	25.5	100.0	27.8	100.0
GPT-4.1-Nano		66.2	97.5	62.6	98.7	29.8	98.6	28.0	98.4	18.98	92.8	170.78	76.1	28.9	99.5	34.1	97.8
<b>Open-Source Models</b>																	
Llama-3.1-70B-Instruct		56.4	100.0	59.6	100.0	23.4	100.0	28.9	99.9	24.67	92.8	238.98	97.0	20.3	96.8	17.7	97.4
Mistral-Small-3.1-24B-Ins		64.8	100.0	69.2	100.0	24.6	100.0	25.8	100.0	17.28	72.0	<b>120.97</b>	43.4	<a href="#">44.7</a>	100.0	<a href="#">42.9</a>	100.0
Llama-3.1-8B-Instruct		36.6	46.5	32.1	46.8	-	3.7	-	1.9	27.68	52.91	186.80	29.8	7.4	28.7	16.2	42.9
Mistral-7B-v0.3		40.5	92.2	34.7	87.6	29.0	86.0	26.9	82.6	-	5.3	-	0.0	24.3	94.2	16.7	96.7
Qwen2.5-Instruct-7B		48.5	100.0	42.8	100.0	21.6	99.8	26.3	100.0	23.28	53.1	146.12	55.46	25.5	100.0	24.9	100.0
<b>Time Series Language Models</b>																	
Time-MQA	Llama3-8B	32.2	29.5	25.1	32.6	30.1	44.3	31.2	37.2	-	1.4	-	0.4	12.0	13.3	11.6	15.8
Time-MQA	Mistral-7B-v0.3	15.1	21.5	27.8	22.1	8.4	50.2	4.0	52.2	-	0.2	-	0.0	5.4	36.1	10.0	47.3
Time-MQA	Qwen2.5-7B	25.0	14.0	37.5	22.7	29.5	33.0	30.5	32.0	19.76	12.2	-	6.5	23.8	58.0	26.4	44.3
ChatTS		-	6.0	-	6.9	18.2	30.1	18.6	26.7	-	0.0	-	0.0	5.8	27.1	11.1	27.1
<b>Time Series Reasoning Models</b>																	
Time-R1	Qwen2.5-Instruct-7B	30.9	94.0	34.0	92.5	<a href="#">30.2</a>	53.8	31.4	48.9	17.61	38.7	-	6.3	27.8	95.7	32.2	93.1
<b>Ours</b>																	
TIMEOMNI-1	Qwen2.5-Instruct-7B	<b>90.7</b>	97.5	<a href="#">87.7</a>	98.3	<b>69.3</b>	99.8	<b>64.0</b>	99.8	<a href="#">14.30</a>	93.8	<a href="#">145.53</a>	82.3	<b>47.9</b>	100	<b>58.9</b>	100

# Further Analysis

Table 2: Comparison across training stages and training strategies. **Red**: the best, **Blue**: the 2nd best.

Training Stage	Training Strategy	Scenario Understanding				Causality Discovery				Event-aware Forecasting				Decision Making			
		ID		OOD		ID		OOD		ID		OOD		ID		OOD	
		ACC	SR%	ACC	SR%	ACC	SR%	ACC	SR%	MAE	SR%	MAE	SR%	ACC	SR%	ACC	SR%
<b>Base Model</b>	<i>Qwen2.5-Instruct-7B</i>	48.5	100.0	42.8	100.0	21.6	99.8	26.3	100.0	23.28	53.1	<b>146.12</b>	55.5	25.5	100.0	24.9	100.0
<b>ANS-SFT</b>	Single-task	77.5	100.0	73.9	100.0	35.7	100.0	33.8	100.0	23.87	39.7	150.42	0.6	20.2	100.0	24.2	100.0
	Multi-task Joint	80.5	100.0	76.6	100.0	30.5	100.0	34.6	100.0	<b>13.23</b>	48.3	156.78	38.9	<b>51.0</b>	100.0	58.2	100.0
<b>CoT-SFT</b>	Single-task	73.9	100.0	85.6	83.9	66.3	96.0	<u>61.8</u>	92.4	15.10	64.6	157.21	34.5	39.4	98.40	47.3	94.87
	Multi-task Joint	<u>89.8</u>	88.0	85.8	88.2	<u>67.7</u>	98.8	61.1	97.0	15.53	71.8	151.41	55.7	40.9	100.0	52.7	100.0
<b>CoT-SFT+RL</b>	Single-task	82.5	100.0	<u>87.0</u>	98.7	67.5	99.6	61.7	99.0	16.76	79.2	169.88	66.0	40.9	100.0	<b>63.6</b>	99.6
	Multi-task Joint	<b>90.7</b>	97.5	<b>87.7</b>	98.3	<b>69.3</b>	99.8	<b>64.0</b>	99.8	<u>14.30</u>	93.8	<u>145.53</u>	82.3	<u>47.9</u>	100.0	<u>58.9</u>	100.0

# Four Key Findings

1

## Finding 1

Time series reasoning ability need not be innate; it can be effectively cultivated via SFT on a small set of high-quality, curated reasoning traces.

2

## Finding 2

Human-guided traces establish decomposition priors critical for time series reasoning.

3

## Finding 3

Reinforcement learning works reliably only once the base model is anchored with fundamental temporal priors, which prevent collapse into spurious exploration.

4

## Finding 4

Joint training turns perception, extrapolation, and decision-making from silos into complementary capabilities, supporting a “train-once, use-across-tasks” paradigm for TSRMs.

# Four Key Findings

## Finding 1

Time series reasoning ability need not be innate; it can be effectively cultivated via supervised fine-tuning on a small set of high-quality, curated reasoning traces.

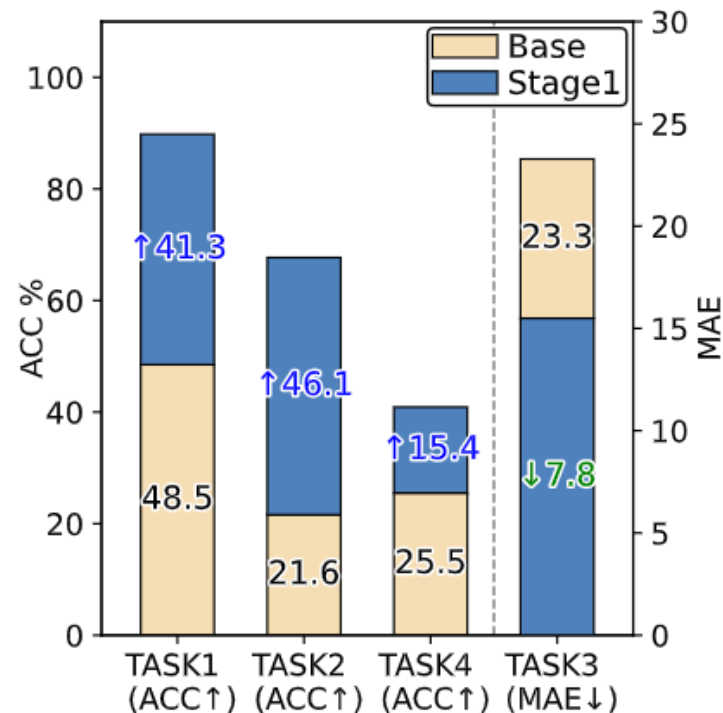


Figure 4: Stage 1 boosts accuracy; Base model at chance.

# Four Key Findings

## Finding 2

Human-guided traces establish decomposition priors critical for time series reasoning.

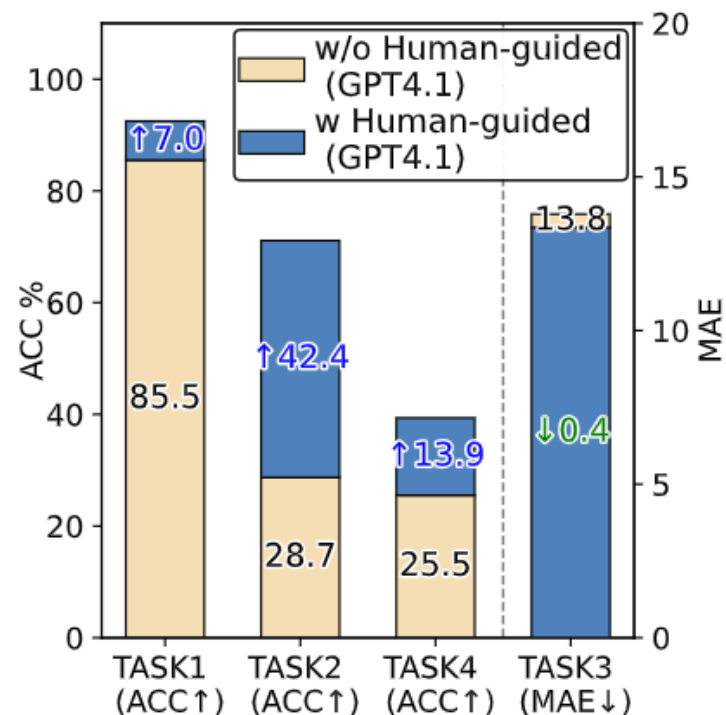


Figure 5: Human-guided templates are critical for priors.

# Four Key Findings

## Finding 3

Reinforcement learning works reliably only once the base model is anchored with fundamental temporal priors, which prevent collapse into spurious exploration.

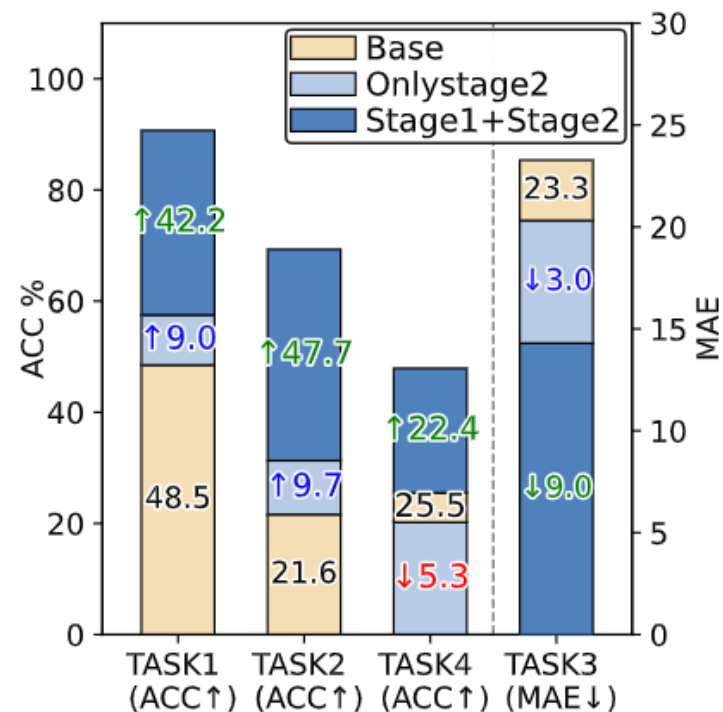


Figure 6: Stage 2 RL alone fails without Stage 1 priors.

# Four Key Findings

## Finding 4

Joint training turns perception, extrapolation, and decision-making from silos into complementary capabilities, supporting a train-once, use-across-tasks paradigm for TSRMs.

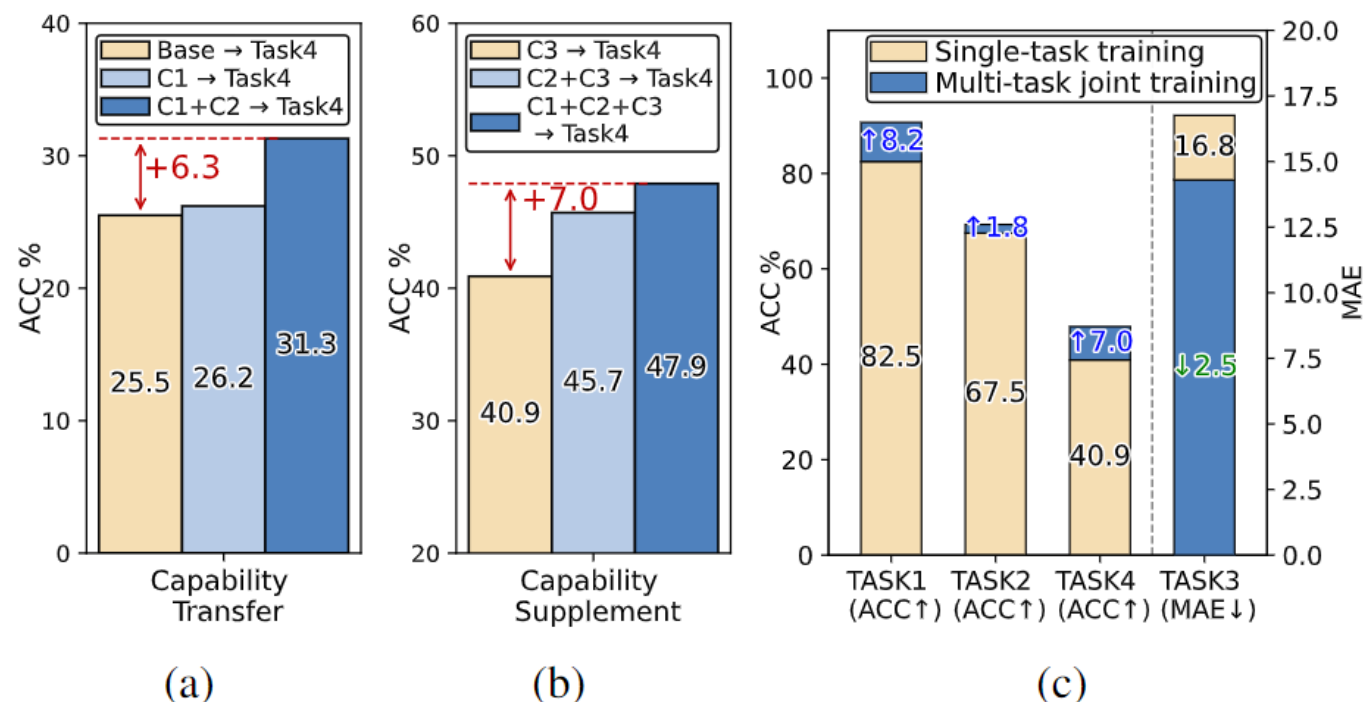


Figure 7: Joint training results: **(a)** Zero-shot capability transfer from precursor tasks (C1: Perception, tasks 1, task 2; C2: Extrapolation, task 3) to decision-making; **(b)** Supervised capability supplement by progressively incorporating tasks (C3: Decision-Making); **(c)** Joint training outperforms single-task training.

# Takeaways

---

## **TSR-Suite**

The first comprehensive time series reasoning suite formalizing three core capabilities: **perception**, **extrapolation**, and **decision-making**.

## **TimeOmni-1**

**The first generalized reasoning model** on time series data.

## **New Paradigm**

A “**train-once, use-across-tasks**” paradigm for future time series reasoning models.

# Thank you



Paper

Published as a conference paper at ICLR 2026



## TIMEOMNI-1: INCENTIVIZING COMPLEX REASONING WITH TIME SERIES IN LARGE LANGUAGE MODELS

Tong Guan<sup>1,2</sup> Zijie Meng<sup>2</sup> Dianqi Li Shiyu Wang Chao-Han Huck Yang<sup>3</sup>  
Qingsong Wen<sup>4</sup> Zuozhu Liu<sup>2</sup> Sabato Marco Siniscalchi<sup>5,6</sup> Ming Jin<sup>1\*</sup> Shirui Pan<sup>1\*</sup>

<sup>1</sup>Griffith University <sup>2</sup>Zhejiang University <sup>3</sup>NVIDIA <sup>4</sup>Squirrel Ai Learning  
<sup>5</sup>University of Palermo <sup>6</sup>Norwegian University of Science and Technology

TimeOmni-1 (Based on Qwen-2.5-Instruct): 7B

anton-hugging/TimeOmni-1-7B

Text Generation • 8B • Updated Feb 13 • 2.77k • 5

Model

