TSAQA: Time Series Analysis Question And Answering Benchmark

Anonymous ACL submission

Abstract

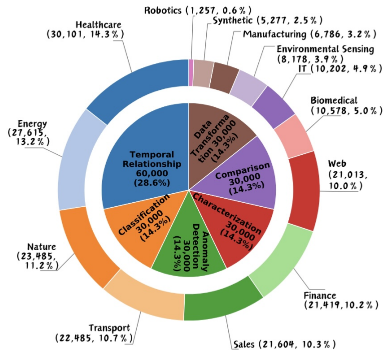
Time series data are integral to critical applications across domains such as finance, healthcare, transportation, and environmental science. While recent work has begun to explore multi-task time series question answering (QA), current benchmarks remain limited to forecasting and anomaly detection tasks. We introduce TSAQA, a novel unified benchmark designed to broaden task coverage and evaluate diverse temporal analysis capabilities. TSAQA integrates six diverse tasks under a single framework ranging from *conventional analysis*, including anomaly detection and classification, to *advanced analysis*, such as characterization, comparison, data transformation, and temporal relationship analysis. Spanning 210k samples across 13 domains, the dataset employs diverse formats, including *true-or-false (TF)*, *multiple-choice (MC)*, and a novel *puzzling (PZ)*, to comprehensively assess time series analysis. Zero-shot evaluation demonstrates that these tasks are challenging for current Large Language Models (LLMs): the best-performing commercial LLM, Gemini-2.5-Flash, achieves an average score of only 65.08. Although instruction tuning boosts open-source performance: the best-performing open-source model, LLaMA-3.1-8B, shows significant room for improvement, highlighting the complexity of temporal analysis for LLMs. The data are available in GitHub: `https://anonymous.4open.science/r/TSAQA-Benchmark-B86B`.

1 Introduction

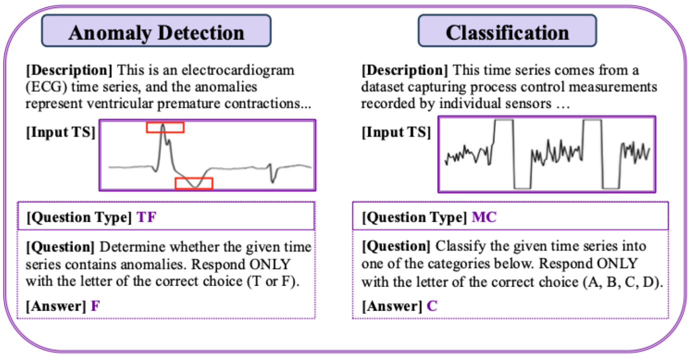
Effective analysis over temporal patterns of time series data is essential for real-world decision-making. Traditionally, research in time series has concentrated on a narrow set of tasks, most notably forecasting future values, anomaly detection, imputation, and classification (Torres et al., 2021; Lim and Zohren, 2021; Wen et al., 2022). While these problems have important applications, the scope of temporal analysis extends far beyond these settings, demanding deeper understanding of fundamental characteristics and patterns of time series.

Recent advances in Large Language Models (LLMs) have revolutionized natural language processing and multimodal learning (OpenAI et al., 2024; Grattafiori et al., 2024; Team et al., 2025b; Yang et al., 2025). This progress has inspired a growing interest in applying LLMs to time series analysis. Early studies have explored leveraging LLMs for traditional time series tasks, such as forecasting and anomaly detection (Zeng et al., 2023; Jin et al., 2023; Zhou and Yu, 2024; Zhang et al., 2024), leaving open to the question of whether LLMs can develop stronger temporal analysis abilities, such as understanding contextual information, and relationships across multiple time series.

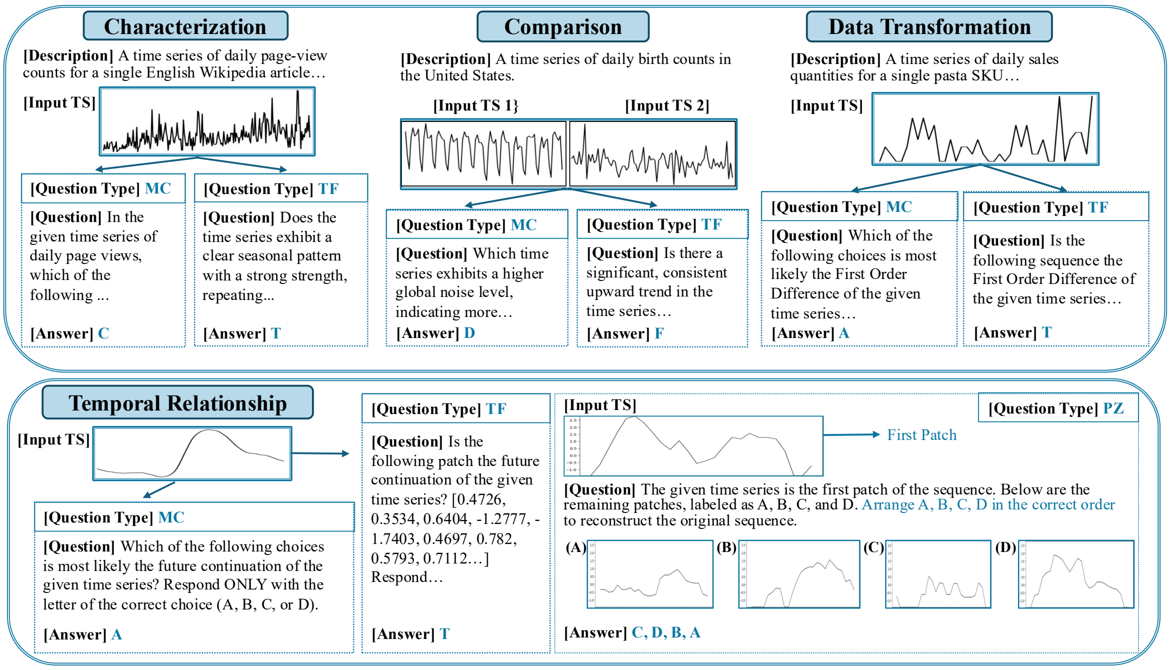
Time series question answering (QA) has recently emerged as a promising paradigm for pushing the boundaries of time series modeling beyond traditional tasks (Merrill et al., 2024; Uddin et al., 2025; Xu et al., 2025a; Zhong et al., 2025a; Wang et al., 2025a; Kong et al., 2025). By reformulating time series tasks through natural language queries, Time series QA enables models to tackle more complex questions about temporal patterns and dynamics, evaluating models' analytical capabilities. For example, ChatTS (Xie et al., 2025) generates situational questions based on *synthetic* time-series attributes. ITFormer (Wang et al., 2025b) introduces EngineMT-QA, which is a *domain-specific* dataset for aero engine time series. Mtbench (Chen et al., 2025) proposes a QA benchmark mainly for *forecasting* tasks. Time-MQA (Kong et al., 2025) constructs question–answer pairs that span both numeric tasks and open-ended QA tasks. While these efforts mark important progress, they are constrained by synthetic or domain-specific data and narrowly scoped tasks. Moreover, questions requiring open-ended answers remain difficult to standardize objectively, limiting fair evaluation across models.



(a) Domain and task distribution of TSAQA.



(b) Illustration of conventional tasks.



(c) Illustration of advanced tasks.

Figure 1: Data distribution and tasks of TSAQA.

In this paper, we introduce TSAQA, Time Series Analysis Question and Answering Benchmark, a large-scale unified benchmark that addresses these limitations by covering diverse domains and tasks, while also providing standardized evaluation protocols. A direct comparison between TSAQA and existing datasets is provided in Table 1. We curate and annotate 210k high-quality samples from 13 domains, as shown in Figure 1a. TSAQA integrates 6 distinct tasks grouped into two complementary categories: (1) *Conventional Analysis*: *Anomaly detection* and *Classification*. (2) *Advanced Analysis*: *Characterization*, *Comparison*, *Data Transformation*, and *Temporal Relationship*. All tasks are cast into a unified QA format with three question types: *true-or-false (TF)*, *multiple-choice (MC)*, and a novel *puzzling (PZ)*. We also carefully detail the *data collection process*, *benchmark construction*, *dataset statistics*, and *evaluation protocol*, ensuring rigorous transparency and reproducibility. Our benchmark provides a standardized platform to evaluate various LLMs (OpenAI et al., 2024; cla; Team et al., 2025b; Grattafiori et al., 2024; Yang et al., 2025). Initial empirical studies demonstrate that existing models struggle across several tasks, particularly in structural and relational reasoning, highlighting substantial future directions for improvement.

In summary, our main contributions are threefold. (1) We introduce TSAQA, a novel large-scale benchmark comprising 210k samples across 13 domains, covering 6 tasks and 3 types of questions. (2) We provide a detailed description of the benchmark's construction along with comprehen-

Table 1: Comparison of time series question answering datasets and benchmarks.

Dataset	Tasks Scope	# Analytical Tasks	# Question Type	# Domain	Size
TS-Insights (Zhang et al., 2023)	Captioning	1	1	7	100k
TSandLanguage (Merrill et al., 2024)	Forecasting	3	2	10	230k
CiK (Williams et al., 2025)	Forecasting	1	1	7	2.9k
MTBench (Chen et al., 2025)	Forecasting	4	3	2	42k
TimeSeriesExam (Cai et al., 2024)	Various	5	1	1	0.7k
ChatTS (Xie et al., 2025)	Various	4	5	4	2.2k
ITFormer (Wang et al., 2025b)	Various	4	2	1	11k
Time-MQA (Kong et al., 2025)	Various	5	4	12	200k
TSAQA (ours)	Various	**6**	**3**	**13**	**210k**

sive statistics. (3) We conduct extensive evaluations of TSAQA using a wide range of popular commercial and open-source LLMs, accompanied by an in-depth analysis of their performance.

2 Related Work

Time Series Analysis: From Numbers to Narratives. Traditional research on time series has primarily focused on numerical sequences, enabling core tasks such as forecasting (Torres et al., 2021), imputation (Wang et al., 2024), and classification (Mohammadi Foumani et al., 2024), often treating them as isolated numeric signals (Hamilton, 2020). In practice, however, time series are rarely independent of their surrounding context. They frequently interact with external information—such as textual reports, domain expertise, or heterogeneous side signals—that shapes or enriches their dynamics (Jiang et al., 2025; Xu et al., 2025b; Liu et al., 2025, 2024; Li et al., 2025). Recognizing this, recent work has moved beyond purely numeric modeling to incorporate multimodal signals across domains including healthcare (Johnson et al., 2016, 2023), finance (Li et al., 2024a; Dong et al., 2024), retail (Skenderi et al., 2024), and transportation (Li et al., 2024b). While much of this research leverages external modalities to improve numeric predictions on predefined tasks, a growing body of work instead positions *natural language as a richer interface* for time series, using language as the medium for querying, reasoning, and interpreting temporal patterns (Merrill et al., 2024; Williams et al., 2025; Wang et al., 2025b; Chen et al., 2025; Xie et al., 2025; Kong et al., 2025). Together, these efforts define the emerging direction of *time series question answering*.

Large Models on Time Series. Advances in large language models (LLMs) (Vaswani et al., 2017) have recently enabled general question answering over time series. A growing line of work integrates LLMs with time series for downstream tasks (Chang et al., 2023; Alnegheimish et al., 2024; Yu et al., 2023; Jin et al., 2023), with extensions to multimodal language models as well (Zhong et al., 2025b; Merrill et al., 2024; Moon et al., 2022). Given their strong generalization ability through natural language interfaces, comprehensive evaluation is critical to ensure the transparency and reliability of LLMs in time series applications.

3 TSAQA Benchmark

In this section, we introduce the proposed TSAQA benchmark, which is designed to provide a benchmark for time series question answering. We begin by formulating the tasks and defining question types in Section 3.1. Next, Section 3.2 describes the data sources and preprocessing procedures. Section 3.3 then details the construction of the benchmark, including its structure and design considerations. Data statistics are discussed in Section 3.4. Finally, Section 3.5 outlines the evaluation protocols used to assess model performance.

3.1 Task Formulation

Task Taxonomy. As shown in Table 2 and Figure 1, the proposed TSAQA benchmark encompasses two groups of tasks with six diverse tasks designed to evaluate a model's ability of understanding the fundamental properties of time series data. The first group, *Conventional Analysis*, includes fundamental tasks in time series analysis: (1) *Anomaly Detection* — identifies irregular or unexpected patterns in time series; (2) *Classification* — recognizes the distinguishable semantic pattern of a time series that is pertinent to a class. The second group, *Advanced Analysis*, consists of novel analytical tasks about intrinsic properties of time series: (3) *Characterization*— infers fundamental properties such as trend, seasonality, and dispersion; (4) *Comparison*— analyzes relative similarities and differences between two time series; (5)

Table 2: Tasks of TSAQA. TF, MC, and PZ denote true-or-false, multiple-choice, and puzzling.

Group	Task	Description	Question Type
Conventional Analysis	Anomaly Detection	Determine whether the input contains anomalies.	TF
	Classification	Classify the input time series.	MC
Advanced Analysis	Characterization	Determine the characteristics of the time series.	TF & MC
	Comparison	Compare the characteristics of two time series.	TF & MC
	Data Transformation	Identify the relationship between raw and transformed data.	TF & MC
	Temporal Relationship	Determine the temporal relationship of patches.	TF & MC & PZ

Data Transformation— understands relationships between original and transformed time series, e.g., Fourier transform; and (6) *Temporal Relationship*— captures the chronological dependencies among time series patches. These advanced tasks push the boundaries of conventional time series modeling, fostering the development of models that can grasp cognitive concepts of time series and analyze over human questions.

To bring all tasks under a single umbrella, we formulate them in a unified Question-Answering (QA) format. Every instance is converted into a time series input X paired with contextual information C and a question Q, and the model is expected to provide the correct answer A, where C and Q are expressed by natural language. Let f denote the model, then the TSAQA problem is formulated as:

$$A = f(X, C, Q). \quad (1)$$

Question Types. Our TSAQA benchmark encompasses a wide variety of question types, such as *true-or-false (TF)*, *multiple-choice (MC)*, and *puzzling (PZ)* questions. A *TF* question requires the model to determine whether a claim about the input time series is True (T) or False (F). A *MC* question requires the model to select the correct claim about the input. In addition, we introduce a novel *puzzling (PZ)* question, which are valuable because they represent realistic, human-like problem settings (Fissler et al., 2018) and have been shown to be effective in evaluating models' general cognitive abilities, as demonstrated in computer vision (Noroozi and Favaro, 2016). In this question, the model is given the first patch of a time series, along with the remaining shuffled patches, and is instructed to correct their chronological order.

3.2 Data Collection

To construct the TSAQA benchmark, we collect and preprocess time series data from diverse public sources to ensure broad coverage and representativeness. At its center are the core datasets, which serve as the primary foundation for a wide range of tasks. In addition, the benchmark integrates two specialized sources: classification datasets and anomaly detection datasets. This subsection describes these data sources and the selection criteria. More details can be found in Appendix A.

Core Datasets. We collect high-quality real-world time series data from a wide range of domains, including energy, finance, healthcare, nature, sales, transport, and web, which are used by time series foundation model benchmarks, such as Lotsa (Woo et al., 2024), Time-300B (Shi et al., 2024), and UTSD (Ma et al., 2024). To ensure data quality, we retain only sequences with a minimum length of 1k. We further filter sequences with a missing rate greater than 1% or an outlier rate (the proportion of points lying beyond three times the interquartile range (3×IQR)) exceeding 5% For each dataset, we refer to the original source to gather background information about the time series and provide a concise, one-sentence description. More details are presented in Appendix A.1

Anomaly Detection Datasets. We extract data from multiple time-series anomaly detection benchmarks (Paparrizos et al., 2022; Su et al., 2019), including ECG (Moody and Mark, 2001), SMD (Su et al., 2019), MGAB (Thill et al.) Genesis (von Birgelen and Niggemann, 2018), GHL (Filonov et al., 2016), Occupancy (Candanedo and Feldheim, 2016). These datasets span various domains. For each dataset, we summarize its description and domain information directly from the original papers. More details are presented in Appendix A.2.

Classification Datasets. Our classification data comes from the univariate UCR Archive (Dau et al., 2019a). We select datasets with at most four classes and sequence lengths under 400, and enrich them with textual descriptions from the official documentation. The resulting subset spans diverse domains. More details are presented in Appendix A.3.

3.3 Benchmark Construction

In this subsection, we describe the construction of the benchmark for each task. To maintain balance across tasks, we allocate an equal number of samples (30k) to each task, except for the temporal relationship task, to which we allocate 60k samples since *PZ* is very challenging. Except for classification and anomaly detection, samples for all other tasks are drawn from the *Core Datasets* (Section 3.2) using *Hierarchical Random Sampling* (Algorithm 1) to ensure a balanced distribution across domains, datasets, and sequences. Unless otherwise specified, each sample has a random length within $[32, 256]$, and is z-scored to reduce data bias (Appendix B.2). Finally, each task's samples are randomly partitioned into 70% for training, 10% for validation, and 20% for testing. Next, we describe the construction process for each task. More details and examples can be found in Appendix B-C.

Data Transformation. This task evaluates the model's ability to infer the transformation relationship between the input time series and its transformed counterpart, which is generated from the Fourier transform, wavelet transform, or first-order differencing. We then use predefined templates to formulate the task as either *TF* or *MC* questions. In *TF* questions, the model is asked to determine whether a given sequence is the correct transformation (e.g., the results of the Fourier transform) of the input time series $\mathbf{x}$. In *MC* questions, the model is required to select the correct transformed sequence given the input time series $\mathbf{x}$ and the specified transform operation (e.g., Fourier transform). All transformations are computed using professional libraries (Harris et al., 2020; Virtanen et al., 2020). The correct transformation is computed directly from the input $\mathbf{x}$, whereas incorrect transformations are generated from other randomly sampled time series $\mathbf{x}'$. For more details, please refer to Appendix B.5.

Temporal Relationship. The temporal relationship task evaluates the model's ability to infer the temporal structure among time series patches, testing 3 core capabilities: *Structural Continuity*, *Chronological Reasoning*, and *Contextual Discrimination*. This task is formulated as *TF*, *MC*, or *PZ* questions. Given the first chronological patch $\mathbf{x}$, a *TF* question asks the model to determine whether a candidate patch $\mathbf{y}$ is the immediate successor of $\mathbf{x}$, while an *MC* question asks the model to choose the correct next patch from candidates $[\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4]$. The false candidates are randomly sampled from the full dataset, but from sequences different from that of $\mathbf{x}$. A *PZ* question presents four shuffled successor patches of $\mathbf{x}$ and asks the model to arrange them in the correct chronological order. All questions are generated using predefined templates. See Appendix B.6 for more details.

Characterization. The characterization task assesses the model's capability to analyze fundamental properties of time series, including trend, seasonality, and dispersion. Questions are posed as *TF* or *MC*, and final answers are determined through multi-LLM consensus.

Given a sample $\mathbf{x}$ and its meta data, we first instruct GPT-4o (Hurst et al., 2024) to generate question and answer pairs based on a randomly selected subset of one to three topics (from Table 7), and question type (*TF* or *MC*). Briefly, the process involves the following steps: (1) We instruct GPT to generate captions for the input and randomly select a sub-topic for each topic (e.g., selecting the sub-topic "trend direction" under the topic "trend"); (2) GPT is instructed to generate a QA pair based on the inputs, captions, sub-topics, and the specified question type; (3) GPT performs a self-check of the generated QA pair and provides a confidence score, where only QA pair with a high confidence is retained; (4) We further leverage other powerful LLMs, including GPT-4.1, Gemini-2.5-Flash, and Claude-3.5-Sonnet, along with the answer given by GPT-4o to produce a consensus answer and reduce model bias. For more details, please refer to Appendix B.3.

Comparison. The comparison task assesses the model's ability to analyze the relative characteristics of the two time series, such as shape and correlation. Similar to the characterization task, this task is also formulated as *TF* or *MC* questions. We first obtain an anchor sample $\mathbf{x}$ from domain M, dataset D, and sequence S. Given the anchor $\mathbf{x}$, we construct a set of 10 comparison samples $\{\mathbf{x}'_1 \ldots \mathbf{x}'_{10}\}$ with the same length as $\mathbf{x}$. Among which, one is drawn from the sequence S, two from other sequences within dataset D, three from other datasets within domain M, and four from other domains. We also use a process similar to the characterization task to generate QA pairs. More details can be found in Appendix B.4.

Anomaly Detection. The anomaly detection task evaluates the model's ability to recognize anomalous patterns in the input time series, which is formulated as a *TF* question. Each sample $\mathbf{x}$ is randomly cropped from a full sequence of the anomaly detection dataset. Since anomalous samples are much fewer than normal ones, we downsample the

Table 3: Main results. A.D. denotes anomaly detection, CLS denotes classification. MC, TF, and PZ denote multiple-choice, true-or-false, and puzzling, respectively. SFT stands for supervised fine-tuning. The best and second-best results are highlighted in **bold** and underlined, respectively.

Group	Task	A.D.	CLS	Characterization		Comparison		Data Transform		Temporal Relation			Overall
	Question Type	MC	MC	TF	MC	TF	MC	TF	MC	TF	MC	PZ	
Zero Shot	GPT-4.1	55.85	50.38	**92.97**	**89.36**	83.57	76.99	54.36	51.13	65.90	79.09	45.77	62.82
	GPT-4o	54.32	47.20	88.15	84.15	78.61	69.07	60.66	53.24	62.25	75.58	45.61	60.73
	Claude-3.5-Sonnet	51.27	41.23	74.39	78.45	66.59	74.14	65.79	57.07	82.05	82.15	54.56	61.19
	Gemini-2.5-Flash	52.08	49.07	85.48	81.08	77.79	72.21	63.62	60.17	75.05	84.49	60.84	65.08
	Qwen3-8B	50.60	50.52	77.35	66.87	71.04	63.21	52.43	34.46	65.22	67.14	21.93	51.04
	LLaMA3.1-8B	54.92	50.20	68.10	62.26	67.84	49.98	51.90	36.56	54.82	40.95	6.80	44.93
	Ministral-8B	53.35	34.08	71.06	63.93	47.54	52.90	50.70	25.28	50.58	33.88	30.77	44.65
	Qwen3-0.6B	50.40	35.83	62.00	48.78	58.03	37.51	49.03	23.62	51.99	37.33	13.38	39.06
	LLaMA3.2-1B	49.47	39.48	63.74	52.55	61.02	36.82	48.87	4.20	48.97	5.44	6.76	35.70
	Gemma3-1B	49.15	49.83	63.74	47.71	61.19	43.37	49.37	24.88	49.42	25.84	23.97	43.03
Instruction Tuning	Qwen3-8B	87.70	90.05	92.37	85.42	86.55	79.08	89.84	84.99	96.84	**97.56**	66.21	84.29
	LLaMA3.1-8B	**91.02**	**91.27**	92.44	83.68	**86.72**	**79.31**	**90.17**	**86.62**	**96.94**	97.41	**67.68**	**85.26**
	Ministral-8B	71.56	74.28	91.31	80.78	84.14	74.63	75.15	71.61	94.07	94.15	56.82	74.74
	Qwen3-0.6B	83.68	85.78	89.38	74.87	80.65	64.84	80.51	73.28	93.92	93.79	63.34	78.32
	LLaMA3.2-1B	83.08	83.83	87.71	74.37	78.61	60.88	68.09	51.67	91.39	88.81	57.53	73.48
	Gemma3-1B	83.10	84.05	87.88	72.54	78.61	59.31	64.06	45.23	91.00	88.05	42.92	69.70

normal samples to balance the classes at a 1:1 ratio. The questions are composed using a predefined template. See Appendix B.7 for more details.

Classification. The classification task evaluates the model's ability to categorize input time series based on their patterns and characteristics. We reformulate the classification task into the *MC* question format, where the original numeric class, labels, e.g., 0 or 1, are converted into informative textual choices, e.g., "Cabernet Sauvignon" or "Shiraz". See Appendix B.8 for more details.

3.4 Data Statistics

Figure 1a shows the distribution of domains and tasks, and Figure 2d shows the distribution of question types. Samples are nearly balanced across tasks, question types, and major domains. Figures 2a–2c present the histograms of time series length, description length, and question length, all of which exhibit long-tail distributions.

3.5 Evaluation Protocol

The TSAQA benchmark includes three question types, each with a specific evaluation metric. *TF* and *MC* questions are evaluated using accuracy. *PZ* questions are scored by comparing each predicted position with the ground truth and computing the proportion of correct matches. For example, with a ground truth A, B, C, D and prediction B, A, C, D, only the last two match, yielding 50% accuracy.

4 Experiments

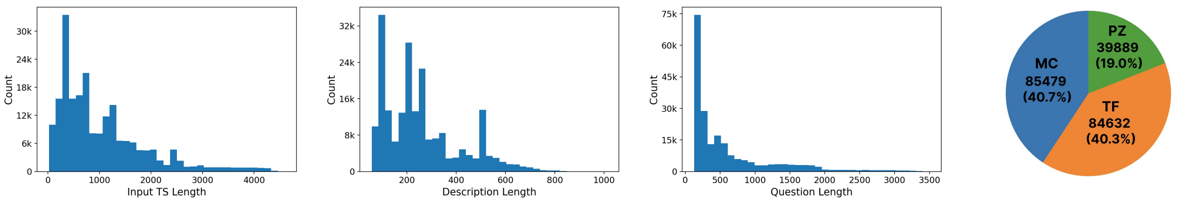
In this section, we present experimental results of both commercial and open-source LLMs on our TSAQA benchmark, and provide analysis of the results.

4.1 Main Results

We evaluate *zero-shot* performance of (1) commercial LLMs: GPT-4.1, GPT-4o, Claude-3.5-Sonnet and Gemini-2.5-Flash; (2) medium size open-source LLMs: Qwen3-8B (Yang et al., 2025), LLaMA3.1-8B (Dubey et al., 2024), Ministral-8B; (3) small size open-source LLMs: Qwen3-0.6B (Yang et al., 2025), LLaMA3.2-1B (Dubey et al., 2024), Gemma3-1B (Team et al., 2025a). We further apply *instruction tuning* (Peng et al., 2023) to the open-source methods using LoRA (Hu et al., 2022). We set LoRA rank as 16, fix learning rate as 10^{-5} with cosine schedule, and train models for 2 epochs on a single A100 GPU.

Overall Results. The rightmost column in Table 3 presents averaged results over all the samples (not simply over each row). (1) *Zero-shot*: Commercial LLMs consistently outperform open-source LLMs, and medium-sized (8B) open-source models outperform small (1B) ones. (2) After *instruction tuning*: All open-source models improve substantially; notably, Gemma3-1B (69.70) surpasses Gemini-2.5-Flash (65.08). These results indicate that instruction tuning can markedly enhance open-source models, narrowing the performance gap with and even outperform commercial LLMs.

Task-Level Results. (1) *Conventional Analysis.* In zero-shot settings, both commercial and open-source LLMs perform poorly on anomaly detection and classification, but open-source models improve markedly after instruction tuning (e.g., LLaMA-3.1-8B reaches 91.02 and 91.27). (2) *Ad-*



(a) Time series length. (b) Description length. (c) Question length. (d) Question type dist.

Figure 2: Histograms of time series, description and question lengths, and question type distribution.

vanced Analysis. For characterization and comparison, commercial models outperform medium-sized open-source models, likely due to broader pretraining exposure. Data transformation and temporal relationship, especially *PZ* questions, remain difficult for all models. Instruction tuning boosts open-source performance, but there are still considerable room to improve. For example, for comparison task, best performing LLaMA-3.1-8B with instruction tuning only achieves 86.72 and 79.31.

Question Type-Level Results. Across the three question types (*TF*, *MC*, *PZ*), open-source models perform best on *TF*, worse on *MC*, and poorest on *PZ*. Performance on *PZ* is substantially lower than on *TF* and *MC*, in both zero-shot and tuned settings. Considerable room for improvement remains, e.g., the best *PZ* score is only 67.68 (LLaMA-3.1-8B after instruction tuning).

4.2 Analysis

We use best performing commercial LLMs, i.e., Geimini-2.5-Flash and GPT-4.1, and open-source LLMs, i.e., LLaMA3.1-8B and Qwen3-8B to conduct further analysis. To examine their analytical ability on the proposed TSAQA Benchmark, we cover three key perspectives: *Accuracy Correlate Analysis*, *Task-Specific Analysis*, and *Case Study*.

4.2.1 Accuracy Correlate Analysis

Input Lengths. Figure 3 in Appendix) illustrates the relationship between input length and model accuracy. Across all six models and five tasks, excluding the Temporal Relationship task, we observe a consistent trend that performance declines as input length increases, indicating that longer inputs correspond to more difficult questions. However, the Temporal Relationship task exhibits the opposite behavior, where accuracy improves with increasing input length. The analysis is shown in Figure 4, which the newly proposed *PZ* question exhibits the opposite trend. The fact that *PZ* performance scales positively with length proves that models are actively utilizing global context to deduce the correct chronological order, proving that *PZ* question is a rigorous probe for *Global Causal Reasoning*. More details are provided in Appendix D.1.

Topics & Subtopics v.s. Accuracy. Tasks such as Characterization and Comparison include questions with different selected topics and subtopics from a predefined list 7. To understand how the complexity of topics and subtopics influences model performance, we analyzed the relationship between the number of topics and subtopics used in each question and the corresponding model accuracy. Based on Table 8, we observe that the complexity of questions with varying number of topics and subtopics doesn't have direct impact on model accuracy, which indicates that the TSAQA Benchmark is largely unbiased. We further analyzed the difficulty of individual topics by examining how different topic combinations influence model performance across tasks. The results of questions from Comparison Task are visualized in Figure 5. We found that questions with topics such as seasonality, autocorrelation, dispersion, and noise are harder to model with lower average accuracy. More details are provided in Appendix D.1.

Domain v.s. Accuracy. We conducted an in-depth analysis of how domain variation impacts overall model accuracy on TSAQA Benchmark. The results are summarized in Table 9. Our analysis reveals that questions from domains including Synthetic, IT, Robotics, and Web pose greater challenges to models under the zero-shot setting, while questions from Sales and Web domains remain the most difficult after instruction tuning. More details are provided in Appendix D.1.

4.2.2 Task Specific Analysis

Data Transformation. We analyze model performance on the Data Transformation task, which is designed to evaluate a model's understanding of three transformation operators: Fourier Transform (FT), Wavelet Transform (WT), and First-Order

Differencing (FOD). For each operator, we assess performance by measuring the accuracy on both *MC* and *TF* question formats. As shown in Table 10, for zero-shot evaluation, our key finding highlights a limitation in which both commercial and open-source models fail to provide accurate answers, except of FOD. More details are provided in Appendix D.2.

Temporal Relationship. Beyond the input length analysis in Section 4.2.1, we further examined how domain-level information influences model performance on *PZ* questions. The results are summarized in Table 11, which Web and Sales domains remain the most challenging across both zero-shot and instruction-tuning settings. To identify the cause, we analyzed the boundary consistency of incorrect predictions and identified a significant *Smoothness Bias*. As shown in Table 12, models consistently attempt to repair legitimate discontinuities by predicting transitions that are smoother than the ground truth. This failure highlights the critical utility of the *PZ* task: since legitimate volatility varies by domain, *PZ* acts as a rigorous discriminator for Temporal Fidelity. It penalizes models that rely on generic smoothing priors and rewards those that capture the specific, irregular structural dynamics of the target domain. More details are provided in Appendix D.2.

Comparison. We analyze model performance on the Comparison task, specifically investigating whether providing explicit domain-level context affects model accuracy. The task requires comparing two input time series, which we test under two conditions: (1) when both series originate from the same domain and (2) when they are from different domains. In both scenarios, the corresponding domain names are provided to the model as textual description. As shown in Table 13, we observe no significant performance difference between the same-domain and different-domain settings across either *MC* or *TF* questions. This suggests that the Comparison Task is domain invariant. More details are provided in Appendix D.2.

4.2.3 Case Study.

First Letter Distribution. To explore potential biases in model behavior, we analyzed the distribution of the first letters in model responses for *PZ* questions. We visualize these distributions in Figure 6 with an interesting pattern emerges: for questions that were answered incorrectly or received partial credit, Qwen3-8B tends to output the choice C more frequently, whereas LLaMA3.1-8B shows a stronger tendency to output A. More details are provided in Appendix D.3.

Incorrect Output Format. While models generally demonstrate a strong understanding of the expected response format for common question types such as *MC* and *TF*, we further include sample responses from each model that highlight unique or unexpected behaviors—cases where the models' answers do not strictly adhere to the specified instructions for the *PZ* question format. More details are provided in Appendix D.3.

4.3 Human Evaluation

We conduct human evaluations of the multi-LLM consensus labels for characterization and comparison (Section 3.3). Six Ph.D.-level experts manually annotate 600 questions (300 each), serving as ground truth. Uncertain or problematic QA pairs are flagged, multiple answers allowed when valid, and explanations provided for disagreements with the benchmark.

Our evaluation yields two main findings. (1) Question quality: Uncertainty rates are low (5% for characterization, 7% for comparison), showing that most questions are clear. (2) Answer accuracy: For unambiguous cases, benchmark answers align with human judgments in 91.2% of characterization and 87.4% of comparison. These results indicate that the automatic pipeline produces reliable QA pairs, though comparison remains harder, with lower agreement and higher uncertainty (Figure 7). More details are provided in Appendix E.

5 Conclusion

TSAQA establishes a large-scale benchmark for time series question answering with 210k samples curated from 13 domains, covering 6 tasks and 3 types of questions, extending evaluation beyond classical tasks, i.e., anomaly detection and classification, to advanced analytical tasks, i.e., characterization, comparison, data transformation, temporal relationship. By spanning diverse domains, tasks, and question types, it offers a unified platform to probe the strengths and limitations of both commercial and open-source LLMs. Our results highlight that, despite progress with instruction tuning, substantial challenges remain, particularly for advanced reasoning and puzzling questions, underscoring the need for further research into models capable of deeper time series understanding.

Limitations

While TSAQA provides a unified and large-scale benchmark for time series question answering, several limitations remain. First, although the benchmark spans 13 domains and 6 analytical tasks, it does not fully capture the diversity of real-world temporal processes. Many application settings involve irregular sampling, strong exogenous drivers, or domain-specific structures, which are only partially reflected in our datasets. Expanding U-TSAQA toward irregular, mixed-frequency, and exogenous-aware scenarios would further improve realism. Second, our task taxonomy primarily focuses on analytical capabilities that can be expressed through structured questions. However, real systems often require richer forms of temporal reasoning, suggesting opportunities to design tasks that more directly probe these behaviors. Third, while the newly proposed puzzling question encourages global structural reasoning, it may penalize models biased toward locally smooth transitions and introduces higher computational costs. Future extensions could incorporate complementary formats that disentangle local continuity, long-range consistency, and domain-specific volatility. Finally, although TSAQA covers multiple domains, the benchmark remains static, whereas real deployments face evolving distributions and emerging domains. Building dynamic extensions that evaluate adaptation and robustness under distribution shifts represents an important next step.

Ethical Considerations

All datasets and language models used in this work are publicly available. The TSAQA dataset was constructed from established, publicly accessible time series benchmarks and synthetic data generation followed ethical guidelines to minimize biases and ensure data quality.

Acknowledgments

We leverage Large Language Models (LLMs) from two perspectives: (1) Polishing the writing, where LLMs are used to refine the clarity, fluency, and consistency of the paper; and (2) Labeling, where LLMs assist in generating high-quality question-answer (QA) pairs and providing preliminary annotations, which are then validated or aggregated through consensus to create reliable ground-truth labels.

References

The claude 3 model family: Opus, sonnet, haiku.

Sarah Alnegheimish, Linh Nguyen, Laure Berti-Equille, and Kalyan Veeramachaneni. 2024. Large language models can be zero-shot anomaly detectors for time series? *arXiv preprint arXiv:2405.14755*.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. 2024a. Chronos: Learning the language of time series. *Preprint*, arXiv:2403.07815.

Abdul Fatir Ansari, Lorenzo Stella, Caner Türkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. 2024b. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*.

Anthony J. Bagnall and Luke M. Davis. 2014. Predictive modelling of bone age through classification and regression of bone shapes. *CoRR*, abs/1406.4781.

Romain Briandet, E. Katherine Kemsley, and Reginald H. Wilson. 1996. Discrimination of Arabica and Robusta in instant coffee by Fourier transform infrared spectroscopy and chemometrics. *Journal of Agricultural and Food Chemistry*, 44(1):170–174.

Yifu Cai, Arjun Choudhry, Mononito Goswami, and Artur Dubrawski. 2024. Timeseriesexam: A time series understanding exam. *Preprint*, arXiv:2410.14752.

Caltrans. 2025. Performance measurement system (pems) data source. `https://dot.ca.gov/programs/traffic-operations/mpr/pems-source`. Accessed 2025-09-24.

Luis M Candanedo and Véronique Feldheim. 2016. Accurate occupancy detection of an office room from light, temperature, humidity and co2 measurements using statistical learning models. *Energy and buildings*, 112:28–39.

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

Jialin Chen, Aosong Feng, Ziyu Zhao, Juan Garza, Gaukhar Nurbek, Cheng Qin, Ali Maatouk, Leandros Tassiulas, Yifeng Gao, and Rex Ying. 2025. Mtbench: A multimodal time series benchmark for temporal reasoning and question answering. *Preprint*, arXiv:2503.16858.

City of Melbourne. 2017. Pedestrian counting system (counts per hour). https://data.melbourne.vic.gov.au/explore/dataset/pedestrian-counting-system-monthly-counts-per-hour/information/. Accessed 2025-09-24.

Hoang Anh Dau, Anthony Bagnall, Kaveh Kamgar, Chin-Chia Michael Yeh, Yan Zhu, Shaghayegh Gharghabi, Chotirat Ann Ratanamahatana, and Eamonn Keogh. 2019a. The ucr time series archive. *IEEE/CAA Journal of Automatica Sinica*, 6(6):1293–1305.

Hoang Anh Dau, Anthony Bagnall, Kaveh Kamgar, Chin-Chia Michael Yeh, Yan Zhu, Shaghayegh Gharghabi, Chotirat Ann Ratanamahatana, and Eamonn Keogh. 2019b. The ucr time series archive. *Preprint*, arXiv:1810.07758.

Zihan Dong, Xinyu Fan, and Zhiyuan Peng. 2024. Fnspid: A comprehensive financial news dataset in time series. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4918–4927.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Pavel Filonov, Andrey Lavrentyev, and Artem Vorontsov. 2016. Multivariate industrial time series with cyber-attack simulation: Fault detection using an lstm-based predictive data model. *CoRR*, abs/1612.06676.

Patrick Fissler, Olivia Caroline Küster, Daria Laptinskaya, Laura Sophia Loy, Christine AF Von Arnim, and Iris-Tatjana Kolassa. 2018. Jigsaw puzzling taps multiple cognitive abilities and is a potential protective factor for cognitive aging. *Frontiers in aging neuroscience*, 10:408085.

FiveThirtyEight. 2015. Uber tlc foil response: Uber pickups in new york city. https://www.kaggle.com/datasets/fivethirtyeight/uber-pickups-in-new-york-city. Accessed 2025-09-24.

Pierre Geurts. May 2002. *Contributions to decision tree induction: bias/variance tradeoff and time series classification*. Ph.D. thesis, ULiège - Université de Liège.

Rakshitha Godahewa, Christoph Bergmeir, Geoffrey I. Webb, Rob J. Hyndman, and Pablo Montero-Manso. 2021. Monash time series forecasting archive. In *NeurIPS Datasets and Benchmarks Track*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

James D Hamilton. 2020. *Time series analysis*. Princeton university press.

Charles R Harris, K Jarrod Millman, Stéfan J Van Der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, and 1 others. 2020. Array programming with numpy. *nature*, 585(7825):357–362.

James K. Holland, E. Katherine Kemsley, and Reginald H. Wilson. 1998. Use of fourier transform infrared spectroscopy and partial least squares regression for the detection of adulteration of strawberry purées. *Journal of the Science of Food and Agriculture*, 76(2):263–269.

Tao Hong, Pierre Pinson, Shu Fan, and 1 others. 2014. Global energy forecasting competition 2012. *International Journal of Forecasting*, 30(2):357–363.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Xiaohui Huang, Yunming Ye, Liyan Xiong, Raymond Y.K. Lau, Nan Jiang, and Shaokai Wang. 2016. Time series k-means. *Inf. Sci.*, 367(C):1–13.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Alexander Ihler, Jon Hutchins, and Padhraic Smyth. 2006. Adaptive event detection with time-varying poisson processes. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, page 207–216, New York, NY, USA. Association for Computing Machinery.

Yushan Jiang, Kanghui Ning, Zijie Pan, Xuyang Shen, Jingchao Ni, Wenchao Yu, Anderson Schneider, Haifeng Chen, Yuriy Nevmyvaka, and Dongjin Song. 2025. Multi-modal time series analysis: A tutorial and survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 6043–6053.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. 2024. Time-llm: Time series forecasting by reprogramming large language models. *Preprint*, arXiv:2310.01728.

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. 2023. Time-llm: Time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2310.01728*.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, and 1 others. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Eamonn Keogh, Li Wei, Xiaopeng Xi, Stefano Lonardi, Jin Shieh, and Scott Sirowy. 2006. Intelligent icons: Integrating lite-weight data mining and visualization into gui operating systems. In *Sixth International Conference on Data Mining (ICDM'06)*, pages 912–916.

Yaxuan Kong, Yiyuan Yang, Yoontae Hwang, Wenjie Du, Stefan Zohren, Zhangyang Wang, Ming Jin, and Qingsong Wen. 2025. Time-mqa: Time series multi-task question answering with context enhancement. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 29736–29753. Association for Computational Linguistics.

Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. 2018. Modeling long- and short-term temporal patterns with deep neural networks. *arXiv preprint arXiv:1703.07015*.

Xiang Li, Zhenyu Li, Chen Shi, Yong Xu, Qing Du, Mingkui Tan, Jun Huang, and Wei Lin. 2024a. Alphafin: Benchmarking financial analysis with retrieval-augmented stock-chain framework. *arXiv preprint arXiv:2403.12582*.

Zhonghang Li, Lianghao Xia, Jiabin Tang, Yong Xu, Lei Shi, Long Xia, Dawei Yin, and Chao Huang. 2024b. Urbangpt: Spatio-temporal large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5351–5362.

Zihao Li, Xiao Lin, Zhining Liu, Jiaru Zou, Ziwei Wu, Lecheng Zheng, Dongqi Fu, Yada Zhu, Hendrik F. Hamann, Hanghang Tong, and Jingrui He. 2025. Language in the flow of time: Time-series-paired texts weaved into a unified temporal narrative. *CoRR*, abs/2502.08942.

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

Haoxin Liu, Harshavardhan Kamarthi, Zhiyuan Zhao, Shangqing Xu, Shiyu Wang, Qingsong Wen, Tom Hartvigsen, Fei Wang, and B Aditya Prakash. 2025. How can time series analysis benefit from multiple modalities? a survey and outlook. *arXiv preprint arXiv:2503.11835*.

Haoxin Liu, Shangqing Xu, Zhiyuan Zhao, Lingkai Kong, Harshavardhan Prabhakar Kamarthi, Aditya Sasanur, Megha Sharma, Jiaming Cui, Qingsong Wen, Chao Zhang, and 1 others. 2024. Time-mmd: Multi-domain multimodal dataset for time series analysis. *Advances in Neural Information Processing Systems*, 37:77888–77933.

Xiangkai Ma, Xiaobin Hong, Wenzhong Li, and Sanglu Lu. 2024. Utsd: Unified time series diffusion model. *Preprint*, arXiv:2412.03068.

Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. 2022. M5 accuracy competition: Results, findings, and conclusions. *International Journal of Forecasting*, 38(4):1346–1364.

Paolo Mancuso, Veronica Piccialli, and Antonio M. Sudoso. 2021. A machine learning approach for forecasting hierarchical time series. *Expert Systems with Applications*, 182:115102.

Michael W. McCracken and Serena Ng. 2016. Fred-md: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, 34(4):574–589.

Mike A. Merrill, Mingtian Tan, Vinayak Gupta, Thomas Hartvigsen, and Tim Althoff. 2024. Language models still struggle to zero-shot reason about time series. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 3512–3533. Association for Computational Linguistics.

Clayton Miller, Anjukan Kathirgamanathan, Bianca Picchetti, Pandarasamy Arjunan, June Young Park, Zoltán Nagy, Paul Raftery, Brodie Hobson, Zixiao Shi, Forrest Meggers, and 1 others. 2020. The building data genome project 2, 1,636 energy meter readings from the ashrae great energy predictor iii competition. *Scientific Data*, 7(1):368.

Navid Mohammadi Foumani, Lynn Miller, Chang Wei Tan, Geoffrey I Webb, Germain Forestier, and Mahsa Salehi. 2024. Deep learning for time series classification and extrinsic regression: A current survey. *ACM Computing Surveys*, 56(9):1–45.

George B Moody and Roger G Mark. 2001. The impact of the mit-bih arrhythmia database. *IEEE engineering in medicine and biology magazine*, 20(3):45–50.

Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Alireza Dirafzoon, Aparajita Saraf, Amy Bearman, and Babak Damavandi. 2022. Imu2clip: Multimodal contrastive learning for imu motion sensors from egocentric videos and text. *arXiv preprint arXiv:2210.14395*.

Soukayna Mouatadid, Paulo Orenstein, Genevieve Flaspohler, Miruna Oprescu, Judah Cohen, Franklyn Wang, Sean Knight, Maria Geogdzhayeva, Sam Levang, Ernest Fraenkel, and Lester Mackey. 2024. Subseasonalclimateusa: A dataset for subseasonal forecasting and benchmarking. *arXiv preprint arXiv:2109.10399*. V4.

Abdullah Mueen, Eamonn Keogh, and Neal Young. 2011. Logical-shapelets: an expressive primitive for time series classification. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, page 1154–1162, New York, NY, USA. Association for Computing Machinery.

David Murray. 2015. A data management platform for personalised real-time energy feedback. *Proc. 8th Int. Conf. Energy Efficiency Domestic Appl. Lighting (EEDAL)*, pages 1–15.

Mehdi Noroozi and Paolo Favaro. 2016. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer.

Robert Thomas Olszewski, Roy Maxion, and Dan Siewiorek. 2001. *Generalized feature extraction for structural pattern recognition in time-series data*. Ph.D. thesis, USA. AAI3040489.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

John Paparrizos, Yuhao Kang, Paul Boniol, Ruey S Tsay, Themis Palpanas, and Michael J Franklin. 2022. Tsb-uad: an end-to-end benchmark suite for univariate time-series anomaly detection. *Proceedings of the VLDB Endowment*, 15(8):1697–1711.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Chotirat Ratanamahatana and Eamonn Keogh. 2005. Three myths about dynamic time warping data mining.

Naoki Saito and Ronald R. Coifman. 1994. Local feature extraction and its applications using a library of bases.

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. *arXiv preprint arXiv:2409.16040*.

Geri Skenderi, Christian Joppi, Matteo Denitto, and Marco Cristani. 2024. Well googled is half done: Multimodal forecasting of new fashion product sales with image-based google trends. *Journal of Forecasting*, 43(6):1982–1997.

Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, and Dan Pei. 2019. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, pages 2828–2837. ACM.

Jimeng Sun, Spiros Papadimitriou, and Christos Faloutsos. 2005. Online latent variable detection in sensor networks. In *Proceedings of the 21st International Conference on Data Engineering*, ICDE '05, page 1126–1127, USA. IEEE Computer Society.

Chang Wei Tan, Christoph Bergmeir, François Petitjean, and Geoffrey I. Webb. 2020. Monash university, uea, ucr time series extrinsic regression archive. *arXiv preprint arXiv:2006.10996*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025a. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025b. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Markus Thill, Wolfgang Konen, and Thomas Bäck. Markusthill/mgab: The mackey-glass anomaly benchmark, apr 2020. *URL https://doi.org/10.5281/zenodo*, 3760086.

José F Torres, Dalil Hadjout, Abderrazak Sebaa, Francisco Martínez-Álvarez, and Alicia Troncoso. 2021. Deep learning for time series forecasting: a survey. *Big data*, 9(1):3–21.

Md Nayem Uddin, Amir Saeidi, Divij Handa, Agastya Seth, Tran Cao Son, Eduardo Blanco, Steven R. Corman, and Chitta Baral. 2025. Unseentimeqa: Time-sensitive question-answering beyond llms' memorization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 1873–1913. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, and 1 others. 2020. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272.

Alexander von Birgelen and Oliver Niggemann. 2018. Anomaly detection and localization for cyber-physical production systems with self-organizing maps. In *IMPROVE-Innovative Modelling Approaches for Production Systems to Raise Validatable Efficiency: Intelligent Methods for the Factory of the Future*, pages 55–71. Springer Berlin Heidelberg Berlin, Heidelberg.

Jun Wang, Wenjie Du, Yiyuan Yang, Linglong Qian, Wei Cao, Keli Zhang, Wenjia Wang, Yuxuan Liang, and Qingsong Wen. 2024. Deep learning for multivariate time series imputation: A survey. *arXiv preprint arXiv:2402.04059*.

Yilin Wang, Peixuan Lei andm Jie Song, Yuzhe Hao, Tao Chen, Yuxuan Zhang, Lei Jia, Yuanxiang Li, and Zhongyu Wei. 2025a. Itformer: Bridging time series and natural language for multi-modal QA with large-scale multitask dataset. *CoRR*, abs/2506.20093.

Yilin Wang, Peixuan Lei, Jie Song, Yuzhe Hao, Tao Chen, Yuxuan Zhang, Lei Jia, Yuanxiang Li, and Zhongyu Wei. 2025b. Itformer: Bridging time series and natural language for multi-modal qa with large-scale multitask dataset. *Preprint*, arXiv:2506.20093.

Qingsong Wen, Tian Zhou, Chaoli Zhang, Weiqi Chen, Ziqing Ma, Junchi Yan, and Liang Sun. 2022. Transformers in time series: A survey. *arXiv preprint arXiv:2202.07125*.

Andrew Robert Williams, Arjun Ashok, Étienne Marcotte, Valentina Zantedeschi, Jithendaraa Subramanian, Roland Riachi, James Requeima, Alexandre Lacoste, Irina Rish, Nicolas Chapados, and Alexandre Drouin. 2025. Context is key: A benchmark for forecasting with essential textual information. *Preprint*, arXiv:2410.18959.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. 2024. Unified training of universal time series forecasting transformers. *arXiv preprint arXiv:2402.02592*.

Zhe Xie, Zeyan Li, Xiao He, Longlong Xu, Xidao Wen, Tieying Zhang, Jianjun Chen, Rui Shi, and Dan Pei. 2025. Chatts: Aligning time series with llms via synthetic data for enhanced understanding and reasoning. *Preprint*, arXiv:2412.03104.

Xiongxiao Xu, Haoran Wang, Yueqing Liang, Philip S. Yu, Yue Zhao, and Kai Shu. 2025a. Can multimodal llms perform time series anomaly detection? *CoRR*, abs/2502.17812.

Xiongxiao Xu, Yue Zhao, S Yu Philip, and Kai Shu. 2025b. Beyond numbers: A survey of time series analysis in the era of multimodal llms. *Authorea Preprints*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Lexiang Ye and Eamonn Keogh. 2009. Time series shapelets: a new primitive for data mining. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, page 947–956, New York, NY, USA. Association for Computing Machinery.

Lexiang Ye and Eamonn Keogh. 2011. Time series shapelets: a novel technique that allows accurate, interpretable and fast classification. *Data Mining and Knowledge Discovery*, 22(1):149–182.

Xinli Yu, Zheng Chen, Yuan Ling, Shujing Dong, Zongyi Liu, and Yanbin Lu. 2023. Temporal data meets llm–explainable financial time series forecasting. *arXiv preprint arXiv:2306.11025*.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. 2023. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 11121–11128.

Xiyuan Zhang, Ranak Roy Chowdhury, Rajesh K Gupta, and Jingbo Shang. 2024. Large language models for time series: A survey. *arXiv preprint arXiv:2402.01801*.

Yunkai Zhang, Yawen Zhang, Ming Zheng, Kezhen Chen, Chongyang Gao, Ruian Ge, Siyuan Teng, Amine Jelloul, Jinmeng Rao, Xiaoyuan Guo, Chiang-Wei Fang, Zeyu Zheng, and Jie Yang. 2023. Insight miner: A large-scale multimodal model for insight mining from time series. In *NeurIPS 2023 AI for Science Workshop*.

Siru Zhong, Weilin Ruan, Ming Jin, Huan Li, Qingsong Wen, and Yuxuan Liang. 2025a. Time-vlm: Exploring multimodal vision-language models for augmented time series forecasting. *CoRR*, abs/2502.04395.

Siru Zhong, Weilin Ruan, Ming Jin, Huan Li, Qingsong Wen, and Yuxuan Liang. 2025b. Time-vlm: Exploring multimodal vision-language models for augmented time series forecasting. *arXiv preprint arXiv:2502.04395*.

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

A Data Collection

In this section, we detail the data sources, including *core datasets* (Appendix A.1), *anomaly detection datasets* (Appendix A.2), and *classification datasets* (Appendix A.3).

A.1 Core Datasets

Table 4: Summary of the core datasets.

dataset_name	total_data_point	domain
AustralianElectricityDemand	1,153,584	energy
BDG-2 Rat	4,728,288	energy
GEF12	788,280	energy
ExchangeRate	56,096	finance
FRED MD	76,612	finance
BIDMC32HR	8,000,000	healthcare
PigArtPressure	624,000	healthcare
USBirths	7,275	healthcare
Sunspot	73,924	nature
Saugeenday	23,711	nature
SubseasonalPrecip	9,760,426	nature
HierarchicalSales	212,164	sales
m5	58,327,370	sales
PedestrianCounts	3,130,762	transport
PEMS03	9,382,464	transport
UberTLCHourly	1,129,444	transport
WikiDaily100k	274,099,872	web

We extract data from multiple time-series datasets including: Australian Electricity Demand (Godahewa et al., 2021), BDG-2 Rat (Miller et al., 2020), GEF12 (Hong et al., 2014), ExchangeRate(Lai et al., 2018), FRED MD(McCracken and Ng, 2016), BIDMC32HR (Tan et al., 2020), PigArtPressure (Dau et al., 2019a), USBirths (Godahewa et al., 2021), Sunspot (Godahewa et al., 2021), Saugeenday (Godahewa et al., 2021), SubseasonalPrecip (Mouatadid et al., 2024), HierarchicalSales (Mancuso et al., 2021), M5 (Makridakis et al., 2022), PedestrianCounts (City of Melbourne, 2017), PEMS03 (Caltrans, 2025), UberTLCHourly (FiveThirtyEight, 2015),WikiDaily100k (Ansari et al., 2024b). Below are some more detailed descriptions on those datasets.

Australian Electricity Demand. A single long time series from the Monash Time Series Archive representing half-hourly electricity demand for Victoria, Australia in 2014 (17,520 observations), extracted from the R package `fpp2` (dataset name: "elecdemand"). Temperatures corresponding to each demand value are available in the original dataset.

BDG-2 Rat. From The Building Data Genome Project 2 (MIT License), consisting of measurements from 3,053 meters across 1,636 commercial buildings over 2016–2017. One or more meters per building measured total electrical, heating and cooling water, steam, solar energy, water, and irrigation usage. We use the whole-building electricity meter measurements from the Bear, Fox, Panther, and Rat sites, totaling 611 buildings (from the CSV file `electricity_cleaned.csv`).

GEF12. A benchmark compiled from the Global Energy Forecasting Competition 2012 (load forecasting tracks), containing 20 aggregated-level hourly load series and 11 temperature series from 2004-01-01 00:00 to 2008-06-30 05:00. Because the one-to-one correspondence between temperature and load series is not clearly defined, a common strategy is to use a single temperature series for all loads (here, the second temperature series). The dataset is competition-grade and was used without additional preprocessing; visualizations show obvious periodicity and seasonality in the aggregated loads.

ExchangeRate. Daily exchange rates for currencies of eight countries—Australia, United Kingdom, Canada, Switzerland, China, Japan, New Zealand, and Singapore—covering 1990 to 2016.

FRED-MD. 107 monthly time series of macroeconomic indicators from the Federal Reserve Bank, starting from 1959-01-01, extracted from the FRED-MD database.

BIDMC32HR. Derived from BIDMC ICU recordings: PPG and respiratory signals/IP (sampling rate 125 Hz) from 53 adult patients, with breath annotations used to form reference targets in the source dataset. Following the adaptation in subsequent work, PPG and ECG are converted into 32-second sliding-window time series; the average heart rate (HR) in each 32 s window is the target. The datasets are split by randomly selecting 30% as test, yielding 5,550 training and 2,399 test time series.

PigArtPressure. Based on a source dataset from 52 pigs with three vital signs monitored before and after an induced injury. Three datasets are created: AirwayPressure (airway pressure), ArtPressure (arterial blood pressure), and CVP (central venous pressure).

US Births. A single long daily time series of the number of births in the United States from 1969-01-01 to 1988-12-31 (7,305 observations), extracted from the R package `mosaicData`.

Sunspot. A single long daily time series of sunspot numbers from 1818-01-01 onward, with additional related series (monthly means, smoothed series, yearly totals, hemispheric series) in the original source. The repository used here contains the daily series from 1818-08-01 to 2020-05-31 and includes both the raw data (with missing values) and an LOCF-imputed version.

Saugeen. A single long daily time series of the Saugeen River mean flow at Walkerton (in cubic meters per second) from 1915-01-01 to 1979-12-31 (23,741 observations), extracted from the R package `deseasonalize` (dataset name: "SaugeenDay").

Subseasonal Precipitation. Extracted from SubseasonalClimateUSA: daily precipitation measurements (millimeters) for a single $1.5° \times 1.5°$ latitude–longitude grid cell, covering 1948–1978.

Hierarchical Sales. 118 daily time series of SKU-level sales for four national pasta brands from 2014-01-01 to 2018-12-31, including a binary indicator for promotion. The series can be organized into a three-level hierarchy.

M5. The M5 "Accuracy" competition dataset requiring point forecasts for 30,490 bottom-level daily series that aggregate to 42,840 time series representing hierarchical unit sales for Walmart. The competition paper details the implementation, results, top methods, and implications for forecasting research.

Pedestrian Counts. Hourly pedestrian counts from 66 sensors in Melbourne starting from May 2009. The original data are updated monthly; the repository snapshot used here contains counts up to 2020-04-30.

PEMS03. Datasets sourced from Caltrans PeMS, which collects 30-second traffic readings and aggregates them into 5-minute intervals (288 time steps per day). Road network structure is derived from connectivity status and actual distances between sensors.

Uber TLC Daily. Counts of Uber pick-ups from various New York City locations between January and June 2015, obtained from FiveThirtyEight's "uber-tlc-foil-response" repository and aggregated at hourly and daily resolutions.

WikiDaily10k. Daily traffic data for 10,000 Wikipedia pages.

A.2 Anomaly Detection Dataset

We extract data from multiple time-series anomaly detection benchmarks (Paparrizos et al., 2022; Su et al., 2019), including ECG (Moody and Mark, 2001), SMD (Su et al., 2019), MGAB (Thill et al.) Genesis (von Birgelen and Niggemann, 2018), GHL (Filonov et al., 2016), Occupancy (Candanedo and Feldheim, 2016). These datasets span various domains, including healthcare (ECG), mathematical biology (MGAB), spacecraft telemetry (Genesis), industrial control system (GHL), environmental sensing (Occupancy), cyber-security on IT Operations (SMD). The statistics of these datasets are shown in Table 5. To address class imbalance, we count the number of anomalous sequences and randomly select an equal number of normal sequences, resulting in a balanced dataset. Below are the meta information for each dataset.

Name	# Samples	Domain
ECG	17,862	Healthcare
SMD	58,888	Cyber-security on IT Operations
MGAB	376	mathematical biology
Genesis	274	Spacecraft Telemetry
GHL	768	Industrial Control System
Occupancy	8,178	Environmental Sensing

Table 5: Summary of anomaly detection datasets.

MGAB. This dataset is composed of Mackey-Glass time series with non-trivial anomalies. Mackey-Glass time series exhibit chaotic behavior that is difficult for the human eye to distinguish.

ECG. This dataset is a standard electrocardiogram dataset and the anomalies represent ventricular premature contractions. The ECG recordings were made using Del Mar Avionics model 445 two-channel reel-to-reel Holter recorders, and the analog signals were recreated for digitization using a Del Mar Avionics model 660 playback unit. The digitization rate (360 samples per second per channel) was chosen to accommodate the use of simple digital notch filters to remove 60 Hz (mains frequency) interference.

Genesis. This dataset is a portable pick-and-place demonstrator which uses an air tank to supply all the gripping and storage units. Data samples were taken through an OPC connection with a resolution of 50 milliseconds for a total of 42 production cycles. The first 38 production cycles contain only normal behavior and were used to train the selforganizing map for both experiments shown in this section. Two of the 4 remaining cycles contain anomalous behavior and are used for the anomaly detection.

GHL. This dataset is a Gasoil Heating Loop Dataset and contains the status of 3 reservoirs such as the temperature and level. Anomalies indicate changes in max temperature or pump frequency. Type of cyber attack to the normal process logic is the unauthorized change of max Receiving Tank level. By changing the time of attack and the value of the hacked max Receiving Tank level, we generated many anomalous data sets used for fault detection.

Occupancy. This dataset contains experimental data of room occupancy, such as temperature, humidity, light, and CO2. Ground-truth occupancy was obtained from time stamped pictures that were taken every minute.

SMD. SMD (Server Machine Dataset) is collected from a large Internet company. The data is sampled every 5 seconds. Labels denote whether a point is an anomaly and the dimensions contribute to every anomaly.

A.3 Classification Dataset

We extract data from the UCR Archive (Dau et al., 2019a). To create a focused subset for our study, we applied two primary selection criteria: we included only datasets with four or fewer classes and time series with a sequence length of 400 time points or less. Through our selection, we extract data from 37 benchmarks in the UCR Archive, including SonyAIBORobotSurface1 & SonyAIBORobotSurface2 (Mueen et al., 2011), FreezerRegularTrain & FreezerSmallTrain (Murray, 2015), ToeSegmentation1 & ToeSegmentation2 (Ye and Keogh, 2011), TwoPatterns (Geurts, May 2002), CBF (Saito and Coifman, 1994), Wafer & ECG200 (Olszewski et al., 2001), TwoLeadECG, ECGFiveDays, DistalPhalanxOutlineCorrect & MiddlePhalanxOutlineCorrect & ProximalPhalanxOutlineCorrect & DistalPhalanxOutlineAgeGroup & MiddlePhalanxOutlineAgeGroup & ProximalPhalanxOutlineAgeGroup & PhalangesOutlinesCorrect (Bagnall and Davis, 2014), MoteStrain (Sun et al., 2005), GunPointMaleVersusFemale & GunPointOldVersusYoung & GunPointAgeSpan & GunPoint (Ratanamahatana and Keogh, 2005), Strawberry (Holland et al., 1998), ItalyPowerDemand (Keogh et al., 2006), Chinatown, BME, PowerCons, DodgersLoopWeekend & DodgersLoopGame (Ihler et al., 2006), DiatomSizeReduction, SmoothSubspace (Huang et al., 2016), UMD, Wine, Coffee (Briandet et al., 1996), and ArrowHead (Ye and Keogh, 2009). These datasets span various domains, including robotics, energy, healthcare, synthetic, manufacturing, nature, and transport. The statistics of these datasets are shown in Table 6.

Name	# Samples	# Classes	Domain
SonyAIBORobotSurface1	486	2	Robotics
SonyAIBORobotSurface2	771	2	Robotics
FreezerRegularTrain	2,404	2	Energy
FreezerSmallTrain	2,353	2	Energy
ToeSegmentation1	210	2	Healthcare
ToeSegmentation2	129	2	Healthcare
TwoPatterns	3,999	4	Synthetic
CBF	757	3	Synthetic
Wafer	5,744	2	Manufacturing
ECG200	159	2	Healthcare
TwoLeadECG	923	2	Healthcare
ECGFiveDays	704	2	Healthcare
DistalPhalanxOutlineCorrect	690	2	Healthcare
MiddlePhalanxOutlineCorrect	731	2	Healthcare
ProximalPhalanxOutlineCorrect	688	2	Healthcare
DistalPhalanxOutlineAgeGroup	423	3	Healthcare
MiddlePhalanxOutlineAgeGroup	435	3	Healthcare
ProximalPhalanxOutlineAgeGroup	485	3	Healthcare
PhalangesOutlinesCorrect	2,076	2	Healthcare
MoteStrain	1,012	2	Nature
GunPointMaleVersusFemale	362	2	Healthcare
GunPointOldVersusYoung	356	2	Healthcare
GunPointAgeSpan	368	2	Healthcare
GunPoint	169	2	Healthcare
Strawberry	786	2	Nature
ItalyPowerDemand	890	2	Energy
Chinatown	293	2	Transport
BME	137	3	Synthetic
PowerCons	294	2	Energy
DodgersLoopWeekend	111	2	Transport
DodgersLoopGame	115	2	Transport
DiatomSizeReduction	248	4	Nature
SmoothSubspace	236	3	Synthetic
UMD	148	3	Synthetic
Wine	85	2	Nature
Coffee	48	2	Nature
ArrowHead	175	3	Nature

Table 6: Classification data used in our experiments.

B Benchmark Construction

In this section, we provide extra content about the construction process for each task and provide examples of each task.

B.1 Hierarchical Uniform Sampling

Algorithm 1: Hierarchical Random Sampling

Input: Domains $\mathcal{M}$;
Datasets $\mathcal{D}(m)$ for each domain $m \in \mathcal{M}$;
Sequences $\mathcal{S}(d)$ for each dataset $d \in \mathcal{D}$;
Segment length l
Output: Segment $s_{t:t+l-1}$

$m \leftarrow$ UniformPick($\mathcal{M}$); `// Randomly select a domain`
$d \leftarrow$ UniformPick($\mathcal{D}(m)$); `// Randomly select a dataset in the domain`
$s \leftarrow$ UniformPick($\mathcal{S}(d)$); `// Randomly select a seq. from the dataset`
$t \leftarrow$ UniformPick$\{1, \dots, |s| - l + 1\}$; `// Randomly select a start index`
return $s_{t:t+l-1}$; `// Return the segment`

For all the advanced reasoning tasks, including characterization, comparison, data transformation and temporal relationship, all the input time series are sampled from the *core dataset* (Appendix A.1). To ensure a balanced distribution over domains, datasets and sequences, we use *Hierarchical Uniform Sampling* presented in Algorithm 1 to obtain samples.

B.2 Data Bias

Unless otherwise specified, all samples have a random length in $[32, 256]$, and are z-scored to reduce data bias. The term data bias refers specifically to scale-based shortcuts or magnitude variance across heterogeneous domains, rather than semantic or sampling bias. We justify the use of z-score normalization on 2 main grounds: (1) *Preventing Magnitude-Based Shortcuts*, (2) *Standard Practice and Task Alignment*.

Preventing Magnitude-Based Shortcuts: TSAQA is a unified benchmark that aggregates data from 13 distinct domains, each possessed of vastly different magnitudes and units. Without normalization, large language models (LLMs) could exploit these scale differences as shortcuts to identify the source domain or dataset without performing genuine temporal reasoning. Normalization prevents this risk, forcing the model to rely on structural reasoning rather than memorizing absolute value ranges.

Standard Practice and Task Alignment: While real-world data is indeed not standardized, normalization is a ubiquitous and necessary preprocessing step in the time series literature to ensure numerical stability and cross-domain comparability. This approach aligns with established protocols in widely used benchmarks such as the UCR Archive (Dau et al., 2019b), and recent time series foundation model studies like Time-LLM (Jin et al., 2024) and Chronos (Ansari et al., 2024a), which consistently utilize normalization or scaling to handle distribution shifts. Additionally, the core objective of TSAQA is to evaluate reasoning capabilities. Z-score normalization is a linear transformation that preserves the fundamental properties required for these tasks while removing the confounding factor of arbitrary absolute magnitudes.

B.3 Characterization

The characterization task assesses the model's capability to analyze fundamental properties of time series, including trend, seasonality, and dispersion. Questions are posed as *TF* or *MC*, and final answers are determined through multi-LLM consensus.

Each instance consists of a univariate time series sample $\mathbf{x}$ with associated metadata (text description, domain, dataset). Given a sample $\mathbf{x}$ and its metadata, we instruct GPT-4o (Hurst et al., 2024) to generate one QA pair per instance using a randomly selected subset of one to three topics (from Table 7) and a question type (TF or MC). The process is as follows.

Step 1: Captioning & sub-topic selection. GPT first produces a short, neutral caption summarizing visible patterns (e.g., "gradual upward drift with weak weekly oscillation"). For each chosen topic, a sub-topic is sampled uniformly at random, e.g., trend, seasonality and dispersion.

Step 2: QA synthesis. GPT generates a TF or MC question grounded in $\mathbf{x}$, the caption, and the selected sub-topics.

Step 3: Self-verification. GPT performs a self-check and outputs a confidence score in [0,1]. We retain QA pairs only if confidence ≥ 0.95.

Step 4: Multi-LLM consensus. We query GPT-4.1, Gemini-2.5-Flash, and Claude-3.5-Sonnet using the same prompt, which includes the generated question along with its allowed answer choices (for both TF and MC formats), and collect their responses. To determine the final label, we adopt a weighted majority voting scheme among these three models and GPT-4o's original answer. Specif-

Topic	Sub-Topics
Trend	trend directions, trend types, trend shapes, trend strength, structural breaks, global and local trends
Seasonality	seasonality period, seasonal strength, multiple seasonality patterns, changing seasonality
Cyclicity	amplitude, peaks and trough, duration
Noise	noise level, global and local noise
Stationarity	stationarity strength, global and local stationarity, types of non-stationarity
Autocorrelation	types of autocorrelation, autocorrelation structures, lags, mean-reversion, persistence of autocorrelation
Dispersion	basic measures of variability (variance level), relative measures (signal-to-noise ratio level), coefficient of variation level, time-varying dispersion (volatility, heteroskedasticity), entropy, multi-scale dispersion
Shape	global shapes, local shapes, shapelets, motifs, curves, change points, pattern complexity
Irregularity	mean shift, variance shift, trend shift, seasonality irregularity, cyclic shift, distributional change, structural breaks, autocorrelation change
Correlation (Comparison only)	causal relationship, correlation strength, correlation types, correlation direction, cross-correlation, time-varying correlation (rolling correlation), lagged correlation, global and local correlations, correlation of decomposed components

Table 7: Topics and Sub-Topics for Time Series Analysis

ically, GPT-4.1 and Gemini-2.5-Flash are assigned higher weights of 1.5 each, reflecting their superior performance in preliminary evaluations, while Claude-3.5-Sonnet and GPT-4o are each assigned a weight of 1.0. The option with the highest total weighted vote is selected as the consensus answer. If a tie occurs—i.e., two or more answers receive the same highest weighted score—the corresponding QA pair is discarded to avoid introducing ambiguity or noise into the dataset. This ensemble-based strategy mitigates single-model biases, smooths out random errors, and produces more reliable and stable labels, which are crucial for ensuring the benchmark's quality.

Here's the *system* prompt template.

```
You are an expert of time series analysis.
1. Generate a meta_caption solely based on the
   meta information within 50 words.
2. Generate a detailed_caption based on both
   meta information and time series within 100
   words.
3. Generate a {} based on the time series,
   meta_caption, detailed_caption and the more
   detailed question instructions.
4. Generate a correct answer {} for your
   question.
5. A successful generation must meet the
   following conditions:
(1) there is only one correct answer;
(2) the question stricktly follows the
    instructions;
(3) the answer of the question cannot be easily
    derived from the meta_caption;
(4) the question should be about the time series
    itself without invovling external knowledge
   ;
(5) do not repeat the input time series in
   questions or answers.
6. Show your confidence of your determination of
    success within 0-1.
```

Here's the *user* prompt template.

```
The time series is {}.
Its meta information is {}.
The question must be about all these topics: {}.
The sub-topics of {} includes but not limited to
    {}.
First think about the all possible sub-topics
   and their taxonomy.
Then randomly pick a sub-topic from each topic
   ({}) to generate the question and answer
   pairs.
```

B.4 Comparison

The comparison task assesses the model's ability to analyze the relative characteristics of two time series, such as overall shape, temporal alignment, and correlation patterns. Similar to the characterization task, this task is also formulated as either *TF* or *MC* questions, where the model must identify similarities or differences between the given pair of sequences. The characteristics evaluated in the task are directly drawn from the standardized taxonomy of Topics and Sub-topics (from Table 7), which is shared with the Characterization task.

To construct the comparison set, we first obtain an anchor sample $\mathbf{x}$ from a specific domain M, dataset D, and sequence S. Given this anchor $\mathbf{x}$, we generate a set of ten comparison samples $\mathbf{x}'1, \ldots, \mathbf{x}'10$, each having the same length as $\mathbf{x}$. These samples are drawn in a structured manner to represent varying degrees of similarity: one from the same sequence S, two from different sequences within the same dataset D, three from other datasets within the same domain M, and four from entirely different domains. This tiered sampling strategy creates a natural hierarchy of difficulty, challenging the model to distinguish between subtle intra-sequence similarities and broader cross-domain differences.

Finally, we apply a process similar to the characterization task to generate QA pairs, where GPT-based models produce questions and candidate answers. The questions are then refined and validated through multi-LLM consensus to ensure accuracy and reduce bias, resulting in high-quality, reliable evaluation data for this task.

B.5 Data Transformation

The data transformation task evaluates the model's ability to infer and analyze the transformation relationship between an input time series and its transformed counterpart. These transformations are generated using well-established signal processing techniques, including the Fourier transform, wavelet transform, and first-order differencing, which are widely used in time series analysis to reveal underlying structures or remove trends. This task is particularly challenging because it requires the model to not only recognize the patterns in the raw input series but also to understand how specific mathematical operations alter these patterns.

We use predefined templates to formulate the task as either *TF* or *MC* questions. For *TF* questions, the model is asked to determine whether a given candidate sequence is indeed the correct transformation of the input time series $\mathbf{x}$ (e.g., whether it is the Fourier transform result of $\mathbf{x}$). For *MC* questions, the model must select the correct transformed sequence from multiple candidates, given both the input series $\mathbf{x}$ and the specified transformation operation (e.g., Fourier transform).

To ensure accuracy and consistency, all transformations are computed using professional and reliable scientific libraries (Harris et al., 2020; Virtanen et al., 2020). The correct transformation is generated directly from the input $\mathbf{x}$, while distractor sequences are created by applying the same transformation to randomly sampled, unrelated time series $\mathbf{x}'$. This setup forces the model to carefully analyze the relationship between the input and its transformation rather than relying on superficial similarities, providing a robust evaluation of its reasoning ability.

Here's the template to construct question.

```
The time series is {}.
Its meta information is {}.
The question must be about all these topics: {}.
The sub-topics of {} includes but not limited to
     {}.
First think about the all possible sub-topics
    and their taxonomy.
Then randomly pick a sub-topic from each topic
    ({}) to generate the question and answer
    pairs.
```

B.6 Temporal Relationship

The Temporal Relationship task is a discriminative sequence-level reasoning task, rather than a generative forecasting task. The task evaluates a model's ability to infer and analyze the temporal structure among sequential patches of a time series. Specifically, the task evaluates whether a model can understand the structural continuity and chronological dependencies of time series patches, testing 3 core capabilities: *Structural Continuity*, *Chronological Reasoning*, and *Contextual Discrimination*. (1) *Structural Continuity* tests whether the model can identify which candidate segment shares the underlying temporal dynamics required to validly continue a given trajectory. (2) *Chronological Reasoning* tests whether the model can reconstruct the correct temporal order of shuffled patches. (3) *Contextual Discrimination* tests the model's ability to distinguish the true continuation from "plausible" but incorrect alternatives that may share similar global statistics but lack local continuity. This task is formulated as *true-or-false (TF)*, *multiple-choice (MC)*, or *puzzling (PZ)* questions.

Given the first chronological patch $\mathbf{x}$: (1) A TF question asks the model to determine whether a candidate patch $\mathbf{y}$ is the immediate successor of $\mathbf{x}$. (2) An MC question requires the model to select the correct next patch from four candidates $[\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4]$.

The false candidates in both TF and MC settings are randomly sampled from the full dataset but are guaranteed to come from sequences different from that of $\mathbf{x}$, preventing the model from simply memorizing patterns. For PZ questions, the model is presented with four shuffled successor patches of $\mathbf{x}$ and must reconstruct their correct chronological order, which poses a greater challenge as it requires deeper temporal reasoning. All questions are generated using predefined templates to ensure consistency and diversity.

We use the following question template to construct questions.

```
Which of the following choices is most likely
    the future continuation of the given time
    series?
Respond ONLY with the letter of the correct
    choice (A, B, C, or D)

Choices:
A: {}
B: {}
C: {}
D: {}
```

```
Is the following patch the future continuation
    of the given time series?
{}
Respond ONLY with the letter of the correct
    choice (T or F).

Choices:
T: True.
F: False.
```

B.7 Anomaly Detection

First, all time-series data are standardized using z-score normalization to remove scale effects across different features. Next, we randomly sample a subsequence of length T, where $T \in [32, 256]$, from each time-series instance to capture varying temporal dynamics. To address class imbalance, we count the number of anomalous sequences and randomly select an equal number of normal sequences, resulting in a balanced dataset. Finally, we enrich each sample with meta information, domain information, the normalized time-series subsequence, and its corresponding label.
Here's the question template.

```
Determine whether the given time series contains
     anomalies.
Respond ONLY with the letter of the correct
    choice (T or F).

Choices:
T: True.
F: False.
```

B.8 Classification

Information about the time series and the task is given in the text description. Here's the template to construct questions.

```
Classify the given time series into one of the
    categories below.
Respond ONLY with the letter of the correct
    choice (A, B).

Choices:
A: {}
B: {}
```

C Examples

In this section, we show some examples of the constructed QA pairs.

TSAQA — Sample 1

Time Series Info

domain: sales

Question & Answer

question: Which statement best describes the overall characteristics of this time series with regard to its motif, variability, and trend? Respond *ONLY* with the letter of the correct choice (A, B, C, or D).
Choices:

A: The time series has a simple motif, low variability, and lacks a consistent upward or downward trend.

B: The time series features a complex motif, high variability, and an upward trend through the series.

C: The time series has a constant value, no motif, and displays a strong downward trend.

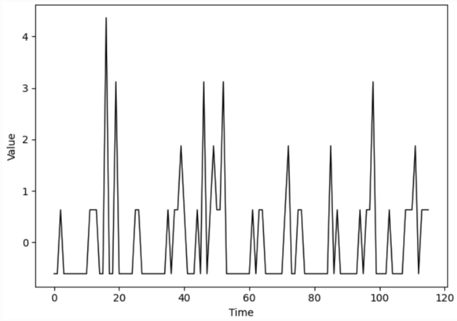
D: The time series contains random patterns with high variability and a significant upward trend.

question_type: multiple_choices **task:** characterization

answer: A

TSAQA — Characterization Sample 1

Time Series Info



Description: A numerical sequence of hourly aggregated Uber pickup counts (integer values) for a single New York City taxi zone, where each element represents the total number of pickups recorded in that zone during one hour.
Question Type: TF **Domain:** transport **Dataset:** uber_tlc_hourly

Question & Answer

Question: Does the time series exhibit constant variance throughout, indicating no change in volatility, along with a clearly defined global shape? Respond ONLY with the letter of the correct choice (T or F).
Choices:

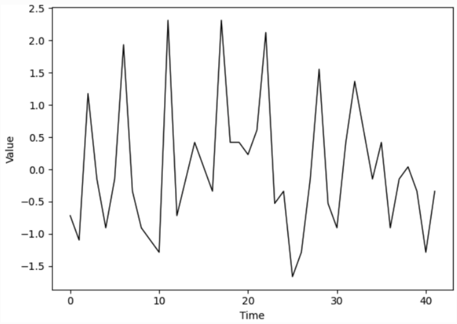
T: True.

F: False.

Answer: F

TSAQA — Characterization Sample 2

Time Series Info



Description: A time series of daily page-view counts for a single English Wikipedia article.
Question Type: MC **Domain:** web **Dataset:** wiki_daily_100k

Question & Answer

Question: Which of the following best describes a prominent motif and cyclic feature observed in the time series? Respond ONLY with the letter of the correct choice (A, B, C, or D).
Choices:

A: A recurring increase in values every few days with a noticeable peak at day 11.

B: A consistent downward trend over the entire period without any peaks.

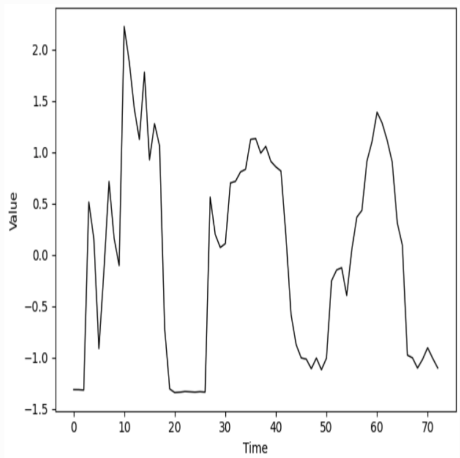
C: An irregular pattern with no identical sequences or cycles.

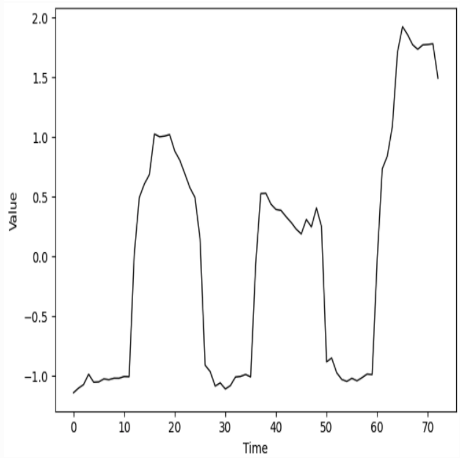
D: A repeating cycle of gradual increase and sudden drop every ten days.

Answer: **A**

TSAQA — Comparison Sample 1

Time Series Info





Description 1: A time series of hourly electricity consumption measurements for the Rat building, representing one building2019s power usage.
Description 2: A time series of hourly electricity consumption measurements for the Rat building, representing one building2019s power usage.
Question Type: TF **Domain:** energy **Dataset:** bdg2_rat

Question & Answer

Question: Does time series 1 display any global upward trend over the entire period?
Choices:

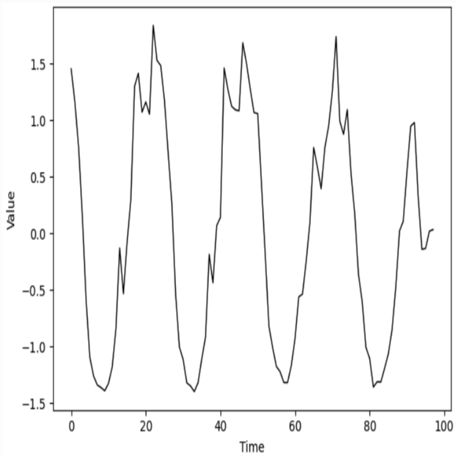
T: True.

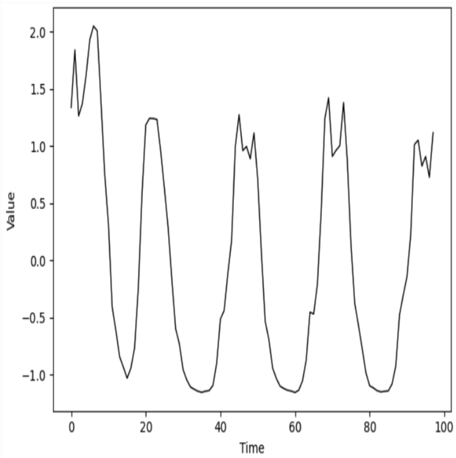
F: False.

Answer: **F**

TSAQA — Comparison Sample 2

Time Series Info





Description 1: A time series of hourly pedestrian count measurements from a single sensor in Melbourne.
Description 2: A time series of hourly pedestrian count measurements from a single sensor in Melbourne.
Question Type: MC **Domain:** transport **Dataset:** pedestrian_counts

Question & Answer

Question: Which time series displays stronger global stationarity, evident from its overall pattern smoothness without clear seasonal strength? Respond ONLY with the letter of the correct choice (A, B, C, or D).
Choices:

A: Time series 1

B: Time series 2

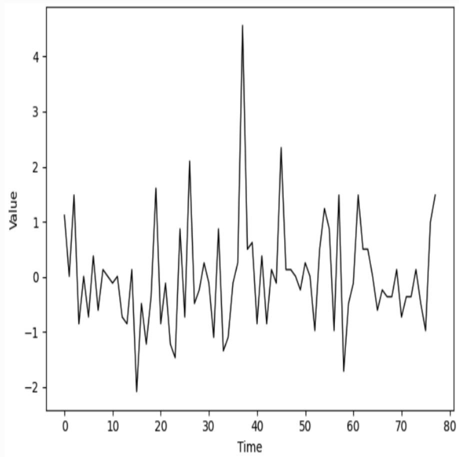
C: Both have similar global stationarity

D: Neither has strong global stationarity

Answer: B

TSAQA — Data Transformation Sample 1

Time Series Info

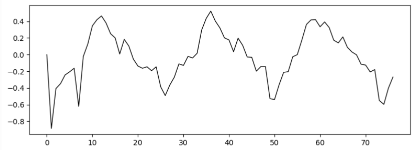


Description: A time series of daily page2010view counts for a single English Wikipedia article.
Question Type: MC **Domain:** web, energy, finance, finance **Dataset:** wiki_daily_100k, gfc12_load, exchange_rat, fred_md

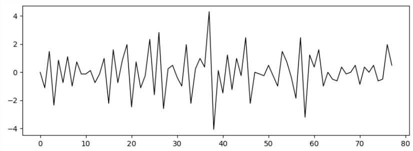
Question & Answer

Question: Which of the following choices is most likely the First Order Difference of the given time series? Respond ONLY with the letter of the correct choice (A, B, C, or D).
Choices:

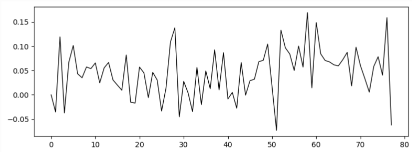
A:



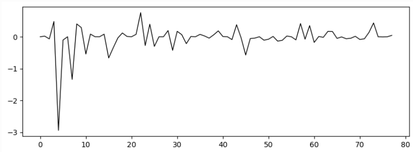
B:



C:



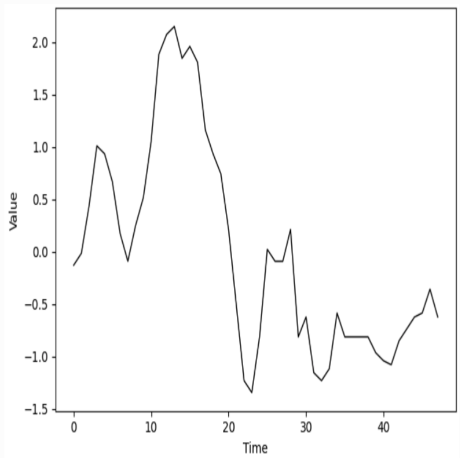
D:



Answer: B

TSAQA — Data Transformation Sample 2

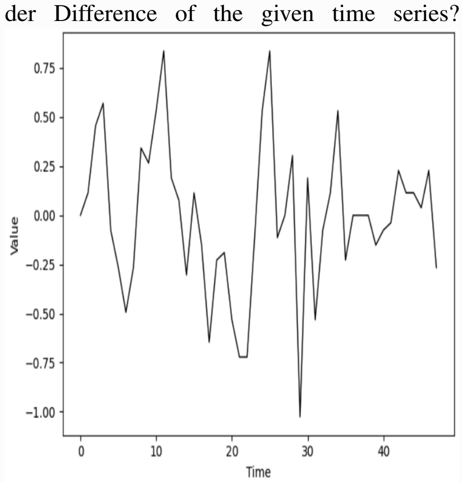
Time Series Info



Description: A time series of daily relative sunspot number measurements, where each value represents the quantified count of sunspot activity on the Sun2019s visible disk.
Question Type: TF **Domain:** nature, energy
Dataset: Nature_sunspot, gfc12_load

Question & Answer

Question:
Is the following sequence the First Order Difference of the given time series?



Respond ONLY with the letter of the correct choice (T or F).
Choices:

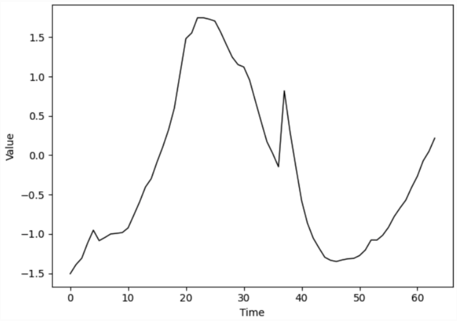
T: True.

F: False.

Answer: T

TSAQA — Temporal Relationship Sample 1

Time Series Info

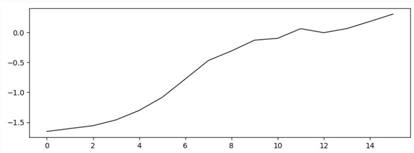


Question Type: MC

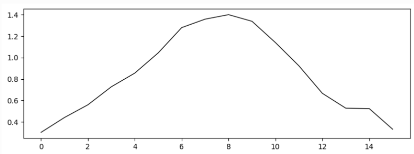
Question & Answer

Question: Which of the following choices is most likely the future continuation of the given time series? Respond ONLY with the letter of the correct choice (A, B, C, or D).
Choices:

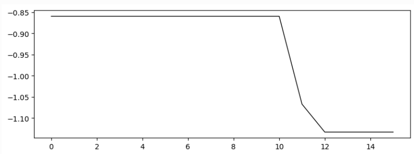
A:



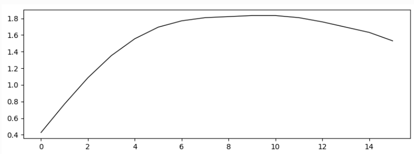
B:



C:



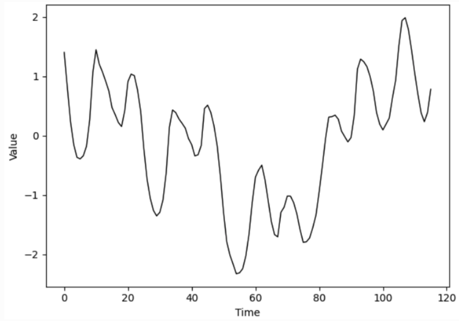
D:



Answer: B

TSAQA — Temporal Relationship Sample 2

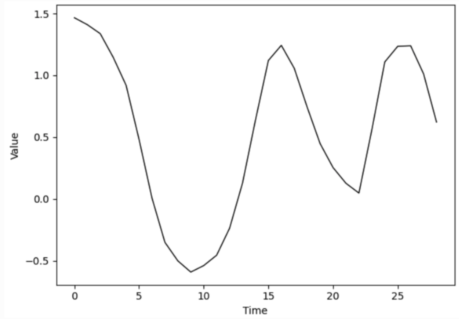
Time Series Info



Question Type: TF

Question & Answer

Question:
Is the following patch the future continuation of the given time series?



Respond ONLY with the letter of the correct choice (T or F).
Choices:

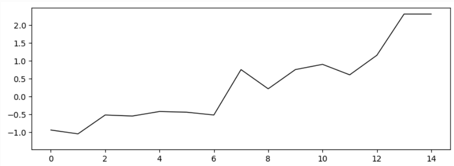
T: True.

F: False.

Answer: T

TSAQA — Temporal Relationship Sample 3

Time Series Info

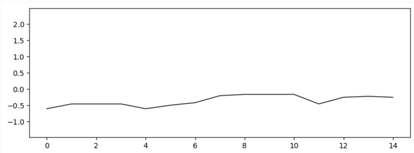


Question Type: PZ **Domain:** finance

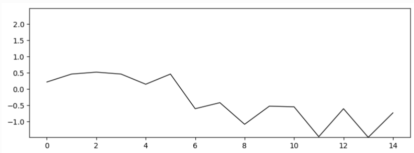
Question & Answer

Question: The given time series is the first patch of the sequence. Below are the remaining patches, labeled as A, B, C, and D. Arrange A, B, C, D in the correct order to reconstruct the original sequence.
Choices:

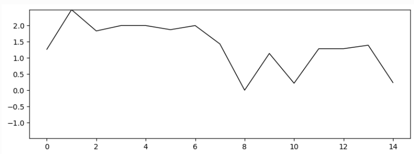
A:



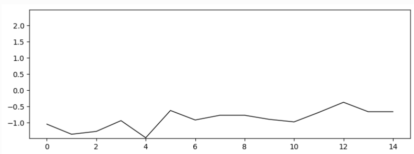
B:



C:



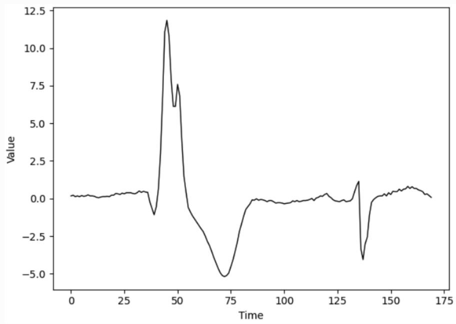
D:



Answer: C,B,D,A

TSAQA — Anomaly Detection Sample 1

Time Series Info



Description: This is an electrocardiogram (ECG) time series, and the anomalies represent ventricular premature contractions. The ECG recordings were made using Del Mar Avionics model 445 two-channel reel-to-reel Holter recorders, and the analog signals were recreated for digitization using a Del Mar Avionics model 660 playback unit. The digitization rate (360 samples per second per channel) was chosen to accommodate the use of simple digital notch filters to remove 60 Hz (mains frequency) interference.
Question Type: TF **Domain:** Healthcare **Dataset:** ECG

Question & Answer

Question: Determine whether the given time series contains anomalies. Respond ONLY with the letter of the correct choice (T or F).
Choices:

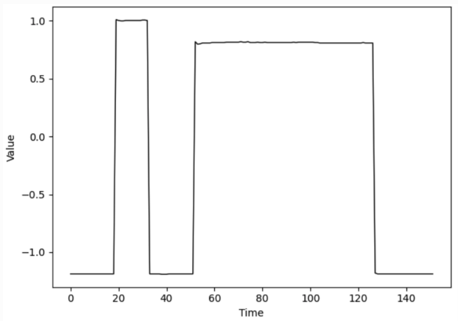
T: True.

F: False.

Answer: T

TSAQA — Classification Sample 1

Time Series Info



meta_info: This time series comes from a dataset capturing process control measurements recorded by individual sensors during the fabrication of silicon wafers in semiconductor manufacturing, providing data for monitoring and classifying normal and abnormal production processes.
Question Type: MC **Domain:** manufacturing
Dataset: UCR_Classification_Wafer

Question & Answer

Question: Classify the given time series into one of the categories below. Respond ONLY with the letter of the correct choice (A, B).
Choices:

A: normal process

B: abnormal process

Answer: B

D Experiment Analysis

We conducted an in-depth analysis of results from the selected Large Language Models. Specifically, our analysis is divided into three major categories: **Accuracy Correlate Analysis**, **Task-Specific Analysis**, and **Case Study**. For each analysis, we selected models from both commercial and open-source families. In particular, we chose the two best-performing models from Table 3 evaluated on our TSAQA Benchmark—namely, GPT-4.1, Gemini 2.5 Flash, LLaMA3-8B, and Qwen3-8B. For LLaMA3.1-8B and Qwen3-8B, we analyzed both the zero-shot and instruction tuned models, resulting in a total of six models considered in our analysis.

D.1 Accuracy Correlate Analysis

In this category, we primarily examined how model accuracy or overall score correlates with various factors, including input length, the topics and subtopics used in time-series question generation, and the influence of domain differences on model performance.

Input Length v.s. Accuracy. To understand how input length impacts model accuracy, we conducted a detailed analysis comparing the length of each input with its corresponding accuracy. Specifically, the input length is calculated as *len(ts + description + domain + dataset + task + question_type + question)* with *String* type. The results are visualized in Figure 3. Each plot may contain input length starting and ending at different length as each task contains questions with different lengths. Across all six models and five tasks, excluding the Temporal Relation task, we observe a consistent trend that longer questions with greater input length generally result in lower accuracy and weaker overall model performance. However, the Temporal Relation task exhibits the opposite behavior, where accuracy improves with increasing input length.

To understand this discrepancy, we conducted a detailed analysis of the four advanced analysis tasks *(Characterization, Comparison, Data Transformation, Temporal Relation)* in our proposed TSAQA Benchmark, focusing on how different question types *(MC, TF, PZ)* and their corresponding input lengths correlate with model accuracy. The results are visualized in Figure 4. The results indicate that for all four advanced analysis tasks, MC and TF question types show a decline in accuracy with increasing input length,

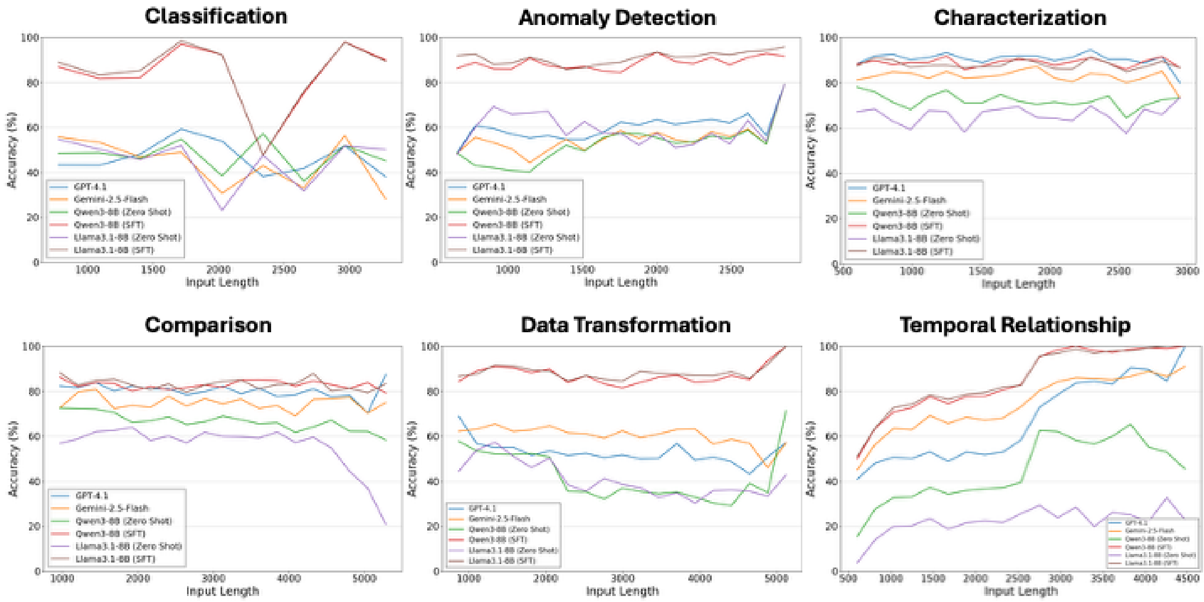


Figure 3: Input lengths vs. Accuracy by Tasks among six models.

Group	Model	Topics	SubTopics	Characterization	Comparison
Zero Shot	GPT-4.1	1	1	89.53	80.39
		2	2	91.75	80.87
		3	3	92.10	79.51
	Gemini-2.5-Flash	1	1	83.68	77.57
		2	2	83.01	74.57
		3	3	83.11	72.62
	Qwen3-8B	1	1	69.14	67.48
		2	2	72.35	66.38
		3	3	74.86	67.08
	LLaMA3.1-8B	1	1	63.85	58.93
		2	2	66.16	58.75
		3	3	65.53	58.71
Instruction Tuning	Qwen3-8B (SFT)	1	1	86.22	82.48
		2	2	89.74	82.50
		3	3	90.57	83.15
	LLaMA3.1-8B (SFT)	1	1	86.02	84.81
		2	2	88.89	81.84
		3	3	89.14	82.18

Table 8: Number of topics and subtopics v.s. Score

whereas the newly proposed PZ type exhibits the opposite trend. This implies that the model is actively using global contexts, all time series segments, to deduce the correct chronological order for answering PZ type question, which confirms that the model is engaging in deductive reasoning rather than local pattern matching. This proves that PZ type question is a rigorous probe for *Global Causal Reasoning*. Consequently, models whose accuracy improves with input length likely demonstrate a stronger ability to reason directly over time-series patterns.

Topics & Subtopics v.s. Accuracy. In our proposed TSAQA Benchmark, tasks such as Characterization and Comparison include questions generated by prompting GPT to select topics and subtopics from a predefined list 7.

To understand how the complexity of topics and subtopics influences model performance, we analyzed the relationship between the number of topics and subtopics used in each question and the corresponding model accuracy. Specifically, we examined how varying topic and subtopic counts affect the model's ability to reason across different levels of conceptual complexity. In our benchmark, each question contains between one to three topics, indicating that one to three distinct topics are considered during question generation, and between one to three associated subtopics depending on the selected topics. The results are summarized in Table 8. Based on the results, we observe that

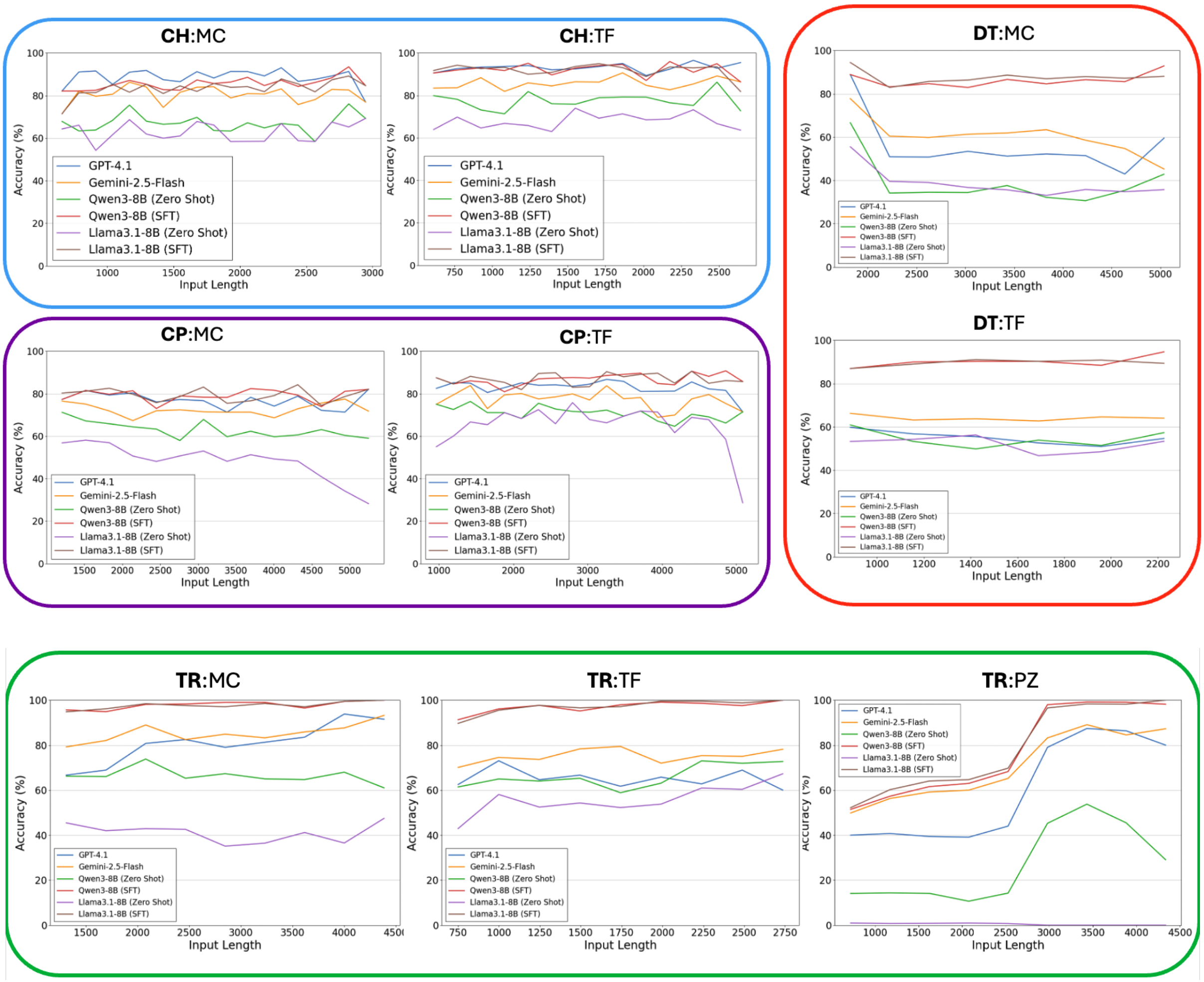


Figure 4: Input length vs. Accuracy by Question Types. CH, CP, DT, and TR denote Characterization, Comparison, Data Transformation, and Temporal Relationship. MC, TF, and PZ denote true-or-false, multiple-choice, and puzzling.

the complexity of questions with varying number of topics and subtopics don't have direct impact on model accuracy. The absence of significant performance differences across varying topics and subtopics indicates that the TSAQA Benchmark is largely unbiased, suggesting that models do not rely on topic-level content from the question, but depend mainly on their time-series analytical capability.

In addition, we further analyzed the difficulty of individual topics by examining how different topic combinations influence model performance across tasks. For questions containing more than one topic, each question was expanded into multiple rows, allowing us to isolate the accuracy associated with each topic across all questions that included it. Comparing the average accuracy of questions containing each topic allowed us to identify which topics posed greater challenges, especially when combined with others. The results of questions from Comparison Task are visualized in Figure 5. Generally, we found that questions with topics such as seasonality, autocorrelation, dispersion, and noise are harder for model.

Domain v.s. Accuracy. In the TSAQA Benchmark, both questions and descriptions are generated from the associated dataset and domain.

To evaluate the influence of such contextual information, we conducted an in-depth analysis of how domain variation impacts overall model accuracy on TSAQA Benchmark. The results are summarized in Table 9. Our analysis reveals that questions from domains including Synthetic, IT, Robotics, and Web pose greater challenges to models under the zero-shot setting, while questions from Sales and Web domains remain the most difficult after instruction tuning. Notably, the questions from Synthetic domain, which initially produced

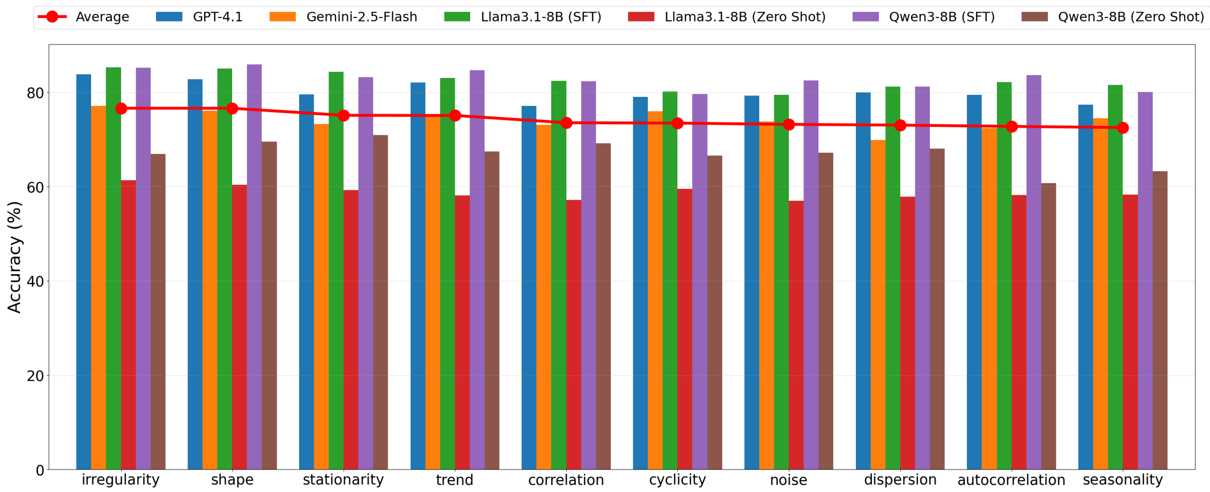


Figure 5: Topics vs. Accuracy of Comparison Task.

Group	Model	Mf	ES	Rbt	Bm	It	Eg	Hc	TP	Nt	Fc	S	W	Sc
Zero Shot	GPT-4.1	84.29	72.31	38.91	49.93	48.56	62.91	64.18	66.07	64.82	67.49	68.73	67.21	**26.95**
	Gemini-2.5-Flash	76.58	59.70	57.59	49.78	48.28	69.44	70.40	71.55	69.52	73.88	70.12	66.02	**14.55**
	Qwen3-8B	83.85	53.09	58.37	50.02	48.56	50.45	50.91	49.78	50.60	51.08	50.99	50.47	**15.49**
	LLaMA3.1-8B	73.45	62.42	59.92	56.48	47.75	42.39	42.18	41.74	41.16	40.86	40.79	40.07	**21.88**
Instruction Tuning	Qwen3-8B	95.05	94.93	84.05	85.44	86.21	85.26	85.15	84.84	83.32	85.25	**82.48**	82.49	98.31
	LLaMA3.1-8B	96.22	96.23	87.94	87.34	92.00	86.63	85.80	85.4	84.28	85.83	**82.23**	83.01	98.03

Table 9: Domain v.s. Accuracy. Mf denotes Manufacturing. ES denotes Environment Sensing. Rbt denotes Robotics. Bm denotes Biomedical. Eg denotes Energy. Hc denotes Healthcare. Tp denotes Transport. Nt denotes Nature. Fc denotes Finance. S denotes Sales. W denotes Web. Sc denotes Synthetic. The lowest and second-lowest results for each model are highlighted in **bold** and underlined, respectively.

the lowest accuracies across all models, show the most substantial improvement after instruction tuning, achieving the highest scores among all domains. However, Web-related questions persist as difficult cases, indicating that domain-specific complexities in this category are not fully mitigated by instruction tuning.

D.2 Task Specific Analysis

In this category, we examined how each model performed across the tasks proposed in our TSAQA Benchmark. Specifically, we focused on the 3 analysis tasks: Data Transformation, Temporal Relationship, and Comparison.

Data Transformation. We analyze model performance on the Data Transformation task, which is designed to evaluate a model's understanding of three transformation operators: Fourier Transform (FT), Wavelet Transform (WT), and First-Order Differencing (FOD). For each operator, we assess performance by measuring the accuracy on both MC and TF question formats. As shown in Table 10, for zero-shot evaluation, our key finding highlights a limitation in which both commercial and open-source models fail to provide accurate answers, except of FOD. In contrast, our instruction-tuned models show a better performance, achieving high accuracy across all tasks. However, FT is still very challenging even after instruction tuning.

To explain our findings, we attribute this systematic performance disparity to two primary factors: the scope of temporal dependency and arithmetic complexity. As shown in Table 10, there is a clear performance degradation trend ($FOD > WT > FT$). This performance degradation is likely due to 3 reasons. (1) FOD relies solely on adjacent time steps ($x_t - x_{t-1}$), aligning well with the local attention capabilities of Transformers. (2) WT requires reasoning over localized windows in both time and frequency. As the dependency scope widens beyond immediate neighbors, model performance drops. (3) FT necessitates aggregating information from the entire sequence to determine frequency components. This global arithmetic reasoning is inherently challenging for LLMs' next-token prediction paradigm, resulting in the lowest performance. The results systematically validate that current LLMs struggle with tasks requiring global aggregation and complex arithmetic compared to robust local

Group	Model	MC			TF		
		FT	WT	FOD	FT	WT	FOD
Zero Shot	GPT-4.1	26.32	35.39	91.90	51.36	51.64	59.81
	Gemini-2.5-Flash	27.97	53.19	100.00	50.25	53.59	85.90
	Qwen3-8B	9.06	28.40	66.4	52.57	52.05	52.66
	LLaMA3.1-8B	24.07	23.87	61.70	52.17	48.87	54.50
Instruction Tuning	Qwen3-8B	67.93	87.55	100.00	80.02	99.90	89.14
	LLaMA3.1-8B	71.83	88.79	99.70	82.54	89.24	98.36

Table 10: Analysis of Data Transformation Task. MC and TF denote multiple-choice and true-or-false, respectively. FT, WT, and FOD denote Fourier Transform, Wavelet Transform, and First-Order Differencing. We evaluate the accuracy on MC and TF questions from Data Transformation Task for each of the three transform operators.

Group	Model	Finance	Healthcare	Transport	Sales	Energy	Nature	Web
Zero Shot	GPT-4.1	62.22	57.75	55.86	52.62	52.53	48.87	**46.53**
	Gemini-2.5-Flash	76.59	80.12	76.65	66.54	76.95	72.46	**63.51**
	Qwen3-8B	27.81	27.76	24.36	22.87	24.32	21.43	**17.90**
	LLaMA3.1-8B	**0.77**	0.78	0.98	0.94	0.88	1.25	0.92
Instruction Tuning	Qwen3-8B	73.31	77.11	72.54	61.03	74.05	68.92	**58.61**
	LLaMA3.1-8B	75.25	77.50	72.22	61.80	75.80	71.16	**60.86**

Table 11: Domain v.s. Accuracy of the PZ question type in the Temporal Relationship task. The lowest and second-lowest results for each model are highlighted in **bold** and underlined, respectively.

pattern matching, which also explains the results shown in Table 10.

Temporal Relationship. We analyzed model performance on the Temporal Relationship task, focusing specifically on our newly proposed Puzzling (PZ) question type. Beyond the input length versus accuracy analysis previously presented in Figure 4, we further examined how domain-level information influences model performance on Puzzling questions and the nature of model errors.

Domain-Level Analysis. The results are summarized in Table 11. The results show that the Web domain remains the most challenging for Puzzling questions across both zero-shot and instruction-tuning settings. Sales and Nature also exhibit lower accuracies, with Sales remaining difficult even after instruction-tuning. This indicates that domains such as Web and Sales impose greater temporal analysis difficulty on models, which is consistent with our findings from the overall domain vs. score analysis presented in Table 9.

Error Analysis: Smoothness Bias. To understand why models struggle in these specific domains (Web, Sales), we further analyzed the boundary consistency of incorrect predictions using the instruction-tuned LLaMA3.1-8B and Qwen3-8B. We calculated the "Smoothness Gap", defined as the difference between the boundary distance of the Ground Truth sequence D_{gt} and the model Predicted sequence D_{pred}. Here, the boundary distance is calculated as the Euclidean distance between the last time step of a preceding patch and the first time step of its succeeding patch. We found that in the challenging domains identified above, the models consistently constructed sequences where the boundary transitions were smoother than the ground truth (i.e., $D_{gt} > D_{pred}$) as shown in Table 12. This reveals that current models suffer from an smoothness bias. They fail in volatile domains because they attempt to repair legitimate discontinuities by selecting patches that connect more seamlessly. The error signifies models' failures to grasp the specific physical dynamics of targeting domains and highlights the critical utility of the PZ type question: it acts as a discriminator for temporal fidelity, penalizing models that rely on generic smoothing priors and rewarding those that can capture the specific irregular structural dynamics of the target domain.

Comparison. We analyze model performance on the Comparison task, specifically investigating whether providing explicit domain-level context affects model accuracy. The task requires comparing two input time series, which we test under two conditions: (1) when both series originate from the same domain and (2) when they are from different domains. In both scenarios, the corresponding domain names are provided to the model as textual description. As shown in Table 13, we observe no significant performance difference between the same-domain and different-domain settings across either MC or TF questions. This finding suggests that our Comparison task is domain-invariant. Additionally, combining the results from our

Group	Model	Metric	Finance	Healthcare	Transport	Sales	Energy	Nature	Web
Instruction Tuning	Qwen3-8B	D_{gt}	1.21	1.84	1.48	2.50	0.93	0.90	2.57
		D_{pred}	1.67	1.81	1.56	2.00	1.28	1.40	2.31
		Gap	-0.46	0.03	-0.08	**0.50**	-0.35	-0.50	<u>0.26</u>
	LLaMA3.1-8B	D_{gt}	1.30	1.97	1.51	2.52	0.94	0.91	2.59
		D_{pred}	1.51	1.84	1.46	2.02	1.18	1.20	2.14
		Gap	-0.21	0.13	0.05	**0.50**	-0.24	-0.29	<u>0.45</u>

Table 12: Analysis of Smoothness Bias in PZ question type. For each domain, we report the boundary distance of Ground Truth (D_{gt}), Predicted (D_{pred}), and the Smoothness Gap (Gap). The largest and second-largest Gap for each model are highlighted in **bold** and <u>underlined</u>, indicating the model is over-smoothing.

analysis on number of topics & subtopics vs. scores from Table 8, these results indicate that the model's performance is notably stable, which again proves the quality of the proposed dataset. Consequently, to answer correctly, models must reason based on the intrinsic patterns of the time series data itself, rather than relying on the textual context as a simple heuristic.

Group	Model	Same Domain MC	Same Domain TF	Different Domain MC	Different Domain TF
Zero Shot	GPT-4.1	76.27	83.62	78.06	83.48
	Gemini-2.5-Flash	70.97	77.90	74.06	77.63
	Qwen3-8B	62.99	70.67	63.54	71.60
	LLaMA3.1-8B	49.43	67.13	50.82	68.93
Instruction Tuning	Qwen3-8B	77.04	85.64	82.14	87.95
	LLaMA3.1-8B	78.02	86.32	81.24	87.35

Table 13: Analysis of Comparison tasks.

D.3 Case Study

In this category, we analyze selected findings from model outputs, focusing on interesting behaviors observed in our newly proposed Puzzling question type. The insights from this analysis may provide useful implications for future work. Specifically, we examine Puzzling (PZ) questions that models answered incorrectly and present several representative case examples.

First Letter Distribution. To explore potential biases in model behavior, we analyzed the distribution of the first letters in model responses for Puzzling questions. Each question provides the first time-series slice and requires reordering of the remaining four. We visualize these distributions in Figure 6, using the instruction-tuned LLaMA3.1-8B and Qwen3-8B, whose outputs adhere more strictly to the expected format. Based on the figure, we observe that for both models, the output distributions of questions that received full credit appear approximately uniform, with similar counts

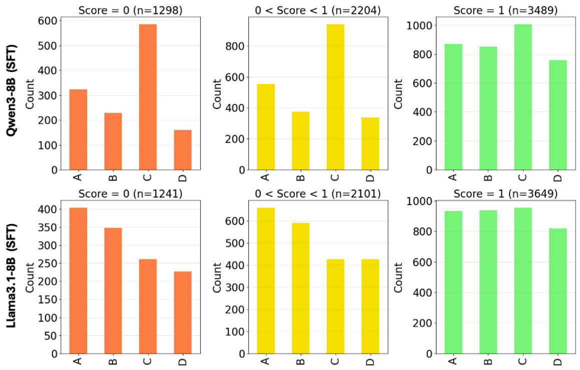


Figure 6: First-letter distribution of model outputs for LLaMA3.1-8B and Qwen3-8B on PZ questions

across all possible choices. However, for questions that were answered incorrectly or received partial credit, an interesting pattern emerges: Qwen3-8B tends to output the choice C more frequently, whereas LLaMA3.1-8B shows a stronger tendency to output A.

Incorrect Output Format. While models generally demonstrate a strong understanding of the expected response format for common question types such as MC and TF, we further analyzed several examples from the closed-source models GPT-4.1, GPT-4o, and Gemini 2.5 Flash on PZ questions. We include sample responses from each model that highlight unique or unexpected behaviors—cases where the models' answers do not strictly adhere to the specified instructions for the PZ question format.

Sample responses from GPT-4o:

```
Question: 'The given time series is the first
    patch of the sequence.\nBelow are the
    remaining patches, labeled as A, B, C, and D
    .\nArrange A, B, C, D in the correct order
    to reconstruct the original sequence.\n\
    nPatches:\nA: [-1.6371, 0.2207, 0.9432,
    0.6299, 0.7266, 0.1741, -1.2879, -1.9146,
    0.2064, 0.7534, 0.5448, 0.4051, 0.4329,
    -1.5225, -1.86, 0.0201, 0.4759]\nB:
    [-1.3166, -1.8215, 0.7561, 0.9665, 0.7722,
    0.8286, 1.125, -1.262, -1.7866, -1.4097,
```

```
    0.6872, 1.3175, 0.9611, 1.2673, -0.9737,
    -1.3918, 0.6585]\nC: [0.2207, 0.4803,
    0.5153, -1.3291, -1.7803, 0.398, 0.6684,
    0.4141, 0.5305, 0.7472, -1.3524, -1.6084,
    0.5636, 0.9396, 0.1759, 0.6039, 0.7579]\nD:
    [0.4938, 0.6209, -1.3506, -1.7982, 0.2323,
    0.7454, -0.1356, 0.6317, 0.6102, -1.3828,
    -2.1053, 0.3595, 0.86, 0.5752, 0.3998,
    0.8241, -1.1912]\n'.

Model Response: 'Input TS -> C -> D -> A -> B'.

Correct Answer: 'D,A,C,B'.
```

Sample responses from GPT-4.1:

```
Question: 'The given time series is the first
    patch of the sequence.\nBelow are the
    remaining patches, labeled as A, B, C, and D
    .\nArrange A, B, C, D in the correct order
    to reconstruct the original sequence.\n\
    nPatches:\nA: [0.783, 0.918, 0.9904, 1.0754,
     1.1464, 1.2123, 1.216, 1.3001, 1.2585,
    1.2408, 1.2431]\nB: [-0.0222, -0.1759,
    -0.3978, -0.5977, -0.6948, -0.8242, -0.9167,
     -0.9671, -1.0311, -1.1072, -1.1483]\nC:
    [-1.2175, -1.3207, -1.3356, -1.38, -1.4384,
    -1.4477, -1.4655, -1.5178, -1.5384, -1.5865,
     -1.628]\nD: [1.251, 1.2076, 1.1889, 1.1529,
     1.0628, 0.9376, 0.8265, 0.6709, 0.4682,
    0.3388, 0.1427]\n'.

Model Response: 'B, C, [-0.1936, -0.1581,
    -0.0984, -0.0119, 0.0745, 0.1174, 0.2594,
    0.3837, 0.4481, 0.6023, 0.7041], A, D'.

Correct Answer: 'A,D,B,C'.
```

Sample responses from Gemini-2.5-Flash:

```
Question: 'The given time series is the first
    patch of the sequence.\nBelow are the
    remaining patches, labeled as A, B, C, and D
    .\nArrange A, B, C, D in the correct order
    to reconstruct the original sequence.\n\
    nPatches:\nA: [0.9363, -0.4796, 0.7003,
    -0.0077, -1.4236, -1.4236, -1.1876, -1.1876,
     0.4643, -0.2436, -0.4796, 0.9363, -0.7156,
    -0.2436, -2.1316, 0.2283, 0.4643, -0.7156,
    0.9363, -0.2436, -0.2436, 0.2283, 2.3523,
    -0.2436, 0.2283, -0.9516, 0.7003, -1.6596,
    0.2283, 0.4643, -0.4796, 0.7003, 0.7003,
    -0.2436, -0.0077, -0.9516, -1.1876]\nB:
    [0.2283, 0.9363, -0.2436, 1.6443, 0.2283,
    0.7003, 0.4643, -0.9516, 0.2283, -0.0077,
    -0.2436, -0.7156, 0.7003, -0.7156, 0.4643,
    -0.7156, -0.0077, 3.0603, 0.7003, -1.4236,
    0.9363, -0.7156, -1.6596, -0.4796, -1.1876,
    0.7003, -0.2436, -0.9516, 0.7003, 0.4643,
    -0.4796, -1.8956, -0.7156, 1.6443, 1.8803,
    -1.1876, -0.0077]\nC: [-0.2436, -1.4236,
    -0.7156, -1.1876, -0.0077, -0.0077, -0.0077,
     1.4083, -0.2436, -0.7156, -0.9516, 0.2283,
    0.4643, 1.6443, 0.9363, 3.7682, 0.4643,
    -0.7156, 0.2283, -0.0077, -0.0077, -0.7156,
    -1.6596, 1.6443, -0.4796, -0.4796, -1.6596,
    -2.1316, -0.7156, 0.4643, 2.5883, 2.5883,
    0.2283, 1.1723, -1.6596, 1.4083, 1.4083]\nD:
     [-0.0077, -0.0077, -1.4236, -0.4796,
    -0.2436, 0.4643, -0.2436, -1.6596, 1.4083,
    -1.4236, -1.1876, -0.7156, 0.7003, 1.4083,
    -0.0077, -0.2436, 0.9363, -0.2436, -0.2436,
    1.1723, 1.1723, 0.4643, -0.4796, -0.2436,
    -0.0077, -0.0077, 0.2283, -1.1876, -0.4796,
    -0.4796, 0.2283, 0.4643, 0.2283, -0.7156,
    0.9363, -0.0077, -0.9516]\n'.

Model Response: '{\n "answer": "The provided
    time series patches do not exhibit any
    standard prefix-suffix overlaps. A common
    approach for reconstructing a sequence from
    patches is to identify the longest
    overlapping segments between the end of one
    patch and the beginning of another. After
    performing a comprehensive search for such
    overlaps (ranging from 1 to 37 elements)
    ........the default alphabetical order is
    chosen."\n}'.

Correct Answer: 'A,B,C,D'.
```

Based on the sample responses, we observe that the Puzzling question type often causes models to struggle in following the question instructions. This finding highlights the continued importance of developing and designing more reasoning-based time-series questions beyond traditional formats. Such question types better evaluate a model's ability to perform complex reasoning and manipulate temporal information, providing a more comprehensive assessment of time-series understanding.

E Human Evaluation

We further examine annotators' explanations in cases of disagreement. In the single-series benchmark, the largest source of mismatches is ambiguous questions (43%). Among well-formed cases, 24% involve trends, while volatility-, stability-, and periodicity-related issues each account for 10%. A small fraction (5%) reflects residual annotator uncertainty. In the multi-series benchmark, mismatches are more strongly tied to stochastic properties: volatility-related issues dominate (23%), followed by stability (13%). Periodicity- and lag-related issues each contribute 7%, while trend-related mismatches are rare (3%). Nearly half of the disagreements (47%) again arise from ambiguous questions, underscoring the greater interpretive difficulty of the multi-series setting. (See Figure 7)

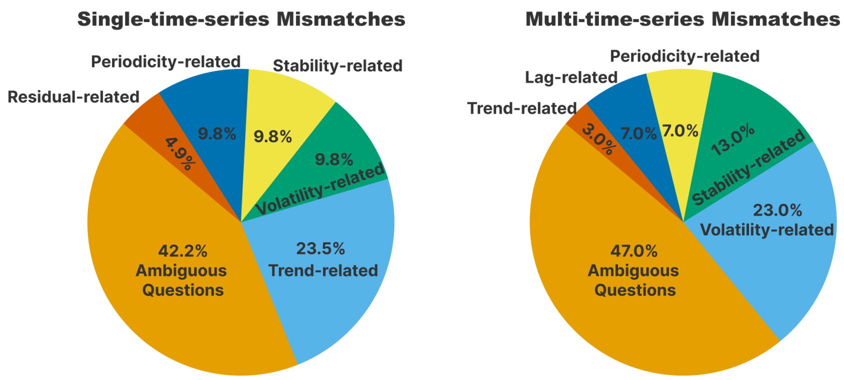


Figure 7: Human explanations for answer mismatches in TSAQA