

000 TRQA: TIME SERIES REASONING QUESTION AND 001 002 ANSWERING BENCHMARK 003 004

005 **Anonymous authors**

006 Paper under double-blind review

007 008 ABSTRACT 009

011 Time series data underpin critical applications across domains such as finance,
012 healthcare, transportation, and environmental science. While recent work has be-
013 gun to explore multi-task time series question answering (QA), current bench-
014 marks remain limited in scope, with an emphasis largely on forecasting and
015 anomaly detection tasks. We introduce TRQA, a novel time series QA bench-
016 mark that substantially broadens task coverage and provides a unified setting for
017 evaluating diverse temporal reasoning abilities. TRQA unifies six diverse tasks
018 under a single framework, organized into two complementary groups: (1) *con-
019 ventional reasoning tasks*, including anomaly detection and classification, and (2)
020 *advanced reasoning tasks*, such as characterization, comparison, data transfor-
021 mation, and temporal relationship reasoning. These tasks span multiple ques-
022 tion types, such as *true-or-false (TF)*, *multiple-choice (MC)*, and a novel *puzzling
023 (PZ)*, enabling a more comprehensive evaluation of diverse aspects of time se-
024 ries reasoning. We curated a large-scale dataset with 210k samples, covering
025 a diverse 13 domains, 6 tasks, and 3 types of questions. Each sample consists
026 of one or more time series, an accompanying question, contextual information
027 about the time series, and a corresponding answer. Zero-shot evaluation demon-
028 strates that these tasks are challenging for both commercial and open-source Large
029 Language Models (LLMs). For example, the best-performing commercial LLM,
030 Gemini-2.5-Flash, achieves an average score of only 65.08. While open-source
031 LLMs show notable performance gains after instruction tuning, there remains con-
032 siderable room for improvement. For instance, the best-performing open-source
033 model, LLaMA-3.1-8B, reaches an average score of 85.26, suggesting that these
034 tasks are still non-trivial and pose ongoing challenges for current models. The
035 data are available in GitHub: [https://anonymous.4open.science/r/
036 TRQA_benchmark-6737](https://anonymous.4open.science/r/TRQA_benchmark-6737).

037 1 INTRODUCTION 038

039 Time series data are ubiquitous, arising naturally in domains such as financial markets, electronic
040 health records, environmental monitoring, and energy management. Effectively reasoning over tem-
041 poral patterns is therefore essential for real-world decision-making. Traditionally, research in time
042 series has concentrated on a relatively narrow set of tasks, most notably forecasting future values,
043 anomaly detection, imputation, and classification (Torres et al., 2021; Lim & Zohren, 2021; Wen
044 et al., 2022). While these problems are fundamental and have important applications, the scope
045 of temporal reasoning extends far beyond these settings, encompassing a richer set of queries that
046 demand deeper understanding of fundamental characteristics and patterns of time series.

047 Recent advances in Large Language Models (LLMs) have revolutionized natural language process-
048 ing and multimodal learning, demonstrating remarkable capabilities in understanding, reasoning,
049 and generating across diverse domains (OpenAI et al., 2024; Grattafiori et al., 2024; Team et al.,
050 2025b; Yang et al., 2025). This progress has inspired a growing interest in applying LLMs to time
051 series analysis. Early studies have explored leveraging LLMs for classical time series tasks, such
052 as forecasting and anomaly detection (Zeng et al., 2023; Jin et al., 2023; Zhou & Yu, 2024; Zhang
053 et al., 2024), often by converting temporal data into textual descriptions or prompt-based represen-
tations. However, most existing approaches primarily focus on numeric prediction, leaving open the

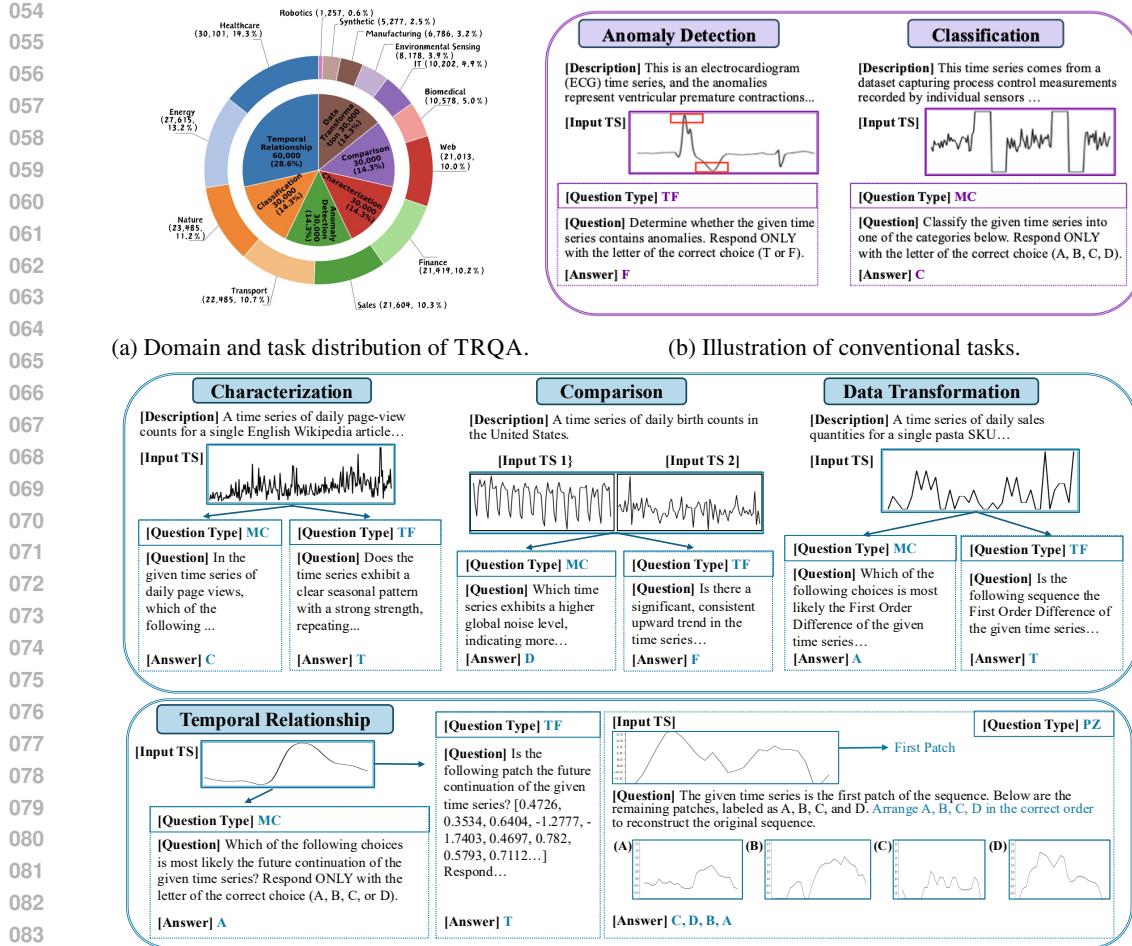


Figure 1: Data distribution and tasks of TRQA.

question of whether LLMs can develop stronger temporal reasoning 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). Rather than being limited to traditional tasks like forecasting, time series QA reformulates time series tasks through natural language queries, enabling models to tackle richer and more complex questions about temporal patterns and dynamics. This paradigm opens the door to evaluating a model’s reasoning capabilities, such as understanding intrinsic characteristics of time series and identifying relationships across multiple sequences. 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 reasoning tasks and open-ended QA tasks, but its *open-ended* answers are difficult to evaluate objectively. While these efforts mark important progress, they are constrained by synthetic or domain-specific data and narrowly scoped tasks. Moreover, some evaluation protocols—particularly for open-ended answers—remain difficult to standardize, limiting fair comparison across models.

In this paper, we introduce TRQA, a large-scale benchmark that addresses these limitations by covering diverse domains and tasks, while also providing standardized evaluation protocols. A direct

108
109
110 Table 1: Comparison of time series question answering datasets and benchmarks.
111
112
113
114
115
116
117

Dataset	Tasks Scope	# Reason Tasks	# Question Type	Human Eval	# 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
TRQA (ours)	Various	6	3	✓	13	210k

120 comparison between TRQA and existing datasets is provided in Table 1. We curate and annotate
 121 210k high-quality samples from 13 domains, as shown in Figure 1a. TRQA integrates 6 distinct
 122 tasks grouped into two complementary categories: (1) *Conventional Tasks: anomaly detection and*
 123 *classification.* (2) *Advanced Tasks: characterization, comparison, data transformation, and temporal*
 124 *relationship reasoning.* All tasks are cast into a unified QA format with three question types:
 125 *true-or-false (TF), multiple-choice (MC),* and a novel *puzzling (PZ).* PZ questions resemble human-
 126 like problem settings (Fissler et al., 2018) and, as evidenced in computer vision (Noroozi & Favaro,
 127 2016), provide a strong probe of general cognition. For time series, they are particularly valuable
 128 in evaluating models’ ability to reason about temporal order and relational structure. Beyond task
 129 design, we carefully detail the *data collection process, benchmark construction, dataset statistics,*
 130 *and evaluation protocol*, ensuring rigorous transparency and reproducibility. Our benchmark pro-
 131 vides a standardized platform to evaluate various LLMs (OpenAI et al., 2024; cla; Team et al.,
 132 2025b; Grattafiori et al., 2024; Yang et al., 2025). Initial empirical studies demonstrate that existing
 133 models struggle across several tasks, particularly in structural and relational reasoning, highlighting
 134 substantial future directions for improvement.

135 In summary, our main contributions are threefold. (1) We introduce TRQA, a novel large-scale
 136 benchmark comprising 210k samples across 13 domains, covering 6 tasks and 3 types of questions.
 137 (2) We provide a detailed description of the benchmark’s construction along with comprehensive
 138 statistics. (3) We conduct extensive evaluations of TRQA using a wide range of popular commercial
 139 and open-source LLMs, accompanied by an in-depth analysis of their performance.

140 141 2 RELATED WORK 142

143
144 **Time Series Analysis: From Numbers to Narratives.** Traditional research on time series has pri-
 145 marily focused on numerical sequences, enabling core tasks such as forecasting (Torres et al., 2021),
 146 imputation (Wang et al., 2024), and classification (Mohammadi Foumani et al., 2024), often treat-
 147 ing them as isolated numeric signals (Hamilton, 2020). In practice, however, time series are rarely
 148 independent of their surrounding context. They frequently interact with external information—such
 149 as textual reports, domain expertise, or heterogeneous side signals—that shapes or enriches their
 150 dynamics (Jiang et al., 2025; Xu et al., 2025b; Liu et al., 2025; 2024; Li et al., 2025). Recognizing
 151 this, recent work has moved beyond purely numeric modeling to incorporate multimodal signals
 152 across domains including healthcare (Johnson et al., 2016; 2023), finance (Li et al., 2024a; Dong
 153 et al., 2024), retail (Skenderi et al., 2024), and transportation (Li et al., 2024b). While much of
 154 this research leverages external modalities to improve numeric predictions on predefined tasks, a
 155 growing body of work instead positions *natural language as a richer interface* for time series, using
 156 language as the medium for querying, reasoning, and interpreting temporal patterns (Merrill et al.,
 157 2024; Williams et al., 2025; Wang et al., 2025b; Chen et al., 2025; Xie et al., 2025; Kong et al.,
 158 2025). Together, these efforts define the emerging direction of *time series question answering.*

159
160 **Large Models on Time Series.** Advances in large language models (LLMs) (Vaswani et al., 2017)
 161 have recently enabled general question answering over time series. A growing line of work integrates
 162 LLMs with time series for downstream tasks (Chang et al., 2023; Alnegheimish et al., 2024; Yu et al.,
 163 2023; Jin et al., 2023), with extensions to multimodal language models as well (Zhong et al., 2025b;
 164 Merrill et al., 2024; Moon et al., 2022). Given their strong generalization ability through natural

162 Table 2: Tasks of TRQA. TF, MC, and PZ denote true-or-false, multiple-choice, and puzzling.
163

164 Group	165 Task	166 Description	167 Question Type
168 Conventional Tasks	169 Anomaly Detection Classification	170 Determine whether the input contains anomalies. 171 Classify the input time series.	172 TF 173 MC
174 Advanced Tasks	175 Characterization Comparison Data Transformation Temporal Relationship	176 Determine the characteristics of the time series. 177 Compare the characteristics of two time series. 178 Identify the relationship between raw and transformed data. 179 Determine the temporal relationship of patches.	180 TF & MC 181 TF & MC 182 TF & MC 183 TF & MC & PZ

172 language interfaces, comprehensive evaluation is critical to ensure the transparency and reliability
173 of LLMs in time series applications.

175 3 TRQA BENCHMARK

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

184 3.1 TASK FORMULATION

185 **Task Taxonomy.** As shown in Table 2 and Figure 1, the proposed TRQA benchmark encompasses
186 two groups of tasks with six diverse tasks designed to evaluate a model’s ability of understanding
187 the fundamental properties of time series data. The first group, *Conventional Tasks*, is comprised
188 of the tasks widely studied in traditional time series analysis: (1) *Anomaly Detection*, identifying
189 irregular or unexpected patterns in time series; (2) *Classification*, reasoning about the relationship
190 between a time series and its underlying conceptual category. The second group, *Advanced Tasks*,
191 consists of novel analytical tasks about intrinsic properties of time series: (3) *Characterization*, in-
192 ferring fundamental properties such as trend, seasonality, and dispersion; (4) *Comparison*, reasoning
193 about relative similarities and differences between two time series; (5) *Data Transformation*, under-
194 standing relationships between original and transformed time series, e.g., Fourier transform; and (6)
195 *Temporal Relationship*, capturing the chronological dependencies among time series patches. These
196 advanced tasks push the boundaries of conventional time series modeling, fostering the development
197 of models that can grasp cognitive concepts of time series and reason over human questions.

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

$$202 \quad A = f(X, C, Q). \quad (1)$$

203 **Question Types.** Our TRQA benchmark encompasses a wide variety of question types, such as
204 *true-or-false* (TF), *multiple-choice* (MC), and *puzzling* (PZ) questions. A TF question requires the
205 model to determine whether a claim about the input time series is True (T) or False (F). A MC
206 question requires the model to select the correct claim about the input. In addition, we introduce a
207 novel *puzzling* (PZ) question to the time series QA community. PZ questions are valuable because
208 they represent realistic, human-like problem settings Fissler et al. (2018) and have been shown to
209 effectively evaluate models’ general cognitive abilities, as demonstrated in computer vision Noroozi
210 & Favaro (2016). In this question, the model is given the first patch of a time series, along with the
211 remaining shuffled patches, and tasked with arranging them in the correct chronological order.

213 3.2 DATA COLLECTION

215 To construct the TRQA benchmark, we collect and preprocess time series data from diverse public
216 sources, spanning domains such as healthcare, transportation, and finance, to ensure broad coverage

216 and representativeness. At its center are the core datasets, which serve as the primary foundation
 217 for a wide range of tasks. In addition, the benchmark integrates two specialized sources: classification
 218 datasets and anomaly detection datasets. This subsection describes these data sources and the
 219 selection criteria. More details can be found in Appendix B.

220 **Core Datasets.** We collect high-quality real-world time series data from a wide range of domains,
 221 including energy, finance, healthcare, nature, sales, transport, and web, which are used by time series
 222 foundation model benchmarks, such as Lotsa (Woo et al., 2024), Time-300B (Shi et al., 2024), and
 223 UTSD (Ma et al., 2024). To ensure data quality, we retain only sequences with a minimum length of
 224 1k. We further filter sequences with a missing rate greater than 1% or an outlier rate (the proportion
 225 of points lying beyond three times the interquartile range (3×IQR) exceeding 5%). For each dataset,
 226 we refer to the original source to gather background information about the time series and provide a
 227 concise, one-sentence description. More details are presented in Appendix B.1

228 **Anomaly Detection Datasets.** We extract data from multiple time-series anomaly detection benchmarks
 229 (Paparrizos et al., 2022; Su et al., 2019), including ECG (Moody & Mark, 2001), SMD (Su
 230 et al., 2019), MGAB (Thill et al.) Genesis (von Birgelen & Niggemann, 2018), GHL (Filonov et al.,
 231 2016), Occupancy (Candanedo & Feldheim, 2016). These datasets span various domains, including
 232 healthcare, mathematical biology, spacecraft telemetry, industrial control systems, environmental
 233 sensing, cybersecurity on IT Operations. For each dataset, we summarize its description and do-
 234 main information directly from the original papers. More details are presented in Appendix B.2.

235 **Classification Datasets.** Our classification data comes from the univariate UCR Archive (Dau et al.,
 236 2019a). We select datasets with at most four classes and sequence lengths under 400, and enrich
 237 them with textual descriptions from the official documentation. The resulting subset spans diverse
 238 domains, including robotics, energy, healthcare, synthetic, manufacturing (see Appendix B.3).

240 3.3 BENCHMARK CONSTRUCTION

241 In this subsection, we describe the construction of the benchmark for each task. To maintain balance
 242 across tasks, we allocate an equal number of samples (30k) to each task, except for temporal rela-
 243 tionship, which we allocate 60k samples since PZ is very challenging. Except for classification and
 244 anomaly detection, samples for all other tasks are drawn from the *Core Datasets* (Section 3.2) us-
 245 ing *Hierarchical Random Sampling* (Algorithm 1) to ensure a balanced distribution across domains,
 246 datasets, and sequences. Unless otherwise specified, all samples have a random length in [32, 256],
 247 and are z-scored to reduce data bias. The term data bias refers specifically to scale-based shortcuts
 248 or magnitude variance across heterogeneous domains, rather than semantic or sampling bias (Ap-
 249 pendix C.2). Finally, each task’s samples are randomly partitioned into 70% for training, 10% for
 250 validation, and 20% for testing. The remainder of this section describes the construction process for
 251 each task. More details and examples can be found in Appendix C-D.

252 **Characterization.** The characterization task assesses the model’s capability to reason about funda-
 253 mental properties of time series, including trend, seasonality, and dispersion. Questions are posed
 254 as TF or MC, and final answers are determined through multi-LLM consensus.

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

265 **Comparison.** The comparison task assesses the model’s ability to reason about the relative char-
 266 acteristics of the two time series, such as shape and correlation. Similar to the characterization task,
 267 this task is also formulated as TF or MC questions. We first obtain an anchor sample \mathbf{x} from domain
 268 M , dataset D , and sequence S . Given the anchor \mathbf{x} , we construct a set of 10 comparison samples
 269 $\{\mathbf{x}'_1 \dots \mathbf{x}'_{10}\}$ with the same length as \mathbf{x} . Among which, one is drawn from the sequence S , two from

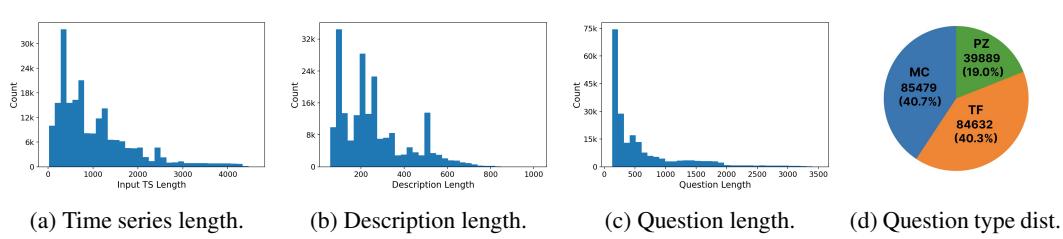


Figure 2: Histograms of time series, description and question lengths, and question type distribution.

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 C.4.

Data Transformation. The data transformation 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 pre-defined 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 x . In MC questions, the model is required to select the correct transformed sequence given the input time series 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 x , whereas incorrect transformations are generated from other randomly sampled time series x' . For more details, please refer to Appendix C.5.

Temporal Relationship. The temporal relationship task evaluates the model’s ability to infer the temporal structure among time series patches, testing 3 core reasoning capabilities: *Structural Continuity*, *Chronological Reasoning*, and *Contextual Discrimination*. This task is formulated as TF, MC, or PZ questions. Given the first chronological patch x , a TF question asks the model to determine whether a candidate patch y is the immediate successor of x , while an MC question asks the model to choose the correct next patch from candidates $[y_1, y_2, y_3, y_4]$. The false candidates are randomly sampled from the full dataset, but from sequences different from that of x . A puzzling question presents four shuffled successor patches of x and asks the model to arrange them in the correct chronological order. All questions are generated using predefined templates. See Appendix C.6 for more details.

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 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 normal samples to balance the classes at a 1:1 ratio. The questions are composed using a predefined template. See Appendix C.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 C.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 TRQA 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

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

328 Group	329 Task Question Type	A.D.		CLS		Characterization		Comparison		Data Transform		Temporal Relation			Overall
		330 MC	331 MC	TF	332 MC	TF	333 MC	TF	334 MC	TF	335 MC	TF	336 MC	PZ	
330 Zero 331 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		
	Minstral-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		
337 Instruction 338 Tuning	Qwen3-8B	87.70	<u>90.05</u>	92.37	85.42	86.55	<u>79.08</u>	89.84	<u>84.99</u>	96.84	97.56	<u>66.21</u>	84.29		
	LLaMA3.1-8B	91.02	91.27	92.44	83.68	86.72	79.31	90.17	86.62	96.94	<u>97.41</u>	67.68	85.26		
	Minstral-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		

343 position with the ground truth and computing the proportion of correct matches. For example, with a
 344 ground truth A, B, C, D and prediction B, A, C, D, only the last two match, yielding 50% accuracy.

345 346 347 4 EXPERIMENTS

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

352 4.1 MAIN RESULTS

354 We evaluate *zero-shot* performance of (1) commercial LLMs: GPT-4.1, GPT-4o, Claude-3.5-
 355 Sonnet and Gemini-2.5-Flash; (2) medium size open-source LLMs: Qwen3-8B (Yang et al., 2025),
 356 LLaMA3.1-8B (Dubey et al., 2024), Minstral-8B; (3) small size open-source LLMs: Qwen3-
 357 0.6B (Yang et al., 2025), LLaMA3.2-1B (Dubey et al., 2024), Gemma3-1B (Team et al., 2025a).
 358 We further apply *instruction tuning* Peng et al. (2023) to the open-source methods using LoRA Hu
 359 et al. (2022). We set LoRA rank as 16, fix learning rate as 10^{-5} with cosine schedule, and train
 360 models for 2 epochs on a single A100 GPU.

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

367 **Task-Level Results.** (1) *Conventional Tasks*. In zero-shot settings, both commercial and open-
 368 source LLMs perform poorly on anomaly detection and classification, but open-source models im-
 369 prove markedly after instruction tuning (e.g., LLaMA-3.1-8B reaches 91.02 and 91.27). (2) *Ad-
 370 vanced Tasks*. For characterization and comparison, commercial models outperform medium-sized
 371 open-source models, likely due to broader pretraining exposure. Data transformation and temporal
 372 relationship, especially PZ questions, remain difficult for all models. Instruction tuning boosts open-
 373 source performance, but there are still considerable room to improve. For example, for comparison
 374 task, best performing LLaMA-3.1-8B with instruction tuning only achieves 86.72 and 79.31.

375 **Question Type-Level Results.** Across the three question types (TF, MC, PZ), open-source models
 376 perform best on TF, worse on MC, and poorest on PZ. Performance on PZ is substantially lower than
 377 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).

378 4.2 ANALYSIS
379380 We use best performing commercial LLMs, i.e., Geimini-2.5-Flash and GPT-4.1, and open-source
381 LLMs, i.e., LLaMA3.1-8B and Qwen3-8B to conduct further analysis. To examine their reasoning
382 ability on the proposed TRQA Benchmark, we cover two key perspectives: *Accuracy Correlate*
383 *Analysis* and *Task-Specific Analysis*.
384385 4.2.1 ACCURACY CORRELATE ANALYSIS
386387 **Input Lengths.** Figures 3 illustrate the relationship between input length and the number of correct
388 predictions. The input length is calculated as $len(ts + description + domain + dataset + task + ques-$
389 $tion_type + question)$ with *String* type. Across all six models and five tasks, excluding the Temporal
390 Relationship task, we observe a consistent trend that performance declines as input length increases,
391 indicating that longer inputs correspond to more difficult questions. However, the Temporal Rela-
392 tionship task exhibits the opposite behavior, where accuracy improves with increasing input length.
393 The analysis is shown in Figure 4, which the newly proposed PZ type question exhibits the opposite
394 trend. The fact that PZ performance scales positively with length proves that the model is actively
395 utilizing global context, all time series segments, to deduce the correct chronological order. This
396 confirms the model is engaging in deductive reasoning rather than local pattern matching, proving
397 that PZ type question is a rigorous probe for *Global Causal Reasoning*. More details are provided
398 in Appendix E.1.
399400 4.2.2 TASK SPECIFIC ANALYSIS
401402 **Comparison.** We analyze model performance on the Comparison task, specifically investigating
403 whether providing explicit domain-level context affects model accuracy. The task requires compar-
404 ing two input time series, which we test under two conditions: (1) when both series originate from
405 the same domain and (2) when they are from different domains. In both scenarios, the corresponding
406 domain names are provided to the model as textual description. As shown in Table 8, we observe no
407 significant performance difference between the same-domain and different-domain settings across
408 either MC or TF questions. This suggests that the Comparison Task is domain invariant. More
409 details are provided in Appendix E.2.
410411 **Data Transformation.** We analyze model performance on the Data Transformation task, which is
412 designed to evaluate a model’s understanding of three transformation operators: Fourier Transform
413 (FT), Wavelet Transform (WT), and First-Order Differencing (FOD). For each operator, we assess
414 performance by measuring the accuracy on both MC and TF question formats. As shown in Ta-
415 ble 9, for zero-shot evaluation, our key finding highlights a limitation in which both commercial and
416 open-source models fail to provide accurate answers, except of FOD. More details are provided in
417 Appendix E.2.
418419 **Temporal Relationship.** Beyond the input length analysis in Section 4.2.1 demonstrating that PZ
420 questions require *Global Causal Reasoning*, we further examined how domain-level information
421 influences model performance on PZ questions. The results are summarized in Table 10, which
422 Web and Sales domains remain the most challenging across both zero-shot and instruction-tuning
423 settings. More details are provided in Appendix E.2.
424425 4.3 HUMAN EVALUATION
426427 We conduct human evaluations of the multi-LLM consensus labels for characterization and compa-
428 rison (Section 3.3). Six Ph.D.-level experts manually annotate 600 questions (300 each), serving as
429 ground truth. Uncertain or problematic QA pairs are flagged, multiple answers allowed when valid,
430 and explanations provided for disagreements with the benchmark.
431432 Our evaluation yields two main findings. (1) Question quality: Uncertainty rates are low (5% for
433 characterization, 7% for comparison), showing that most questions are clear. (2) Answer accuracy:
434 For unambiguous cases, benchmark answers align with human judgments in 91.2% of characteriza-
435 tion and 87.4% of comparison. These results indicate that the automatic pipeline produces reliable
436 QA pairs, though comparison remains harder, with lower agreement and higher uncertainty (Fig-
437 ure 5). More details are provided in Appendix F.
438

432

5 CONCLUSION

434 TRQA establishes a large-scale comprehensive benchmark for time series question answering with
 435 210k samples curated from 13 domains, covering 6 tasks and 3 types of questions, extending evalua-
 436 tion beyond traditional tasks, i.e., anomaly detection and classification, to advanced reasoning tasks,
 437 i.e., characterization, comparison, data transformation, temporal relationship reasoning. By span-
 438 ning diverse domains, tasks, and question types, it offers a unified platform to probe the strengths and
 439 limitations of both commercial and open-source LLMs. Our results highlight that, despite progress
 440 with instruction tuning, substantial challenges remain—particularly for advanced reasoning and puz-
 441 zling questions—underscoring the need for further research into models capable of deeper time
 442 series understanding.

444

6 REPRODUCIBILITY STATEMENT

446 We provide detailed descriptions of the TRQA benchmark to ensure reproducibility. The formula-
 447 tion of tasks and question types is introduced in 3.1, with data sources and preprocessing procedures
 448 explained in 3.2 and Appendix B. Benchmark construction, including sampling strategy, data parti-
 449 tioning, and design considerations, is described in 3.3 and Appendices C–D, while dataset statistics
 450 are summarized in 3.4. Evaluation protocols for different question types (TF, MC, PZ) are given
 451 in 3.5. The experimental setup, including model configurations, hyperparameters, and instruction-
 452 tuning details, is provided in 4. Human evaluation methodology and results are discussed in 4.3
 453 and F. All data and code required to reproduce our results are available at the anonymous repository:
 454 https://anonymous.4open.science/r/TRQA_benchmark-6737.

456

7 ETHICS STATEMENT

458 This research was conducted in accordance with the ICLR Code of Ethics. The study did not involve
 459 human participants or animal subjects, and all datasets employed are publicly accessible and were
 460 utilized in strict compliance with their respective licensing agreements and data usage policies. The
 461 data contains no personally identifiable information (PII), and our experimental design inherently
 462 mitigates privacy and security risks.

464

REFERENCES

466 The claude 3 model family: Opus, sonnet, haiku. URL <https://api.semanticscholar.org/CorpusID:268232499>.

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

472 Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen,
 473 Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor,
 474 Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, An-
 475 drew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the
 476 language of time series, 2024a. URL <https://arxiv.org/abs/2403.07815>.

477 Abdul Fatir Ansari, Lorenzo Stella, Caner Türkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen,
 478 Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor,
 479 Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, An-
 480 drew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the lan-
 481 guage of time series. *arXiv preprint arXiv:2403.07815*, 2024b. doi: 10.48550/arXiv.2403.07815.
 482 URL <https://arxiv.org/abs/2403.07815>.

484 Anthony J. Bagnall and Luke M. Davis. Predictive modelling of bone age through classification and
 485 regression of bone shapes. *CoRR*, abs/1406.4781, 2014. URL <http://arxiv.org/abs/1406.4781>.

486 Romain Briandet, E. Katherine Kemsley, and Reginald H. Wilson. Discrimination of Arabica and
 487 Robusta in instant coffee by Fourier transform infrared spectroscopy and chemometrics. *Journal*
 488 *of Agricultural and Food Chemistry*, 44(1):170–174, 1996. doi: 10.1021/jf950305a. URL
 489 <https://hal.science/hal-01606904>.

490 Yifu Cai, Arjun Choudhry, Mononito Goswami, and Artur Dubrawski. Timeseriesexam: A time
 491 series understanding exam, 2024. URL <https://arxiv.org/abs/2410.14752>.

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

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

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

498 Jialin Chen, Aosong Feng, Ziyu Zhao, Juan Garza, Gaukhar Nurbek, Cheng Qin, Ali Maatouk,
 499 Leandros Tassiulas, Yifeng Gao, and Rex Ying. Mtbench: A multimodal time series benchmark
 500 for temporal reasoning and question answering, 2025. URL <https://arxiv.org/abs/2503.16858>.

501 City of Melbourne. Pedestrian counting system (counts per hour).
 502 <https://data.melbourne.vic.gov.au/explore/dataset/pedestrian-counting-system-monthly-counts-per-hour/information/>,
 503 2017. Accessed 2025-09-24.

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

507 Hoang Anh Dau, Anthony Bagnall, Kaveh Kamgar, Chin-Chia Michael Yeh, Yan Zhu, Shaghayegh
 508 Gharghabi, Chotirat Ann Ratanamahatana, and Eamonn Keogh. The ucr time series archive,
 509 2019b. URL <https://arxiv.org/abs/1810.07758>.

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

513 Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha
 514 Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models.
 515 *arXiv e-prints*, pp. arXiv–2407, 2024.

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

519 Patrick Fissler, Olivia Caroline Küster, Daria Laptinskaya, Laura Sophia Loy, Christine AF
 520 Von Arnim, and Iris-Tatjana Kolassa. Jigsaw puzzling taps multiple cognitive abilities and is
 521 a potential protective factor for cognitive aging. *Frontiers in aging neuroscience*, 10:408085,
 522 2018.

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

525 Pierre Geurts. *Contributions to decision tree induction: bias/variance tradeoff and time series*
 526 *classification*. PhD thesis, ULiège - Université de Liège, May 2002. URL <http://www.montefiore.ulg.ac.be/services/stochastic/pubs/2002/Geu02>.

527 Rakshitha Godahewa, Christoph Bergmeir, Geoffrey I. Webb, Rob J. Hyndman, and Pablo Montero-
 528 Manso. Monash time series forecasting archive. In *NeurIPS Datasets and Benchmarks Track*,
 529 2021. URL <https://arxiv.org/abs/2105.06643>.

540 Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad
 541 Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan,
 542 Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Ko-
 543 reniev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava
 544 Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux,
 545 Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret,
 546 Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius,
 547 Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary,
 548 Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab
 549 AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco
 550 Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind That-
 551 tai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Kore-
 552 vaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra,
 553 Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Ma-
 554 hadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu,
 555 Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jong-
 556 so Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasudevan Alwala,
 557 Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid
 558 El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren
 559 Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin,
 560 Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi,
 561 Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew
 562 Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Ku-
 563 mar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoy-
 564 chev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan
 565 Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajwal Bhargava, Pratik Dubal, Praveen Krishnan,
 566 Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ra-
 567 mon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Ro-
 568 hit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan
 569 Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell,
 570 Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparth, Sheng
 571 Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer
 572 Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman,
 573 Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mi-
 574 haylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor
 575 Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei
 576 Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang
 577 Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Gold-
 578 schlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning
 579 Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh,
 580 Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolf Victoria,
 581 Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein,
 582 Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, An-
 583 drew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, An-
 584 nie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel,
 585 Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leon-
 586 hardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu
 587 Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Mon-
 588 talvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao
 589 Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia
 590 Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide
 591 Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le,
 592 Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily
 593 Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smo-
 thers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni,
 Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia
 Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan,
 Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harri-
 son Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj,

594 Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James
 595 Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jen-
 596 nifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang,
 597 Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Jun-
 598 jie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy
 599 Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang,
 600 Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell,
 601 Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa,
 602 Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias
 603 Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L.
 604 Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike
 605 Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari,
 606 Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan
 607 Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong,
 608 Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent,
 609 Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar,
 610 Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Ro-
 611 driguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy,
 612 Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin
 613 Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon,
 614 Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ra-
 615 maswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha,
 616 Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal,
 617 Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satter-
 618 field, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj
 619 Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo
 620 Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook
 621 Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Ku-
 622 mar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov,
 623 Wei Li, Wencheng Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiao-
 624 jian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia,
 625 Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao,
 626 Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhao-
 627 duo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL
 628 <https://arxiv.org/abs/2407.21783>.

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

630 Charles R Harris, K Jarrod Millman, Stéfan J Van Der Walt, Ralf Gommers, Pauli Virtanen, David
 631 Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, et al. Array program-
 632 ming with numpy. *nature*, 585(7825):357–362, 2020.

633 James K. Holland, E. Katherine Kemsley, and Reginald H. Wilson. Use of fourier transform infrared
 634 spectroscopy and partial least squares regression for the detection of adulteration of strawberry
 635 purées. *Journal of the Science of Food and Agriculture*, 76(2):263–269, February 1998. ISSN
 0022-5142. doi: 10.1002/(SICI)1097-0010(199802)76:2<263::AID-JSFA943>3.0.CO;2-F.

636 Tao Hong, Pierre Pinson, Shu Fan, et al. Global energy forecasting competition 2012. *In-
 637 ternational Journal of Forecasting*, 30(2):357–363, 2014. doi: 10.1016/j.ijforecast.2013.
 638 07.001. URL <https://www.sciencedirect.com/science/article/abs/pii/S0169207013000745>.

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

641 Xiaohui Huang, Yunming Ye, Liyan Xiong, Raymond Y.K. Lau, Nan Jiang, and Shaokai Wang.
 642 Time series k-means. *Inf. Sci.*, 367(C):1–13, November 2016. ISSN 0020-0255. doi: 10.1016/j.
 643 ins.2016.05.040. URL <https://doi.org/10.1016/j.ins.2016.05.040>.

644 Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Os-
 645 trow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint
 646 arXiv:2410.21276*, 2024.

648 Alexander Ihler, Jon Hutchins, and Padhraic Smyth. Adaptive event detection with time-varying
 649 poisson processes. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowl-*
 650 *edge Discovery and Data Mining*, KDD '06, pp. 207–216, New York, NY, USA, 2006. Associa-
 651 *tion for Computing Machinery*. ISBN 1595933395. doi: 10.1145/1150402.1150428. URL
 652 <https://doi.org/10.1145/1150402.1150428>.

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

657 Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yux-
 658 *uan Liang, Yuan-Fang Li, Shirui Pan, et al. Time-llm: Time series forecasting by reprogramming*
 659 *large language models. arXiv preprint arXiv:2310.01728*, 2023.

660 Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen,
 661 Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. Time-llm: Time series forecasting
 662 by reprogramming large language models, 2024. URL <https://arxiv.org/abs/2310.01728>.

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

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

669 Eamonn Keogh, Li Wei, Xiaopeng Xi, Stefano Lonardi, Jin Shieh, and Scott Sirowy. Intelligent
 670 icons: Integrating lite-weight data mining and visualization into gui operating systems. In *Sixth*
 671 *International Conference on Data Mining (ICDM'06)*, pp. 912–916, 2006. doi: 10.1109/ICDM.
 672 2006.90.

673 Yaxuan Kong, Yiyuan Yang, Yoontae Hwang, Wenjie Du, Stefan Zohren, Zhangyang Wang, Ming
 674 Jin, and Qingsong Wen. Time-mqa: Time series multi-task question answering with context en-
 675 hancement. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar
 676 (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*
 677 (*Volume 1: Long Papers*), ACL 2025, Vienna, Austria, July 27 - August 1, 2025, pp. 29736–
 29753. Association for Computational Linguistics, 2025. URL <https://aclanthology.org/2025.acl-long.1437/>.

678 Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long- and short-term
 679 temporal patterns with deep neural networks. *arXiv preprint arXiv:1703.07015*, 2018. URL
 680 <https://arxiv.labs.arxiv.org/html/1703.07015>.

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

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

687 Zihao Li, Xiao Lin, Zhining Liu, Jiaru Zou, Ziwei Wu, Lecheng Zheng, Dongqi Fu, Yada Zhu,
 688 Hendrik F. Hamann, Hanghang Tong, and Jingrui He. Language in the flow of time: Time-series-
 689 paired texts weaved into a unified temporal narrative. *CoRR*, abs/2502.08942, 2025. doi: 10.
 690 48550/ARXIV.2502.08942. URL <https://doi.org/10.48550/arXiv.2502.08942>.

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

702 Haoxin Liu, Shangqing Xu, Zhiyuan Zhao, Lingkai Kong, Harshavardhan Prabhakar Kamarthi,
 703 Aditya Sasanur, Megha Sharma, Jiaming Cui, Qingsong Wen, Chao Zhang, et al. Time-mmd:
 704 Multi-domain multimodal dataset for time series analysis. *Advances in Neural Information Pro-*
 705 *cessing Systems*, 37:77888–77933, 2024.

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

710 Xiangkai Ma, Xiaobin Hong, Wenzhong Li, and Sanglu Lu. Utsd: Unified time series diffusion
 711 model, 2024. URL <https://arxiv.org/abs/2412.03068>.

712 Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. M5 accuracy competi-
 713 tion: Results, findings, and conclusions. *International Journal of Forecasting*, 38(4):1346–1364,
 714 2022. doi: 10.1016/j.ijforecast.2021.11.013. URL <https://www.sciencedirect.com/science/article/pii/S0169207021001874>.

715 Paolo Mancuso, Veronica Piccialli, and Antonio M. Sudoso. A machine learning approach
 716 for forecasting hierarchical time series. *Expert Systems with Applications*, 182:115102,
 717 2021. doi: 10.1016/j.eswa.2021.115102. URL <https://www.sciencedirect.com/science/article/pii/S0957417421005431>.

718 Michael W. McCracken and Serena Ng. Fred-md: A monthly database for macroeconomic research.
 719 *Journal of Business & Economic Statistics*, 34(4):574–589, 2016. doi: 10.1080/07350015.2015.
 720 1086655. URL <https://www.tandfonline.com/doi/full/10.1080/07350015.2015.1086655>.

721 Mike A. Merrill, Mingtian Tan, Vinayak Gupta, Thomas Hartvигsen, and Tim Althoff. Lan-
 722 guage models still struggle to zero-shot reason about time series. In Yaser Al-Onaizan, Mohit
 723 Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguis-
 724 tics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pp. 3512–3533. Association
 725 for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-EMNLP.201. URL
 726 <https://doi.org/10.18653/v1/2024.findings-emnlp.201>.

727 Clayton Miller, Anjukan Kathirgamanathan, Bianca Picchetti, Pandarasamy Arjunan, June Young
 728 Park, Zoltán Nagy, Paul Raftery, Brodie Hobson, Zixiao Shi, Forrest Meggers, et al. The building
 729 data genome project 2, 1,636 energy meter readings from the ashrae great energy predictor iii
 730 competition. *Scientific Data*, 7(1):368, 2020. doi: 10.1038/s41597-020-00712-x. URL <https://www.nature.com/articles/s41597-020-00712-x>.

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

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

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

739 Soukayna Mouatadid, Paulo Orenstein, Genevieve Flaspholer, Miruna Oprescu, Judah Cohen,
 740 Franklyn Wang, Sean Knight, Maria Geogdzhayeva, Sam Levang, Ernest Fraenkel, and Lester
 741 Mackey. Subseasonalclimateusa: A dataset for subseasonal forecasting and benchmarking.
 742 *arXiv preprint arXiv:2109.10399*, 2024. doi: 10.48550/arXiv.2109.10399. URL <https://arxiv.org/abs/2109.10399>. v4.

743 Abdullah Mueen, Eamonn Keogh, and Neal Young. Logical-shapelets: an expressive primitive for
 744 time series classification. In *Proceedings of the 17th ACM SIGKDD International Conference on
 745 Knowledge Discovery and Data Mining*, KDD ’11, pp. 1154–1162, New York, NY, USA, 2011.
 746 Association for Computing Machinery. ISBN 9781450308137. doi: 10.1145/2020408.2020587.
 747 URL <https://doi.org/10.1145/2020408.2020587>.

756 David Murray. A data management platform for personalised real-time energy feedback. *Proc. 8th*
 757 *Int. Conf. Energy Efficiency Domestic Appl. Lighting (EEDAL)*, pp. 1–15, 08 2015.

758

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

761

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

764

765 OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Floren-
 766 cia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red
 767 Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Moham-
 768 mad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher
 769 Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brock-
 770 man, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann,
 771 Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis,
 772 Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey
 773 Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux,
 774 Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila
 775 Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix,
 776 Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gib-
 777 son, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan
 778 Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hal-
 779 lacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan
 780 Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu,
 781 Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun
 782 Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Ka-
 783 mali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook
 784 Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel
 785 Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen
 786 Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel
 787 Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez,
 788 Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv
 789 Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney,
 790 Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick,
 791 Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel
 792 Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Ra-
 793 jeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O’Keefe,
 794 Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel
 795 Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe
 796 de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny,
 797 Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl,
 798 Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra
 799 Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders,
 800 Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Sel-
 801 sam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor,
 802 Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky,
 803 Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang,
 804 Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Pre-
 805 ston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vi-
 806 jayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan
 807 Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng,
 808 Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Work-
 809 man, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming
 Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao
 Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL
<https://arxiv.org/abs/2303.08774>.

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

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

812

813 Chotirat Ratanamahatana and Eamonn Keogh. Three myths about dynamic time warping data min-
 814 ing. 04 2005. doi: 10.1137/1.9781611972757.50.

815

816 Naoki Saito and Ronald R. Coifman. Local feature extraction and its applications using a library of
 817 bases. 1994. URL <https://api.semanticscholar.org/CorpusID:117929906>.

818

819 Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. Time-
 820 moe: Billion-scale time series foundation models with mixture of experts. *arXiv preprint
 821 arXiv:2409.16040*, 2024. URL <https://arxiv.org/abs/2409.16040>.

822

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

826

827 Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, and Dan Pei. Robust anomaly detection for
 828 multivariate time series through stochastic recurrent neural network. In Ankur Teredesai, Vipin
 829 Kumar, Ying Li, Rómer Rosales, Evimaria Terzi, and George Karypis (eds.), *Proceedings of the
 830 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD
 831 2019, Anchorage, AK, USA, August 4-8, 2019*, pp. 2828–2837. ACM, 2019.

832

833 Jimeng Sun, Spiros Papadimitriou, and Christos Faloutsos. Online latent variable detection in sensor
 834 networks. In *Proceedings of the 21st International Conference on Data Engineering, ICDE '05*,
 835 pp. 1126–1127, USA, 2005. IEEE Computer Society. ISBN 0769522858. doi: 10.1109/ICDE.
 836 2005.100. URL <https://doi.org/10.1109/ICDE.2005.100>.

837

838 Chang Wei Tan, Christoph Bergmeir, François Petitjean, and Geoffrey I. Webb. Monash university,
 839 uea, ucr time series extrinsic regression archive. *arXiv preprint arXiv:2006.10996*, 2020. doi:
 840 10.48550/arXiv.2006.10996. URL <https://arxiv.org/abs/2006.10996>.

841

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

845

846 Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej,
 847 Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas
 848 Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Cas-
 849 bon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xi-
 850 aohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Cole-
 851 man, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry,
 852 Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi,
 853 Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe
 854 Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa
 855 Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András
 856 György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia
 857 Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini,
 858 Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel
 859 Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivaku-
 860 mar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eu-
 861 gene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna
 862 Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian
 863 Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wi-
 864 eting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh,
 865 Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine,
 866 Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael
 867 Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Ni-
 868 lay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Ruben-
 869 stein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya
 870 Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu,
 871 Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti

864 Sheth, Siim Pöder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi
 865 Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry,
 866 Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein
 867 Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat
 868 Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas
 869 Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Bar-
 870 rral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam
 871 Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena
 872 Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier
 873 Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot.
 874 Gemma 3 technical report, 2025b. URL <https://arxiv.org/abs/2503.19786>.

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

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

879

880 Md Nayem Uddin, Amir Saeidi, Divij Handa, Agastya Seth, Tran Cao Son, Eduardo Blanco,
 881 Steven R. Corman, and Chitta Baral. Unseentimeqa: Time-sensitive question-answering be-
 882 yond llms' memorization. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mo-
 883 hammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association
 884 for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27
 885 - August 1, 2025*, pp. 1873–1913. Association for Computational Linguistics, 2025. URL
 886 <https://aclanthology.org/2025.acl-long.94/>.

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

890 Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau,
 891 Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: funda-
 892 mental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.

893

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

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

901 Yilin Wang, Peixuan Lei andm Jie Song, Yuzhe Hao, Tao Chen, Yuxuan Zhang, Lei Jia, Yuanxiang
 902 Li, and Zhongyu Wei. Itformer: Bridging time series and natural language for multi-modal QA
 903 with large-scale multitask dataset. *CoRR*, abs/2506.20093, 2025a. doi: 10.48550/ARXIV.2506.
 904 20093. URL <https://doi.org/10.48550/arXiv.2506.20093>.

905

906 Yilin Wang, Peixuan Lei, Jie Song, Yuzhe Hao, Tao Chen, Yuxuan Zhang, Lei Jia, Yuanxiang Li,
 907 and Zhongyu Wei. Itformer: Bridging time series and natural language for multi-modal qa with
 908 large-scale multitask dataset, 2025b. URL <https://arxiv.org/abs/2506.20093>.

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

911

912 Andrew Robert Williams, Arjun Ashok, Étienne Marcotte, Valentina Zantedeschi, Jithendaraa Sub-
 913 ramanian, Roland Riachi, James Requeima, Alexandre Lacoste, Irina Rish, Nicolas Chapados,
 914 and Alexandre Drouin. Context is key: A benchmark for forecasting with essential textual infor-
 915 mation, 2025. URL <https://arxiv.org/abs/2410.18959>.

916 Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sa-
 917 hoo. Unified training of universal time series forecasting transformers. *arXiv preprint
 918 arXiv:2402.02592*, 2024. URL <https://arxiv.org/abs/2402.02592>.

918 Zhe Xie, Zeyan Li, Xiao He, Longlong Xu, Xidao Wen, Tieying Zhang, Jianjun Chen, Rui Shi, and
919 Dan Pei. Chatts: Aligning time series with llms via synthetic data for enhanced understanding
920 and reasoning, 2025. URL <https://arxiv.org/abs/2412.03104>.

921 Xiongxiao Xu, Haoran Wang, Yueqing Liang, Philip S. Yu, Yue Zhao, and Kai Shu. Can multimodal
922 llms perform time series anomaly detection? *CoRR*, abs/2502.17812, 2025a. doi: 10.48550/
923 ARXIV.2502.17812. URL <https://doi.org/10.48550/arXiv.2502.17812>.

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

926 An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang
927 Gao, Chengan Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu,
928 Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin
929 Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang,
930 Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui
931 Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang
932 Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger
933 Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan
934 Qiu. Qwen3 technical report, 2025. URL <https://arxiv.org/abs/2505.09388>.

935 Lexiang Ye and Eamonn Keogh. Time series shapelets: a new primitive for data mining. In
936 *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery
937 and Data Mining*, KDD '09, pp. 947–956, New York, NY, USA, 2009. Association for Com-
938 puting Machinery. ISBN 9781605584959. doi: 10.1145/1557019.1557122. URL <https://doi.org/10.1145/1557019.1557122>.

939 Lexiang Ye and Eamonn Keogh. Time series shapelets: a novel technique that allows ac-
940 curate, interpretable and fast classification. *Data Mining and Knowledge Discovery*, 22(1):
941 149–182, 2011. doi: 10.1007/s10618-010-0179-5. URL <https://doi.org/10.1007/s10618-010-0179-5>.

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

944 Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series
945 forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp.
946 11121–11128, 2023.

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

949 Yunkai Zhang, Yawen Zhang, Ming Zheng, Kezhen Chen, Chongyang Gao, Ruian Ge, Siyuan
950 Teng, Amine Jelloul, Jinmeng Rao, Xiaoyuan Guo, Chiang-Wei Fang, Zeyu Zheng, and Jie
951 Yang. Insight miner: A large-scale multimodal model for insight mining from time series. In
952 *NeurIPS 2023 AI for Science Workshop*, 2023. URL <https://openreview.net/forum?id=E1khscdUdH>.

953 Siru Zhong, Weilin Ruan, Ming Jin, Huan Li, Qingsong Wen, and Yuxuan Liang. Time-vlm:
954 Exploring multimodal vision-language models for augmented time series forecasting. *CoRR*,
955 abs/2502.04395, 2025a. doi: 10.48550/ARXIV.2502.04395. URL <https://doi.org/10.48550/arXiv.2502.04395>.

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

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

961

972 A THE USE OF LARGE LANGUAGE MODELS
973974 We leverage Large Language Models (LLMs) from two perspectives: (1) Polishing the writing,
975 where LLMs are used to refine the clarity, fluency, and consistency of the paper; and (2) Labeling,
976 where LLMs assist in generating high-quality question-answer (QA) pairs and providing preliminary
977 annotations, which are then validated or aggregated through consensus to create reliable ground-
978 truth labels.980 B DATA COLLECTION
981982 In this section, we detail the data sources, including *core datasets* (Appendix B.1), *anomaly detec-*
983 *tion datasets* (Appendix B.2), and *classification datasets* (Appendix B.3).985 B.1 CORE DATASETS
986987 We extract data from multiple time-series datasets including: Australian Electricity Demand (Godahewa
988 et al., 2021), BDG-2 Rat (Miller et al., 2020), GEF12 (Hong et al., 2014), ExchangeRate(Lai
989 et al., 2018), FRED MD(McCracken & Ng, 2016), BIDMC32HR (Tan et al., 2020), PigArtPres-
990 sure (Dau et al., 2019a), USBirths (Godahewa et al., 2021), Sunspot (Godahewa et al., 2021),
991 Saugeenday (Godahewa et al., 2021), SubseasonalPrecip (Mouatadid et al., 2024), Hierarchical-
992 Sales (Mancuso et al., 2021), M5 (Makridakis et al., 2022), PedestrianCounts (City of Melbourne,
993 2017), PEMSO3 (Caltrans, 2025), UberTLCHourly (FiveThirtyEight, 2015),WikiDaily100k (Ansari
994 et al., 2024b). Below are some more detailed descriptions on those dataaets.995 **Australian Electricity Demand.** A single long time series from the Monash Time Series Archive
996 representing half-hourly electricity demand for Victoria, Australia in 2014 (17,520 observations),
997 extracted from the R package fpp2 (dataset name: “elecdemand”). Temperatures corresponding to
998 each demand value are available in the original dataset.999 **BDG-2 Rat .** From The Building Data Genome Project 2 (MIT License), consisting of measure-
1000 ments from 3,053 meters across 1,636 commercial buildings over 2016–2017. One or more meters
1001 per building measured total electrical, heating and cooling water, steam, solar energy, water, and
1002 irrigation usage. We use the whole-building electricity meter measurements from the Bear, Fox,
1003 Panther, and Rat sites, totaling 611 buildings (from the CSV file electricity_cleaned.csv).1004 **GEF12.** A benchmark compiled from the Global Energy Forecasting Competition 2012 (load fore-
1005 casting tracks), containing 20 aggregated-level hourly load series and 11 temperature series from
1006 2004-01-01 00:00 to 2008-06-30 05:00. Because the one-to-one correspondence between tempera-
1007 ture and load series is not clearly defined, a common strategy is to use a single temperature series
1008 for all loads (here, the second temperature series). The dataset is competition-grade and was used
1009 without additional preprocessing; visualizations show obvious periodicity and seasonality in the
1010 aggregated loads.1011 **ExchangeRate.** Daily exchange rates for currencies of eight countries—Australia, United Kingdom,
1012 Canada, Switzerland, China, Japan, New Zealand, and Singapore—covering 1990 to 2016.1013 **FRED-MD.** 107 monthly time series of macro-economic indicators from the Federal Reserve Bank,
1014 starting from 1959-01-01, extracted from the FRED-MD database.1015 **BIDMC32HR.** Derived from BIDMC ICU recordings: PPG and respiratory signals/IP (sampling
1016 rate 125 Hz) from 53 adult patients, with breath annotations used to form reference targets in the
1017 source dataset. Following the adaptation in subsequent work, PPG and ECG are converted into 32-
1018 second sliding-window time series; the average heart rate (HR) in each 32 s window is the target.
1019 The datasets are split by randomly selecting 30% as test, yielding 5,550 training and 2,399 test time
1020 series.1021 **PigArtPressure.** Based on a source dataset from 52 pigs with three vital signs monitored before and
1022 after an induced injury. Three datasets are created: AirwayPressure (airway pressure), ArtPressure
1023 (arterial blood pressure), and CVP (central venous pressure).1024 **US Births.** A single long daily time series of the number of births in the United States from 1969-
1025 01-01 to 1988-12-31 (7,305 observations), extracted from the R package mosaicData.

1026 **Sunspot.** A single long daily time series of sunspot numbers from 1818-01-01 onward, with ad-
 1027 ditional related series (monthly means, smoothed series, yearly totals, hemispheric series) in the
 1028 original source. The repository used here contains the daily series from 1818-08-01 to 2020-05-31
 1029 and includes both the raw data (with missing values) and an LOCF-imputed version.

1030 **Saugeen.** A single long daily time series of the Saugeen River mean flow at Walkerton (in cubic
 1031 meters per second) from 1915-01-01 to 1979-12-31 (23,741 observations), extracted from the R
 1032 package `deseasonalize` (dataset name: “`SaugeenDay`”).

1033 **Subseasonal Precipitation.** Extracted from `SubseasonalClimateUSA`: daily precipitation measure-
 1034 ments (millimeters) for a single $1.5^\circ \times 1.5^\circ$ latitude–longitude grid cell, covering 1948–1978.

1035 **Hierarchical Sales.** 118 daily time series of SKU-level sales for four national pasta brands from
 1036 2014-01-01 to 2018-12-31, including a binary indicator for promotion. The series can be organized
 1037 into a three-level hierarchy.

1038 **M5.** The M5 “Accuracy” competition dataset requiring point forecasts for 30,490 bottom-level daily
 1039 series that aggregate to 42,840 time series representing hierarchical unit sales for Walmart. The
 1040 competition paper details the implementation, results, top methods, and implications for forecasting
 1041 research.

1042 **Pedestrian Counts.** Hourly pedestrian counts from 66 sensors in Melbourne starting from May
 1043 2009. The original data are updated monthly; the repository snapshot used here contains counts up
 1044 to 2020-04-30.

1045 **PEMS03.** Datasets sourced from Caltrans PeMS, which collects 30-second traffic readings and
 1046 aggregates them into 5-minute intervals (288 time steps per day). Road network structure is derived
 1047 from connectivity status and actual distances between sensors.

1048 **Uber TLC Daily.** Counts of Uber pick-ups from various New York City locations between January
 1049 and June 2015, obtained from FiveThirtyEight’s “`uber-tlc-foil-response`” repository and aggregated
 1050 at hourly and daily resolutions.

1051 **WikiDaily10k.** Daily traffic data for 10,000 Wikipedia pages.

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

B.2 ANOMALY DETECTION DATASET

1078 We extract data from multiple time-series anomaly detection benchmarks Paparrizos et al. (2022);
 1079 Su et al. (2019), including ECG (Moody & Mark, 2001), SMD (Su et al., 2019), MGAB (Thill et al.)
 Genesis (von Birgelen & Niggemann, 2018), GHL (Filonov et al., 2016), Occupancy (Candanedo &

1080 Feldheim, 2016). These datasets span various domains, including healthcare (ECG), mathematical
 1081 biology (MGAB), spacecraft telemetry (Genesis), industrial control system (GHL), environmental
 1082 sensing (Occupancy), cyber-security on IT Operations (SMD). The statistics of these datasets are
 1083 shown in Table 5. To address class imbalance, we count the number of anomalous sequences and
 1084 randomly select an equal number of normal sequences, resulting in a balanced dataset. Below are
 1085 the meta information for each dataset.

1086 **MGAB.** This dataset is composed of Mackey-Glass time series with non-trivial anomalies. Mackey-
 1087 Glass time series exhibit chaotic behavior that is difficult for the human eye to distinguish.
 1088

1089 **ECG.** This dataset is a standard electrocardiogram dataset and the anomalies represent ventricular
 1090 premature contractions. The ECG recordings were made using Del Mar Avionics model 445 two-
 1091 channel reel-to-reel Holter recorders, and the analog signals were recreated for digitization using
 1092 a Del Mar Avionics model 660 playback unit. The digitization rate (360 samples per second per
 1093 channel) was chosen to accommodate the use of simple digital notch filters to remove 60 Hz (mains
 1094 frequency) interference.

1095 **Genesis.** This dataset is a portable pick-and-place demonstrator which uses an air tank to supply
 1096 all the gripping and storage units. Data samples were taken through an OPC connection with a
 1097 resolution of 50 milliseconds for a total of 42 production cycles. The first 38 production cycles
 1098 contain only normal behavior and were used to train the selforganizing map for both experiments
 1099 shown in this section. Two of the 4 remaining cycles contain anomalous behavior and are used for
 1100 the anomaly detection.

1101 **GHL.** This dataset is a Gasoil Heating Loop Dataset and contains the status of 3 reservoirs such
 1102 as the temperature and level. Anomalies indicate changes in max temperature or pump frequency.
 1103 Type of cyber attack to the normal process logic is the unauthorized change of max Receiving Tank
 1104 level. By changing the time of attack and the value of the hacked max Receiving Tank level, we
 1105 generated many anomalous data sets used for fault detection.

1106 **Occupancy.** This dataset contains experimental data of room occupancy, such as temperature, hu-
 1107 midity, light, and CO₂. Ground-truth occupancy was obtained from time stamped pictures that were
 1108 taken every minute.

1109 **SMD.** SMD (Server Machine Dataset) is collected from a large Internet company. The data is
 1110 sampled every 5 seconds. Labels denote whether a point is an anomaly and the dimensions contribute
 1111 to every anomaly.

1112
 1113 Table 5: Summary of anomaly detection datasets.
 1114

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

1122
 1123 **B.3 CLASSIFICATION DATASET**

1124 We extract data from the UCR Archive (Dau et al., 2019a). To create a focused subset for our
 1125 study, we applied two primary selection criteria: we included only datasets with four or fewer
 1126 classes and time series with a sequence length of 400 time points or less. Through our selection,
 1127 we extract data from 37 benchmarks in the UCR Archive, including SonyAIBORobotSurface1 &
 1128 SonyAIBORobotSurface2 (Mueen et al., 2011), FreezerRegularTrain & FreezerSmallTrain (Mur-
 1129 ray, 2015), ToeSegmentation1 & ToeSegmentation2 (Ye & Keogh, 2011), TwoPatterns (Geurts, May
 1130 2002), CBF (Saito & Coifman, 1994), Wafer & ECG200 (Olszewski et al., 2001), TwoLeadECG,
 1131 ECGFiveDays, DistalPhalanxOutlineCorrect & MiddlePhalanxOutlineCorrect & ProximalPhalanx-
 1132 OutlineCorrect & DistalPhalanxOutlineAgeGroup & MiddlePhalanxOutlineAgeGroup & Proximal-
 1133 PhalanxOutlineAgeGroup & PhalangesOutlinesCorrect (Bagnall & Davis, 2014), MoteStrain (Sun

et al., 2005), GunPointMaleVersusFemale & GunPointOldVersusYoung & GunPointAgeSpan & GunPoint (Ratanamahatana & 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 & 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.

Table 6: Classification data used in our experiments.

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

1179
1180
1181
1182
1183
1184
1185
1186
1187

1188
1189 **Algorithm 1:** Hierarchical Random Sampling
1190 **Input:** Domains \mathcal{M} ;
1191 Datasets $\mathcal{D}(m)$ for each domain $m \in \mathcal{M}$;
1192 Sequences $\mathcal{S}(d)$ for each dataset $d \in \mathcal{D}$;
1193 Segment length l
1194 **Output:** Segment $s_{t:t+l-1}$
1195 $m \leftarrow \text{UniformPick}(\mathcal{M})$; // Randomly select a domain
1196 $d \leftarrow \text{UniformPick}(\mathcal{D}(m))$; // Randomly select a dataset in the domain
1197 $s \leftarrow \text{UniformPick}(\mathcal{S}(d))$; // Randomly select a seq. from the dataset
1198 $t \leftarrow \text{UniformPick}\{1, \dots, |s| - l + 1\}$; // Randomly select a start index
1199 $\text{return } s_{t:t+l-1}$; // Return the segment

C BENCHMARK CONSTRUCTION

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

C.1 HIERARCHICAL UNIFORM SAMPLING

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 B.1). To ensure a balanced distribution over domains, datasets and sequences, we use *Hierarchical Uniform Sampling* presented in Algorithm 1 to obtain samples.

C.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: TRQA 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 TRQA 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.

C.3 CHARACTERIZATION

The characterization task assesses the model’s capability to reason about 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.

1242
1243
1244 Table 7: Topics and Sub-Topics for Time Series Analysis
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295

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

1271 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)
1272 to generate one QA pair per instance using a randomly selected subset of one to three topics (from
1273 Table 7) and a question type (TF or MC). The process is as follows.

1274 *Step 1: Captioning & sub-topic selection.* GPT first produces a short, neutral caption summarizing
1275 visible patterns (e.g., “gradual upward drift with weak weekly oscillation”). For each chosen topic,
1276 a sub-topic is sampled uniformly at random, e.g., trend, seasonality and dispersion.

1277 *Step 2: QA synthesis.* GPT generates a TF or MC question grounded in \mathbf{x} , the caption, and the
1278 selected sub-topics.

1279 *Step 3: Self-verification.* GPT performs a self-check and outputs a confidence score in [0,1]. We
1280 retain QA pairs only if confidence ≥ 0.95 .

1281 *Step 4: Multi-LLM consensus.* We query GPT-4.1, Gemini-2.5-Flash, and Claude-3.5-Sonnet using
1282 the same prompt, which includes the generated question along with its allowed answer choices (for
1283 both TF and MC formats), and collect their responses. To determine the final label, we adopt a
1284 weighted majority voting scheme among these three models and GPT-4o’s original answer. Specif-
1285 ically, GPT-4.1 and Gemini-2.5-Flash are assigned higher weights of 1.5 each, reflecting their su-
1286 perior performance in preliminary evaluations, while Claude-3.5-Sonnet and GPT-4o are each as-
1287 signed a weight of 1.0. The option with the highest total weighted vote is selected as the consensus
1288 answer. If a tie occurs—i.e., two or more answers receive the same highest weighted score—the
1289 corresponding QA pair is discarded to avoid introducing ambiguity or noise into the dataset. This
1290 ensemble-based strategy mitigates single-model biases, smooths out random errors, and produces
1291 more reliable and stable labels, which are crucial for ensuring the benchmark’s quality.

1292 Here’s the *system* prompt template.

1293
1294 | You are an expert of time series analysis.

1296 1. Generate a `meta_caption` solely based on the meta information within 50
 1297 words.
 1298 2. Generate a `detailed_caption` based on both meta information and time
 1299 series within 100 words.
 1300 3. Generate a `{}` based on the time series, `meta_caption`, `detailed_caption`
 1301 and the more detailed question instructions.
 1302 4. Generate a correct answer `{}` for your question.
 1303 5. A successful generation must meet the following conditions:
 1304 (1) there is only one correct answer;
 1305 (2) the question strictly follows the instructions;
 1306 (3) the answer of the question cannot be easily derived from the
 1307 `meta_caption`;
 1308 (4) the question should be about the time series itself without involving
 1309 external knowledge;
 1310 (5) do not repeat the input time series in questions or answers.
 1311 6. Show your confidence of your determination of success within 0-1.

1312 Here's the *user prompt template*.

1313 The time series is `{}`.
 1314 Its meta information is `{}`.
 1315 The question must be about all these topics: `{}`.
 1316 The sub-topics of `{}` includes but not limited to `{}`.
 1317 First think about the all possible sub-topics and their taxonomy.
 1318 Then randomly pick a sub-topic from each topic `({})` to generate the
 1319 question and answer pairs.

1320 C.4 COMPARISON

1321 The comparison task assesses the model's ability to reason about the relative characteristics of two
 1322 time series, such as overall shape, temporal alignment, and correlation patterns. Similar to the
 1323 characterization task, this task is also formulated as either TF or MC questions, where the model
 1324 must identify similarities or differences between the given pair of sequences. The characteristics
 1325 evaluated in the task are directly drawn from the standardized taxonomy of Topics and Sub-topics
 1326 (from Table 7), which is shared with the Characterization task.

1327 To construct the comparison set, we first obtain an anchor sample \mathbf{x} from a specific domain M ,
 1328 dataset D , and sequence S . Given this anchor \mathbf{x} , we generate a set of ten comparison samples
 1329 $\mathbf{x}'1, \dots, \mathbf{x}'10$, each having the same length as \mathbf{x} . These samples are drawn in a structured manner to
 1330 represent varying degrees of similarity: one from the same sequence S , two from different sequences
 1331 within the same dataset D , three from other datasets within the same domain M , and four from
 1332 entirely different domains. This tiered sampling strategy creates a natural hierarchy of difficulty,
 1333 challenging the model to distinguish between subtle intra-sequence similarities and broader cross-
 1334 domain differences.

1335 Finally, we apply a process similar to the characterization task to generate QA pairs, where GPT-
 1336 based models produce questions and candidate answers. The questions are then refined and validated
 1337 through multi-LLM consensus to ensure accuracy and reduce bias, resulting in high-quality, reliable
 1338 evaluation data for this task.

1340 C.5 DATA TRANSFORMATION

1341 The data transformation task evaluates the model's ability to infer and reason about the transformation
 1342 relationship between an input time series and its transformed counterpart. These transformations
 1343 are generated using well-established signal processing techniques, including the Fourier transform,
 1344 wavelet transform, and first-order differencing, which are widely used in time series analysis to re-
 1345 veal underlying structures or remove trends. This task is particularly challenging because it requires
 1346 the model to not only recognize the patterns in the raw input series but also to understand how
 1347 specific mathematical operations alter these patterns.

1348 We use predefined templates to formulate the task as either TF or MC questions. For TF questions,
 1349 the model is asked to determine whether a given candidate sequence is indeed the correct transfor-

1350 mation of the input time series x (e.g., whether it is the Fourier transform result of x). For MC
 1351 questions, the model must select the correct transformed sequence from multiple candidates, given
 1352 both the input series x and the specified transformation operation (e.g., Fourier transform).

1353 To ensure accuracy and consistency, all transformations are computed using professional and reliable
 1354 scientific libraries (Harris et al., 2020; Virtanen et al., 2020). The correct transformation is generated
 1355 directly from the input x , while distractor sequences are created by applying the same transformation
 1356 to randomly sampled, unrelated time series x' . This setup forces the model to carefully analyze the
 1357 relationship between the input and its transformation rather than relying on superficial similarities,
 1358 providing a robust evaluation of its reasoning ability.

1359 Here's the template to construct question.

1360 The time series is {}.
 1361 Its meta information is {}.
 1362 The question must be about all these topics: {}.
 1363 The sub-topics of {} includes but not limited to {}.
 1364 First think about the all possible sub-topics and their taxonomy.
 1365 Then randomly pick a sub-topic from each topic ({{}}) to generate the
 1366 question and answer pairs.

1368 C.6 TEMPORAL RELATIONSHIP

1369 The Temporal Relationship task is a discriminative sequence-level reasoning task, rather than a generative
 1370 forecasting task. The task evaluates a model's ability to infer and reason about the temporal
 1371 structure among sequential patches of a time series. Specifically, the task evaluates whether a model
 1372 can understand the structural continuity and chronological dependencies of time series patches, test-
 1373 ing 3 core reasoning capabilities: *Structural Continuity*, *Chronological Reasoning*, and *Contextual*
 1374 *Discrimination*. (1) *Structural Continuity* tests whether the model can identify which candidate
 1375 segment shares the underlying temporal dynamics required to validly continue a given trajectory.
 1376 (2) *Chronological Reasoning* tests whether the model can reconstruct the correct temporal order of
 1377 shuffled patches. (3) *Contextual Discrimination* tests the model's ability to distinguish the true con-
 1378 tinuation from "plausible" but incorrect alternatives that may share similar global statistics but lack
 1379 local continuity. This task is formulated as true-or-false (TF), multiple-choice (MC), or puzzling
 1380 (PZ) questions.

1381 Given the first chronological patch x : (1) A TF question asks the model to determine whether a
 1382 candidate patch y is the immediate successor of x . (2) An MC question requires the model to select
 1383 the correct next patch from four candidates $[y_1, y_2, y_3, y_4]$.

1384 The false candidates in both TF and MC settings are randomly sampled from the full dataset but
 1385 are guaranteed to come from sequences different from that of x , preventing the model from simply
 1386 memorizing patterns. For PZ questions, the model is presented with four shuffled successor patches
 1387 of x and must reconstruct their correct chronological order, which poses a greater challenge as
 1388 it requires deeper temporal reasoning. All questions are generated using predefined templates to
 1389 ensure consistency and diversity.

1390 We use the following question template to construct questions.

1391 Which of the following choices is most likely the future continuation of
 1392 the given time series?
 1393 Respond ONLY with the letter of the correct choice (A, B, C, or D)
 1394
 1395 Choices:
 1396 A: {}
 1397 B: {}
 1398 C: {}
 1399 D: {}

1400 Is the following patch the future continuation of the given time series?
 1401 {}
 1402 Respond ONLY with the letter of the correct choice (T or F).

1404
 1405 Choices:
 1406 T: True.
 1407 F: False.

1408

1409

1410 C.7 ANOMALY DETECTION

1411 First, all time-series data are standardized using z-score normalization to remove scale effects across
 1412 different features. Next, we randomly sample a subsequence of length T , where $T \in [32, 256]$,
 1413 from each time-series instance to capture varying temporal dynamics. To address class imbalance,
 1414 we count the number of anomalous sequences and randomly select an equal number of normal
 1415 sequences, resulting in a balanced dataset. Finally, we enrich each sample with meta information,
 1416 domain information, the normalized time-series subsequence, and its corresponding label.

1417 Here's the question template.
 1418

1419 Determine whether the given time series contains anomalies.
 1420 Respond ONLY with the letter of the correct choice (T or F).

1421 Choices:
 1422 T: True.
 1423 F: False.

1424

1425

1426 C.8 CLASSIFICATION

1427 Information about the time series and the task is given in the text description. Here's the template to
 1428 construct questions.
 1429

1430 Classify the given time series into one of the categories below.
 1431 Respond ONLY with the letter of the correct choice (A, B).

1432 Choices:
 1433 A: {}
 1434 B: {}

1435

1436

1437

1438

1439

1440

1441

1442

1443

1444

1445

1446

1447

1448

1449

1450

1451

1452

1453

1454

1455

1456

1457

1458 D EXAMPLES

1459

1460

1461

1462

1463 In this section, we show some examples of the constructed QA pairs.

1464

1465

1466

1467

1468

1469 **TRQA — Sample 1**

1470

1471

1472

1473

1474

1475

1476

1477

1478

1479

1480

1481

1482

1483

1484

1485

1486

1487

1488

1489

1490

1491

1492

1493

1494

1495

1496

1497

1498

1499

1500

1501

1502

1503

1504

1505

1506

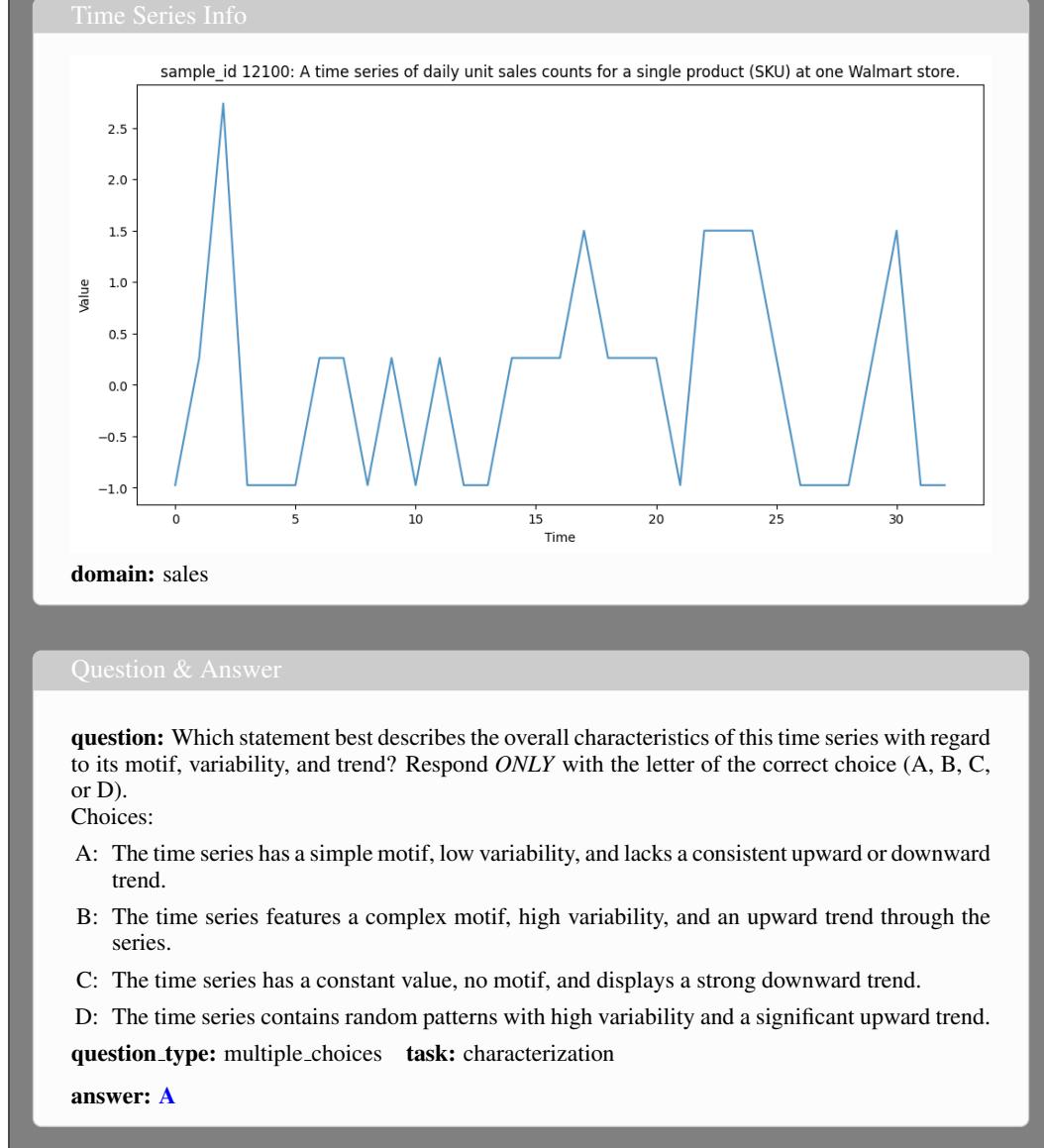
1507

1508

1509

1510

1511



1512

TRQA — Characterization Sample 1

1513

1514

1515

1516

1517

1518

1519

1520

1521

1522

1523

1524

1525

1526

1527

1528

1529

1530

1531

1532

1533

1534

1535

1536

1537

1538

1539

1540

1541

1542

1543

1544

1545

1546

1547

1548

1549

1550

1551

1552

1553

1554

1555

1556

1557

1558

1559

1560

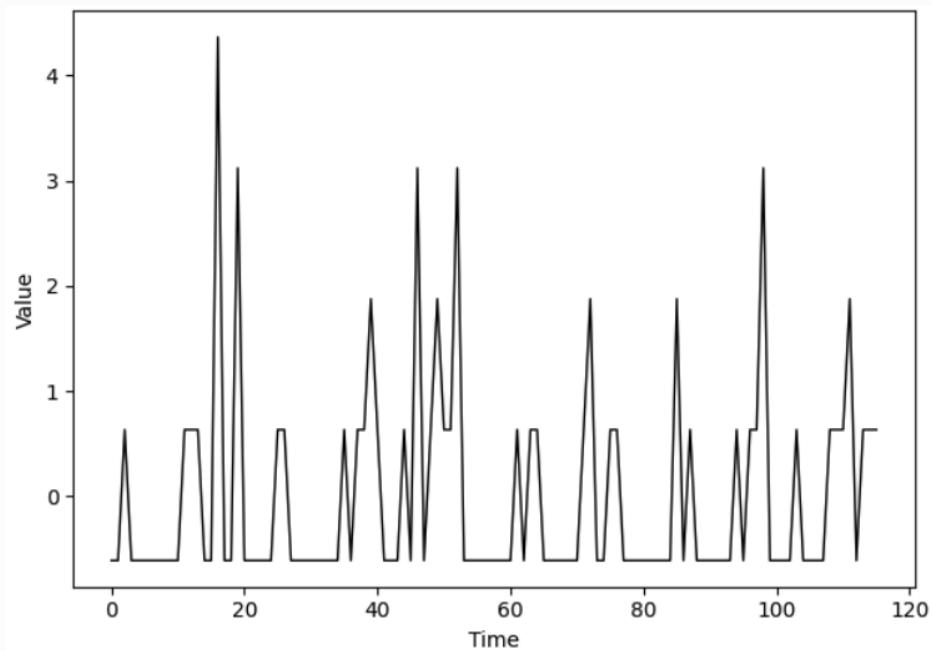
1561

1562

1563

1564

1565

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**

1620

TRQA — Comparison Sample 1

1621

1622

1623

1624

1625

1626

1627

1628

1629

1630

1631

1632

1633

1634

1635

1636

1637

1638

1639

1640

1641

1642

1643

1644

1645

1646

1647

1648

1649

1650

1651

1652

1653

1654

1655

1656

1657

1658

1659

1660

1661

1662

1663

1664

1665

1666

1667

1668

1669

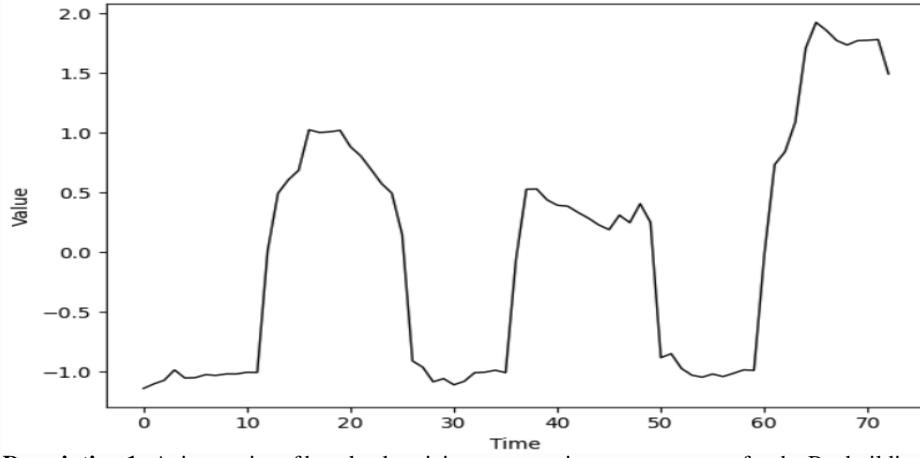
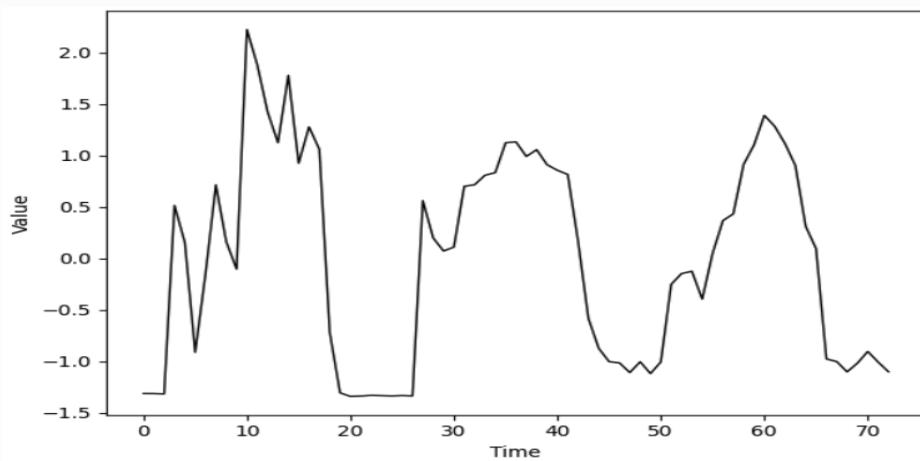
1670

1671

1672

1673

Time Series Info



Description 1: A time series of hourly electricity consumption measurements for the Rat building, representing one building's power usage.

Description 2: A time series of hourly electricity consumption measurements for the Rat building, representing one building's 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

1674

TRQA — Comparison Sample 2

1675

1676

1677

1678

1679

1680

1681

1682

1683

1684

1685

1686

1687

1688

1689

1690

1691

1692

1693

1694

1695

1696

1697

1698

1699

1700

1701

1702

1703

1704

1705

1706

1707

1708

1709

1710

1711

1712

1713

1714

1715

1716

1717

1718

1719

1720

1721

1722

1723

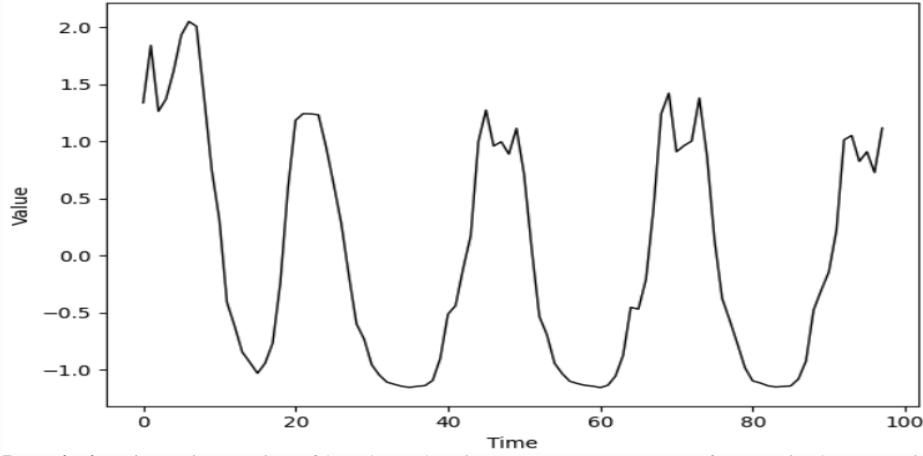
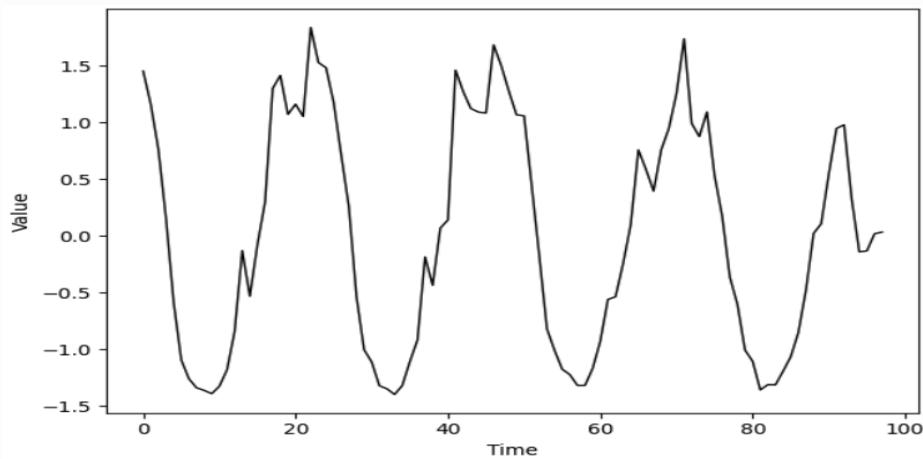
1724

1725

1726

1727

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

1782

1783

1784

1785

1786

1787

1788

1789

1790

1791

1792

1793

1794

1795

1796

1797

1798

1799

1800

1801

1802

1803

1804

1805

1806

1807

1808

1809

1810

1811

1812

1813

1814

1815

1816

1817

1818

1819

1820

1821

1822

1823

1824

1825

1826

1827

1828

1829

1830

1831

1832

1833

1834

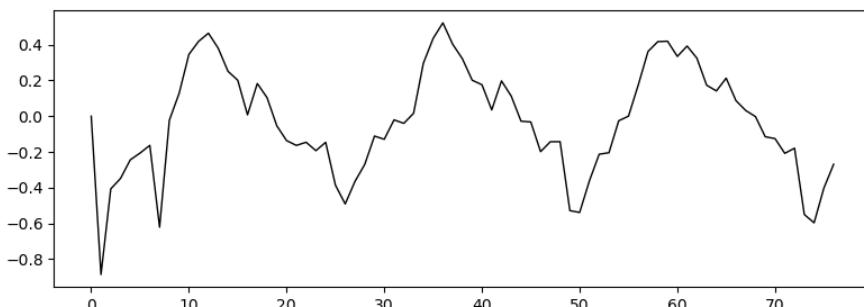
1835

Question & Answer

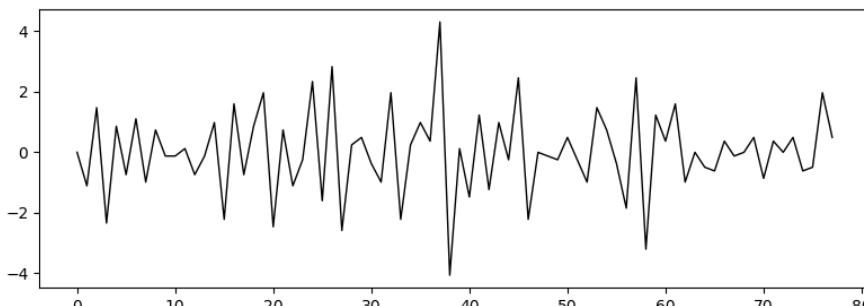
Question: Which of the following choices is most likely the First Order Difference of the given time series? Respond ONLY with the letter of the correct choice (A, B, C, or D).

Choices:

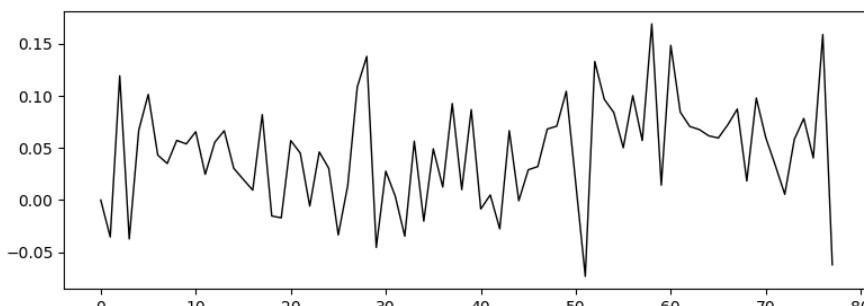
A:



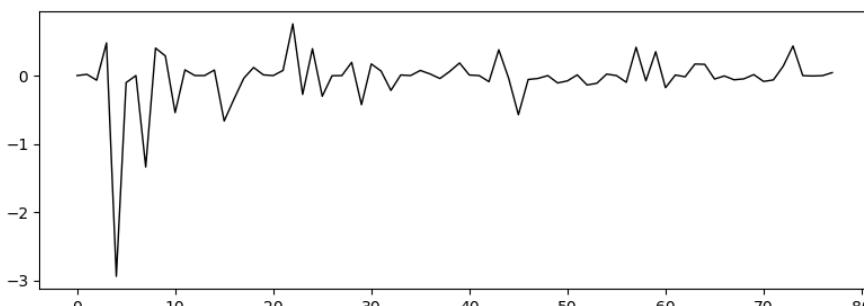
B:

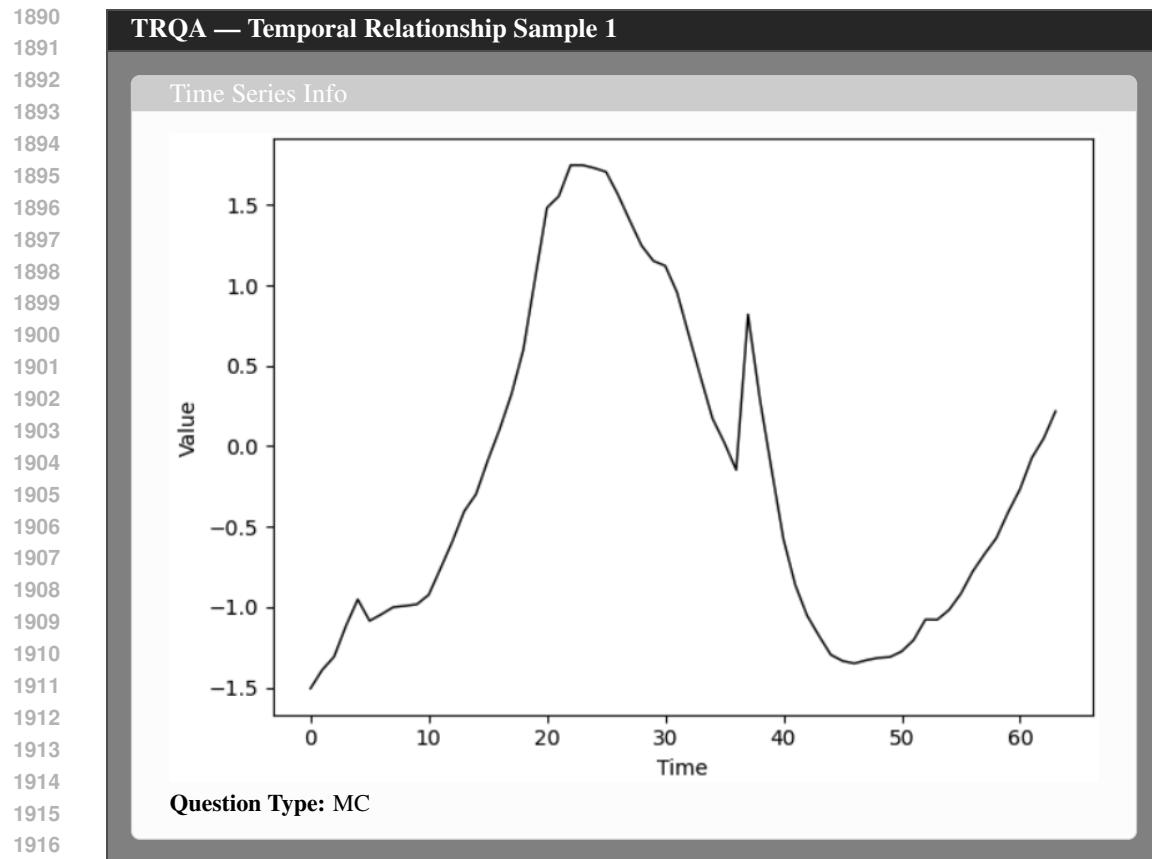


C:



D:

**Answer: B**



1944

1945

1946

1947

1948

1949

1950

1951

1952

1953

1954

1955

1956

1957

1958

1959

1960

1961

1962

1963

1964

1965

1966

1967

1968

1969

1970

1971

1972

1973

1974

1975

1976

1977

1978

1979

1980

1981

1982

1983

1984

1985

1986

1987

1988

1989

1990

1991

1992

1993

1994

1995

1996

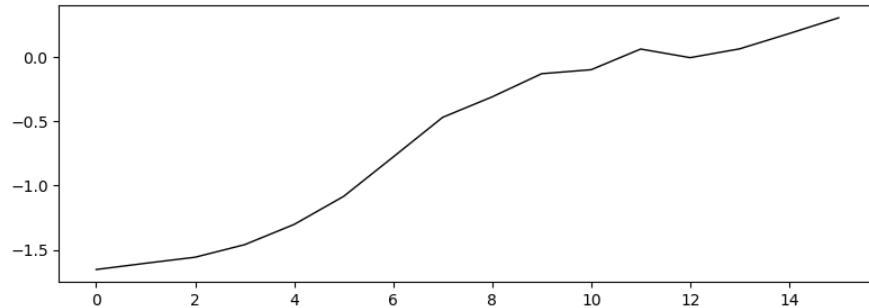
1997

Question & Answer

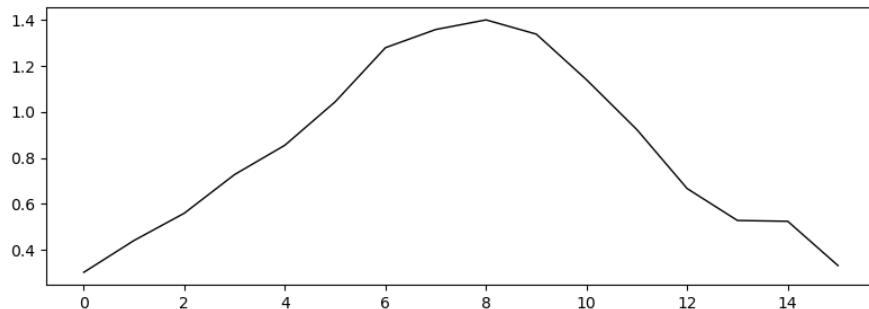
Question: Which of the following choices is most likely the future continuation of the given time series? Respond ONLY with the letter of the correct choice (A, B, C, or D).

Choices:

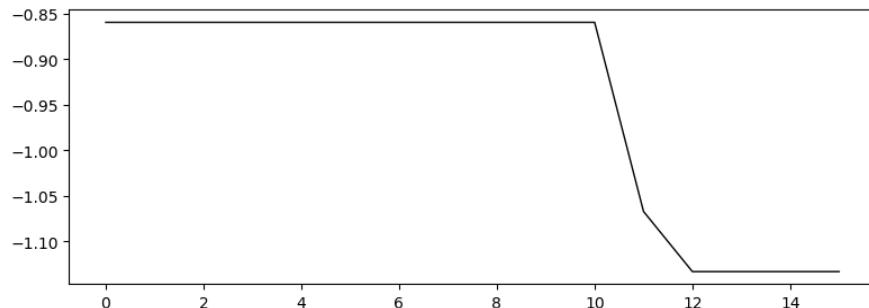
A:



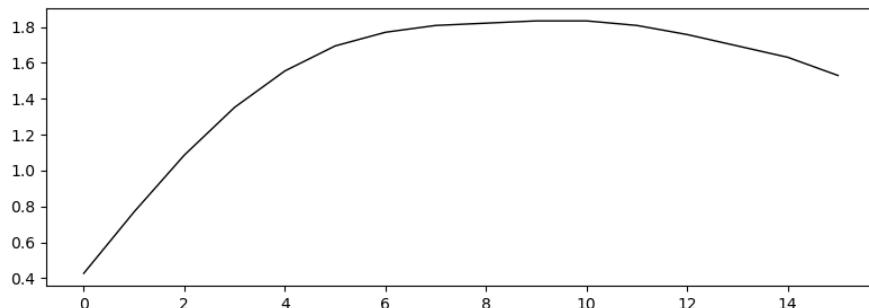
B:

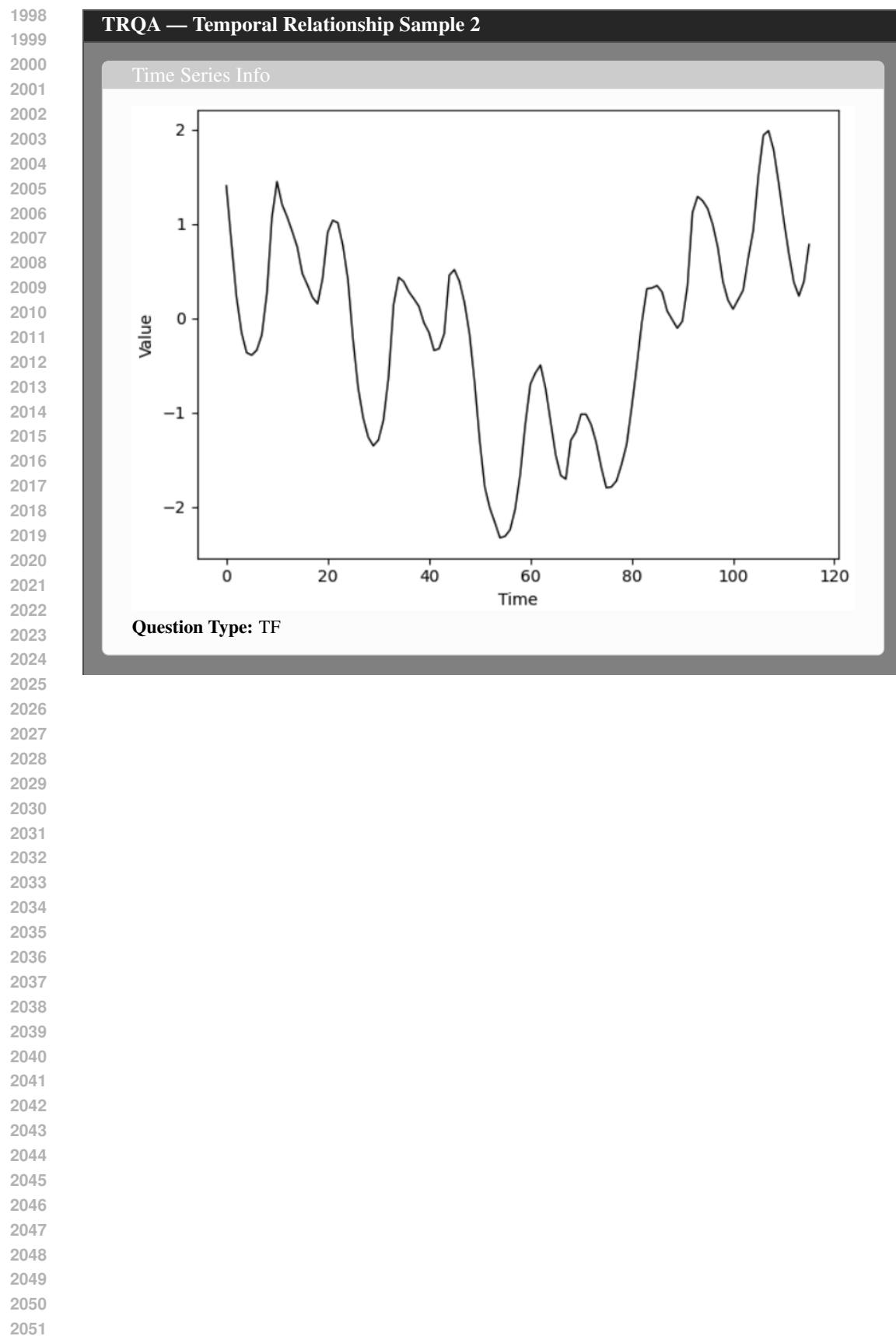


C:



D:

**Answer: B**



2052

2053

2054

2055

2056

2057

2058

2059

2060

2061

2062

2063

2064

2065

2066

2067

2068

2069

2070

2071

2072

2073

2074

2075

2076

2077

2078

2079

2080

2081

2082

2083

2084

2085

2086

2087

2088

2089

2090

2091

2092

2093

2094

2095

2096

2097

2098

2099

2100

2101

2102

2103

2104

2105

Question & Answer

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

Value

Time

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

Choices:

T: True.

F: False.

Answer: T

TRQA — Temporal Relationship Sample 3

Time Series Info

2.0

1.5

1.0

0.5

0.0

-0.5

-1.0

0 2 4 6 8 10 12 14

Question Type: PZ **Domain:** finance

2106

2107

2108

2109

2110

2111

2112

2113

2114

2115

2116

2117

2118

2119

2120

2121

2122

2123

2124

2125

2126

2127

2128

2129

2130

2131

2132

2133

2134

2135

2136

2137

2138

2139

2140

2141

2142

2143

2144

2145

2146

2147

2148

2149

2150

2151

2152

2153

2154

2155

2156

2157

2158

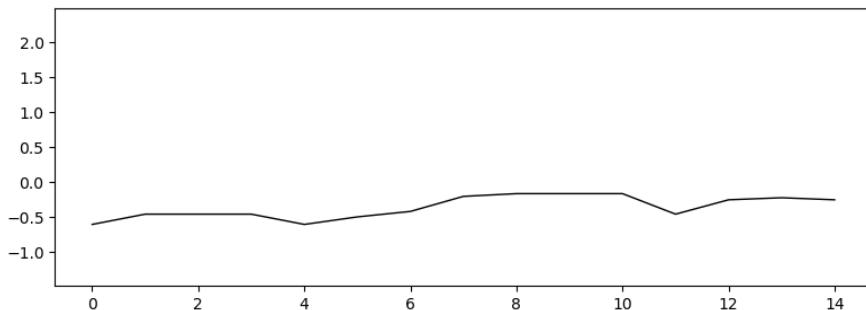
2159

Question & Answer

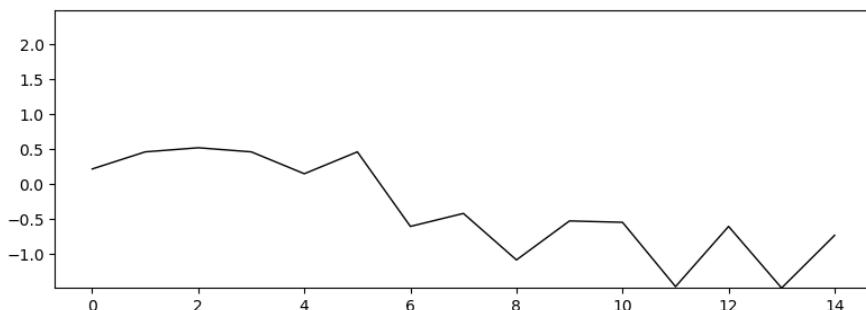
Question: The given time series is the first patch of the sequence. Below are the remaining patches, labeled as A, B, C, and D. Arrange A, B, C, D in the correct order to reconstruct the original sequence.

Choices:

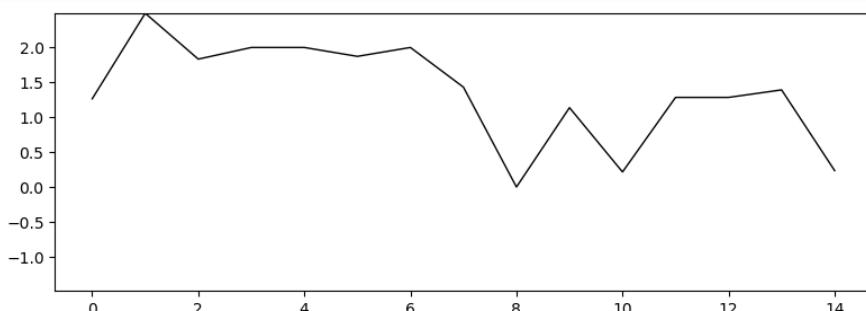
A:



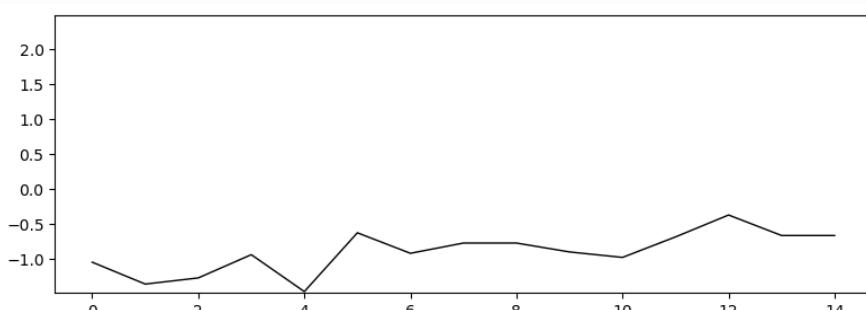
B:



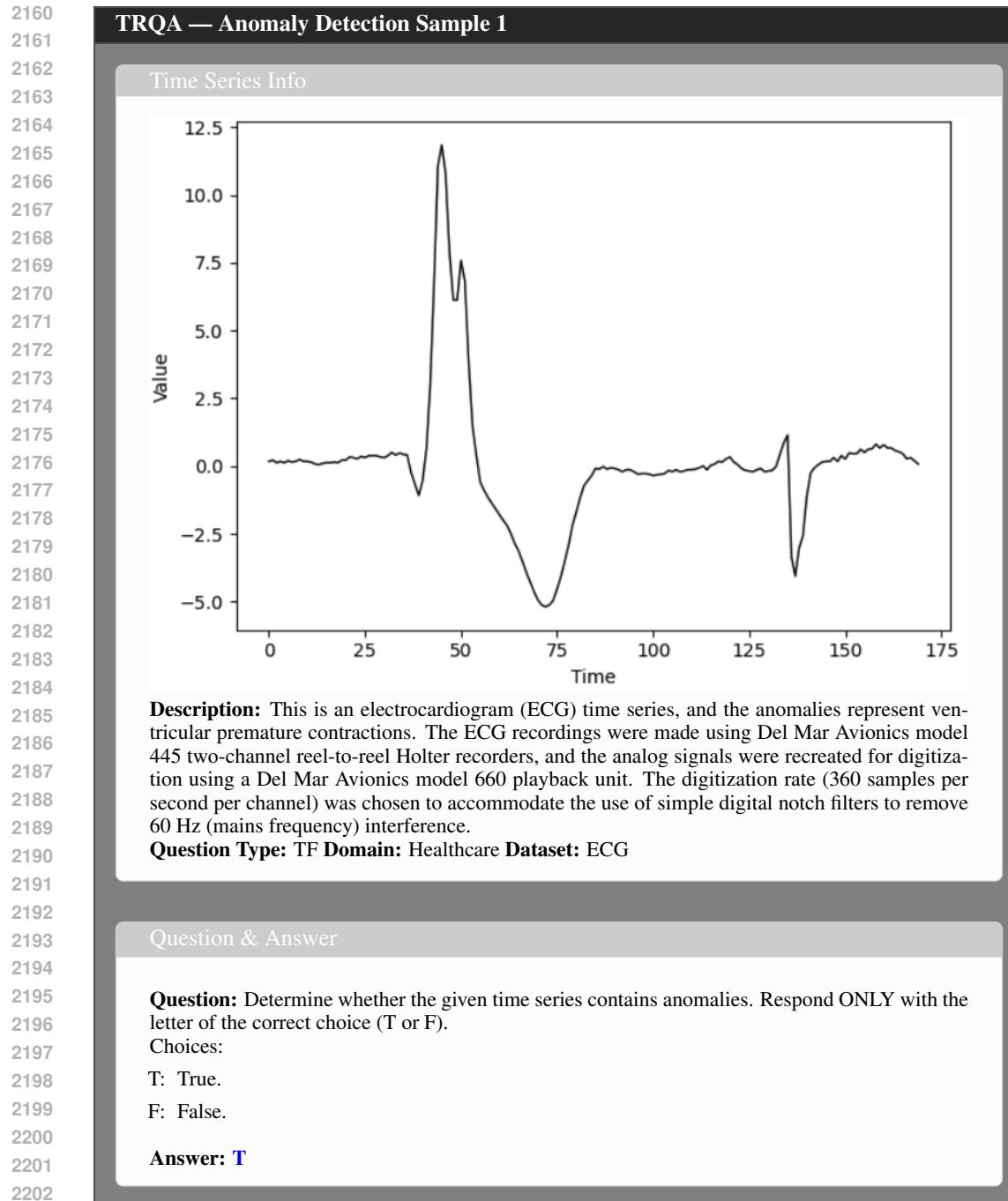
C:

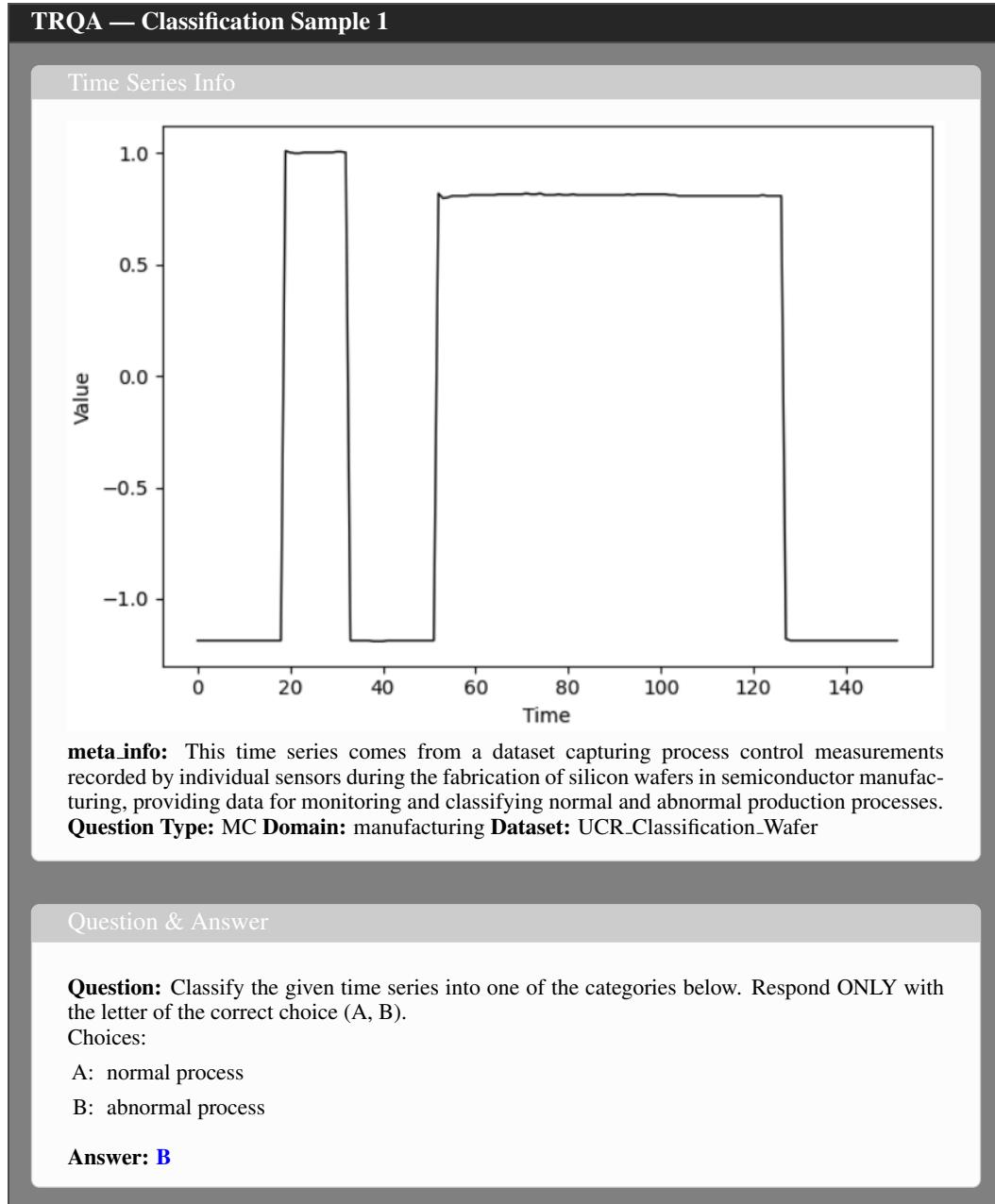


D:



Answer: **C,B,D,A**





2268

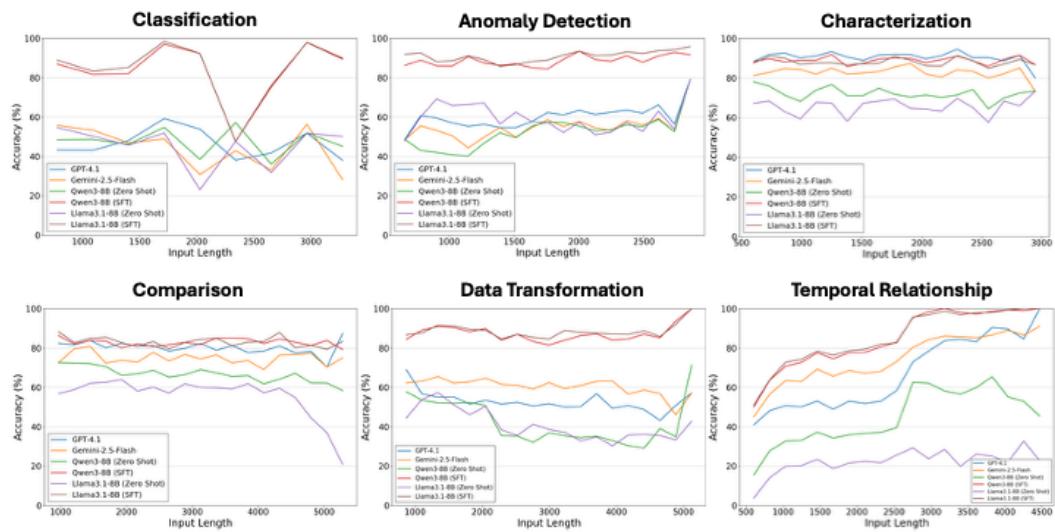
E EXPERIMENT ANALYSIS

2269
 2270 We conducted an in-depth analysis of results from the selected Large Language Models. Specifically,
 2271 our analysis is divided into two major categories: **Accuracy Correlate Analysis** and **Task-Specific**
 2272 **Analysis**. For each analysis, we selected models from both commercial and open-source families. In
 2273 particular, we chose the two best-performing models from Table 3 evaluated on our TRQA Bench-
 2274 mark—namely, GPT-4.1, Gemini 2.5 Flash, LLaMA3-8B, and Qwen3-8B. For LLaMA3.1-8B and
 2275 Qwen3-8B, we analyzed both the zero-shot and instruction tuned models, resulting in a total of six
 2276 models considered in our analysis.

2277

E.1 ACCURACY CORRELATE ANALYSIS

2278 In this category, we examined how model accuracy or overall score correlates with input length.



2301 Figure 3: Input lengths vs. Accuracy by Tasks among six models.

2302
 2303 **Input Length v.s. Accuracy.** To understand how input length impacts model accuracy, we
 2304 conducted a detailed analysis comparing the length of each input with its corresponding accuracy.
 2305 Specifically, the input length is calculated as $len(ts + description + domain + dataset + task +$
 $2306 question_type + question)$ with *String* type. The results are visualized in Figure 3. Each plot may
 2307 contain input length starting and ending at different length as each task contains questions with
 2308 different lengths. Across all six models and five tasks, excluding the Temporal Relation task, we
 2309 observe a consistent trend that longer questions with greater input length generally result in lower
 2310 accuracy and weaker overall model performance. However, the Temporal Relation task exhibits
 2311 the opposite behavior, where accuracy improves with increasing input length. To understand
 2312 this discrepancy, we conducted a detailed analysis of the four reasoning tasks (*Characterization*,
 2313 *Comparison*, *Data Transformation*, *Temporal Relation*) in our proposed TRQA Benchmark,
 2314 focusing on how different question types (*MC*, *TF*, *PZ*) and their corresponding input lengths
 2315 correlate with model accuracy. The results are visualized in Figure 4. The results indicate that for
 2316 all four reasoning tasks, *MC* and *TF* question types show a decline in accuracy with increasing
 2317 input length, whereas the newly proposed *PZ* type exhibits the opposite trend. This implies that the
 2318 model is actively using global contexts, all time series segments, to deduce the correct chronological
 2319 order for answering *PZ* type question, which confirms that the model is engaging in deductive
 2320 reasoning rather than local pattern matching. This proves that *PZ* type question is a rigorous probe
 2321 for *Global Causal Reasoning*. Consequently, models whose accuracy improves with input length
 likely demonstrate a stronger ability to reason directly over time-series patterns.

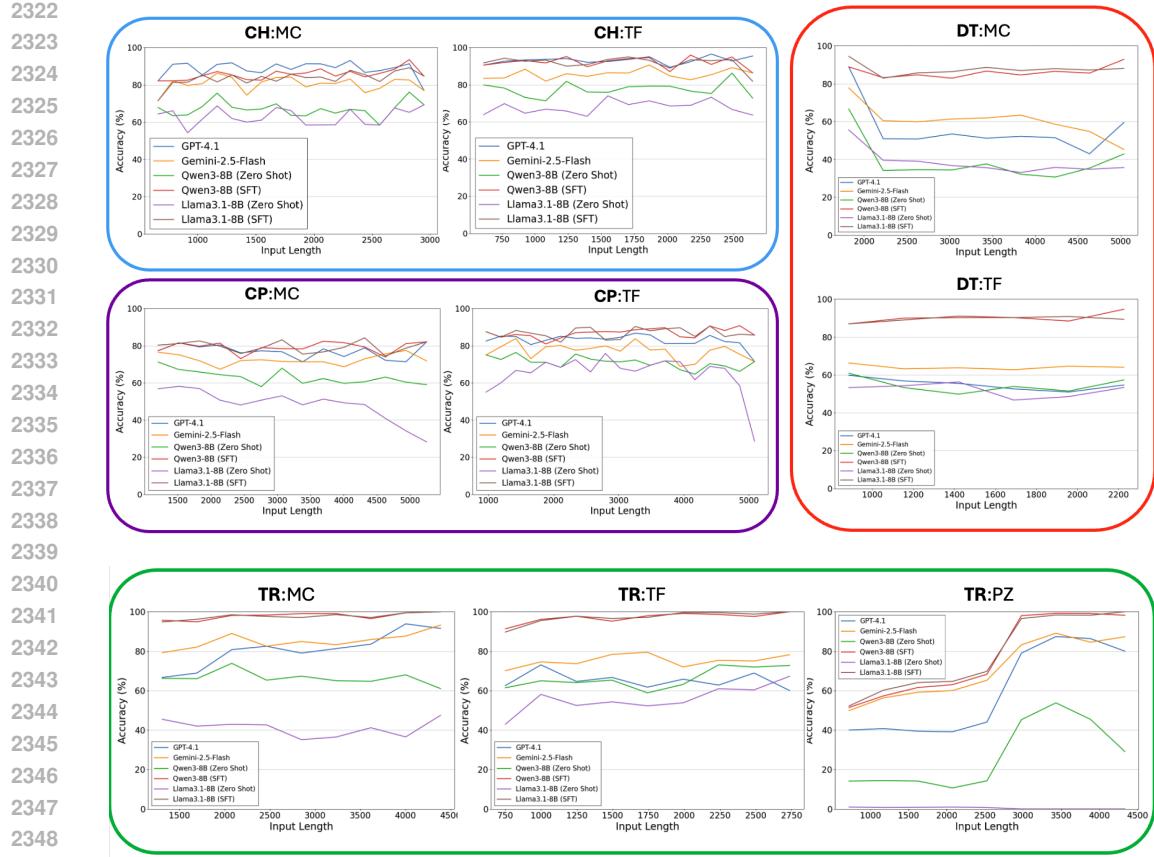


Figure 4: Input length vs. Accuracy by Question Types. CH, CP, DT, and TR denote Characterization, Comparison, Data Transformation, and Temporal Relationship. MC, TF, and PZ denote true-or-false, multiple-choice, and puzzling.

E.2 TASK SPECIFIC ANALYSIS

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

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 8, 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, we also analyzed how question complexity affects GPT-4.0’s performance by varying the number of topics and subtopics used for generation. The model achieved an accuracy of 74.07% on questions derived from a single topic and subtopic, 72.17% for two, and 75.08% for three. These results indicate that the model’s performance is notably stable, which again proves the quality of the proposed benchmark. 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.

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 Trans-

Table 8: Analysis of Comparison tasks.

Group	Model	Same Domain		Different Domain	
		MC	TF	MC	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

form (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 9, 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 9, 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 pattern matching, which also explains the results shown in Table 9.

Table 9: 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	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

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.

Domain-Level Analysis. The results are summarized in Table 10. 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 reasoning difficulty on models.

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	<u>66.54</u>	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	<u>61.03</u>	74.05	68.92	58.61
	LLaMA3.1-8B	75.25	77.50	72.22	<u>61.80</u>	75.80	71.16	60.86

Table 10: 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.

F 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 5)

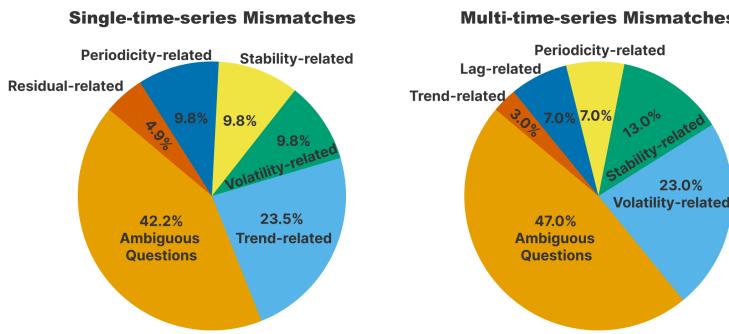


Figure 5: Human explanations for answer mismatches in TRQA