TIME-RA: Towards Time Series Reasoning for Anomaly Diagnosis with LLM Feedback

Anonymous ACL submission

Abstract

Time series anomaly detection (TSAD) has traditionally focused on binary classification and often lacks the fine-grained categorization and explanatory reasoning required for transparent decision-making. To address these limitations, we propose Time-series Reasoning for Anomaly (TIME-RA), a novel task that reformulates TSAD from a discriminative into a generative, reasoning-intensive paradigm. To facilitate this, we introduce RATS40K, the first real-world large-scale multimodal benchmark with ~40,000 samples across 10 domains, integrating raw time series, textual context, and visual plots with structured reasoning annotations. Extensive benchmarking shows that while supervised fine-tuning and visual representations boost diagnostic accuracy and reasoning consistency, performance varies across complex scenarios. Notably, fine-tuned models demonstrate strong "plug-and-play" transferability, outperforming traditional baselines on unseen real-world datasets. Our work establishes a foundation for interpretable, multimodal time series analysis. All code and the RATS40K dataset are fully open-sourced to facilitate future research[1].

1 Introduction

Time series anomaly detection (TSAD) is critical across diverse domains, including finance, healthcare, AIOps, and industrial systems, where timely identification of anomalies prevents severe operational disruptions and economic losses (Wang et al., 2024; Cook et al., 2019; Ren et al., 2019; Khan and Alkhathami, 2024; Liu et al., 2024c). With the rapid development of artificial intelligence, especially deep learning (DL) and large language models (LLMs), significant advancements have been achieved in modeling complex temporal patterns and anomaly detection tasks (Jin et al., 2024). However, the reasoning and detailed categorization of anomalies in time series remain underexplored, limiting our ability to reason diagnoses and hindering deeper understanding and informed decision-making based on identifying underlying causes beyond mere detection. In addition, existing TSAD benchmarks often lack explanatory reasoning and fine-grained anomaly categories for comprehensive diagnosis, creating a bottleneck for further advancement of TSAD (Zhou and Yu, 2024).

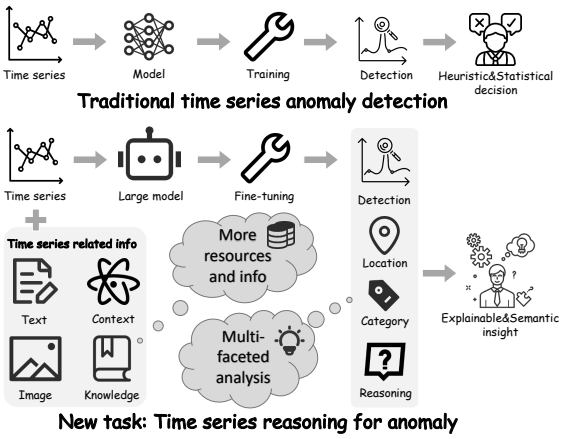


Figure 1: Comparison between traditional TSAD task and the proposed time series reasoning for anomaly task.

As shown in Figure 1, the traditional TSAD task (Yang et al., 2023) focuses mainly on binary anomaly classification (anomalous vs. normal). It fails to provide the specific anomaly categories and diagnostic reasoning necessary for root-cause analysis. Also, existing researches on the TSAD task mainly focus on anomaly detection, neglecting the critical task of uncovering the underlying reasons behind these anomalies. Understanding these root causes is essential for future comprehensive decision-making processes, as it provides actionable insights and fosters interpretability for stakeholders (Chow et al., 2024; Liu et al.). The absence of in-depth causal analysis substantially limits its practical utility, particularly in scenarios where preventive or corrective actions depend on anomaly origins, highlighting the need for a shift toward

[1] https://anonymous.4open.science/r/C6CD

more comprehensive anomaly diagnosis.

Secondly, the scarcity of real-world datasets integrating multiple modalities (numeric time series data, textual explanations, and visual representations) hinders the advancement of multimodal TSAD and reasoning methods. Existing multimodal datasets are typically synthetic or limited to narrow contexts, inadequately capturing real-world complexity and variability (Liu et al., 2025b; Kong et al., 2025b). Moreover, the reasoning capabilities of recent multimodal LLMs (MLLMs) remain underexplored due to a lack of high-quality annotated reasoning data. This gap limits the potential for explainable and interpretable anomaly detection outcomes, highlighting the pressing need for comprehensive, publicly available multimodal datasets and standardized benchmarks (Chen et al., 2025).

To address these limitations, we define a brand new task, **Time** series **R**easoning for **A**nomaly (TIME-RA), as shown in Figure 1. This task converts discriminative models into generative models, e.g., LLMs and MLLMs, and performs domain-specific learning with multimodal inputs. Moreover, TIME-RA reformats these inputs as structured prompts and fine-tunes LLMs, guiding them through a human-like diagnostic process structured into Observation, Thought, and Action stages. The task needs to output multi-objective results, including not only binary detection, but specific classes of anomalies and their reasons. This pipeline ensures clarity, consistency, and interpretability of the explanations generated by models.

To support TIME-RA task, we introduce **R**easoning for **A**nomaly in **T**ime **s**eries **40K** (RATS40K), the first real-world comprehensive multimodal dataset. It uniquely integrates numeric time series, contextual text, and visual representations, covering 10 real-world scenarios. It includes 14 univariate and 6 multivariate anomaly types, each with formal definitions, illustrative examples, and real-world scenarios for clarity and reproducibility. In addition, the initial annotations generated by a pool of the strongest LLMs currently available are subsequently refined through a rigorous AI-driven feedback process. This process results in annotations that are sufficiently accurate and suitable for supporting the TIME-RA task.

Leveraging RATS40K dataset, we extensively fine-tune and evaluate multiple advanced LLMs, assessing their anomaly detection performance and reasoning capabilities across diverse anomaly categories and modalities. Through rigorous experimental evaluations, we derive meaningful insights into the strengths, limitations, and potential areas of improvement for current models, demonstrating the dataset's capability to support comprehensive benchmarking and foster substantial advancements in time series reasoning for anomaly task. In summary, our key contributions include:

- **Task Reformulation (TIME-RA)**: We define TIME-RA, a novel task that elevates traditional binary detection into a generative diagnosis paradigm requiring joint detection, fine-grained categorization, and causal explanation.
- **Multimodal Dataset (RATS40K)**: We construct RATS40K, the first real-world multimodal reasoning dataset for TIME-RA with ~40K samples across 10 domains, integrating raw time series, textual context, and visual plots with expert-aligned diagnostic labels.
- **AI-Feedback Alignment Pipeline**: We develop a structured prompting and AI-assisted feedback pipeline that utilizes a diverse LLM pool and GPT-4 refinement to ensure high-quality, interpretable reasoning annotations.
- **Benchmarking and Generalization**: We provide extensive benchmarks of SOTA (M)LLMs, demonstrating that while fine-tuning and visual data boost reasoning quality, the task remains a complex frontier.

2 A New Lens for Anomaly Detection

In this section, we introduce the Time series Reasoning for Anomaly (TIME-RA) task, with an overview of its workflow presented in Figure 2. We begin by curating a large-scale and diversified dataset, RATS40K, which includes ground-truth of category and reason, detailed in the Section 3. This universal anomaly detection dataset is specifically formatted using our prompt engineering approach with a comprehensive anomaly definition, preparing it for model fine-tuning. Subsequently, the fine-tuned language model acquires enhanced expertise in time series analysis, enabling it to perform downstream anomaly detection tasks such as anomaly reasoning. Overall, TIME-RA establishes a multi-tasking workflow for time series analysis, extending beyond the conventional binary anomaly detection problem (Blázquez-García et al., 2021) to address complex scenarios.

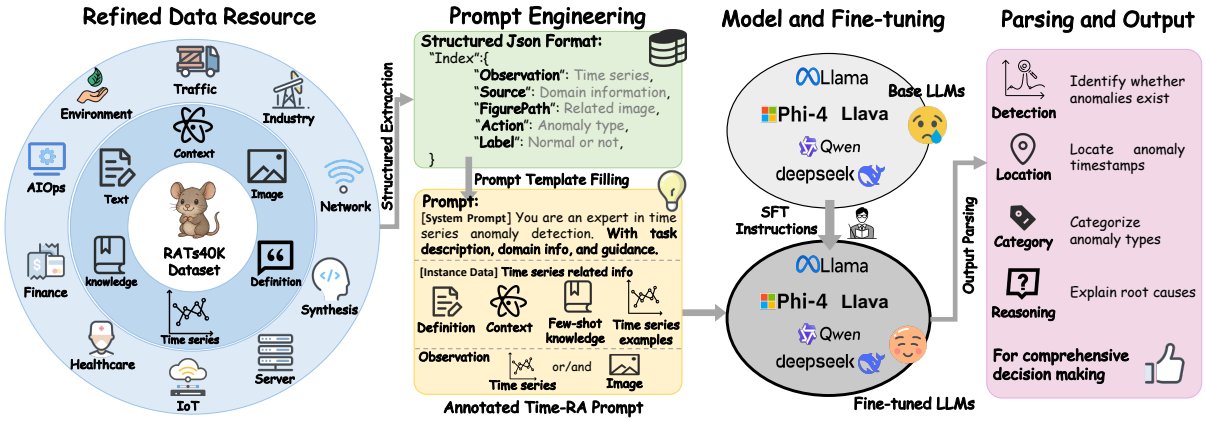


Figure 2: The end-to-end TIME-RA workflow highlights how multimodal inputs are structured and jointly optimized for detection, categorization, and reasoning.

2.1 Task Definition

The TIME-RA task extends traditional time series analysis by integrating multimodal reasoning and causal analysis. Formally, given multimodal input $\{T, D, V\}$, where T is univariate/multivariate time series data, D represents contextual textual metadata (e.g., domain descriptions), and V is a visual representation. This task is multi-objective and requires a (vision) language model π to perform:

- **Anomaly Detection**: Identify whether the time series T contains anomalous segments, i.e, determine an anomaly detection label $y_l = \pi_{\text{detect}}(\cdot|T, D, V) \in \{0, 1\}$,
- **Fine-grained Classification**: For anomalous segments, the model should also identify the specific class of anomaly so that we can invoke it as a model action. An action $a = \pi_{\text{classify}}(\cdot|T, D, V) \in \mathcal{C}_{\text{uni}} \cup \mathcal{C}_{\text{multi}}$, where $\mathcal{C}_{\text{uni}}$ and $\mathcal{C}_{\text{multi}}$ denote our comprehensive anomaly taxonomies in the Section 2.2.
- **Model Thoughts**: The model also needs to generate human-understandable explanations, which usually contain the location of the anomaly and the model's thoughts for justifying both the anomaly presence and category assignment. Mathematically, a model's thought $r = \pi_{\text{reason}}(\cdot|T, D, V)$.

Based on the task definition, we first introduce expert-defined anomaly classifications and corresponding interpretations that provide a detailed reference for prompt engineering of anomaly data.

2.2 Anomaly Category Definition

Traditional anomaly detection datasets mostly treat detection as a binary task (normal vs. anomalous) without distinguishing *anomaly types*. In fact, identifying the category of an anomaly in practice is crucial for root cause analysis and effective decision-making. We therefore define a fine-grained taxonomy of anomaly classes and label each anomalous segment with its specific category. After surveying the literature and consolidating existing taxonomies (Chandola et al., 2009; Schmidl et al., 2022; Blázquez-García et al., 2021; Liu et al., 2025a), we select *14 univariate anomaly types* and *6 multivariate anomaly types* as our categorical labels. The univariate anomaly classes focus on time-localized deviations, such as point outliers, trend drifts, and nonlinear pattern anomalies. The multivariate anomaly classes consider both temporal and cross-series aberrations, for example, a trend deviation anomaly where one variable's trend diverges significantly from others, or a joint-context anomaly where an otherwise acceptable pattern becomes anomalous when multiple variables are considered together. Each data segment in our dataset is thus annotated with both a binary label (normal or anomalous) and a specific anomaly category (or "normal" category for non-anomalous cases). In the Appendix, we provide detailed definitions for each anomaly type, along with example time series and real-world scenario descriptions to illustrate these categories in Tables 6 and 7.

2.3 Fine-tuning and Evaluation

Empowering LLMs. Traditional anomaly detection methods, typically rooted in machine learning or deep learning, often face challenges in accurately classifying anomalies due to the scarcity of trustworthy labels and in providing meaningful causal analysis. These limitations hinder their practical application in complex, real-world sce-

narios. More recently, the advent of generative pre-trained models such as LLMs has opened new avenues for time series anomaly detection (Zhou and Yu, 2024; Xu et al., 2025). Yet many of these initial explorations primarily focus on demonstrating LLMs' inherent capabilities without fully leveraging real-world, domain-specific data for enhanced learning. In TIME-RA task, we address these limitations by fine-tuning LLMs with our unique RATS40K dataset. This approach empowers the LLMs with the ability to analyze time series anomalies more effectively and to generalize their understanding across diverse domains.

Mathematically, given a (vision) language model π and after formatting an input $x = \{T, D, V\}$, we perform supervised fine-tuning as follows:

$$\max_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{RATS40K}}}[\log P_\theta(y|x)].$$

Here, θ represents the parameters of the language model π, (x, y) are input-output pairs from the RATS40K, and $y = \{y_l, a, r\}$. This fine-tuning process allows the LLM to learn the intricate patterns and contextual information present in our real-world dataset, thereby improving its ability to identify and interpret anomalies.

To thoroughly assess the performance of our fine-tuned TIME-RA model, we adopt a multifaceted evaluation approach, considering three distinct aspects of anomaly detection: binary anomaly classification, multi-class anomaly type identification, and textual reasoning and explanation generation.

Evaluation. To thoroughly assess the performance of our fine-tuned models, we consider three types of evaluation separately: numerical binary label y_l for determining whether a point is anomalous, numerical multi-class action a for identifying the type of anomaly, and textual reasoning r for explanation. For the first two quantitative evaluations, we treat them as classification problems and employ standard metrics such as precision (*P*), recall (*R*), and F1-score (*F1*) for evaluation. For the evaluation of reasoning, we employ text similarity-based metrics, including *Cosine* Similarity based on ELECTRA (Clark et al., 2020), *TF-IDF* Similarity, *Levenshtein* Similarity, *Token* Sequence Similarity, and Reasoning Consistency Score (*RCS*). For RCS, we employ `GPT-4` as an evaluator to score the logical integrity of the generated reasoning on a scale of 1 to 5, where a higher score indicates superior logical consistency and fewer hallucinations.

3 A New Dataset for TIME-RA Task

To complete the TIME-RA task, we establish a first anomaly reasoning dataset RATS40K, including scalability and diversity of anomalies for model generalization. In this section, we integrate these objectives into all three phases: data collection, generating completions, and preference annotation inspired by the data engineering principles (Cui et al., 2024; Chiang et al., 2023). As illustrated in Figure 3 and compared with existing datasets in Table 1, RATS40K emphasizes abnormal sample collection, enabling language models to capture heterogeneous anomaly types and their underlying causes. The details are provided in Appendix I.

3.1 Dataset Collection

We first collect multimodal time series anomaly data, incorporating numeric time series, descriptive text, and visualization images. The description of data sources is deferred to the Table 13. Numerous time series segments of different lengths were extracted from these sources and labeled for anomalies. We then attach a brief textual description to each sample, providing its application context and explaining the meaning of each feature in the multivariate series. Furthermore, each time series segment is rendered as a visual chart to facilitate comparative analysis of temporal patterns across variables. Finally, we obtain nearly 40K segments from various modalities across multiple fields, aiming to replicate how human analysts combine data sources for interpretation.

3.2 Reason Completion Sampling

To generate initial reasoning labels for our dataset, we design a structured prompting strategy that guides an LLM through a step-by-step anomaly analysis. The prompt adopts a role-based instruction, explicitly casting the model as an "expert in time series anomaly detection". It formalizes the task into three stages, **Observation**, **Thought**, and **Action**, mimicking the diagnostic reasoning process of human analysts. In the {Observation} stage, the prompt presents the time series data along with its domain knowledge. The model is then asked to articulate its {Thought}: a detailed reasoning process examining the time series behavior, relationships among variables, and any deviations from normal patterns. Finally, the model must output an {Action}, which in this context is the decided anomaly category from our predefined list. The

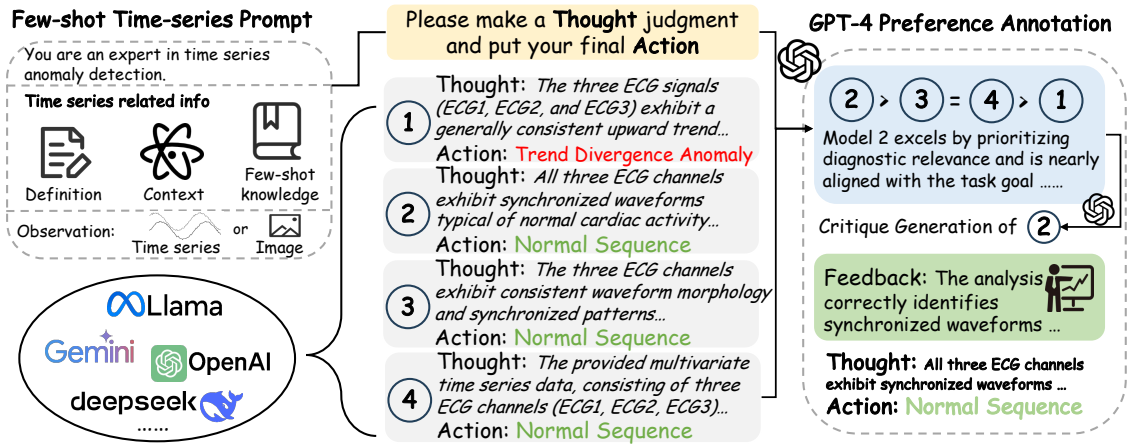


Figure 3: RATs40K dataset construction pipeline. We ensure diversity by sampling anomalous thoughts and decision actions from a large model pool, then we query `GPT-4` with detailed definitions and prompts for preference selection and correction to generate high-quality, fine-grained annotations.

prompt includes the full list of anomaly category options with brief natural-language definitions, also a few exemplary question-answer cases serving as in-context demonstrations. The resultant structured prompt for univariate/multivariate anomaly detection is formalized in Appendix H.

Using the above prompt setup, we leverage LLMs to automatically annotate each time series segment with a reasoning explanation and an anomaly category. More specifically, we built a model pool that consists of 4 powerful models in current arena leaderboards: `GPT-4o` (Achiam et al., 2023), `Gemini-2.5` (Team et al., 2023), `DeepSeek-R1` (Guo et al., 2025), and `Llama-3.3-70B-Instruct` (Grattafiori et al., 2024). Then, the models are given the Observation and domain context, and they produce a Thought r and an Action a of the time series segment, where statistics of responses are in the Appendix E. This yielded a preliminary set of anomaly annotations with human-readable explanations. In the subsequent steps, we further refine these LLM-generated annotations through an iterative feedback process to ensure the final labels are of high quality.

3.3 AI Feedback Annotation

After generating 158,340 model completions from 36,311 univariate and 3,274 multivariate instructions, we use `GPT-4` to provide two kinds of feedback for each: quantitative rankings that pinpoint a model's placement in the model pool, and textual critiques that provide detailed suggestions for the thought of anomaly detection. We leverage `GPT-4` for feedback generation due to the limited scalability of human evaluation and the potential subjectivity introduced by human annotators. In total, this resulted in more than 150,000 feedback data points. To ensure the quality of the model responses in the pool, we identify preferences and conduct a critique, selecting the top-ranked results as the final dataset. Furthermore, to validate the reliability of the LLM-generated feedback, we conduct expert evaluations on a subset of the results in the Section 4.2, which confirmed the alignment between `GPT-4` assessments and human judgments. We put the `GPT-4` annotation and critique instruction in Appendix H.

Preference Annotation. To enhance the reliability of GPT-4's annotations and minimize subjectivity and randomness, we implement three key techniques (Cui et al., 2024): (i) Reference: For each type of time series anomaly, we provide expert classification and corresponding reasons in the instruction as the model's few-shot, which helps reduce randomness. (ii) Standardization: For each of these aspects, we gave `GPT-4` a clear Likert scale with scores from 1 to 5 as a reference. (iii) Reasoning: In addition to ranking the models, `GPT-4` also provides a reason to explain how to rate each model based on abnormal time series. Combining all techniques, we ultimately obtain scalar scores and reasons for each model response. Then, we also use `GPT-4` for preference ranking in the model pool and select the anomaly detection answer with the best score as a label for each sample.

Critique Generation. To complement our scalar scores, we leverage `GPT-4` for textual critiques. Our goal is to enable the LLM to serve as a mentor, providing specific and detailed advice for the top-ranked model in each ranking to guide its improvement and placement in the final dataset, rather than

Table 1: Statistics of existing time series anomaly datasets. "-" denotes that the corresponding data does not exist or missing classification and reasoning labels.

Dataset	# Sample	Modalities	# Domain	Is Real-World?	Time series Dimension	Anomaly Ratio	# Anomaly Category	Thought Length	Annotation
AnomLLM (Zhou and Yu, 2024)	3,200	Time+Text	-	✗	U	64.5%	8	-	Synthetic
LLMAD (Liu et al., 2024b)	37,000	Time+Text+Image	3	✓	U	0.71%- 2.35%	8	-	100 by Human
VisualTimeAnomaly (Xu et al., 2025)	12,400	Time+Text+Image	-	✗	U & M	-	9	-	Synthetic
RATs40K (Ours)	39,574	Time+Text+Image	10	✓	U & M	83.7%	14 + 6	101.378	AI feedback

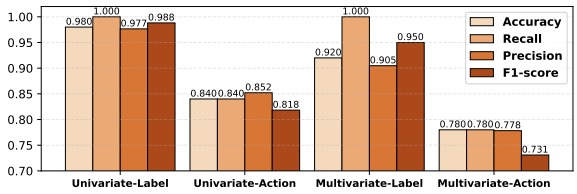
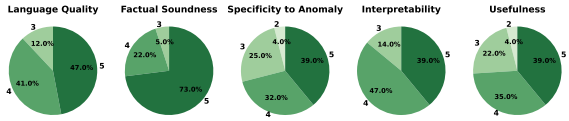
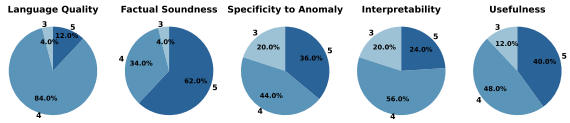


Figure 4: Comparison of LLM-generated labels with expert annotations.



(a) Univariate quality for the textual reasoning.



(b) Multivariate quality for the textual reasoning.

Figure 5: Textual reasoning quality based on Likert scale (Based on 100 uni- and 50 multivariate samples).

simply providing direct preference rankings. These prompts are unique for each anomaly time series and are generated from the overall perspective of the model pool, thus serving as the optimal reasoning for anomalies. After generating feedback through `GPT-4`, we merge the feedback results with the top-ranked results in the model pool, and place them in the final RATs40K dataset to improve the label quality of anomaly reason.

4 Experiments

In this section, we first conduct an in-depth quality analysis of RATs40K. Then, we present the performances of LLMs/MLLMs on it and provide insights to guide future work on the TIME-RA task.

4.1 Settings

We evaluate some widely used open-source LLMs/MLLMs under both zero-shot and supervised fine-tuning (SFT) settings. The detailed list of models with their links can be found in Appendix F and Table 10. For the SFT, each sample is formatted based on a fixed instruction template, and we use LoRA (Hu et al., 2022) for parameter-efficient fine-tuning. To evaluate the model outputs, we design regular expressions to automatically extract the predicted Thought and Action, which will then be compared against the ground truth. Evaluation metrics follow the definition of the TIME-RA task, with the best results in **bold** and second-best underline.

4.2 Reliability of the RATs40K Dataset

Following LLMAD (Liu et al., 2024b), we assess label quality by comparing 7 human experts' average annotations (both binary and action classification) with LLM-based generated ranking results as shown in Figure 4. For both univariate and multivariate samples, LLM-based labels show high consistency with expert annotations. Binary classification results demonstrate strong agreement with high accuracy and F1-scores, while action classification get slightly lower scores, reflecting the increased complexity and uncertainty of fine-grained anomaly categorization. As for the quality of the textual reasoning, we evaluate five key dimensions inspired by human explanation quality criteria based on a well-designed Likert scale. A more detailed description can be found in Appendix G. For the univariate results from Figure 5a, the average scores across these dimensions range from 4.04 to 4.58 (out of 5), indicating consistent clarity, factual alignment, and actionable insights. The multivariate subset shows similarly high quality, with average scores from 4.08 to 4.28, as shown in Figure 5b. These results further demonstrate the high quality of our annotated data.

4.3 Results and Discussion

For the results and evaluation, we focus on four research questions for the TIME-RA task as follows:

RQ 1: Can LLMs adapt to the number of variates? Based on the prompt design and input architecture discussed previously, the results show in Tables 2, 3, 4, and 5 indicate that the LLMs can effectively adapt to the varying number of variates. Specifically, models such as `Qwen2.5` show comparable or better F1-scores on multivariate tasks (~0.85) compared to univariate tasks (~0.9). Besides, LLM can adapt to different numbers of variables (from 2 to 9) in Tables 2 and 3 and even the

Table 2: Performance on time-series-only anomalies based on LLM. Pure color rows are the machine learning or LLM result, while gray rows represent the performance after fine-tuning. F1 means weighted-F1 score.

		Label Matching				ActionID Matching				Thought Matching				
	Model	P	R	F1	Rank	P	R	F1	Rank	Cosine	TFIDF	Lev.	Token	RCS
Univariate Time Series Anomaly Detection	DeepSeek-7B	0.8175	0.3295	0.4697	10th	0.1190	0.1407	0.0703	10th	0.8780	0.2026	0.1731	0.1001	2.1746
	DeepSeek-7B	0.8598	0.2729	0.4143	12th	0.1532	0.1613	0.0771	9th	0.8971	0.2298	0.1773	0.1039	2.3524
	Llama-3-8B	0.9098	0.3148	0.4678	11th	0.1767	0.1754	0.0897	8th	0.9205	0.3154	0.2569	0.1533	2.2330
	Llama-3-8B	0.8576	0.7605	0.8061	8th	0.3887	0.1650	0.1511	**1st**	0.9242	0.3065	0.2484	0.1511	2.6780
	Llama-3.2-3B	0.8510	0.3946	0.5391	9th	0.2353	0.1518	0.0916	7th	0.9127	0.2752	0.2295	0.1467	2.1363
	Llama-3.2-3B	0.8458	0.8447	0.8453	5th	0.1846	0.1038	0.0986	6th	0.9264	0.3175	0.2524	0.1613	2.3111
	Phi-4-mini	0.8631	0.8062	0.8337	7th	0.2141	0.1428	0.1309	3rd	0.9166	0.3208	0.2187	0.1525	3.2840
	Phi-4-mini	0.8582	0.8158	0.8364	6th	0.2035	0.1482	0.1397	2nd	0.9168	0.3154	0.2163	0.1518	3.2633
	Qwen2.5-3B	0.8296	0.9834	0.9000	**1st**	0.2016	0.0605	0.0301	12th	0.9108	0.2841	0.2129	0.1306	2.0446
	Qwen2.5-3B	0.8299	0.9768	0.8974	2nd	0.1592	0.0623	0.0371	11th	0.9118	0.2849	0.2131	0.1316	2.3682
	Qwen2.5-7B	0.8465	0.9352	0.8886	3rd	0.2436	0.1112	0.1064	5th	0.9253	0.3059	0.2435	0.1619	3.2422
	Qwen2.5-7B	0.8452	0.9295	0.8854	4th	0.2532	0.1229	0.1100	4th	0.9270	0.3160	0.2432	0.1624	3.2120
Multivariate Time Series Anomaly Detection	DeepSeek-7B	0.6980	0.4419	0.5411	8th	0.0626	0.1224	0.0821	12th	0.9344	0.3420	0.2305	0.1190	2.2534
	DeepSeek-7B	0.7112	0.4264	0.5331	9th	0.0658	0.1344	0.0874	11th	0.9344	0.3358	0.2267	0.1207	2.2811
	Llama-3-8B	0.7736	0.3067	0.4393	12th	0.4491	0.2114	0.1303	9th	0.9398	0.3869	0.2604	0.1481	2.1926
	Llama-3-8B	0.8176	0.3668	0.5064	11th	0.5934	0.1803	0.1151	10th	0.9455	0.3887	0.2625	0.1459	2.2610
	Llama-3.2-3B	0.8345	0.4515	0.5860	7th	0.5008	0.2842	0.2583	7th	0.9580	0.4479	0.2805	0.1698	2.8624
	Llama-3.2-3B	0.8232	0.4963	0.6193	6th	0.4838	0.3016	0.2893	6th	0.9578	0.4481	0.2788	0.1682	2.8934
	Phi-4-mini	0.7676	0.8237	0.7947	3rd	0.4834	0.4056	0.4339	2nd	0.9385	0.3842	0.2433	0.1319	2.4657
	Phi-4-mini	0.7622	0.8195	0.7899	4th	0.4627	0.4230	0.4372	**1st**	0.9376	0.3776	0.2399	0.1313	2.5389
	Qwen2.5-3B	0.6339	0.4551	0.5299	10th	0.2397	0.2180	0.2209	8th	0.8822	0.2901	0.1808	0.0892	2.3759
	Qwen2.5-3B	0.6545	0.6039	0.6281	5th	0.3563	0.2876	0.2933	5th	0.8762	0.2893	0.1738	0.0937	2.6421
	Qwen2.5-7B	0.7567	0.9801	0.8541	2nd	0.5256	0.4084	0.4127	4th	0.9532	0.3792	0.2676	0.1499	2.7454
	Qwen2.5-7B	0.7557	0.9826	0.8544	**1st**	0.5675	0.4322	0.4300	3rd	0.9554	0.3805	0.2689	0.1504	2.7033

Table 3: Performance on visualized time series anomalies based on MLLMs. Pure color rows are the model's observations solely through images, while gray rows are observations that include both images and time series.

		Label Matching				ActionID Matching				Thought Matching				
	Model	P	R	F1	Rank	P	R	F1	Rank	Cosine	TFIDF	Lev.	Token	RCS
Univariate Time Series Anomaly Detection	Llava-v1.5-7B	0.8272	0.9489	0.8839	2nd	0.1722	0.0358	0.0231	8th	0.9274	0.3244	0.2299	0.1450	1.9572
	Llava-v1.5-7B	0.8261	0.9668	0.8909	**1st**	0.1933	0.0579	0.0580	7th	0.9195	0.2876	0.2197	0.1242	2.7730
	Llava-v1.5-13B	0.8268	0.7601	0.7921	7th	0.1717	0.1931	0.1693	**1st**	0.9083	0.2995	0.2386	0.1143	2.1874
	Llava-v1.5-13B	0.8216	0.9165	0.8665	4th	0.1781	0.1572	0.1128	2nd	0.9355	0.3436	0.2725	0.1301	2.3602
	Llama-3.2-11B-v	0.8322	0.8050	0.8183	6th	0.1866	0.1030	0.0863	4th	0.8921	0.2483	0.1586	0.0956	2.1240
	Llama-3.2-11B-v	0.8216	0.9165	0.8665	4th	0.1725	0.0944	0.0856	5th	0.9155	0.2834	0.1954	0.1072	2.7210
	Qwen2.5-vl-7B	0.8281	0.3977	0.5373	8th	0.0413	0.1566	0.0651	6th	0.8998	0.2088	0.1698	0.1177	1.9824
	Qwen2.5-vl-7B	0.8426	0.8973	0.8691	3rd	0.1171	0.1128	0.0930	3rd	0.9011	0.2685	0.1873	0.1297	2.7376
Multivariate Time Series Anomaly Detection	Llava-v1.5-7B	0.9174	0.2532	0.3968	8th	0.5970	0.2618	0.1569	4th	0.9130	0.3755	0.2642	0.1125	2.1634
	Llava-v1.5-7B	0.7340	0.4340	0.5455	7th	0.4715	0.1238	0.1150	6th	0.9359	0.3959	0.2529	0.1266	2.2750
	Llava-v1.5-13B	0.7287	0.6919	0.7098	5th	0.3716	0.4216	0.3944	**1st**	0.9451	0.3628	0.2525	0.1160	2.3619
	Llava-v1.5-13B	0.8058	0.8682	0.8358	2nd	0.6308	0.2036	0.2880	3rd	0.9449	0.3869	0.2479	0.1267	2.3647
	Llama-3.2-11B-v	0.8012	0.4164	0.5480	6th	0.2402	0.1986	0.1300	5th	0.9254	0.3088	0.2200	0.1108	2.1667
	Llama-3.2-11B-v	0.8571	0.7500	0.8000	3rd	0.4091	0.2727	0.2922	2nd	0.9350	0.3429	0.2219	0.1129	3.3636
	Qwen2.5-vl-7B	0.7381	1.0000	0.8493	**1st**	0.0039	0.0568	0.0073	8th	0.9307	0.3079	0.2348	0.1296	1.5842
	Qwen2.5-vl-7B	0.7627	0.7816	0.7721	4th	0.3728	0.1245	0.1077	7th	0.9526	0.3994	0.2651	0.1419	2.5886

17 and 19 dimensions in Table 4. Overall, these quantitative insights suggest that an appropriate prompt engineering design may help LLMs better leverage the additional complexity inherent in multivariate time series. It indicates promising adaptability across varying numbers of variables.

RQ 2: Can SFT enhance LLMs' performance on TIME-RA task using RATS40K? Analysis of Tables 2 and 3 indicates that SFT generally improves performance, with most models achieving higher F1-scores and better semantic alignment in Thought Matching after tuning. However, these gains are not universal. In some complex multivariate scenarios, performance remains stagnant or even slightly regresses. This inconsistency underscores the inherent difficulty of the TIME-RA task, particularly when dealing with the noise and complexity of real-world data. These results suggest that while SFT is a beneficial baseline, it may not be sufficient to fully resolve the intricacies of the task. The gap between LLMs and specialized supervised models highlights the need for more sophisticated and specialized adaptation methods to achieve robust reliability in practical applications.

RQ 3: Can visual representation enhance the TIME-RA task? Table 3 demonstrates that visual representation is a powerful catalyst for the TIME-RA task, particularly in enhancing the depth and coherence of model reasoning. The obvious gains are observed in Thought Matching metrics, where visual plots consistently help models align their internal logic with task-specific actions, achieving higher semantic scores. While the impact on classification varies by model, the integration of visual

Table 4: Performances of the recent and RATS40K fine-tuned LLM based on other real-world datasets (in-domain and out of domain) for the anomaly detection task. GHL has 19 channels. CATSv2 has 17 channels.

Dataset	Model	Univariate TS			Dataset	Multivariate TS		
		P	R	F1		P	R	F1
NEK (In)	KNN	0.9167	0.1571	0.2683	GHL (In)	0.5625	0.1023	0.1731
	LOF	0.8333	0.0714	0.1316		0.4000	0.0682	0.1165
	AE1SVM	0.3333	0.0571	0.0976		0.5000	0.0795	0.1373
	XGBoost	0.9429	0.9429	0.9429		0.9053	0.9773	0.9399
	LSTM	0.6796	1.0000	0.8092		0.5252	0.8295	0.6432
	TimesNet	0.7130	0.8034	0.7556		0.4843	0.7998	0.6034
	AT	0.7531	0.8998	0.8200		0.5000	0.8129	0.6193
	Timer	0.7370	0.7175	0.7272		0.5276	0.5998	0.5615
	Chronos	0.7579	0.7680	0.7630		0.5778	0.6921	0.6299
	Qwen2.5-7B	0.5397	0.9855	0.6974		0.8923	0.6591	0.7582
SED (Out)	KNN	1.0000	0.1569	0.2712	CATSv2 (Out)	0.7333	0.1000	0.1760
	LOF	0.9583	0.2255	0.3651		0.5600	0.1273	0.2074
	AE1SVM	0.9444	0.1667	0.2833		0.6364	0.1273	0.2121
	XGBoost	0.9806	0.9902	0.9854		0.7500	0.8182	0.7826
	LSTM	0.9886	0.8529	0.9158		0.4864	0.9727	0.6485
	TimesNet	0.8129	0.8541	0.8331		0.4680	0.7137	0.5654
	AT	0.8234	0.7943	0.8087		0.5014	0.7083	0.5873
	Timer	0.7870	0.7596	0.7732		0.4984	0.7470	0.5980
	Chronos	0.8038	0.7981	0.8011		0.5032	0.7331	0.5969
	Qwen2.5-7B	0.5714	1.0000	0.7273		0.4977	0.9818	0.6606

Table 5: Performances of the recent algorithms and RATS40K fine-tuned LLM for the anomaly detection.

Training Type	Model	Univariate TS			Multivariate TS		
		P	R	F1	P	R	F1
Unsupervised Learning	KNN	0.9040	0.1129	0.2008	0.9019	0.1141	0.2026
	LOF	0.9509	0.1161	0.2070	0.9069	0.0967	0.1748
	AE1SVM	0.9160	0.1177	0.2087	0.8627	0.1091	0.1938
	DeepSVDD	0.9165	0.1141	0.2030	0.9298	0.1315	0.2304
	Qwen2.5-7B	0.8465	0.9352	0.8886	0.7567	0.9801	0.8541
Supervised Learning	XGBoost	0.9074	0.9682	0.9368	0.9162	0.9776	0.9459
	LightGBM	0.9023	0.9758	0.9376	0.9316	0.9801	0.9552
	LSTM	0.8300	0.9976	0.9061	0.7840	1.0000	0.8790
	TimesNet	0.8521	0.9683	0.9066	0.8234	0.9890	0.8987
	AT	0.8781	0.9702	0.9219	0.8823	0.9918	0.9339
	Qwen2.5-7B	0.8452	0.9295	0.8854	0.7557	0.9826	0.8544
Pre-trained Model	TimesFM	0.7872	0.6258	0.6974	0.8435	0.7432	0.7903
	Timer	0.8135	0.7295	0.7693	0.8257	0.7324	0.7764
	Chronos	0.7976	0.7814	0.7895	0.8814	0.7945	0.8358
	MOMENT	0.8005	0.6834	0.7374	0.8471	0.7775	0.8109

and raw data allows models like `Llama-3.2-11B-v` to reach peak performance. These results suggest that visualization effectively bridges the gap between raw data and high-level reasoning. Although the complexity of multimodal fusion remains a challenge, the clear improvement in reasoning quality confirms that visual representation is instrumental in guiding MLLMs toward more accurate and interpretable anomaly diagnostics.

RQ 4: Can fine-tuned LLMs ready for plug-and-play TIME-RA task? We evaluate the fine-tuned `Qwen2.5-7B` models on real-world datasets from other domains and compare the results with those of common TSAD models. Tables 4 and 5 demonstrate that the fine-tuned `Qwen2.5-7B` possesses strong cross-domain transferability, effectively serving as a "plug-and-play" solution for the TIME-RA task. When applied directly to new datasets without additional tuning, `Qwen2.5-7B` consistently outperforms classic unsupervised methods like KNN, LOF, and AE1SVM[2], and remains competitive with recent state-of-the-art deep learning and time series foundation models such as TimesNet, Anomaly Transformer (AT), Timer, and Chronos[3]. Notably, the model achieves exceptionally high recall across both in-domain (NEK, GHL) and out-of-domain (SED, CATSv2) datasets (Liu and Paparrizos, 2024), indicating a robust capability to detect anomalies in diverse real-world environments. While specialized supervised models like XGBoost[2] may maintain a lead in precision as shown in Table 5, `Qwen2.5-7B` bridges the gap between traditional zero-shot approaches and supervised benchmarks. These results conclude that fine-tuned LLMs are ready for practical deployment in scenarios where labeled data is scarce, providing reliable diagnostic performance across different domains without the need for task-specific retraining.

Further Analysis. We provide qualitative case studies across diverse domains in Appendix D, demonstrating the model's diagnostic logic for both univariate and multivariate anomalies. We also conduct a failure case analysis in Appendix K, identifying limitations in detecting subtle drifts or overlapping patterns. Furthermore, ablation studies in Appendix J show that prompt engineering, specifically few-shot examples and Chain-of-Thought reasoning, enhances reasoning coherence and categorization accuracy.

5 Related Work

Due to space limitation, we put the related work in Appendix A.

6 Conclusion

In this paper, we introduced Time-series Reasoning for Anomaly, TIME-RA, a new reasoning-focused anomaly detection task addressing critical gaps in traditional approaches by integrating detection, fine-grained classification, and causal explanation. To support this task, we constructed RATS40K, the first real-world multimodal dataset with detailed high-quality annotations generated through structured prompting and `GPT-4` refinement. Extensive evaluations demonstrate that structured fine-tuning enhances anomaly detection and interpretability. Our work establishes a foundation for future research, underscoring the potential of generative, reasoning-enhanced anomaly detection models in real-world applications.

[2]Evaluate by PyOD.

[3]Evaluate by Time-Series-Library.

Limitations

While TIME-RA and the RATS40K dataset facilitate progress in anomaly diagnosis, several limitations persist. Currently, the framework primarily identifies a single dominant anomaly category per sequence and may overlook secondary, co-occurring patterns. Furthermore, processing extremely long time-series sequences remains challenging due to LLM token constraints, and while visual representations provide a bridge, they do not fully replace the need for specialized long-context architectural adaptations. Finally, although fine-tuned models exhibit strong transferability, maintaining high diagnostic precision across vastly different domains without task-specific retraining remains a frontier for future exploration. Future work may extend TIME-RA to multi-label or hierarchical time series anomaly diagnosis.

Ethical considerations

The construction of RATS40K adheres to ethical research standards, utilizing publicly available, open-source repositories where sensitive data (e.g., healthcare ECG) has been de-identified at the source. This study involves human subjects as annotators. We have provided all participants with clear instructions, the details of which are documented in the Section 4.2 and Appendix G (recruitment, compensation, and data consent protocols were not applicable for this study). We emphasize that the diagnostic reasoning generated by our system is designed to serve as a decision-support tool for human experts, not as a replacement for professional judgment in high-stakes environments like medical diagnosis or industrial safety. Users should be mindful of potential model hallucinations in complex scenarios and are encouraged to deploy these tools responsibly to ensure human-in-the-loop verification. Overall, our approach does not involve any personally identifiable information (PII) or sensitive data, and we adhere to responsible AI practices by following the ACL Code of Ethics.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Subutai Ahmad, Alexander Lavin, Scott Purdy, and Zuha Agha. 2017. Unsupervised real-time anomaly detection for streaming data. *Neurocomputing*, 262:134–147.

Marc Bachlin, Meir Plotnik, Daniel Roggen, Inbal Maidan, Jeffrey M Hausdorff, Nir Giladi, and Gerhard Troster. 2009. Wearable assistant for parkinson's disease patients with the freezing of gait symptom. *IEEE Transactions on Information Technology in Biomedicine*, 14(2):436–446.

Alexander Bakhtin, Jesse Nyyssölä, Yuqing Wang, Noman Ahmad, Ke Ping, Matteo Esposito, Mika Mäntylä, and Davide Taibi. 2025. Lo2: Microservice api anomaly dataset of logs and metrics. *arXiv preprint arXiv:2504.12067*.

Guillermo Barrenetxea. 2019. Sensorscope data. Data set.

Ane Blázquez-García, Angel Conde, Usue Mori, and Jose A Lozano. 2021. A review on outlier/anomaly detection in time series data. *ACM computing surveys (CSUR)*, 54(3):1–33.

Paul Boniol, Qinghua Liu, Mingyi Huang, Themis Palpanas, and John Paparrizos. 2024. Dive into time-series anomaly detection: A decade review. *arXiv preprint arXiv:2412.20512*.

George EP Box and David A Pierce. 1970. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American statistical Association*, 65(332):1509–1526.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58.

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

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, and 1 others. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, march 2023. *URL https://lmsys. org/blog/2023-03-30-vicuna*, 3(5).

Nwodo Benita Chikodili, Mohammed D Abdulmalik, Opeyemi A Abisoye, and Sulaimon A Bashir. 2020. Outlier detection in multivariate time series data using a fusion of k-medoid, standardized euclidean distance and z-score. In *International Conference on Information and Communication Technology and Applications*, pages 259–271. Springer.

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

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Andrew A Cook, Göksel Mısırlı, and Zhong Fan. 2019. Anomaly detection for iot time-series data: A survey. *IEEE Internet of Things Journal*, 7(7):6481–6494.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2024. ULTRAFEEDBACK: Boosting language models with scaled AI feedback. In *ICML*, pages 9722–9744.

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

Jingkun Gao, Xiaomin Song, Qingsong Wen, Pichao Wang, Liang Sun, and Huan Xu. 2020. Robusttad: Robust time series anomaly detection via decomposition and convolutional neural networks. *arXiv preprint arXiv:2002.09545*.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. 2000. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. 2024. Moment: A family of open time-series foundation models. In *International Conference on Machine Learning*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Scott David Greenwald, Ramesh S Patil, and Roger G Mark. 1990. *Improved detection and classification of arrhythmias in noise-corrupted electrocardiograms using contextual information*. IEEE.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

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

Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Soderstrom. 2018. Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 387–395.

Vincent Jacob, Fei Song, Arnaud Stiegler, Bijan Rad, Yanlei Diao, and Nesime Tatbul. 2020. Exathlon: A benchmark for explainable anomaly detection over time series. *arXiv preprint arXiv:2010.05073*.

Ming Jin, Yifan Zhang, Wei Chen, Kexin Zhang, Yuxuan Liang, Bin Yang, Jindong Wang, Shirui Pan, and Qingsong Wen. 2024. Position: What can large language models tell us about time series analysis. In *Forty-first International Conference on Machine Learning*.

Neha Kant and Manish Mahajan. 2019. Time-series outlier detection using enhanced k-means in combination with pso algorithm. In *Engineering Vibration, Communication and Information Processing: ICoEVCI 2018, India*, pages 363–373. Springer.

Paweł Karczmarek, Adam Kiersztyn, Witold Pedrycz, and Ebru Al. 2020. K-means-based isolation forest. *Knowledge-based systems*, 195:105659.

Maryam Mahsal Khan and Mohammed Alkhathami. 2024. Anomaly detection in iot-based healthcare: machine learning for enhanced security. *Scientific reports*, 14(1):5872.

Yaxuan Kong, Yiyuan Yang, Yoontae Hwang, Wenjie Du, Stefan Zohren, Zhangyang Wang, Ming Jin, and Qingsong Wen. 2025a. Time-mqa: Time series multitask question answering with context enhancement. *arXiv preprint arXiv:2503.01875*.

Yaxuan Kong, Yiyuan Yang, Shiyu Wang, Chenghao Liu, Yuxuan Liang, Ming Jin, Stefan Zohren, Dan Pei, Yan Liu, and Qingsong Wen. 2025b. Position: Empowering time series reasoning with multimodal llms. *arXiv preprint arXiv:2502.01477*.

Kwei-Herng Lai, Daochen Zha, Junjie Xu, Yue Zhao, Guanchu Wang, and Xia Hu. 2021. Revisiting time series outlier detection: Definitions and benchmarks. In *Thirty-fifth conference on neural information processing systems datasets and benchmarks track (round 1)*.

N Laptev, S Amizadeh, and Y Billawala. 2015. S5-a labeled anomaly detection dataset, version 1.0 (16m).

Yuxuan Liang, Haomin Wen, Yuqi Nie, Yushan Jiang, Ming Jin, Dongjin Song, Shirui Pan, and Qingsong Wen. 2024. Foundation models for time series analysis: A tutorial and survey. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 6555–6565.

Chen Liu, Shibo He, Shizhong Li, Zhenyu Shi, and Wenchao Meng. 2025a. Detecting both seen and unseen anomalies in time series. *ACM Transactions on Knowledge Discovery from Data*, 19(4):1–29.

Chen Liu, Shibo He, Qihang Zhou, Shizhong Li, and Wenchao Meng. 2024a. Large language model guided knowledge distillation for time series anomaly detection. *arXiv preprint arXiv:2401.15123*.

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

Jun Liu, Chaoyun Zhang, Jiaxu Qian, Minghua Ma, Si Qin, Chetan Bansal, Qingwei Lin, Saravan Rajmohan, and Dongmei Zhang. 2024b. Large language models can deliver accurate and interpretable time series anomaly detection. *arXiv preprint arXiv:2405.15370*.

Qinghua Liu and John Paparrizos. 2024. The elephant in the room: Towards a reliable time-series anomaly detection benchmark. *Advances in Neural Information Processing Systems*, 37:108231–108261.

Zichuan Liu, Tianchun Wang, Jimeng Shi, Xu Zheng, Zhuomin Chen, Lei Song, Wenqian Dong, Jayantha Obeysekera, Farhad Shirani, and Dongsheng Luo. 2024c. Timex++ learning time-series explanations with information bottleneck. In *ICML*, pages 32062–32082.

Zichuan Liu, Yingying ZHANG, Tianchun Wang, Zefan Wang, Dongsheng Luo, Mengnan Du, Min Wu, Yi Wang, Chunlin Chen, Lunting Fan, and 1 others. Explaining time series via contrastive and locally sparse perturbations. In *ICLR*.

Daehyung Park, Yuuna Hoshi, and Charles C Kemp. 2018. A multimodal anomaly detector for robot-assisted feeding using an lstm-based variational autoencoder. *IEEE Robotics and Automation Letters*, 3(3):1544–1551.

Peter CB Phillips and Sainan Jin. 2021. Business cycles, trend elimination, and the hp filter. *International Economic Review*, 62(2):469–520.

Hansheng Ren, Bixiong Xu, Yujing Wang, Chao Yi, Congrui Huang, Xiaoyu Kou, Tony Xing, Mao Yang, Jie Tong, and Qi Zhang. 2019. Time-series anomaly detection service at microsoft. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3009–3017.

Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. 2018. Deep one-class classification. In *International conference on machine learning*, pages 4393–4402. PMLR.

Mayu Sakurada and Takehisa Yairi. 2014. Anomaly detection using autoencoders with nonlinear dimensionality reduction. In *Proceedings of the MLSDA 2014 2nd workshop on machine learning for sensory data analysis*, pages 4–11.

Sebastian Schmidl, Phillip Wenig, and Thorsten Papenbrock. 2022. Anomaly detection in time series: a comprehensive evaluation. *Proceedings of the VLDB Endowment*, 15(9):1779–1797.

Youjin Shin, Sangyup Lee, Shahroz Tariq, Myeong Shin Lee, Okchul Jung, Daewon Chung, and Simon S Woo. 2020. Itad: integrative tensor-based anomaly detection system for reducing false positives of satellite systems. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 2733–2740.

Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, and Dan Pei. 2019. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2828–2837.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Markus Thill, Wolfgang Konen, and Thomas Bäck. 2020. Markusthill/mgab: the mackey-glass anomaly benchmark. *Version v1. 0.1. Zenodo. doi*, 10.

Luan Tran, Liyue Fan, and Cyrus Shahabi. 2016. Distance-based outlier detection in data streams. *Proceedings of the VLDB Endowment*, 9(12):1089–1100.

Hubert Truchan and Zahra Ahmadi. 2025. Nonastreda: Multimodal Dataset for Identifying Tool Wear Condition. https://doi.org/10.17632/m892d2wtzh.1.

Bingxing Wang, Yuxin Dong, Jianhua Yao, Honglin Qin, and Jiajing Wang. 2024. Exploring anomaly detection and risk assessment in financial markets using deep neural networks. *International Journal of Innovative Research in Computer Science and Technology*, 12(4).

Chengsen Wang, Qi Qi, Jingyu Wang, Haifeng Sun, Zirui Zhuang, Jinming Wu, Lei Zhang, and Jianxin Liao. 2025. Chattime: A unified multimodal time series foundation model bridging numerical and textual data. In *AAAI Conference on Artificial Intelligence*.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. 2023. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *International Conference on Learning Representations*.

Renjie Wu and Eamonn J Keogh. 2021. Current time series anomaly detection benchmarks are flawed and are creating the illusion of progress. *IEEE transactions on knowledge and data engineering*, 35(3):2421–2429.

Jiehui Xu, Haixu Wu, Jianmin Wang, and Mingsheng Long. 2021. Anomaly transformer: Time series anomaly detection with association discrepancy. *arXiv preprint arXiv:2110.02642*.

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

Kun Yang, Samory Kpotufe, and Nick Feamster. 2021. An efficient one-class svm for anomaly detection in the internet of things. *arXiv preprint arXiv:2104.11146*.

Yiyuan Yang, Chaoli Zhang, Tian Zhou, Qingsong Wen, and Liang Sun. 2023. Dcdetector: Dual attention contrastive representation learning for time series anomaly detection. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3033–3045.

Zahra Zamanzadeh Darban, Geoffrey I Webb, Shirui Pan, Charu Aggarwal, and Mahsa Salehi. 2024. Deep learning for time series anomaly detection: A survey. *ACM Computing Surveys*, 57(1):1–42.

Chaoli Zhang, Tian Zhou, Qingsong Wen, and Liang Sun. 2022. Tfad: A decomposition time series anomaly detection architecture with time-frequency analysis. In *Proceedings of the 31st ACM international conference on information & knowledge management*, pages 2497–2507.

Zhijie Zhong, Zhiwen Yu, Yiyuan Yang, Weizheng Wang, and Kaixiang Yang. 2024. Patchad: A lightweight patch-based mlp-mixer for time series anomaly detection. *arXiv preprint arXiv:2401.09793*.

Tian Zhou, Peisong Niu, Liang Sun, Rong Jin, and 1 others. 2023. One fits all: Power general time series analysis by pretrained lm. *Advances in neural information processing systems*, 36:43322–43355.

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

Appendix Contents

A Related Work

Time Series Anomaly Detection. Time series anomaly detection involves identifying patterns in temporal sequences that deviate significantly from expected normal behavior (Blázquez-García et al., 2021; Boniol et al., 2024). Early approaches relied on statistical methods (e.g., Z-score (Chikodili et al., 2020), moving averages, exponential smoothing (Phillips and Jin, 2021), and ARIMA (Box and Pierce, 1970)) and decomposition-based methods (e.g., HP Filter and STL (Gao et al., 2020; Zhang et al., 2022)). However, these methods often struggle with complex and non-stationary patterns (Zamanzadeh Darban et al., 2024). ML-based methods (Kant and Mahajan, 2019; Ruff et al., 2018; Shin et al., 2020; Karczmarek et al., 2020; Yang et al., 2021) improved flexibility but typically require feature engineering or assumptions about the underlying anomaly patterns. DL-based methods, such as autoencoders (Sakurada and Yairi, 2014), variational autoencoders (Park et al., 2018), and recurrent neural networks (Su et al., 2019), learn temporal dependencies and can capture complex nonlinearities. Transformer-based methods have further enhanced long-range dependency modeling and become state-of-the-art in many benchmark datasets (Xu et al., 2021; Yang et al., 2023; Zhong et al., 2024). Besides, approaches designed for time series foundation models, i.e., a unified "one-fits-all" framework, can also be adapted for anomaly detection (Zhou et al., 2023; Wu et al., 2023; Goswami et al., 2024; Liang et al., 2024). Recently, there has been growing interest in applying LLMs or multimodal/vision-language models (MLLMs/VLMs) for the TSAD task (Liu et al., 2024b; Zhou and Yu, 2024; Xu et al., 2025).

Despite these advancements, challenges remain, including handling rare anomaly categories, ensuring robustness under noisy or missing data, and more detailed anomaly categories and attribution (Zamanzadeh Darban et al., 2024; Jin et al., 2024). Recent works have started trying multimodal data fusion with novel datasets and LLM to address these issues (Zhou and Yu, 2024; Kong et al., 2025a).

Multimodal LLMs for Time Series Analysis. LLMs/MLLMs have recently been explored and utilized for combining their powerful sequence modeling and reasoning abilities. As for the TSAD task, early studies show that LLM-based methods can achieve detection performance comparable to DL-based models and provide clear explanations (Liu et al., 2024b). LLMAD enhances interpretability and performance by retrieving similar segments and applying anomaly-aware chain-of-thought prompting (Liu et al., 2024a). Others convert time series into images for VLMs to visually detect patterns and anomalies (Zhou and Yu, 2024; Xu et al., 2025). They concluded that LLMs often understand time series better when it's presented visually rather than as text or numerical data. However, the same work noted that simply prompting LLMs to reason through the time series did not lead to significant performance gains (Zhou and Yu, 2024). For forecasting tasks, researchers (Wang et al., 2025; Kong et al., 2025a) adapt LLMs by formatting time series as text with relative information, allowing general LLMs to perform zero-shot or few-shot predictions.

Despite recent progress, the application of LLMs to time series analysis remains in its early stages and faces several key challenges. LLMs seem limited in detecting subtle or complex anomalies that are not explicitly covered in their training data or prompts, especially in real-world, context-dependent scenarios (Zhou and Yu, 2024). Moreover, effectively combining different data modalities in a unified model is non-trivial, and usually requires careful prompting based on the specifics of each task (Kong et al., 2025b), also lacks real datasets for model training and evaluation (Xu et al., 2025; Liu et al., 2024a). Therefore, we propose the TIME-RA task that devises novel reasoning paradigms to improve the reliability of LLM-driven time series analysis.

Data Resources in Anomaly Detection and Reasoning. A range of public datasets supports research in the TSAD task. Single-modality benchmarks include univariate datasets (e.g., UCR (Dau et al., 2019) and Yahoo (Laptev et al., 2015)) and multivariate datasets (e.g., NASA's MSL and SMAP (Hundman et al., 2018) and NIPS-TS (Lai et al., 2021)). Recently, time series-based multimodal datasets have emerged, combining temporal data with textual logs or images and attracting growing attention. Examples include the AIOps challenge dataset that integrates performance metrics, logs, and traces from microservices (Bakhtin et al., 2025), and the Nonastreda dataset, pairing machine sensor time series with microscope images to detect manufacturing tool anomalies (Truchan and Ahmadi, 2025). There are also some general multimodality time series datasets from multiple

application areas (Kong et al., 2025b). However, comprehensive benchmarks that incorporate numeric, textual, and visual modalities remain rare. Also, the existing limited multimodal data for time series reasoning for anomaly mostly comes from synthetic data, lacking real-world complexity (Xu et al., 2025; Zhou and Yu, 2024).

Notably, current datasets for time series reasoning and anomaly interpretation remain limited. Although synthetic datasets and tasks, such as TimeMQA (time series question answering) (Kong et al., 2025a), have been proposed to evaluate reasoning capabilities, establishing widely accepted real-world labeled datasets for the Time-RA task remains an open research challenge (Chow et al., 2024; Jin et al., 2024; Kong et al., 2025b). To fill this gap, we design the first real-world multimodal dataset RATs40K that moves beyond anomaly detection toward deeper analytical reasoning.

B Boundary and further direction

For the boundaries, there are some worth noting: **(i) Detection of Multiple Anomaly Types in a Single Sequence.** In cases where multiple types of anomalies coexist within a single time series segment, our model tends to identify only the most salient or dominant anomaly type. While the reasoning module may mention secondary or other anomaly patterns, the primary classification output reflects the most prominent category. This limitation arises from the model's training objective, which emphasizes accuracy in identifying the most impactful anomaly rather than exhaustively listing all types. In particular, for our dataset, the time series length is 16-128, so the probability of multiple anomalies occurring simultaneously is not very high. **(ii) Multivariate Anomaly Detection Bias.** Our current approach to multivariate time series focuses primarily on capturing inter-variable relationships. However, we do not explicitly categorize univariate anomalies that may occur independently within each dimension for the multivariate experiment. Although these univariate anomalies are often mentioned in the LLM-generated reasoning, they are not directly reflected in the anomaly type classification, potentially omitting finer-grained insights for each variable. **(iii) Reliability of the RATs40K Dataset.** During the expert verification process for the RATs40K dataset, experts were allowed to select and rank multiple plausible anomaly types per instance. In our evaluation, if the model's predicted anomaly type matched any of the top-ranked expert-annotated types, particularly the most important one, we considered the prediction correct. While this evaluation strategy increases robustness, it may mask subtle discrepancies between model output and expert intent, especially in borderline or ambiguous cases.

Several directions remain for exploration. **(i) Scalability to Longer Time Series.** As time series scales in length, directly inputting the entire sequence for LLM-based reasoning becomes challenging due to the token length constraints. Future work could explore time-series-specific embedding or a hierarchical summarization strategy to compactly represent long sequences while preserving critical temporal patterns. Alternatively, enhancing the weight of the accompanying visual representations (e.g., time series plots) may serve as a complementary approach. **(ii) Improved Modeling of Multiple Coexisting Anomalies.** Current classification is limited to predicting a single dominant anomaly type. Future research could focus on multi-label anomaly categorization that explicitly supports detection and reasoning over multiple overlapping anomaly types within the same sequence, potentially with structured output formats. **(iii) Continual Learning and Domain Adaptation.** Time series distribution may shift over time or across application domains (e.g., finance vs. Machine shown in Figure 6). Developing continual learning frameworks or lightweight domain adaptation strategies compatible with reasoning models remains an important challenge for real-world deployment.

C The Definition of Anomaly Types

To systematically address the challenge of time series anomaly detection, it is essential to first establish a clear and comprehensive taxonomy of anomaly types. Accordingly, we categorize anomalies into two primary classes: univariate and multivariate. These structured definitions of anomaly types are the foundational framework of TIME-RA that guides the design of robust classification metrics.

Univariate anomalies, detailed in Table 6, pertain to abnormality in a single time series. These can range from simple, isolated deviations such as *Point Anomalies* and *Sudden Spikes*, to more complex structural shifts. The classification encompasses changes in statistical properties (e.g., *Change Point*

and *Distributional Change Anomaly*), alterations in long-term behavior (e.g., *Trend Change* and *Drift Anomaly*), and disruptions in cyclical or repetitive patterns (e.g., *Periodic Change* and *Pattern Change Anomaly*). Furthermore, we define anomalies of data acquisition failures, such as *Sudden Flatline* and *Repeated Value* anomalies. For each category, the table provides a formal definition, a representative visual illustration, and a practical domain example to facilitate intuitive understanding.

In contrast, multivariate anomalies arise from the complex interplay between variables. As cataloged in Table 7, these anomalies are often subtle, as individual variables may appear normal when inspected in isolation. The abnormality lies in the violation of their expected joint behavior or inter-variable relationships. Our taxonomy covers several key types, including disruptions in statistical dependencies like the *Covariance Structure Anomaly* and *Collinearity Shift Anomaly*, and deviations in temporal synchronization, defined as *Temporal Dependency Anomaly*. We also identify anomalies where variables diverge from a common pattern (*Trend Divergence Anomaly*) or where their combined state is improbable (*Joint Space Anomaly*). Finally, some anomalies are only detectable in a reduced-dimensional latent space (*Principal Component Space Anomaly*). Similar to the univariate table, each multivariate anomaly type is accompanied by a precise definition, a two-variable visualization, and a real-world scenario to clarify its meaning.

D Example of TIME-RA

This section demonstrates how SFT works for the TIME-RA task through two case studies. Table 8 shows univariate anomaly detection on ECG data, where the model identifies a nonlinear pattern anomaly through step-by-step reasoning. Table 9 presents a multivariate case detecting temporal dependency anomalies across synchronized ECG channels. Supporting visualizations of these cases are shown in Figures 6 and 7, illustrating the model's analytical process and decision-making. In the univariate case (Figure 6), source information is particularly important for accurate action classification and for grounding the relevance of the Thought. In the univariate case, such as the "normal" ECG example, source context is crucial for accurate classification and domain-specific Thought generation. Without it, healthcare terms may be omitted or anomalies misclassified as amplitude anomaly. Similarly, in multivariate cases (Figure 7), source and variable data provide essential context for interpreting Action and grounding Thought. In summary, Action can be correctly interpreted using the available multimodal data, while Thought can rely on both source and variable information for domain-specific attribution and precise anomaly localization.

E Statistics of the Model Pool

We further analyse the statistics of model responses in Figure 8. Since we take as input the label of whether or not it is anomalous, the four strong models have similar anomaly judgements and follow the instruction label. However, as shown in Figure 8b, they differed in their performance on the task of classification of the source of anomalies, especially for the univariate time series. Further leading to differences in the length of the anomaly reasons in Figure 8a.

F LLM Descriptions

Throughout this paper, the experiments use the version of off-the-shelf models detailed in Table 10. All models are selected in their instruction-tuned variants to ensure they can effectively follow prompts and perform the specified tasks. To provide a comprehensive analysis, we organize our selection into three distinct categories: open-source LLMs, VLMs, and state-of-the-art models accessed via APIs.

G Likert Scale for Evaluating LLM-Generated Annotation

To assess the quality of the automatically generated reasoning, we design a structured human evaluation protocol based on a Likert scale. Specifically, we evaluate each explanation along five key dimensions: language quality, factual soundness, specificity to the anomaly, interpretability, and usefulness for downstream tasks. Table 12 defines each dimension along with practical indicators. Table 11 describes the interpretation of Likert scores from 1 (very poor) to 5 (excellent).

H Generation and Annotation Templates

Our data generation and refinement process is governed by a series of structured templates. The initial **Instruction for Anomaly Detection** Template prompts a model to perform time series anomaly

Table 6: Univariate anomaly types with example observations and explanation.

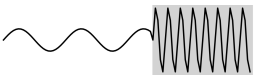
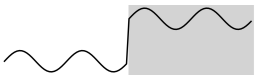
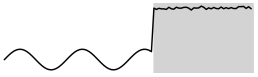
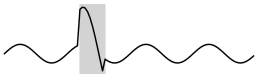
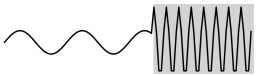
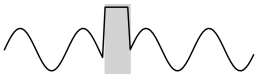
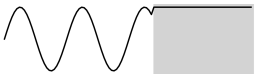
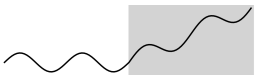
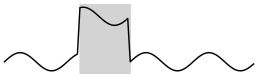
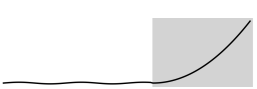
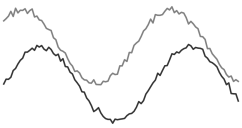
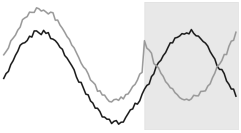
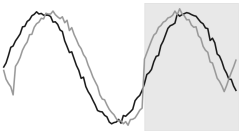
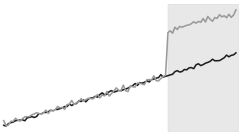
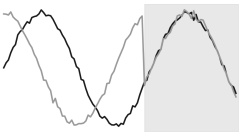
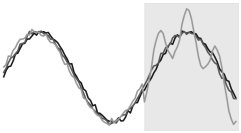
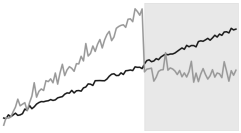
Anomaly Category	**Definition**	**Observation**	**Domain Example**
Normal Sequence	There are no abnormal situations in this time series.		Industrial sensor recording shows normal behavior
Point Anomaly	A single data point significantly deviates from the local or global pattern.		Sudden spike in heart rate during sleep detected by a wearable device.
Periodic Change Anomaly	The original periodic pattern is disrupted, such as broken periods or anomalous amplitude.		Power consumption pattern changes due to a faulty machine affecting periodic load.
Trend Change Anomaly	A sudden change in the long-term trend of the time series.		Sales data shows a sudden increase due to a product launch.
Change Point Anomaly	Statistical properties (e.g., mean, variance) change abruptly at certain points.		Network latency shifts abruptly due to a routing change.
Distributional Change Anomaly	The statistical distribution of the time series changes significantly.		A change in website traffic from normal to bot-generated patterns.
Amplitude Anomaly	The amplitude of data points exceeds the normal upper or lower bounds.		ECG signal exceeds expected amplitude thresholds, indicating potential arrhythmia.
Pattern Change Anomaly	The pattern of the time series suddenly changes from one form to another.		Stock market index changes from oscillating to a steady downward trend.
Sparse Anomaly	Isolated anomalous patterns occasionally appear in a long time series.		Occasional fraudulent transactions in a large sequence of financial data.
Repeated Value Anomaly	Continuous or intermittent repeated values disrupt the normal fluctuation pattern.		Temperature sensor gets stuck and repeatedly reports the same value.
Sudden Flatline Anomaly	The time series suddenly becomes a flat line with no normal fluctuations.		IoT device disconnects and reports a flat signal for several minutes.
Drift Anomaly	The data gradually drifts away from the normal level.		GPS drift causes a slowly diverging location reading over time.
Sudden Spike Anomaly	The data suddenly spikes or drops within a short time and then returns to normal.		Sudden voltage spike in an electrical grid followed by return to normal.
Continuous Segment Anomaly	A continuous segment of data points deviates from the normal pattern.		A segment of network traffic deviates from expected behavior during a DDoS attack.
Nonlinear Pattern Anomaly	Nonlinear changes appear, breaking the original linear rule.		A traffic speed pattern changes from linear increase to nonlinear surge due to congestion.

Table 7: Multivariate anomaly types with example observations and explanation (two variables for example).

Anomaly Category	Definition	Observation	Domain Example
Normal Sequence	From a multivariate view, all variables follow expected patterns over time. Relationships among variables and their dynamics remain stable without any abnormality.		A smart factory's sensors operate consistently: temperature, pressure, and vibration stay within expected ranges and show stable interdependencies.
Covariance Structure Anomaly	The usual covariance or correlation structure among variables changes suddenly, such as reversal or unexpected decorrelation.		In a financial system, the strong positive correlation between two stock prices suddenly breaks, potentially indicating market manipulation or systemic stress.
Temporal Dependency Anomaly	Expected temporal dependencies (e.g., fixed lags and variable response delays) are violated, indicating possible desynchronization or timing failures.		In a manufacturing line, the normal delay between motor start and sensor response disappears, suggesting a sensor fault or a control system failure.
Trend Divergence Anomaly	A subset of variables unexpectedly deviates from a shared trend, suggesting localized failures or partial system faults.		In a power grid, one region's voltage levels begin to drift from the national trend, possibly indicating equipment aging or a localized overload.
Joint Space Anomaly	Although individual variable values may appear normal, their joint configuration is anomalous—suggesting system-level inconsistency in the multivariate space.		In autonomous driving, speed and steering angle are each within normal limits, but their combination implies unsafe turning behavior.
Principal Component Space Anomaly	An anomaly becomes evident only in a lower-dimensional latent space (e.g., PCA), revealing subtle structural deviation across many variables.		In climate modeling, subtle but coordinated changes across dozens of climate indicators (e.g., temperature, pressure, humidity) show up only in the PCA space, signaling early signs of climate shifts.
Collinearity Shift Anomaly	Strong linear dependencies or redundancies between variables suddenly break down, often due to malfunctioning or desynchronized components.		In server monitoring, CPU and memory usage used to be tightly coupled, but suddenly became independent—suggesting possible memory leakage or process isolation failure.

detection, generating a Thought (reasoning) and Action (classification) for a given observation. We then use a **Annotation Template** to have a judge model, such as `GPT-4`, score and rank the outputs from multiple LLMs in the model pool based on a predefined 5-point rubric. Finally, a **GPT-4 Critique Feedback** Template instructs an expert model to review and, if necessary, rewrite a model's response to create a gold-standard data sample, which is then used for further training. This multi-stage process ensures the creation of a high-quality, well-annotated dataset.

I The Details of the RATs40K Dataset

The table 13 lists how many time series segments were extracted from each subdataset. Specifically, we select raw time series from comprehensive open source repositories across diverse domains, including AIOps systems (Laptev et al., 2015; Liu and Paparrizos, 2024), environment (Barrenetxea, 2019; Liu and Paparrizos, 2024), finance (Ahmad et al., 2017; Tran et al., 2016), healthcare (Wu and Keogh, 2021; Bachlin et al., 2009; Goldberger et al., 2000; Greenwald et al., 1990), IoT (Barrenetxea, 2019; Ahmad et al., 2017), industrial sensors (Wu and Keogh, 2021; Liu and Paparri-

Table 8: An example of SFT for a univariate time series.

##User:

You are an expert in time series anomaly detection. We provide a time series (called Observation), you should give us the anomaly type (called Action) and its reasons (called Thought). Thought steps can infer the current abnormal situation of a time series.
The anomaly detection of each time series is divided into three steps: Observation, Thought, and Action. After analyzing each observation, please provide the next Thought and next Action. Here is a time series observation that we need to check for anomaly categories. The observation is from the domain of Healthcare-ECG.
Please make a Thought judgment and put your final Action within \\boxed1{} and \\boxed2{} respectively, where action must just be a category name not id.
Observation: -0.04, -0.05, -0.09, 0.0, 0.32, 0.37, -0.46, -1.56, -1.52, -0.98, -0.51, -0.2, -0.03, 0.0, -0.02, 0.1, 0.02, 0.11, 0.07, 0.13, 0.1, 0.24, 0.15, 0.14, 0.17, 0.24, 0.2, 0.2, 0.21, 0.23, 0.37, 0.24, 0.19, 0.22, 0.16, 0.16, 0.14, 0.07, 0.12, -0.01, 0.06, -0.06, 0.14, 0.03, 0.09, 0.07, -0.01, 0.06, 0.08, 0.03, 0.0, 0.0, -0.02, 0.09, 0.08, 0.1, 0.0, 0.05, 0.16, 0.09, 0.17, 0.13, 0.06, 0.16, 0.14, 0.11, 0.09, 0.19, 0.07, 0.13, 0.06, 0.09, -0.09, -0.18, -0.26, -0.06, 0.46, 1.41, 2.73, 4.34, 4.65, 4.34, 3.18, 2.55, 2.2, 2.87, 2.35, 1.57, 0.51, 0.05, -0.34, -0.48, -0.54, -0.54, -0.66, -0.81, -0.8, -0.97, -0.95, -1.09, -1.21, -1.53, -1.55, -1.68, -1.86, -2.09, -2.08, -2.22, -2.09, -2.05, -2.08, -1.67, -1.53, -1.28, -0.99, -0.82, -0.62, -0.39, -0.22, -0.11, -0.19, 0.06, -0.03, -0.05, 0.09, 0.03, -0.05, 0.1
Thought: \\boxed1{}
Action: \\boxed2{}

##Assistant:

Thought: \\boxed1{The time series exhibits multiple anomalies. Initially, there is a sudden drop and recovery in values (e.g., -0.46 to -1.56 and back to near 0), followed by a gradual drift upward (e.g., from 0.46 to 4.65). This is then followed by a nonlinear pattern of exponential decay (e.g., from 4.65 to -2.22) and oscillations. The sequence shows a combination of drift, nonlinear changes, and potential periodic disruptions, which are characteristic of ECG anomalies. The most prominent anomaly is the nonlinear pattern of exponential growth and decay, which breaks the expected linear or periodic behavior of an ECG signal.}
Action: \\boxed2{Nonlinear Pattern Anomaly}

zos, 2024), server data (Ahmad et al., 2017; Liu and Paparrizos, 2024), traffic (Ahmad et al., 2017), network records (Ahmad et al., 2017; Jacob et al., 2020), and synthetic dataset (Thill et al., 2020; Lai et al., 2021). Each subdataset provides at least 200 samples, with more than 80% of them being anomalous. The sampling imbalance mainly arises from differences in the original subdataset sizes and lengths, as well as inherent differences in anomaly ratios across domains. For example, some of the Healthcare datasets are derived from ECG signals. Although the data length is similar to that of Finance datasets, the anomaly ratio in ECG is higher, leading to imbalanced sampling. Empirically, fine-tuning large models can partially generalize across such sampling imbalances. Moreover, we include the source information as part of our prompt, which can potentially mitigate the effects of imbalance.

Regarding the UCR datasets, although there are 250 subsequences, we treat each subsequence independently. A segment is labeled as anomalous if between 0 and 80% of its original timestamps are annotated as anomalies within a given sampling window. We acknowledge that some segments may contain overlapping timestamps, but we ensure that no two segments are identical. In addition, for each subdataset, we set multiple segment lengths to further increase the number of final samples.

As shown in Listing 1, a typical item based on a time series segment is structured in the RATs40K dataset. The meaning of each key is as follows: "Action" is the anomaly category, "ActionID" is the number corresponding to the anomaly category used for numerical evaluation, "Figurepath" is the image corresponding to the time series segment, "Observation" corresponds to the original time series itself (if it is multivariate, we provide the variable name and a brief explanation),

Table 9: An example of SFT for a multivariate time series.

##User:

You are an expert in time series anomaly detection. We provide a time series (called Observation), you should give us the anomaly type (called Action) and its reasons (called Thought). Thought steps can infer the current abnormal situation of a time series.
The anomaly detection of each time series is divided into three steps: Observation, Thought, and Action. After analyzing each observation, please provide the next Thought and next Action. Here is a time series observation that we need to check for anomaly categories. The observation is from the domain of Medical-ECG.
Please make a Thought judgment and put your final Action within \\boxed1{} and \\boxed2{} respectively, where action must just be a category name not id.
Observation: "ECG1": [0.12, 0.11, 0.1, 0.07, 0.06, 0.04, 0.02, 0.02, 0.01, 0.01, 0.0, -0.01, -0.01, -0.01, 0.02, 0.09, 0.06, -0.26, -0.56, -0.71, -0.79, -0.85, -0.78, -0.62, -0.57, -0.54, -0.53, -0.52, -0.42, -0.3, -0.19, -0.09, -0.02, 0.05, 0.14, 0.23, 0.26, 0.27, 0.27, 0.28, 0.29, 0.3, 0.3, 0.32, 0.34, 0.34, 0.36, 0.38, 0.39, 0.41, 0.41, 0.43, 0.43, 0.42, 0.42, 0.4, 0.37, 0.34, 0.3, 0.26, 0.2, 0.16, 0.14, 0.09]; "ECG2": [0.06, 0.06, 0.06, 0.05, 0.05, 0.04, 0.02, 0.01, 0.01, 0.01, 0.0, -0.01, -0.01, -0.01, -0.01, 0.03, 0.0, -0.15, -0.53, -0.79, -0.81, -0.76, -0.7, -0.58, -0.49, -0.45, -0.39, -0.33, -0.28, -0.18, -0.12, -0.06, -0.01, 0.04, 0.1, 0.16, 0.2, 0.2, 0.2, 0.2, 0.21, 0.21, 0.21, 0.21, 0.23, 0.23, 0.24, 0.24, 0.24, 0.26, 0.26, 0.26, 0.27, 0.27, 0.26, 0.24, 0.23, 0.2, 0.17, 0.13, 0.1, 0.07, 0.04, 0.02]; "ECG3": [0.14, 0.14, 0.14, 0.09, 0.07, 0.06, 0.04, 0.03, 0.01, 0.01, 0.01, -0.01, -0.01, 0.0, 0.0, 0.0, 0.01, 0.05, 0.07, 0.06, -0.06, -0.07, -0.1, -0.1, -0.08, 0.03, 0.01, -0.12, -0.28, -0.38, -0.45, -0.39, -0.26, -0.1, -0.01, 0.01, 0.03, 0.1, 0.13, 0.11, 0.14, 0.13, 0.14, 0.14, 0.14, 0.16, 0.17, 0.17, 0.2, 0.21, 0.2, 0.2, 0.23, 0.24, 0.25, 0.23, 0.22, 0.24, 0.23, 0.2, 0.2, 0.18, 0.12, 0.07]
Thought: \\boxed1{}
Action: \\boxed2{}

##Assistant:

Thought: \\boxed1{The three ECG channels exhibit a shared general structure, but there are notable deviations in their dynamics and relationships. Initially, all channels show a gradual decline, followed by a sharp negative deflection (e.g., around time steps 17–21). However, the magnitude and timing of this deflection differ across channels. For example, ECG1 shows a more pronounced and earlier negative peak compared to ECG2 and ECG3. Additionally, ECG3 exhibits an unusual behavior around time steps 26–30, where it deviates significantly from the patterns of ECG1 and ECG2, showing a delayed and less synchronized recovery. This desynchronization and violation of expected temporal dependencies between the channels suggest a potential anomaly in the timing or coordination of the system, possibly due to sensor misalignment, physiological irregularities, or external interference.}
Action: \\boxed2{Temporal Dependency Anomaly}

"Source" indicates the data source and specific domain, and "Thought" is the ground truth of the reasoning.

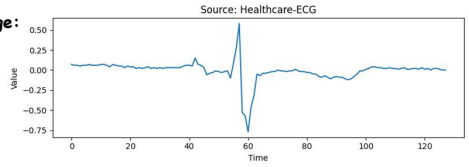
Instruction for Anomaly Detection

Univariate Time Series Instruction: You are an expert in time series anomaly detection. We provide a time series (called Observation), you should give us the anomaly type (called Action) and its reasons (called Thought). Thought steps can infer the current abnormal situation of a time series. Action is an abnormal category with the following 0-14 types, where 0 is a normal category. The explanations of 0-14 actions are as follows:
{UNIVARIATE TIME SERIES ANOMALY CATEGORY}

Multivariate Time Series Instruction: You are an expert in multivariate time series anomaly detection. We provide a multivariate time series (called Observation), where each time point contains multiple variables. Your task is to identify the anomaly type (called Action) and provide detailed reasoning (called Thought). The Thought should analyze the relationships, dynamics, and structures across all variables and time points to infer any abnormal behavior. The Action must be one of the following seven types,

Univariate Time Series Example - Normal

Image:



Label: Normal

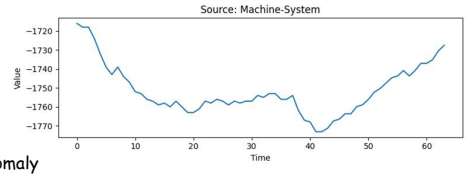
Source: Healthcare-ECG

Action: Normal Sequence

Thought: The provided single-channel time series displays a characteristic ECG waveform. It starts with a relatively stable baseline, followed by a small positive deflection (P-wave), a sharp triphasic complex with an initial negative deflection, a large positive peak, and a deep negative deflection (QRS complex), and then a slower, broader deflection (T-wave) as the signal returns to baseline. The amplitudes and durations of these components appear consistent with a typical healthy heartbeat. There are no sudden extreme spikes, flat lines, erratic fluctuations, or missing components that would suggest an anomaly in the heart's electrical activity or the sensor reading.

Univariate Time Series Example - Anomaly

Image:



Label: Anomaly

Source: Machine-System

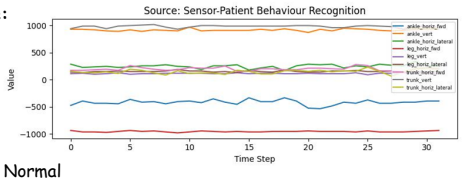
Action: Trend Divergence Anomaly

Thought: The observed single time series displays a relatively stable or slowly decreasing trend for the initial part of the sequence. However, around time step 38 (value -1754.0), the series experiences a distinct and sharp downward deviation, bottoming out at -1773.15. Following this significant dip, the series systematically recovers its value over the subsequent time steps, eventually returning to a level similar to its pre-anomaly state (-1727.55). This sudden, pronounced, and sustained deviation from its established trend, followed by a clear recovery, indicates an unexpected transient event or fault that caused this specific variable to diverge from its expected behavior. While the definition often refers to multiple variables, a single variable deviating from its own characteristic or expected trend fits the spirit of Trend Divergence as an unexpected change in its temporal progression.

Figure 6: Univariate TIME-RA case study.

Multivariate Time Series Example - Normal

Image:



Label: Normal

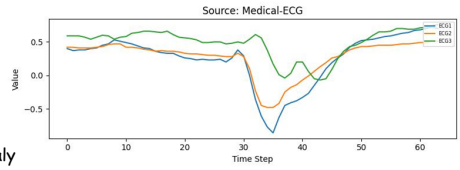
Source: Sensor-Patient Behaviour Recognition

Action: Normal Sequence

Thought: The multivariate time series data across all channels (ankle, leg, and trunk in various directions) exhibit expected variability and dynamics consistent with normal sensor-patient behavior. The values fluctuate within reasonable ranges, and there are no abrupt changes, outliers, or unexpected deviations in trends. The relationships among variables, such as the interplay between horizontal, vertical, and lateral movements, remain stable and consistent over time. Temporal dependencies and correlations across channels are preserved, and no structural anomalies are observed in the joint or latent spaces. Overall, the system behaves as expected for this domain.

Multivariate Time Series Example - Anomaly

Image:



Label: Anomaly

Source: Medical-ECG

Action: Temporal Dependency Anomaly

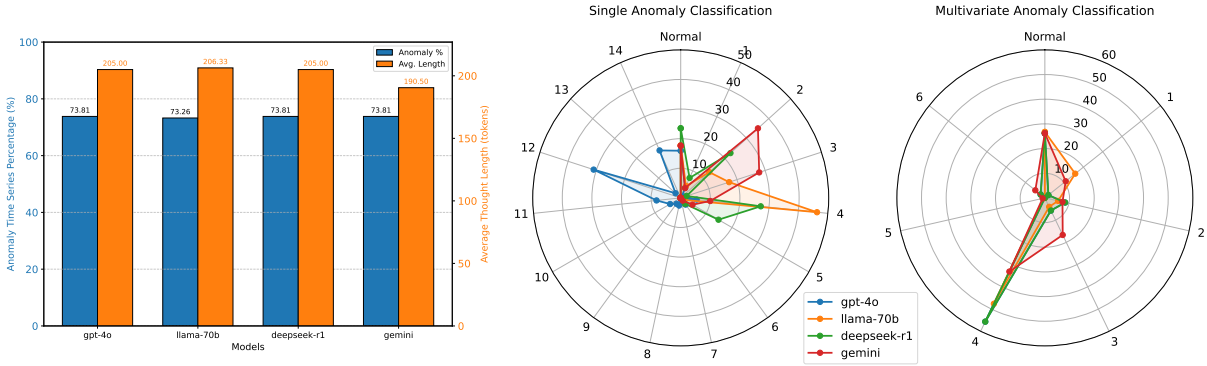
Thought: The three ECG channels exhibit a complex pattern with notable deviations in their dynamics. Initially, all channels show relatively stable oscillations, but around time step 30, there is a significant disruption. ECG1 exhibits a sharp drop to 0 and then negative values, followed by a recovery. ECG2 also shows a drop but with a less pronounced negative excursion. ECG3, on the other hand, shows a different behavior, with a sharp drop followed by a recovery that is not synchronized with the other two channels. This desynchronization and the differing recovery patterns suggest a breakdown in the temporal dependencies between the channels. Such behavior could indicate timing failures, desynchronization in the system, or disturbances in the underlying physiological processes. The anomaly is most consistent with a violation of expected temporal dependencies.

Figure 7: Multivariate TIME-RA case study.

where type 0 means no anomaly. The definitions are:
{MULTIPLE TIME SERIES ANOMALY CATEGORY}

The anomaly detection of each time series is divided into three steps: Observation, Thought, and Action. After analyzing each observation, please provide the next Thought and next Action. Here are some examples:
{SAMPLES OF EXPERT DEFINITIONS}

Here is a univariate/multivariate time series observation that we need to check for anomaly categories. We already know that it is a {ANOMALY LABLE} sequence and from the domain of {SAMPLE SOURCE}. Please make a Thought judgment within \\boxed1{} and put your final Action in \\boxed2{} respectively, where action must just be a category name, not id.
Observation: {SAMPLE OBSERVATION}
Thought: \\boxed1{}
Action: \\boxed2{}



(a) Anomaly rate (blue) and thought length (orange) for different models.

(b) Classification of anomalies in different models with univariate and multivariate time series.

Figure 8: Global statistics of model response in the model pool.

Table 10: Summary of models and links used in our experiments.

Model Name	**Link**
DeepSeek-7B	https://huggingface.co/deepseek-ai/deepseek-llm-7b-chat
Llama-3-8B	https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
Llama-3.2-3B	https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct
Phi-4-mini	https://huggingface.co/microsoft/Phi-4-mini-instruct
Qwen2.5-3B	https://huggingface.co/Qwen/Qwen2.5-3B-Instruct
Qwen2.5-7B	https://huggingface.co/Qwen/Qwen2.5-7B-Instruct
Llava-v1.5-7B	https://huggingface.co/llava-hf/llava-1.5-7b-hf
Llava-v1.5-13B	https://huggingface.co/llava-hf/llava-1.5-13b-hf
Llama-3.2-11B-v	https://huggingface.co/meta-llama/Llama-3.2-11B-Vision
Qwen2.5-vl-7B	https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct
Gemini-2.5-flash	https://deepmind.google/models/gemini/flash/
DeepSeek-R1	https://huggingface.co/deepseek-ai/DeepSeek-R1
Llama-3.3-70B-Instruct	https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct
gpt-4o_2024-11-20	https://platform.openai.com/docs/models/gpt-4o
GPT-4	https://platform.openai.com/docs/models/gpt-4

Table 11: Interpretation of quality scores for time series reasoning for anomaly.

Score	**Interpretation**
1	Very Poor – Invalid, irrelevant, or nonsensical
2	Poor – Major issues or weak alignment
3	Fair – Partially reasonable but lacking clarity or depth
4	Good – Mostly clear, relevant, and logical
5	Excellent – Accurate, specific, useful, and well-written

Annotation Template

A task we have is:
[{EXAMPLE INSTRUCTION FOR ANOMALY DETECTION}]

Now we have the outputs of models, there are:
The model [gpt-4o] output is:
Thought: {GPT-4O THOUGHT}
Action: {GPT-4O ACTION}

Table 12: Evaluation dimensions and criteria for time series reasoning for anomaly.

Dimension	Description	What to Look For
Language Quality	Is the reason grammatically correct, clearly phrased, and structurally complete? The sentence should read fluently, without ambiguity or major grammatical issues.	(1) Proper sentence structure (2) No grammatical or syntactic errors (3) Clear and fluent wording
Factual Soundness	Is the reasoning factually aligned with the time series behavior (e.g., trend, spike, drop) and contextual data? The explanation should be plausible and consistent with the actual data.	(1) Matches visible pattern in the time series (2) Correctly reflects contextual data (3) Avoids hallucinations or random guesses
Specificity to Anomaly	Does the reason specifically address the anomaly point in question? It should not be vague, overly generic, or a templated explanation that could apply to any anomaly.	(1) Mentions a concrete possible cause for a certain anomaly (2) Avoids boilerplate text (3) Shows evidence of localization in time or context
Interpretability	Is the reason easy for a human to understand and follow? It should be logically structured, free from jargon, and ideally follow a cause-effect or descriptive reasoning pattern.	(1) Logical flow of thought (2) Minimal technical jargon (3) Easily understood by analysts or domain experts
Usefulness	Does the explanation provide actionable insight or support further steps such as labeling, filtering, or alerting? A useful reason helps humans make better decisions.	(1) Supports human labeling decisions (2) Offers insight that could influence next steps (3) Helps explain or validate the anomaly for stakeholders

The model [llama-70b] output is:
Thought: {LLAMA-70B THOUGHT}
Action: {LLAMA-70B ACTION}

The model [deepseek-r1] output is:
Thought: {DEEPSEEK-R1 THOUGHT}
Action: {DEEPSEEK-R1 ACTION}

The model [GEMINI] output is:
Thought: {GEMINI THOUGHT}
Action: {GEMINI ACTION}

Please evaluate the consistency between the output of each model and the task intent, and score and provide reasons for the answers of each model. The score is from 1 to 5:
1. **Irrelevant**: No alignment.
2. **Partial Focus**: Poor handling in a certain aspect, such as misclassification of exceptions.
3. **Partial Compliance**: The classification of anomalies is accurate, but there may be slight deviations or neglect of others in the reasons.
4. **Almost There**: Alignment close to expert answers, slight deviation.
5. **Comprehensive Compliance**: Completely consistent with expert answers, meeting all requirements.

Based on the above ratings, please provide me with a ranking to compare the output results from the above models.
Here is an example of the output format:
<|begin|>gpt-4o>gemini>deepseek-r1>llama-70b<|end|>

GPT-4 Critique Feedback

A task we have is:
[{EXAMPLE INSTRUCTION FOR ANOMALY DETECTION}]
Given the model answer to an instruction, your role is to provide specific and constructive feedback for me. When you review the model answer, consider its helpfulness, truthfulness, honesty, and how well it followed the instructions.
The model answer is:
Thought: {THOUGHT}
Action: {ACTION}

I need you to assume the role of an anomaly detection expert. It's essential that your feedback not only highlights areas for improvement but also provides actionable suggestions to help the model understand how to enhance its performance. Please make improvements based on the thought and action of the model and follow the same output. If no improvement is needed, just return **None**.
The following are examples of formats that need to be improved for output:
Thought: {GPT-4 THOUGHT}
Action: {GPT-4 ACTION}

```
{
  "Action": "Joint Space Anomaly",
  "ActionID": 4,
```

Table 13: The details of the proposed RATs40K dataset.

Type	Dataset	Number of Segments	Key of Source
Univariate	TODS and NAB-artificial	220	Synthetic data
	NAB-Cloudwatch	277	AIOps
	MBA-ECG and MITDB	14,394	Healthcare
	NAB-Exchange	254	Finance
	IOPS_KPI	3,171	Server
	MGAB	900	IoT
	SensorScope	4,776	Environment
	NAB-Traffic	220	Traffic
	NAB-Tweets	582	Network
	UCR	2,358	Industrial sensors
	YAHOO	9,148	Server
Multivariate	GECCO	914	IoT
	Daphnet	360	Industrial sensors
	TAO	1,000	Environment
	MITDB, SVDB, and LTDB	1,000	Healthcare

```
  "FigurePath": "figures_multi/train/27.
      jpg",
  "Label": "Anomaly",
  "Observation": {
    "ECG1": "[Real time series data]",
    "ECG2": "[Real time series data]",
    "ECG3": "[Real time series data]"
  },
  "Source": "Medical-ECG",
  "Thought": "The reasoning text"
}
```

Listing 1: A time series segment example in RATs40K dataset

J Prompt Design Ablation Study

In addition to the ablation study of image and observation shown in Table 3 of the main text, we conduct further detailed ablation studies to analyze the effect of prompt design on model performance, focusing on prompt format, the number of in-context examples (zero-, one-, and 7-shot), prompt variants, and the inclusion of chain-of-thought (CoT) reasoning. Table 14 shows that increasing the number of few-shot examples leads to consistent improvements, especially for anomaly category classification and reasoning-related metrics, while gains on label matching are comparatively modest. Adding chain-of-thought prompting provides additional benefits across most settings, with the most noticeable improvements observed in thought matching metrics, suggesting more structured and coherent reasoning outputs. At higher shot numbers (e.g., 14-shot), the performance differences between prompt structures become smaller, indicating that sufficient in-context examples can partially compensate for prompt complexity.

K Failure Case Analysis

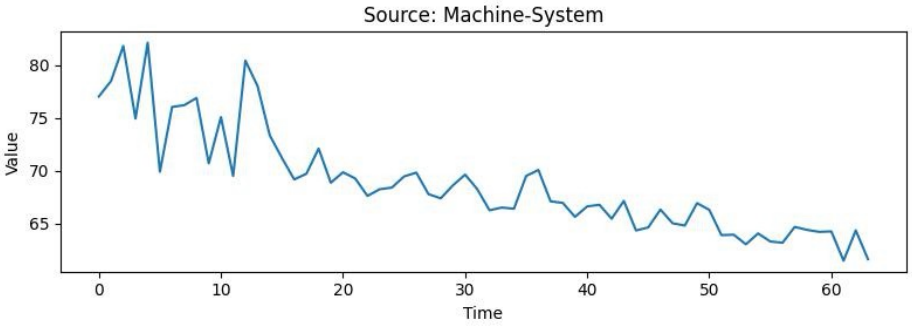
We further analyze representative failure cases to better understand the limitations of the proposed approach. There are three typical failure cases as shown in Figures 9 - 11. These failure cases provide additional insights into the boundaries of our approach rather than fundamental weaknesses. In failure case 1 (Figure 9), the anomaly manifests as a slow and smooth downward drift, which can be difficult to distinguish from normal long-term trends without explicit prior knowledge, leading to occasional under-detection. Failure case 2 (Figure 10) shows that while the model successfully detects anomalous behavior, it may confuse closely related anomaly types (e.g., periodic change versus amplitude variation) when multiple irregular patterns co-occur. Failure case 3 (Figure 11) highlights the challenge of multivariate reasoning under extreme local deviations, where a dominant abnormal channel can obscure or bias the interpretation of cross-variable relationships. Overall, these cases suggest that the model remains effective at identifying anomalous events, but finer-grained anomaly categorization and joint dependency modeling leave room for further improvement.

Table 14: Results of the prompt design ablation study. We report the performance of a fine-tuned `Qwen2.5-7B` model on univariate time series.

Prompt	# Shots	CoT	Label Matching			ActionID Matching			Thought Matching			
			P	R	F1	P	R	F1	Cosine	TFIDF	Lev.	Token
Base	0	–	0.8205	0.9280	0.8710	0.1132	0.1108	0.0743	0.8683	0.2129	0.1783	0.1003
Base	0	✓	0.8474	0.9273	0.8857	0.1206	0.1019	0.0767	0.8812	0.2381	0.1820	0.1036
Base	1	–	0.8199	0.9295	0.8714	0.1490	0.1152	0.0802	0.8927	0.2367	0.1873	0.1063
Base	1	✓	0.8485	0.9159	0.8810	0.1543	0.1108	0.0815	0.9090	0.2518	0.1929	0.1101
Base	7 (half)	–	0.8252	0.9297	0.8744	0.1736	0.1013	0.0909	0.9107	0.2717	0.2232	0.1388
Base	7 (half)	✓	0.8463	0.9323	0.8873	0.1739	0.1042	0.0918	0.9165	0.3008	0.2390	0.1591
Simplified	14 (full)	–	0.8218	0.9785	0.8934	0.1916	0.1292	0.0959	0.9117	0.2725	0.2298	0.1468
Simplified	14 (full)	✓	0.8371	0.9643	0.8963	0.1989	0.1486	0.0991	0.9208	0.2948	0.2401	0.1593
Base	14 (full)	–	0.8286	0.9310	0.8769	0.2498	0.1081	0.1051	0.9198	0.2857	0.2328	0.1519
Base	14 (full)	✓	0.8452	0.9295	0.8854	0.2532	0.1229	0.1100	0.9270	0.3160	0.2432	0.1624

Failure Case1 - Univariate Time Series

Observation:

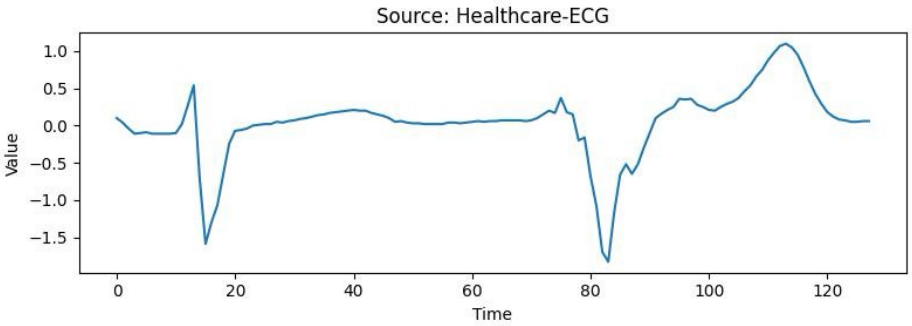


	Ground Truth	Fine-tuned Llama-3-8B	Fine-tuned Qwen2.5-7B
Label	Anomaly	Normal	Anomaly
Action	Drift Anomaly	Normal	Sudden Flatline Anomaly
Thought	The time series exhibits a gradual downward drift in values, starting from around 77.03 and consistently decreasing over time to 61.63. This indicates a systematic deviation from the original level, characteristic of a drift anomaly. The sequence does not return to its original range, and the decline appears to be progressive rather than abrupt.	The sequence exhibits a consistent fluctuation pattern with no extreme outliers or sudden changes, but there is a slight trend of decreasing values towards the end of the sequence. The values are mostly within a range of 60-70, with some minor deviations. However, the overall pattern is not drastically different from the initial trend.	The sequence shows a general downward trend with occasional small fluctuations. However, there are several instances where the values suddenly drop sharply, such as from 69.86 to 67.62, and from 66.78 to 65.46. These sharp drops could indicate anomalies.

Figure 9: Univariate time series failure case 1.

Failure Case2 - Univariate Time Series

Observation:

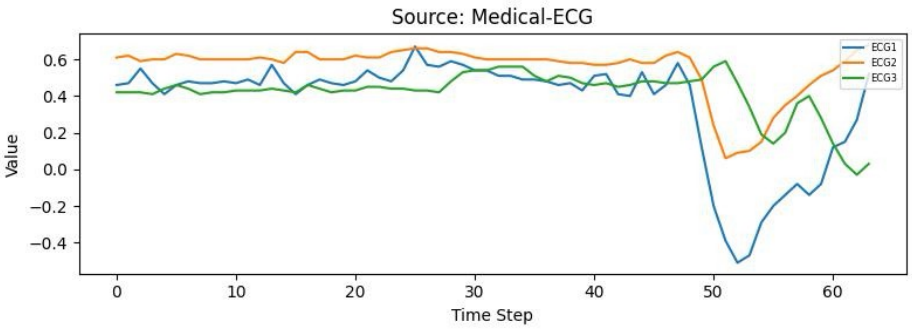


	Ground Truth	Fine-tuned Llama-3-8B	Fine-tuned Qwen2.5-7B
Label	Anomaly	Anomaly	Anomaly
Action	Periodic Change Anomaly	Amplitude Anomaly	Amplitude Anomaly
Thought	The time series exhibits multiple irregularities, including sudden large deviations (e.g., -0.71, -1.59, -1.7, -1.83) and abrupt changes in amplitude and pattern. These deviations disrupt the normal ECG waveform, which typically follows a consistent and periodic pattern. The sequence also shows segments of abnormal fluctuations and transitions, indicating a disruption in the expected periodicity and amplitude of the ECG signal. This suggests a significant anomaly in the periodic pattern of the time series.	The sequence exhibits a consistent pattern of fluctuations, with a mix of positive and negative values, but there is a noticeable deviation from the normal pattern starting at index 16, where the values suddenly drop to -1.59 and then continue to decrease, indicating a disruption in the expected variation.	The time series shows a significant drop in values around index 15, followed by a sharp increase around index 18. This indicates a sudden change in the amplitude of the ECG signal, which could be due to a heartbeat or other physiological events.

Figure 10: Univariate time series failure case 2.

Failure Case3 - Multivariate Time Series

Observation:



	Ground Truth	Fine-tuned phi-4-mini	Fine-tuned Qwen2.5-7B
Label	Anomaly	Anomaly	Anomaly
Action	Joint Space Anomaly	Local Structural Jump Anomaly	Local Structural Jump Anomaly
Thought	The multivariate ECG signals (ECG1, ECG2, and ECG3) exhibit a significant deviation from their typical patterns toward the end of the sequence. Initially, all three signals maintain relatively stable dynamics with small fluctuations, consistent with normal cardiac activity. However, starting around time step 50, there is a sharp and synchronized drop in all three signals, followed by irregular and inconsistent recovery patterns. ECG1 shows a dramatic drop into negative values, which is highly unusual for an ECG signal. ECG2 also exhibits a steep decline, reaching abnormally low values before recovering. ECG3, while less extreme, also shows a marked deviation from its prior stable behavior. The synchronized nature of these deviations across all three channels suggests a system-level inconsistency in the multivariate space, rather than isolated anomalies in individual signals. This points to a breakdown in the joint configuration of the signals, potentially caused by a pathological cardiac event or a systemic failure in the recording apparatus.	ECG1 shows a significant drop to very low values around time steps 49-59, including a sharp negative peak at time step 52 (-0.12, -0.2, -0.39, -0.51, -0.47, -0.29, -0.2, -0.14, -0.08, -0.14, -0.08). ECG2 and ECG3 also show some fluctuations but do not exhibit such extreme values. This sudden and dramatic change in ECG1's values is unusual and could indicate a sensor malfunction, a sudden physiological change, or an artifact in the signal.	ECG1 shows a significant drop to very low values around time steps 44-52, including extreme negative values like -0.51 and -0.67. ECG2 and ECG3 show minor fluctuations but do not exhibit such dramatic changes. This sudden and severe deviation in ECG1 is not mirrored in the other ECG signals, indicating a localized anomaly.

Figure 11: Multivariate time series failure case.