Position: Empowering Time Series Reasoning with Multimodal LLMs

Anonymous ACL submission

Abstract

Understanding time series data is crucial for multiple real-world applications. While large language models (LLMs) show promise in time series tasks, current approaches often rely on numerical data alone, overlooking the multimodal nature of time-dependent information, such as textual descriptions, visual data, and audio signals. Moreover, these methods underutilize LLMs' reasoning capabilities, limiting the analysis to surface-level interpretations instead of deeper temporal and multimodal reasoning. In this position paper, we argue that multimodal LLMs (MLLMs) can enable more powerful and flexible reasoning for time series analysis, enhancing decision-making and real-world applications. We call on researchers and practitioners to leverage this potential by developing strategies that prioritize robust reasoning in MLLMs. Lastly, we highlight key research directions, including novel reasoning paradigms, architectural innovations, and domain-specific applications, to advance time series reasoning with MLLMs.

1 Introduction

Time series analysis has long been a cornerstone of real-world applications across domains such as finance, healthcare, and energy. Prior to the rise of large language models (LLMs), research in this area predominantly focused on classic tasks such as forecasting and anomaly detection (Nie et al., 2022; Yang et al., 2023b). While tasks such as explainable time series and dependency analysis existed before the emergence of LLMs, they primarily relied on numerical data. The key shift with LLMs lies in their ability to incorporate rich contextual information beyond pure numerical representations (Wang et al., 2024b; Niu et al., 2024; Liu et al., 2024a; Aksu et al., 2024; Zhang et al., 2024b). Additionally, researchers continue exploring classic tasks leveraging LLMs to enhance these

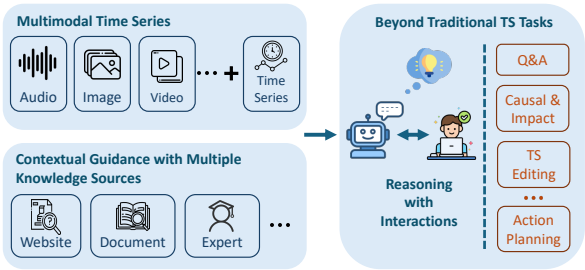

Figure 1: MLLMs integrate multimodal time series and external knowledge, enhancing reasoning and expanding time series tasks.

traditional approaches (Jin et al., 2023a,b). However, these efforts are often limited in scope, focusing narrowly on tasks like forecasting rather than advancing broader reasoning and inference capabilities based on extra contextual information (Jin et al., 2024; Zhou and Yu, 2024; Su et al., 2024).

Deeper reasoning and contextual understanding in time series analysis are critical for identifying patterns, causal relationships, and subtle contextual dynamics (Hamilton, 2020; Fatemi and Gowda, 2024; Hu et al., 2024). These subtle contextual dynamics may include shifts in temporal dependencies, latent external influences, or evolving structural patterns that are not easily discernible through conventional numerical analysis. However, most current research treats time series as purely numerical input, overlooking the inherently multimodal nature of real-world and time-dependent contexts (Zhang et al., 2025). In practice, time series are frequently accompanied by complementary data streams (e.g., text and images) that provide additional layers of information. Most existing systems do not fully exploit this multimodal richness, leaving a considerable gap in achieving robust reasoning for more complex time series tasks (Merrill et al., 2024; Wang et al., 2024c; Zhou et al., 2025).

To bridge this gap, we believe that it is crucial to develop the next generation of multimodal large language model (MLLM) frameworks that can in-

tegrate multiple sources of time-dependent data, thereby unlocking richer insights and more powerful decision-making abilities (Wang et al., 2024d). Figure 1 illustrates this novel integration, where an MLLM fuses multiple modalities and external knowledge to enhance reasoning and tackle various time series tasks.

Our Position. Given the growing need for advanced time series reasoning in real-world applications, we believe that **MLLMs can enable more powerful and flexible reasoning for time series analysis, enhancing decision-making and real-world applications.** To support this central claim, in this position paper, we propose a framework that goes beyond traditional methods and address three key points: ***(1) A New Reasoning Paradigm*** – We define time series reasoning, highlight its essential components, and discuss both current and prospective architectures to enable deeper inference. ***(2) Beyond Traditional Tasks*** – We illustrate how time series reasoning, coupled with MLLMs, opens doors to novel tasks that go beyond the scope of classical tasks, demonstrating the broader real-world relevance. ***(3) Resources and Future Directions*** – We review existing resources, identify unresolved challenges among datasets, benchmarks, and evaluations, and emphasize the need for robust multimodal training strategies to further advance time series reasoning.

2 Time Series Reasoning

2.1 What is Time Series Reasoning?

Time series reasoning refers to the open-ended ability of an MLLM to process and interpret time series with human-like logic. It captures temporal structures, trends, and patterns to generate precise and interpretable results across various time series tasks, delivering insights in clear and natural language. Unlike traditional time series methods focused on specific goals, time series reasoning unifies these tasks into an integrated framework. It combines context-awareness, time series characteristics, and advanced inference to provide deeper insights, enhanced interpretability, and the ability to handle complex tasks requiring external information beyond the time series itself. For further definitions and references, please refer to Appendix A.

2.2 Types of Time Series Reasoning

The role of time series reasoning is to understand how sequence patterns change over time and to explain the mechanisms behind these changes for more informed decision-making. Numerous approaches exist to model temporal dependencies. However, for conducting time series reasoning, it is crucial to consider both how the reasoning process is structured and what the analysis aims to achieve. Therefore, we can naturally categorize time series reasoning based on two perspectives: *Reasoning Structure* and *Task Objective*.

Here, we first formally define time series reasoning. Consider a univariate time series $\mathbf{x} = (x_0, x_1, \ldots, x_{T-1})$ of length T, where $\mathbf{x} \in \mathbb{R}^T$. Let M be an MLLM that takes as input (i) the time series $\mathbf{x}$ and (ii) a sequence of context tokens $\mathbf{c}$, which may encode additional information such as domain knowledge or prompts. The model M produces an output sequence $\mathbf{y}$, which can include numeric tokens representing time series values, textual tokens providing explanations or descriptions of $\mathbf{x}$, and tokens for other modalities. The model defines a probability distribution over the output sequence $\mathbf{y}$, conditioned on the inputs $\mathbf{x}$ and $\mathbf{c}$, as $P_M(\mathbf{y} \mid \mathbf{x}, \mathbf{c}) = \prod_{t=1}^{|\mathbf{y}|} P_M(y_t \mid y_1, \ldots, y_{t-1}, \mathbf{x}, \mathbf{c})$, where each token y_t is predicted based on the preceding tokens $(y_1, \ldots, y_{t-1})$ with time series $\mathbf{x}$ and context $\mathbf{c}$.

Reasoning Structure. A reasoning structure is a set of systematic steps that connects initial observations to final conclusions (Wei et al., 2022; Chu et al., 2023). In time series reasoning, it provides a transparent roadmap, showing how each inference is derived and how input factors interact to shape results. Reasoning structure can be categorized into four types. *(1) End-to-end reasoning* directly maps inputs to outputs, prioritizing efficiency over interpretability by skipping intermediate steps. *(2) Forward reasoning* adopts a bottom-up approach, building solutions step-by-step with explicit intermediate steps, making it suitable for tasks like math problems or trend prediction. *(3) Backward reasoning* uses a top-down approach, breaking problems into smaller sub-problems, often applied in diagnostic tasks to trace the causes of anomalies. Finally, *(4) Forward-backward reasoning* combines both approaches, employing forward reasoning to propose solutions and backward reasoning to validate or refine them, making it particularly useful for iterative tasks like anomaly detection. Detailed descriptions of the above are provided in Appendix B for further clarity.

Table 1: Comparison of reasoning types for task objectives with definition, example, and mathematical formulation.

Reasoning Type	Definition	Example	Mathematical Formulation
Deductive Reasoning	Derives specific, logically consistent conclusions from general principles or rules.	Confirming a 25% sales spike aligns with a known rule that holiday promotions increase sales by 20–30%.	$P_M(\mathbf{y}^+ \mid \mathbf{x}, \mathbf{c}) > P_M(\mathbf{y}^- \mid \mathbf{x}, \mathbf{c})$, where $\mathbf{y}^+$ represents a correct conclusion and $\mathbf{y}^-$ an incorrect one.
Inductive Reasoning	Infers general rules or patterns from specific observations.	Hypothesizing that sales spikes every December indicate seasonality.	$P_M(\mathbf{r}^+ \mid \mathbf{x}, \mathbf{c}) > P_M(\mathbf{r}^- \mid \mathbf{x}, \mathbf{c})$, where $\mathbf{r}^+$ represents a valid generalization and $\mathbf{r}^-$ an invalid one.
Etiological Reasoning	Identifies and explains underlying causes or mechanisms behind observed patterns.	Explaining a sudden drop in energy consumption by analyzing potential causes.	$P_M(\mathbf{d}^+ \mid \mathbf{x}, \mathbf{c}) > P_M(\mathbf{d}^- \mid \mathbf{x}, \mathbf{c})$, where $\mathbf{d}^+$ represents a correct explanation and $\mathbf{d}^-$ an incorrect one.
Causal Reasoning	Measures causal effect relationships and requires clear demonstration of true causal impacts beyond etiological reasoning.	Evaluating if a new pricing strategy caused sustained sales increases by comparing pre- and post-policy data.	$P_M(\mathbf{h}^+ \mid \mathbf{x}, \mathbf{c}) > P_M(\mathbf{h}^- \mid \mathbf{x}, \mathbf{c})$, where $\mathbf{h}^+$ denotes a correct causal statement (such as $\mathbf{x}$ causes $\mathbf{y}$) and $\mathbf{h}^-$ a non-causal or incorrect statement.
Analogical Reasoning	Identifies similarities between different time series or contexts.	Comparing sales patterns across stores to infer shared seasonality.	$P_M(\mathbf{a}^+ \mid \mathbf{x}, \mathbf{x}', \mathbf{c}) > P_M(\mathbf{a}^- \mid \mathbf{x}, \mathbf{x}', \mathbf{c})$, where $\mathbf{a}^+$ is a correct analogy, $\mathbf{a}^-$ a flawed analogy, and $\mathbf{x}'$ the reference time series.
Counterfactual Reasoning	Explores "what-if" scenarios by imagining alternative conditions or interventions.	Analyzing what sales would look like without a promotional discount in June.	$P_M(\mathbf{u}^+ \mid \mathbf{x}, \mathbf{c}) > P_M(\mathbf{u}^- \mid \mathbf{x}, \mathbf{c})$, where $\mathbf{u}^+$ is plausible outcome under counterfactual condition and $\mathbf{u}^-$ an implausible one.
Mathematical Reasoning	Applies mathematical techniques or logical derivations to interpret data.	Proving stationarity by examining bounded partial sums over time.	$P_M(\mathbf{m}^+ \mid \mathbf{x}, \mathbf{c}) > P_M(\mathbf{m}^- \mid \mathbf{x}, \mathbf{c})$, where $\mathbf{m}^+$ is correct mathematical derivation and $\mathbf{m}^-$ incorrect or logically inconsistent one.
Abductive Reasoning	Forms the most plausible explanation for observations when evidence is incomplete.	Proposing a supply chain issue as the cause for an unexplained sales dip.	$P_M(\mathbf{a}^+ \mid \mathbf{x}, \mathbf{c}) > P_M(\mathbf{a}^- \mid \mathbf{x}, \mathbf{c})$, where $\mathbf{a}^+$ is a plausible abductive explanation and $\mathbf{a}^-$ an implausible or less likely one.

Task Objective. In time series reasoning, the primary goal is to extract meaningful and actionable insights from temporal and other complementary multimodal data. Depending on the objective - such as explaining anomalies, forecasting, or identifying causal factors - different reasoning types can be applied. For instance, etiological reasoning identifies causes of sudden shifts, while deductive reasoning confirms hypotheses about periodicity. The task objective determines the appropriate reasoning type and its application. Table 1 highlights commonly used reasoning types along with examples of their application to time series analysis.

2.3 Key Components for Achieving Time Series Reasoning

Robust time series reasoning often demands a comprehensive understanding of temporal patterns and the integration of relevant task-related contextual information. Accordingly, we propose four essential components for achieving effective time series reasoning: *(1) Understanding Time Series Characteristics*, *(2) Contextual Guidance*, *(3) Reasoning Process*, and *(4) Iterative Feedback* (illustrated in Figure 2).

Understanding Time Series Characteristics. One essential aspect of time series reasoning is identifying patterns such as seasonality, trends, and abrupt fluctuations, which are vital for analyses and decision-making. In MLLMs, numerical time series data are often converted into textual tokens, enabling integration with textual or multimodal inputs and leveraging the model's language capabilities. However, this can reduce the ability to recognize temporal patterns due to differences in how LLMs tokenize numerical data (Gruver et al., 2024). A key debate centers on using customized tokenization or encoding methods to represent numerical time series data (Talukder et al., 2024). Customized tokenization offers flexibility by aligning representation with the model architecture and integrating well with language tasks but may introduce repetitive symbols for numerical values. Encoding methods, while preserving local fluctuations and long-range dependencies, can disrupt the natural flow of language processing. Choosing between these approaches requires balancing interpretability, computational efficiency, and the preservation of temporal relationships. Addressing these challenges and improving methods remain critical areas for innovation.

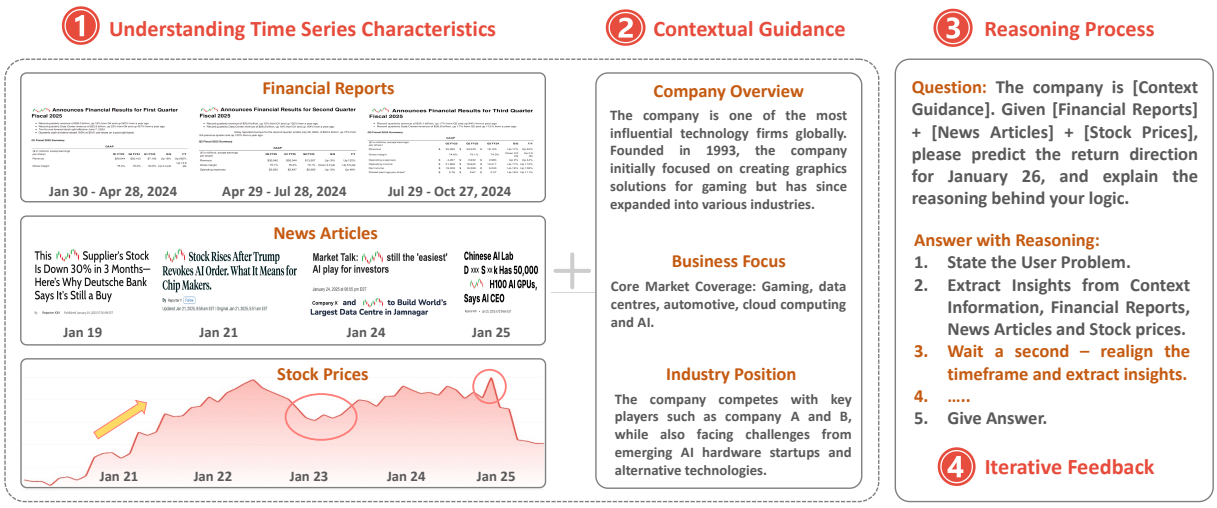

Figure 2: Key components for achieving time series reasoning (illustrated with financial time series example).

To fully leverage multimodal integration in time-dependent data, it is crucial to understand correlated features across multimodal time series. In healthcare, electronic health records (EHR), wearable device data, and patient-reported outcomes often influence each other with time lags. For example, abnormalities in heart rate variability might precede changes in EHR metrics such as blood pressure by days, while self-reported fatigue might occur before alterations in wearable or EHR data. Capturing these lag effects is a key challenge and opportunity for MLLMs. Using mechanisms like temporal attention (Rosin and Radinsky, 2022) or dynamic temporal graph networks (Rossi et al., 2020), MLLMs can infer lagged relationships, enabling proactive decision-making and personalized insights across domains.

Contextual Guidance. Context plays a crucial role in guiding time series reasoning by providing additional knowledge for interpreting patterns and shaping forecasting outcomes. Time series data rarely exists in isolation, and the same data can lead to vastly different predictions depending on the context (Requeima et al., 2024; Williams et al., 2024). This context may come from internal features like seasonality, trends, or anomalies, or external sources such as economic indicators, news events, or environmental factors. Since internal patterns often interact with external influences, incorporating contextual information could significantly improve data comprehension and decision-making. However, integrating external context presents challenges, as factors like policy changes, market shifts, or global events are often sporadic, unstructured, and difficult to quantify. Even when accessible, such data may be inconsistent or incomplete, potentially undermining analysis. Overcoming these obstacles is essential to advancing time series reasoning and enabling more informed decisions.

Reasoning Process. In Section 2.2, we introduced various approaches to time series reasoning, where the choice of method depends on the complexity and goals of the analysis. For instance, combining backward reasoning with etiological reasoning is effective for tracing the causes of anomalies, while forward reasoning with an inductive approach is better suited for identifying recurring seasonal patterns. Moreover, rather than focusing on a single task, time series reasoning aims to unify multiple objectives within an integrated framework, which introduces several challenges. First, maintaining consistent logic across reasoning steps is difficult, as earlier conclusions must remain valid when addressing subsequent goals. Second, incorporating external context without introducing spurious correlations can be complex. Lastly, as reasoning spans multiple tasks, articulating each inference step clearly becomes harder, especially when new information necessitates revising earlier conclusions.

Iterative Feedback. To tackle the above challenges in the reasoning process, iterative feedback and refinement become quite important. It allows the model to incrementally improve its reasoning by identifying inconsistencies, revising intermediate conclusions, and incorporating new information. This process can be facilitated through several methods: leveraging LLM agents to evaluate and critique reasoning steps, embedding self-evaluation

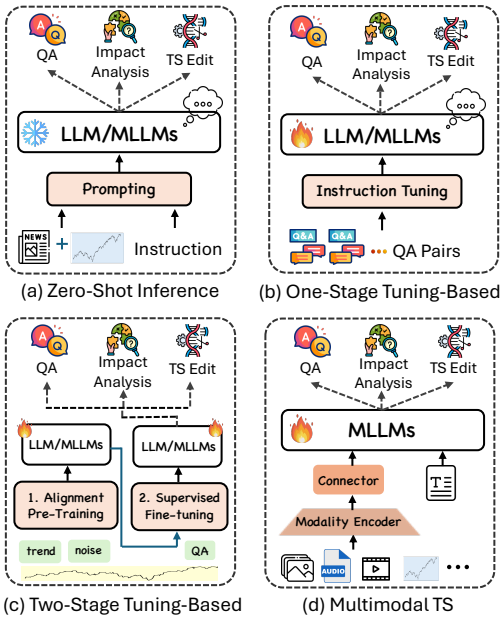

Figure 3: Different categories of advanced time series reasoning task and architectures.

mechanisms within the model to detect and resolve potential errors, or integrating a feedback loop directly into the model's architecture to allow it to adjust based on performance metrics.

2.4 Promising Model Design

There are generally four model design ideas for advanced reasoning tasks that either leverage the built-in reasoning capabilities of LLM/MLLMs, design a time-series MLLM, or fully utilize multimodal inputs and capabilities. We categorize these methods as *Zero-Shot Inference*, *One-Stage Tuning-Based*, *Two-Stage Tuning-Based*, and *Multimodal Time Series* approaches (illustrated in Figure 3).

Zero-Shot Inference. LLM/MLLMs inherently possess zero-shot reasoning abilities, enabling them to generate insights into temporal patterns through direct prompting using their built-in knowledge. Incorporating different types of reasoning structures (as discussed in Section 2.2) within prompts can enhance the quality of reasoning and interpretations. Additionally, the in-context learning approach — providing a small set of question-answer pairs with accompanying rationale — can further refine the model's temporal reasoning capabilities.

One-Stage Tuning-Based. A time-series MLLM is fine-tuned using a systematically compiled dataset of ⟨instruction, response⟩ pairs - often involving time series data - to align model behavior with human objectives. This process, known as instruction tuning, mitigates issues such as untruthful, unhelpful, or unsafe responses by guiding the model to produce accurate and contextually relevant outputs. A key aspect of this strategy involves formulating clear instructions that capture user goals and providing detailed, precise answers. To enhance reasoning capability, the dataset's instructions may sometimes include different types of reasoning structures to guide the thinking process, and the corresponding responses may feature a detailed explanation of how the final answer was derived. The model parameters are updated under a supervised loss function, computed from the generated response tokens. This approach ensures the model can better generalize across diverse tasks as well as strengthen its reasoning and explanatory capabilities.

Two-Stage Tuning-Based. This approach begins by establishing an initial alignment between text and time series modalities, followed by a supervised fine-tuning stage. In the first step, the model is trained to map textual descriptions to corresponding temporal attributes, ensuring a robust linkage between linguistic concepts and time series features. Building on this foundation, the second step focuses on supervised fine-tuning, where the model is optimized for question-answering and reasoning tasks over the aligned modalities. By separating the alignment process from the fine-tuning phase, this two-stage approach aims to both equip the model with a strong multimodal representation of time series data and facilitate contextually accurate, inference-driven responses (Xie et al., 2024).

Multimodal Time Series. Multimodal time series data include not only numerical sequences but also other temporal modalities. Tackling complex tasks that require robust reasoning demands a model design that integrates diverse modalities to enhance the capabilities of an MLLM. Addressing advanced reasoning capabilities in this domain involves three key components. First, a *modality encoder* transforms various raw inputs – such as numerical data, images, or audio - into meaningful embeddings. This step often leverages pre-trained models (e.g., CLIP for images) to efficiently capture domain-specific features (Radford et al., 2021). Second, a *modality interface* aligns these embeddings into a unified, text-like representation, ensuring seamless integration with the language-based reasoning engine. Finally, the *LLM backbone*

serves as the central reasoning system, synthesizing the transformed inputs to perform advanced analysis and decision-making.

Our position (1/3): *We advocate that, to achieve robust time series reasoning, we should leverage MLLMs to capture the unique temporal patterns inherent in sequential data, integrate relevant external knowledge, adopt appropriate multi-step inference strategies, and incorporate iterative feedback loops for continuous refinement. Future directions should seamlessly unifying time series reasoning with the flexible capabilities of MLLMs by addressing four key areas: (1) Understanding Time Series Characteristics to effectively interpret temporal data, (2) Contextual Guidance through the integration of external knowledge, (3) Reasoning Process by adopting suitable reasoning types, and (4) Iterative Feedback for continuous improvement.*

3 Beyond Classical Time Series Tasks

The integration of MLLMs and time series reasoning has inspired new tasks beyond traditional time series tasks (illustrated in Figure 4). These include *question answering*, *causal inference & impact analysis*, and *time series generation & editing*, which focus on reasoning and creative manipulation of time series. This section introduces these tasks, highlighting their distinct views and applications.

3.1 Question Answering

Time series-based Question Answering (QA) represents a shift from classical analysis to high-level reasoning, where a time series serves as the primary input, optionally enriched with multimodal data like text, images, or structured data for alignment (e.g., event timestamps) or context enrichment (e.g., supplementary details). The core of QA lies in answering open-ended questions posed by users based on input multimodal time series. Moreover, other modalities beyond the time series play an indispensable role in this novel task. This is particularly impactful in healthcare, where multimodal integration (e.g., clinical notes, imaging, wearable device data) facilitates holistic patient assessments, enabling real-time critical care insights or chronic disease trend analysis. In Appendix C, we analyze the QA tasks and evaluate the reasoning using different MLLMs for real healthcare data in Figure 5. Besides, integrating meteorological data with satellite imagery helps uncover patterns in extreme weather, while combining sensor data and maintenance logs to identify causes of system failures in industry. Leveraging MLLMs, this approach bridges raw time series data and actionable insights by aligning inputs, modeling temporal dependencies, and interpreting patterns.

3.2 Causal Inference and Impact Analysis

Time series causal inference and impact analysis focus on uncovering causal relationships and quantifying the effects of specific events or interventions (Moraffah et al., 2021). This task often involves integrating time series data with additional modalities, such as text or tabular data, which can provide alignment or supplementary context and uncover nuanced causal relationships that may not be apparent from time series alone. For instance, in finance, combining stock price sequences with company real-time financial reports enables analysis of how financial performance impacts market fluctuations (Kong et al., 2024b,a). In Appendix Figure 6, we compare the causal inference and impact analysis task and evaluate the reasoning using different MLLMs for real finance data with and without financial reports. Similarly, in healthcare, integrating electronic health records with external environmental data helps evaluate the causal effects of interventions, such as new drug treatments, on patient outcomes. Applications extend to marketing, where promotional timelines are analyzed alongside sales data to assess campaign effectiveness, and public policy, where the impacts of reforms on economic indicators like employment rates are quantified.

3.3 Time Series Generation and Editing

Time series generation and editing focus on synthesizing or modifying time series, with inputs optionally enriched by other modalities like text, images, or structured data to provide alignment or supplemental information (Narasimhan et al., 2024; Jing et al., 2024). This task is widely applicable in weather forecasting and urban management. For instance, generating synthetic weather data helps simulate extreme climate scenarios and explore the reasons behind extreme weather, while editing traffic flow data allows urban planners to assess the impact of infrastructure changes. Time series reasoning is fundamental to this process, ensuring that generated or edited data aligns with logical tempo-

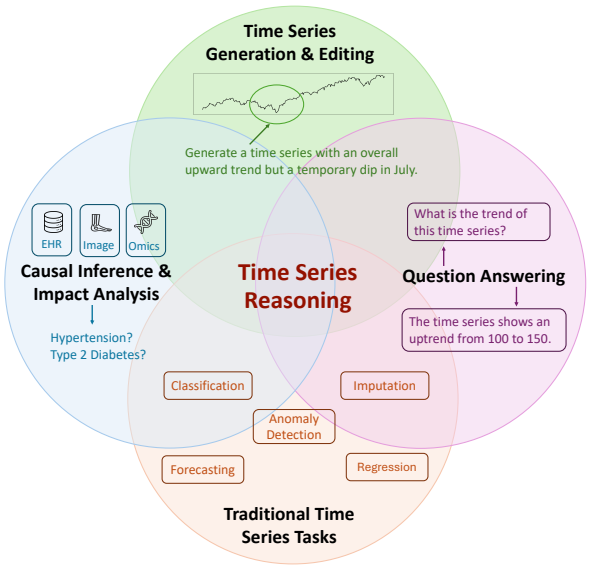

Figure 4: Illustration of diverse time series tasks — including traditional tasks, question answering, causal inference and impact analysis, as well as time series generation and editing — highlighting the emerging era of time series reasoning.

ral dependencies and contextual relationships, such as maintaining seasonal weather patterns or capturing the cause-effect dynamics of urban systems. Multimodal inputs further enhance these tasks by providing additional context-satellite imagery that can guide the generation of weather time series, while map data or demographic statistics inform traffic simulations - ensuring the outputs are realistic and actionable. Moreover, we can leverage generation or editing techniques for improved time series imputation, utilizing conditions provided by information from other modalities. By incorporating reasoning from these modalities, we can achieve more realistic generation and imputation results. We show in Appendix Figure 7 an example of generating electrical data that demonstrates the usefulness of other modes for time series editing and generation.

Our position (2/3): *We advocate for continued innovation in this intersection, particularly in high-stakes domains like healthcare, finance, and industrial systems, where robust reasoning and multimodal integration are critical. Future directions could extend beyond traditional time series tasks to explore more novel applications, including question answering, causal inference and impact analysis, and time series generation and editing, ultimately leading to enhanced time series reasoning with multimodal LLMs.*

4 Resources and Challenges

Dataset and Benchmark. There is a notable shortage of publicly available datasets and codes in this area of research. To address this, we summarize several existing datasets in Table 2. However, several opportunities and challenges remain. Many datasets are artificially generated by GPT models or LLMs, and standard evaluation methods for these generated questions are lacking. Additionally, while most datasets pair numerical time series with textual descriptions, they often lack multimodal representations incorporating additional modalities, limiting their broader applicability. Furthermore, datasets that naturally merge time series into textual information are limited. Finally, existing datasets primarily focus on forward reasoning structure, such as chain-of-thought approaches, leaving opportunities to explore more diverse reasoning processes in future research.

Evaluation Metrics. Reasoning is often intangible and highly subjective, making it relatively difficult to evaluate. Most existing research compares the outcomes of different LLMs and measures their accuracy, which is the approach commonly applied to multiple-choice or true/false questions (Chang et al., 2024). However, the methods for quantifying reasoning vary based on the specific tasks, datasets, and types of reasoning involved. For example, in QA tasks requiring inductive reasoning, answers are evaluated using RAGAS, a keyword-matching approach through LLM-based fuzzy matching (Es et al., 2023). To assess both forecast accuracy and the integration of contextual information, the Region of Interest CRPS (RCRPS) metric was introduced, which prioritizes context-sensitive windows in the prediction and accounts for constraint satisfaction (Williams et al., 2024). At present, there is no standard evaluation metric in the time series reasoning field. Future research should address this gap by designing task-specific metrics that can evaluate not only the accuracy of answers but also the underlying reasoning process.

Training Strategy. Current and potential model designs for trading strategies are detailed in Section 2.4. One potential area for improvement lies in the integration of explicit reasoning processes into the training phase. Currently, most approaches focus on including reasoning structure only within the question-and-answer pairs, without fully embedding reasoning mechanisms into the training

Table 2: The time series reasoning multimodal LLM datasets and benchmarks.

Dataset	Time Series	Text	Image	URL	Application Domain	Task Type	Algorithm	Open-source
Jolt (Cai et al., 2023)	✔	✔	✘	✘	Healthcare	Representation	LLM-based	✘
TS-Insights (Zhang et al., 2023)	✔	✔	✘	✘	General	Trend Classification	LLM-based	✔
Time-series-reasoning (Merrill et al., 2024)	✔	✔	✘	✘	General	Q&A	-	✔
ChatTS (Xie et al., 2024)	✔	✔	✘	✘	Weather&AIOps	Q&A	MLLM-based	✔
TimeMMD (Liu et al., 2024b)	✔	✔	✘	✔	General	Forecasting	LLM-based	✔
Finst (Dong et al., 2024)	✔	✔	✘	✔	Finance	Forecasting	LLM-based	✔
CiK (Williams et al., 2024)	✔	✔	✘	✘	General	Forecasting	LLM-based	✔
ChatTime (Wang et al., 2024a)	✔	✘	✘	✘	Traffic	Q&A	LLM-based	✔
TabPFN (Hollmann et al., 2025)	✔	✘	✘	✘	General	Forecasting	LLM-based	✔
Time-MQA (Kong et al., 2025)	✔	✔	✘	✔	General	Universal	-	✔

process itself. This leaves open opportunities to explore whether the reasoning embedded in these pairs is optimal and how incorporating more detailed and high-quality reasoning could play a greater role in training. Such advancements could enhance model performance and decision-making capabilities, offering promising directions for future development in trading strategy models.

Our position (3/3): *We advocate for continued progress across multimodal datasets, evaluation metrics, and training strategies to advance time series reasoning. Future directions should develop more realistic and multimodal benchmark datasets to support diverse reasoning structures, employ specialized evaluation metrics that assess both outcomes and underlying reasoning processes, and integrate explicit reasoning into training to strengthen model performance, decision-making, and interpretability.*

5 Further Discussion

Hallucination. It is a persistent issue in MLLMs, leading to inaccurate results in time series analysis, especially in critical fields like finance and healthcare. Incorporating reasoning mechanisms into MLLMs offers a solution by enabling the model to understand causal relationships and context, allowing it to cross-check and validate outputs. This reasoning process can also flag and correct hallucinations, ensuring more reliable and accurate analyses.

Environmental and Computational Cost. The integration of MLLMs into time series analysis may introduce challenges related to scalability and computational complexity. Critics highlight the environmental and computational costs, as training and deploying these models require substantial resources, particularly when handling large-scale, high-precision numerical data and in domains like finance and science. Addressing these challenges necessitates strategies to alleviate computational burdens, such as optimizing MLLMs, refining time-series data processing pipelines, and developing efficient alignment and inference mechanisms to enhance scalability while reducing overhead.

Data Confidentiality and Operational Constraints. MLLMs face challenges in time series analysis, including data confidentiality, as sensitive data like financial transactions or patient records cannot be shared with external services. Real-time forecasting, such as in wind power management, demands low latency. Additionally, cloud-based MLLMs also struggle in remote areas due to connectivity issues. Local deployment of open-source models ensures data control, real-time processing, and offline operation. Future research should focus on developing high-quality, accessible open-source MLLMs for time series analysis.

6 Conclusion

This position paper highlights the potential of time series reasoning with MLLMs for both researchers and practitioners. Our central position is that MLLMs can enable more powerful and flexible reasoning for time series analysis, enhancing decision-making and real-world applications. To strengthen our position, we propose novel time series reasoning paradigms and introduce new task frameworks that leverage MLLMs to tackle complex temporal challenges. Despite current limitations - such as scarce datasets and the need for more sophisticated evaluation metrics - the integration of MLLMs with time series reasoning represents a noteworthy advancement. Looking ahead, we encourage researchers to focus on developing innovative architectures, refining training methodologies, and establishing comprehensive benchmarks to unlock MLLMs' potential in real-world time series contexts.

Limitations

Conceptual Orientation and Scope of the Current Work. Even though this paper primarily offers a conceptual perspective on empowering time series reasoning through MLLMs, we aimed to provide a coherent theoretical foundation and practical insights that can guide future empirical work. While large-scale validation and implementation are beyond the present scope, we acknowledge that realizing these ideas in practice may involve addressing significant challenges related to data heterogeneity, computational efficiency, and infrastructure readiness across diverse domains.

Transferability and Domain-specific Considerations. Although we outline broad opportunities for MLLMs across various application areas, this position paper does not exhaustively explore how their effectiveness and adaptability might differ across domains such as healthcare, finance, or environmental science. Nevertheless, by proposing a flexible framework, we intend to offer a foundation that can be tailored and refined through future domain-specific research.

Complexities of Interpretability and Trust in Multimodal Reasoning. While this paper emphasizes the importance of building trustworthy and interpretable multimodal time series reasoning systems, we acknowledge that further exploration is needed to fully understand the dynamics of human–LLM collaboration, effective uncertainty communication, and bias mitigation. Even though these elements were not examined in depth within this work, we consider them essential for translating the proposed conceptual vision into safe, transparent, and practically reliable real-world systems.

Ethics Statement

This position paper does not involve the use of human subjects, personally identifiable information (PII), or proprietary datasets. All discussions are conceptual and based on publicly available research and benchmarks. We recognize the ethical importance of fairness, transparency, and accountability in the development of MLLMs for time series reasoning, and we adhere to responsible AI principles consistent with the ACL Code of Ethics.

References

Taha Aksu, Chenghao Liu, Amrita Saha, Sarah Tan, Caiming Xiong, and Doyen Sahoo. 2024. Xforecast: Evaluating natural language explanations for time series forecasting. *arXiv preprint arXiv:2410.14180*.

Maciej Besta, Julia Barth, Eric Schreiber, Ales Kubicek, Afonso Catarino, Robert Gerstenberger, Piotr Nyczyk, Patrick Iff, Yueling Li, Sam Houliston, et al. 2025. Reasoning language models: A blueprint. *arXiv preprint arXiv:2501.11223*.

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

Yifu Cai, Mononito Goswami, Arjun Choudhry, Arvind Srinivasan, and Artur Dubrawski. 2023. Jolt: Jointly learned representations of language and time-series. In *Deep Generative Models for Health Workshop NeurIPS 2023*.

Defu Cao, Furong Jia, Sercan O Arik, Tomas Pfister, Yixiang Zheng, Wen Ye, and Yan Liu. 2023. Tempo: Prompt-based generative pre-trained transformer for time series forecasting. *arXiv preprint arXiv:2310.04948*.

Ching Chang, Wen-Chih Peng, and Tien-Fu Chen. 2023. Llm4ts: Two-stage fine-tuning for time-series forecasting with pre-trained llms. *arXiv preprint arXiv:2308.08469*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

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*.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2023. A survey of chain of thought reasoning: Advances, frontiers and future. *arXiv preprint arXiv:2309.15402*.

Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, et al. 2024. A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 958–979.

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

Wenjie Du, Jun Wang, Linglong Qian, Yiyuan Yang, Zina Ibrahim, Fanxing Liu, Zepu Wang, Haoxin Liu, Zhiyuan Zhao, Yingjie Zhou, et al. 2024. Tsi-bench: Benchmarking time series imputation. *arXiv preprint arXiv:2406.12747*.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2023. Ragas: Automated evaluation of retrieval augmented generation. *arXiv preprint arXiv:2309.15217*.

Mehdi Fatemi and Sindhu Gowda. 2024. A dynamical view of the question of why. *arXiv preprint arXiv:2402.10240*.

Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew G Wilson. 2024. Large language models are zero-shot time series forecasters. *Advances in Neural Information Processing Systems*, 36.

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

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. 2025. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326.

Chi Hu, Yuan Ge, Xiangnan Ma, Hang Cao, Qiang Li, Yonghua Yang, Tong Xiao, and Jingbo Zhu. 2024. Rankprompt: Step-by-step comparisons make language models better reasoners. *arXiv preprint arXiv:2403.12373*.

Yushan Jiang, Kanghui Ning, Zijie Pan, Xuyang Shen, Jingchao Ni, Wenchao Yu, Anderson Schneider, Haifeng Chen, Yuriy Nevmyvaka, and Dongjin Song. 2025. Multi-modal time series analysis: A tutorial and survey. *arXiv preprint arXiv:2503.13709*.

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

Ming Jin, Qingsong Wen, Yuxuan Liang, Chaoli Zhang, Siqiao Xue, Xue Wang, James Zhang, Yi Wang, Haifeng Chen, Xiaoli Li, et al. 2023b. Large models for time series and spatio-temporal data: A survey and outlook. *arXiv preprint arXiv:2310.10196*.

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*.

Baoyu Jing, Shuqi Gu, Tianyu Chen, Zhiyu Yang, Dongsheng Li, Jingrui He, and Kan Ren. 2024. Towards editing time series. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Gebhard Kirchgässner, Jürgen Wolters, and Uwe Hassler. 2012. *Introduction to modern time series analysis*. Springer Science & Business Media.

Yaxuan Kong, Yuqi Nie, Xiaowen Dong, John M Mulvey, H Vincent Poor, Qingsong Wen, and Stefan Zohren. 2024a. Large language models for financial and investment management: Applications and benchmarks. *Journal of Portfolio Management*, 51(2).

Yaxuan Kong, Yuqi Nie, Xiaowen Dong, John M Mulvey, H Vincent Poor, Qingsong Wen, and Stefan Zohren. 2024b. Large language models for financial and investment management: Models, opportunities, and challenges. *Journal of Portfolio Management*, 51(2).

Yaxuan Kong, Yiyuan Yang, Yoontae Hwang, Wenjie Du, Stefan Zohren, Zhangyang Wang, Ming Jin, and Qingsong Wen. 2025. Time-MQA: Time series multi-task question answering with context enhancement. In *ACL*.

Martha Lewis and Melanie Mitchell. 2024. Using counterfactual tasks to evaluate the generality of analogical reasoning in large language models. *arXiv preprint arXiv:2402.08955*.

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.

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

Haoxin Liu, Chenghao Liu, and B Aditya Prakash. 2024a. A picture is worth a thousand numbers: Enabling llms reason about time series via visualization. *arXiv preprint arXiv:2411.06018*.

Haoxin Liu, Shangqing Xu, Zhiyuan Zhao, Lingkai Kong, Harshavardhan Kamarthi, Aditya B Sasanur, Megha Sharma, Jiaming Cui, Qingsong Wen, Chao Zhang, et al. 2024b. Time-MMD: Multi-domain multimodal dataset for time series analysis. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Mike A Merrill, Mingtian Tan, Vinayak Gupta, Tom Hartvigsen, and Tim Althoff. 2024. Language models still struggle to zero-shot reason about time series. *arXiv preprint arXiv:2404.11757*.

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

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. 2023. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265.

Raha Moraffah, Paras Sheth, Mansooreh Karami, Anchit Bhattacharya, Qianru Wang, Anique Tahir, Adrienne Raglin, and Huan Liu. 2021. Causal inference for time series analysis: Problems, methods and evaluation. *Knowledge and Information Systems*, 63:3041–3085.

Sai Shankar Narasimhan, Shubhankar Agarwal, Oguzhan Akcin, Sujay Sanghavi, and Sandeep Chinchali. 2024. Time weaver: A conditional time series generation model. *arXiv preprint arXiv:2403.02682*.

Yuqi Nie, Yaxuan Kong, Xiaowen Dong, John M Mulvey, H Vincent Poor, Qingsong Wen, and Stefan Zohren. 2024. A survey of large language models for financial applications: Progress, prospects and challenges. *arXiv preprint arXiv:2406.11903*.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. 2022. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*.

Shuai Niu, Jing Ma, Liang Bai, Zhihua Wang, Yida Xu, Yunya Song, and Xian Yang. 2024. Multimodal clinical reasoning through knowledge-augmented rationale generation. *arXiv preprint arXiv:2411.07611*.

Willa Potosnak, Cristian Challu, Mononito Goswami, Michał Wiliński, Nina Żukowska, and Artur Dubrawski. 2024. Implicit reasoning in deep time series forecasting. *arXiv preprint arXiv:2409.10840*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

James Requeima, John Bronskill, Dami Choi, Richard E Turner, and David Duvenaud. 2024. Llm processes: Numerical predictive distributions conditioned on natural language. *arXiv preprint arXiv:2405.12856*.

Guy D Rosin and Kira Radinsky. 2022. Temporal attention for language models. *arXiv preprint arXiv:2202.02093*.

Emanuele Rossi, Ben Chamberlain, Fabrizio Frasca, Davide Eynard, Federico Monti, and Michael Bronstein. 2020. Temporal graph networks for deep learning on dynamic graphs. *arXiv preprint arXiv:2006.10637*.

Xiaoming Shi, Siqiao Xue, Kangrui Wang, Fan Zhou, James Zhang, Jun Zhou, Chenhao Tan, and Hongyuan Mei. 2024. Language models can improve event prediction by few-shot abductive reasoning. *Advances in Neural Information Processing Systems*, 36.

Robert H Shumway, David S Stoffer, and David S Stoffer. 2000. *Time series analysis and its applications*, volume 3. Springer.

Jing Su, Chufeng Jiang, Xin Jin, Yuxin Qiao, Tingsong Xiao, Hongda Ma, Rong Wei, Zhi Jing, Jiajun Xu, and Junhong Lin. 2024. Large language models for forecasting and anomaly detection: A systematic literature review. *arXiv preprint arXiv:2402.10350*.

Sabera Talukder, Yisong Yue, and Georgia Gkioxari. 2024. Totem: Tokenized time series embeddings for general time series analysis. *arXiv preprint arXiv:2402.16412*.

Francis Tang and Ying Ding. 2024. Are large language models useful for time series data analysis? *arXiv preprint arXiv:2412.12219*.

Chengsen Wang, Qi Qi, Jingyu Wang, Haifeng Sun, Zirui Zhuang, Jinming Wu, Lei Zhang, and Jianxin Liao. 2024a. Chattime: A unified multimodal time series foundation model bridging numerical and textual data. *arXiv preprint arXiv:2412.11376*.

Jiahao Wang, Mingyue Cheng, Qingyang Mao, Qi Liu, Feiyang Xu, Xin Li, and Enhong Chen. 2024b. Tabletime: Reformulating time series classification as zero-shot table understanding via large language models. *arXiv preprint arXiv:2411.15737*.

Xinlei Wang, Maike Feng, Jing Qiu, Jinjin Gu, and Junhua Zhao. 2024c. From news to forecast: Integrating event analysis in llm-based time series forecasting with reflection. *arXiv preprint arXiv:2409.17515*.

Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You, and Hongxia Yang. 2024d. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning. *arXiv preprint arXiv:2401.06805*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

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

Andrew Robert Williams, Arjun Ashok, Étienne Marcotte, Valentina Zantedeschi, Jithendaraa Subramanian, Roland Riachi, James Requeima, Alexandre Lacoste, Irina Rish, Nicolas Chapados, et al. 2024. Context is key: A benchmark for forecasting with essential textual information. *arXiv preprint arXiv:2410.18959*.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and S Yu Philip. 2023. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2247–2256. IEEE.

Yu Xia, Rui Wang, Xu Liu, Mingyan Li, Tong Yu, Xiang Chen, Julian McAuley, and Shuai Li. 2024. Beyond chain-of-thought: A survey of chain-of-x paradigms for llms. *arXiv preprint arXiv:2404.15676*.

Zhe Xie, Zeyan Li, Xiao He, Longlong Xu, Xidao Wen, Tieying Zhang, Jianjun Chen, Rui Shi, and Dan Pei. 2024. Chatts: Aligning time series with llms via synthetic data for enhanced understanding and reasoning. *arXiv preprint arXiv:2412.03104*.

Yiyuan Yang, Ming Jin, Haomin Wen, Chaoli Zhang, Yuxuan Liang, Lintao Ma, Yi Wang, Chenghao Liu, Bin Yang, Zenglin Xu, et al. 2024. A survey on diffusion models for time series and spatio-temporal data. *arXiv preprint arXiv:2404.18886*.

Yiyuan Yang, Rongshang Li, Qiquan Shi, Xijun Li, Gang Hu, Xing Li, and Mingxuan Yuan. 2023a. Sgdp: A stream-graph neural network based data prefetcher. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Yiyuan Yang, Yi Li, Taojia Zhang, Yan Zhou, and Haifeng Zhang. 2021a. Early safety warnings for long-distance pipelines: A distributed optical fiber sensor machine learning approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14991–14999.

Yiyuan Yang, Chaoli Zhang, Tian Zhou, Qingsong Wen, and Liang Sun. 2023b. 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.

Yiyuan Yang, Haifeng Zhang, and Yi Li. 2021b. Long-distance pipeline safety early warning: a distributed optical fiber sensing semi-supervised learning method. *IEEE sensors journal*, 21(17):19453–19461.

Wen Ye, Yizhou Zhang, Wei Yang, Lumingyuan Tang, Defu Cao, Jie Cai, and Yan Liu. 2024. Beyond forecasting: Compositional time series reasoning for end-to-end task execution. *arXiv preprint arXiv:2410.04047*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.

Fei Yu, Hongbo Zhang, Prayag Tiwari, and Benyou Wang. 2024. Natural language reasoning, a survey. *ACM Computing Surveys*, 56(12):1–39.

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.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024a. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*.

Hanyu Zhang, Chuck Arvin, Dmitry Efimov, Michael W Mahoney, Dominique Perrault-Joncas, Shankar Ramasubramanian, Andrew Gordon Wilson, and Malcolm Wolff. 2024b. Llmforecaster: Improving seasonal event forecasts with unstructured textual data. *arXiv preprint arXiv:2412.02525*.

Haochuan Zhang, Chunhua Yang, Jie Han, Liyang Qin, and Xiaoli Wang. 2025. Tempogpt: Enhancing temporal reasoning via quantizing embedding. *arXiv preprint arXiv:2501.07335*.

Yunkai Zhang, Yawen Zhang, Ming Zheng, Kezhen Chen, Chongyang Gao, Ruian Ge, Siyuan Teng, Amine Jelloul, Jinmeng Rao, Xiaoyuan Guo, et al. 2023. Insight miner: A time series analysis dataset for cross-domain alignment with natural language. In *NeurIPS 2023 AI for Science Workshop*.

Pengyuan Zhou, Lin Wang, Zhi Liu, Yanbin Hao, Pan Hui, Sasu Tarkoma, and Jussi Kangasharju. 2024. A survey on generative ai and llm for video generation, understanding, and streaming. *arXiv preprint arXiv:2404.16038*.

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

Xin Zhou, Weiqing Wang, Shilin Qu, Zhiqiang Zhang, and Christoph Bergmeir. 2025. Unveiling the potential of text in high-dimensional time series forecasting. *arXiv preprint arXiv:2501.07048*.

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

A Literature Review

A.1 MLLM-based Time Series Analysis

A.1.1 Multimodal Large Language Model Definition

A Multimodal Large Language Model (MLLM) is an advanced AI system that extends the reasoning capabilities of Large Language Models (LLMs) by enabling them to process, interpret, and generate information across multiple modalities, including text, images, audio, and time-series data (Wu et al., 2023). Unlike traditional LLMs that rely solely on textual data, MLLMs integrate multimodal representations through sophisticated deep-learning architectures, allowing them to perceive and reason about complex relationships between different data types. This capability enhances their performance in tasks such as multimodal question answering, image and video captioning, and medical image analysis, where understanding information from multiple sources is essential (Yin et al., 2023). By leveraging advanced fusion mechanisms, MLLMs generate contextually rich and coherent outputs that go beyond text-based reasoning, making them highly effective in applications requiring comprehensive multimodal understanding (Zhang et al., 2024a).

A.1.2 Time Series Definition

A time series refers to a collection of data points organized in chronological order, representing the progression of one or more variables over time. Formally, a univariate time series is denoted as $\mathbf{x} = (x_0, x_1, \ldots, x_{T-1}) \in \mathbb{R}^T$, where T is the total number of time steps, and each $x_t \in \mathbb{R}$ represents the value of the series at time t. This structure captures the temporal evolution of a single variable. In contrast, a multivariate time series extends this definition to multiple dimensions and is represented as $\mathbf{X} = (\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{T-1}) \in \mathbb{R}^{T \times D}$. Here, D represents the number of features, so each $\mathbf{x}_t \in \mathbb{R}^D$ is a vector containing D values at time t, reflecting the simultaneous evolution of multiple interrelated variables. Time series is fundamental in numerous domains, including finance, healthcare, and environmental monitoring, due to its ability to capture temporal dependencies and trends.

A.1.3 Multimodal Time Series

The increasing availability of heterogeneous data has highlighted the need to better capture the complexity of real-world phenomena. Multimodal time series address this challenge by integrating data from multiple modalities, where each modality represents a distinct type of information, such as images, text, audio, or structured numerical data. By extending traditional single-modal time series analysis, this approach enables the incorporation of diverse and complementary data sources, providing a more comprehensive understanding of complex systems. Formally, a multimodal time series can be represented as $\mathbf{X} = \mathbf{X}^{(m)}{}_{m=1}^{M}$, where M denotes the total number of modalities, and each $\mathbf{X}^{(m)}$ corresponds to the time series for modality m. For instance, $\mathbf{X}^{(m)} = (\mathbf{x}_0^{(m)}, \mathbf{x}_1^{(m)}, \ldots, \mathbf{x}_{T-1}^{(m)})$ represents the sequential data of modality m over T time steps, with $\mathbf{x}_t^{(m)}$ varying in form depending on the modality, such as vectors for numerical data, matrices for images, or sequences for text and audio. Multimodal time series analysis enables a more comprehensive understanding of temporal patterns and interactions across modalities by integrating multiple data sources. This is particularly important in applications such as healthcare, autonomous driving, and multimedia analysis (Moor et al., 2023; Cui et al., 2024; Zhou et al., 2024).

A.1.4 Time Series Classical Tasks

Time series analysis and its various classical tasks are widely applied across real-world domains, such as financial forecasting, healthcare monitoring, traffic flow analysis, climate modeling, industrial predictive maintenance, and AIOps (Shumway et al., 2000; Nie et al., 2024; Yang et al., 2021a, 2023a; Liang et al., 2024). Also, time series analysis encompasses a diverse set of tasks aimed at extracting insights and addressing challenges in temporal data (Hamilton, 2020; Kirchgässner et al., 2012). Among them, common tasks include forecasting, which predicts future values based on historical trends and can be divided into short-term and long-term predictions (Wen et al., 2022), and anomaly detection, which identifies unusual patterns or deviations from expected behavior (Zamanzadeh Darban et al., 2024). Imputation addresses missing or corrupted data points to ensure dataset completeness (Du et al., 2024), while generation creates synthetic time series to replicate statistical properties for data augmentation or scenario simulation (Yang et al., 2024). Other tasks include classification, which assigns categorical labels based on patterns, and regression, which predicts continuous target values (Mohammadi Foumani et al., 2024; Yang

et al., 2021b). In recent years, an increasing number of approaches have explored leveraging multimodal data to enhance classical time series analysis tasks, validating the effectiveness and rationality of these methods (Zhou et al., 2023; Liu et al., 2024b; Gruver et al., 2024; Chang et al., 2023; Cao et al., 2023; Yin et al., 2023). For instance, Time-LLM aligns and reprograms LLMs for time series forecasting through textual input alignment (Jin et al., 2023a), while Time-MMD incorporates additional textual data with Transformer-based models using weighted fusion to perform time series forecasting and potentially other tasks (Liu et al., 2024b). Beyond text modalities, medical data employs supplementary image or tabular data to enhance time series analysis using foundation models (Hollmann et al., 2025; Moor et al., 2023). Moreover, methods leveraging generative models, such as diffusion models, have also been proposed. These approaches inject multiple modalities into the conditional space, improving the robustness of temporal tasks and enhancing generative diversity (Yang et al., 2024).

A.2 Reasoning in NLP and Time Series

A.2.1 Reasoning in NLP

In the field of Natural Language Processing (NLP), reasoning refers to the process of deriving conclusions from textual evidence and logical principles (Besta et al., 2025). It involves tasks such as understanding implicit information, performing logical inferences, and applying commonsense knowledge. Reasoning capabilities are crucial for addressing complex language tasks like natural language inference, multi-hop question answering, and commonsense reasoning (Yu et al., 2024). Types of reasoning include Chain-of-Thought (CoT), which breaks problems into intermediate steps for clarity, deductive reasoning which applies general rules to specific cases, and inductive reasoning which generalizes from observations (Xia et al., 2024). Abductive reasoning identifies the most plausible explanations, while analogical reasoning transfers knowledge based on similarities (Shi et al., 2024; Lewis and Mitchell, 2024). Others include commonsense reasoning, probabilistic reasoning, and causal reasoning (Yu et al., 2024). These approaches enhance the interpretability and performance of NLP systems.

A.2.2 Reasoning in Time Series

Time series analysis tasks traditionally focus on narrower objectives — like forecasting or anomaly detection — each addressed by its own specialized model, often relying solely on numerical patterns within the data. In contrast, time series reasoning with logic integrates multiple tasks under a single, context-aware framework with human-like reasoning (Chow et al., 2024). It readily incorporates domain knowledge and external data sources, providing natural language explanations and causal insights rather than mere numerical outputs (Potosnak et al., 2024). This approach allows time series reasoning to adapt to shifting conditions and novel questions, delving into the "why" behind observed patterns and bridging the gap between automated analysis and real-world decision-making.

Furthermore, reasoning in NLP can enhance time series analysis by enabling models to infer complex temporal patterns and relationships, improving interpretability and decision-making (Potosnak et al., 2024; Cai et al., 2024). By incorporating reasoning capabilities, models can better handle ambiguous or incomplete data, leading to more robust predictions and insights. Currently, few studies have explored this area (Chow et al., 2024; Ye et al., 2024; Xie et al., 2024). However, reasoning in time series analysis remains an underexplored yet promising and impactful field.

B Types of Reasoning

We summarize reasoning structure types in Table 3. It includes four types: End-to-end Reasoning, Forward Reasoning, Backward Reasoning, and Forward-Backward Reasoning. Each type is defined, exemplified, and mathematically formulated, highlighting its distinct characteristics and applications. End-to-end reasoning is characterized by its direct mapping from inputs to outputs, encapsulating the reasoning process within hidden states, which makes it less interpretable but effective for tasks requiring concise outputs. Forward Reasoning, on the other hand, adopts a bottom-up approach, explicitly stating intermediate steps, making it suitable for tasks like solving math problems or predicting time series trends sequentially. Backward Reasoning employs a top-down strategy, breaking down the main problem into smaller subproblems, which is particularly useful for diagnostic tasks such as identifying the causes of anomalies. Lastly, Forward-Backward Reasoning com-

bines forward and backward approaches, proposing potential solutions and verifying them, making it ideal for complex tasks like analyzing time series anomalies.

In addition, these reasoning structures can be further organized into chain-, tree-, or graph-based formalisms to represent the reasoning path more explicitly. A *chain-based* approach arranges the reasoning steps in a sequential, linear fashion, making it straightforward to follow how each conclusion is derived. A *tree-based* approach expands reasoning into branches, allowing multiple concurrent paths and a hierarchical breakdown of complex problems. Meanwhile, a *graph-based* approach generalizes these connections and can accommodate interdependencies and cyclic references among different inference steps.

C The Examples of Beyond Classical Time Series Tasks

This section presents a set of figures that demonstrate the capabilities of different LLMs in handling zero-shot open questions across various application domains, going beyond classical time series tasks. The first figure, Figure 5, focuses on healthcare applications. By providing an ECG time series recording and relevant background information, it examines how ChatGPT-o1 and Deepseek respond without and with other modal information. The models are tasked with making statistical judgments, determining the presence of anomalies in the time series, and identifying potential illnesses. The task is related to abductive reasoning, which infers the most plausible diagnoses from the observed ECG anomalies and employs backward reasoning to trace those anomalous patterns back to their underlying causes. ChatGPT-o1 and Deepseek analyze the data from different perspectives, considering factors like amplitude swings, baseline wander, and the shape of the QRS complex. They both acknowledge that while the data shows potential anomalies, further analysis and clinical correlation are necessary to confirm a diagnosis, and noise or artifacts need to be ruled out. In addition, if we add other modal information, such as image information of normal as well as various abnormal ECG species to ChatGPT-o1, then this LLM can generate more logical as well as closer to the doctor's answers.

Figure 6 delves into financial applications. Here, the input is a time series of Nvidia's stock prices from September 9, 2024 to January 24, 2025. The LLMs are required to perform causal inference and impact analysis regarding the stock price rise or fall and prospects. The task is related to etiological reasoning, which involves tracing observed price movements back to their root causes, and forward–backward reasoning, which iteratively combines forward-looking pattern detection with backward-looking causal diagnosis. Without other modal information, ChatGPT-o1 and Deepseek analyze the overall trend, daily fluctuations, and the factors driving the price changes, such as market expectations, earnings reports, and macroeconomic conditions. When ChatGPT-o1 has access to other modal information (i.e., Financial Reports from Nvidia's official website), it incorporates detailed financial data from the company's reports to provide a more comprehensive analysis of the stock's performance and future prospects.

The third figure, Figure 7, is centered around electrical applications. It presents an incomplete time series of energy consumption from the London Smart Meters Dataset, with missing values represented by "X". The LLMs need to provide statistical analysis and impute the missing values. The task is related to abductive reasoning that provides the most plausible explanation for a missing data point given incomplete evidence, and forward–backward reasoning that combines information from both prior and subsequent readings to propose and then confirm the imputed value. ChatGPT-o1 and Deepseek, without other modal information, use linear interpolation to fill the gaps, with explanations about why this method is suitable for maintaining the continuity of the data and supporting subsequent analyses. When ChatGPT-o1 has other modal information from the news of those days and some useful website links about local weather records, specifically knowing that the data was collected during a summer period with high temperatures, it employs quadratic interpolation to better capture the potentially rapid changes in consumption.

Overall, these figures showcase the diverse ways in which LLMs can handle time series data in different real-world application scenarios, providing valuable insights and analysis for various fields. Moreover, if we add more related modal information, the generated answers will be more accurate and realistic.

Table 3: Comparison of reasoning structure types with definitions, examples, and mathematical formulations.

Reasoning Type	Definition	Example	Mathematical Formula
End-to-end Reasoning	Allows the model to learn a direct mapping from the inputs $(\mathbf{x}, \mathbf{c})$ to the outputs $\mathbf{y}$ without revealing intermediate reasoning steps. The logic chain is encapsulated in the model's hidden states, making it less interpretable.	Effective in descriptive tasks such as brief analysis where intermediate reasoning is not required.	$P_M(\mathbf{y} \mid \mathbf{x}, \mathbf{c}) = \prod_{t=1}^{\lvert\mathbf{y}\rvert} P_M(y_t \mid y_1, \ldots, y_{t-1}, \mathbf{x}, \mathbf{c})$
Forward Reasoning	A bottom-up approach that starts with existing knowledge and applies inference rules step-by-step to derive the solution. Intermediate steps ("chain of thought") are explicitly stated.	Solving a math problem step-by-step or predicting time series trends sequentially.	$P_M(\mathbf{r}, \mathbf{y} \mid \mathbf{x}, \mathbf{c}) = \left[\prod_{j=1}^{k} P_M(r_j \mid r_{<j}, \mathbf{x}, \mathbf{c})\right] \times \left[\prod_{t=1}^{\lvert\mathbf{y}\rvert} P_M(y_t \mid y_{<t}, \mathbf{r}, \mathbf{x}, \mathbf{c})\right]$
Backward Reasoning	A top-down approach that begins with the main problem and breaks it into smaller sub-problems, continuing recursively until solvable with existing knowledge.	Diagnosing an anomaly by identifying key data patterns in reverse order (e.g., starting from the anomaly and tracing its cause).	$P_M(\mathbf{r}, \mathbf{y} \mid \mathbf{x}, \mathbf{c}) = \left[\prod_{j=k}^{1} P_M(r_j \mid r_{j+1}, \ldots, r_k, \mathbf{x}, \mathbf{c})\right] \times \left[\prod_{t=1}^{\lvert\mathbf{y}\rvert} P_M(y_t \mid y_{<t}, \mathbf{r}, \mathbf{x}, \mathbf{c})\right]$
Forward-Backward Reasoning	Combines forward and backward reasoning. Forward reasoning proposes potential solutions, and backward reasoning verifies or refines them by analyzing dependencies.	Analyzing time series anomalies by first identifying possible causes (forward) and then verifying them against the data (backward).	$P_M(\mathbf{r}^F, \mathbf{r}^B, \mathbf{y} \mid \mathbf{x}, \mathbf{c}) = \prod_{j=1}^{k} P_M(r_j^F \mid r_{<j}^F, r_{<*}^B, \mathbf{x}, \mathbf{c}) \times \prod_{m=1}^{\ell} P_M(r_m^B \mid r_{<m}^B, r_{<*}^F, \mathbf{x}, \mathbf{c}) \times \prod_{t=1}^{\lvert\mathbf{y}\rvert} P_M(y_t \mid y_{<t}, \mathbf{r}^F, \mathbf{r}^B, \mathbf{x}, \mathbf{c})$

D Potential Evaluation Methods for Time Series Reasoning

Evaluating time series reasoning, particularly in multimodal contexts, remains an underexplored area. Existing benchmarks and metrics are largely designed for either unimodal time series tasks or general-purpose reasoning, leaving a gap in tools specifically tailored for assessing multimodal time series reasoning performance. In this section, we outline potential directions for evaluation, synthesizing insights from current literature and community feedback.

Task Outcome Metrics As a foundational layer, standard performance metrics for task-specific outcomes, such as accuracy, F1 score, mean absolute error (MAE), or area under the curve (AUC), remain essential. These metrics help assess whether a model produces correct predictions or decisions based on time series input. However, they often fail to capture the quality or correctness of the reasoning process behind the outcomes, especially in complex settings involving multi-hop or temporally-aware inference.

Reasoning Quality Assessment To complement task-level evaluation, it is crucial to assess the fidelity and coherence of the model's reasoning process. We propose incorporating the following: (1) Semantic Coherence: Measuring how logically consistent the reasoning is, either via automated metrics (e.g., textual entailment, coherence models) or through human annotation. (2) Alignment with Human Rationales: Comparing the model's explanations or intermediate steps with human-annotated reasoning paths or expert-generated rationales. (3) Temporal Grounding: Evaluating whether the model's reasoning respects temporal dependencies inherent in the data, such as ordering and causality across time points.

Modality Alignment and Temporal Dependency Evaluation In multimodal settings, alignment across modalities (e.g., time series and text) is key (Liu et al., 2025; Jiang et al., 2025). Potential metrics in this direction include: (1) Cross-Modal Attention Alignment Scores: Quantifying whether attention is paid to temporally and semantically relevant regions across modalities. (2) Temporal Alignment Accuracy: Measuring the correctness of aligning timestamps or temporal events between different modalities. We further propose embedding structured temporal cues (e.g., timestamps, event markers) or utilizing temporal-aware architectures (e.g., dynamic temporal graphs, temporal attention mechanisms) to enhance interpretability

and alignment. The effectiveness of these mechanisms could be evaluated by how well they preserve temporal consistency across modalities.

Human-Centered Evaluation via User Studies To better understand aspects of reasoning that are hard to measure automatically, such as whether the reasoning makes sense, is trustworthy, or seems convincing, we suggest using user studies involving real users or domain experts (Kong et al., 2025). These studies can help evaluate how people actually perceive and interpret the model's reasoning process. Examples include: (1) Reasoning Trace Evaluation: Participants read through the model's reasoning steps and give feedback on how clear, accurate, and logical they are. (2) Comparative Judgment: Users are shown outputs from different models and asked to compare them, judging which one is more helpful, trustworthy, or logically sound based on the temporal data. (3) Trust and Usability Surveys: Participants complete short questionnaires (e.g., using a 1–5 scale) to rate how much they trust the model and whether they find it useful in practical settings, such as healthcare or finance. These types of user studies are especially valuable in high-stakes domains where it's not enough for a model to just be accurate; the reasoning behind its decisions also needs to be understandable and reliable to human users.

E In What Ways Can LLMs/MLLMs be Leveraged for Single and Multimodal Learning?

There is ongoing debate about whether LLMs truly outperform classical time series techniques (Tang and Ding, 2024; Jin et al., 2024). For straightforward numerical data (such as stationary, low-volatility time series) methods like ARIMA, exponential smoothing, and tree-based regressors often deliver accurate and interpretable forecasts with minimal computational overhead. However, real-world problems frequently require integrating external context, such as policy updates, news events, or clinical findings, as well as handling unstructured inputs including written reports, medical images, and social media streams.

In these scenarios, LLM/MLLMs could offer distinctive advantages, as they can natively ingest raw time series measurement alongside free-text descriptions, generate human-readable rationales, and, in multimodal configurations, seamlessly fuse image or audio data with numerical inputs to identify complex cross-modal patterns. In this position paper, we do not claim LLM/MLLMs are a universal solution for every time series scenario. Instead, we see them as tools that complement traditional methods, especially when richer reasoning or multimodal data come into play.

Therefore, to guide the selection of appropriate methods for time series analysis, we propose below framework: (1) Modality-based criteria — use classical models for purely numerical data; adopt LLM-centric architectures for datasets combining time series data with textual information; and deploy MLLMs for multimodal datasets. (2) Resource and interpretability trade-offs — balance computational budgets and latency constraints against the necessity for transparent and interpretable reasoning. (3) Data volume considerations - leverage LLMs' transfer-learning capabilities, particularly for short or sparse time series datasets, to overcome data scarcity and enhance predictive performance.

F Alternative Views

Is Single-Modality Data Sufficient to Advance Time Series Reasoning in Real-World Applications? Single-modality numerical data can sometimes be sufficient - particularly in scenarios where the time series is extremely sparse and relies on simple, clear assumptions. However, this approach often fails to capture the rich contextual factors driving real-world phenomena. For instance, if we only observe a small set of yearly sales data for a product and neglect considerations like untapped markets or slowing innovation cycles, even the most sophisticated models are constrained by the narrow assumptions derived from these limited observations.

In contrast, LLMs excel at integrating diverse sources of information - including textual descriptions, background knowledge, and multimodal time series data. By leveraging these varied inputs, LLMs can uncover deeper causal factors that go beyond the time series alone. This capability enables us to integrate richer contextual information and formulate more flexible assumptions. For example, consider how electric vehicles are shaped by shifting government incentives, rapid technological advancements, and evolving consumer attitudes. By factoring in these contextual elements, time series analysis can yield more reliable predictions and insights. Hence, while single-modality data can suffice in tightly constrained scenarios, employing

a multimodal approach and tapping into LLMs' broader reasoning capabilities enables richer, more accurate time series analysis for complex, real-world applications.

Could LLM/MLLMs Truly Contribute to Time Series Reasoning? While LLMs show promise in time series analysis, they have inherent limitations stemming from their design. As text-based sequence predictors, LLMs may lack an intrinsic understanding of mathematical logic, numerical precision, and temporal dynamics. Although they can mimic patterns from training data, they cannot perform true calculations or provide rigorous domain-specific temporal reasoning. Consequently, it is necessary to integrate explicit reasoning mechanisms - such as causal, analogical, and counterfactual reasoning - into MLLMs, as discussed in Sections 2.3 - 2.4.

Another concern is whether real-world time series might be inadvertently included in LLM pretraining datasets - an issue heightened by the opacity of training sources. Although textual-numerical pairs appear in specialized domains, multimodal time series data are usually absent from standard pretraining corpora. Moreover, real-world deployments often rely on proprietary, domain-specific datasets (e.g., healthcare vitals with clinical annotations), which differ significantly from generic pretraining data. This specificity reduces the likelihood of overlap with undisclosed LLM training sets. However, given the lack of transparency in pretraining, practitioners cannot rule out overlaps entirely, emphasizing the need for rigorous evaluation frameworks (see Section 4).

Despite these challenges, MLLMs offer transformative potential by unifying multimodal time series with contextual guidance. This facilitates deeper temporal reasoning - helping to identify causal links, anomalies, and domain insights — rather than mere pattern matching. To achieve this potential, we advocate for strategies that prioritize interpretability and robust reasoning within MLLMs. Through these advancements, MLLMs could ultimately deliver reliable insights in high-stakes domains such as healthcare and industrial systems.

Prompt:

You are an expert in healthcare especially in the diagnosis of heart disease. We provide a time-series recording and some background information, you should answer some open questions and provide reasons based on the time series analysis and your knowledge of the medical field. Besides, reasoning steps can infer the current situations of time-series records. Let's begin!

The following data represents ECG (Electrocardiogram) signals, which record the ECG readings of patients during each heartbeat and serve as an important tool for diagnosing heart diseases. Using a signal sampled at a frequency of 25 Hz, we can determine whether abnormalities are present in the human body and heart. The input recording is [-1.1009, -3.9968, -4.2858, -4.5066, -4.0224, -3.2344, -1.5661, -0.9923, -0.7547, 0.0423, 0.1490, 0.1835, 0.2949, 0.1902, ..., 0.6882, -0.5199, 0.0394, 0.5603, 0.5384, 0.6569, 0.7875, 0.7240, 0.5558, 0.4763, 0.7738, 1.1196, -1.4362].

Please make some statistical judgments and determine whether there are anomalies in this time series given the information above, and also please determine whether there are some potential illnesses.

ChatGPT-o1 without other modal information: More information required

Basic Statistical Observations: The data ranges roughly from -4.51 mV to +2.16 mV, indicating large amplitude swings. **While negative and positive spikes can be normal ECG components (such as S or R waves), the extent of these fluctuations might reflect motion artifacts or electrode noise.** Because the sampling rate is 25 Hz, the resolution for detailed ECG analysis is limited, and additional methods may be needed to confirm whether these peaks are real signals or artifacts.

Possible Signs of Anomalies: Some pronounced negative and positive deflections could imply abnormal QRS complexes, but may also arise from technical interference. **Baseline wander is evident, possibly due to breathing or poor electrode contact, which can obscure ST segments and T waves.** The absence of a clear, regular heartbeat pattern in this snippet raises the possibility of an arrhythmia, yet noise remains a likely explanation that must be ruled out.

Potential Illnesses or Conditions: If real and confirmed, large voltage swings might suggest left ventricular hypertrophy or other conduction abnormalities. However, common causes like electrode noise must be excluded first.

Summary and Takeaways: While there appear to be significant swings and baseline drift, the data alone are insufficient for a definitive diagnosis. If these traces genuinely reflect the patient's ECG rather than artifacts, further evaluation (e.g., repeated ECG, continuous monitoring, echocardiogram) is prudent. Ultimately, clinical symptoms and higher-quality ECG recordings are essential for accurate interpretation and any subsequent medical decisions.

Deepseek-R1 without other modal information: More information required

Overall, the data begins with negative values, transitions to positive values, and then returns to negative values, suggesting the presence of a QRS complex. However, the significant increase in amplitude toward the end of the series raises concerns about potential anomalies. Detecting peaks corresponding to the QRS complex is crucial, as irregularities in its shape, duration, or amplitude may suggest arrhythmias or conduction abnormalities. Additionally, trend analysis can reveal sudden changes or spikes in the signal, **which could be caused by noise, artifacts, or pathological conditions such as ventricular hypertrophy or ischemia**. The data exhibits a notable spike toward the end, with values like 2.1643 and 1.8084, which could indicate either noise or a pathological condition. **Noise might result from external interference or poor electrode contact, while pathological conditions could include ventricular hypertrophy, arrhythmias, or ischemia.** The presence of negative values at the beginning and end of the series is typical, but the magnitude and pattern should be compared to a standard ECG to rule out abnormalities. Potential illnesses suggested by the data include ventricular hypertrophy, **characterized by a high-amplitude QRS complex due to thickened ventricular walls, often caused by high blood pressure**. Arrhythmias, such as ventricular tachycardia or fibrillation, might be indicated by irregularities in the QRS complex or abnormal spikes. Ischemia or infarction could be suggested by abnormal ST segments or T waves, though these are not clearly visible in this dataset. In conclusion, the ECG time-series data shows potential anomalies, particularly the high-amplitude spike toward the end, which could indicate conditions like ventricular hypertrophy, arrhythmias, or ischemia. However, further analysis and clinical correlation are necessary to confirm the diagnosis. Noise or artifacts should also be ruled out to ensure the accuracy of the findings.

ChatGPT-o1 with extra images*: More certain answers and reasoning

The given ECG data ranges from -4.51 mV to +2.16 mV, showing large amplitude swings. **Compared with normal and abnormal ECG references,** the absence of periodic features like clear P waves and regular QRS complexes suggests potential arrhythmias such as atrial fibrillation (AFib) or premature ventricular contractions (PVCs). **The pronounced baseline drift may stem from poor electrode contact or breathing artifacts, complicating the interpretation. If real, these voltage changes might indicate left ventricular hypertrophy (LVH) or electrolyte imbalances like hyperkalemia.** The low sampling rate (25 Hz) limits detailed diagnosis, requiring higher-quality recordings and clinical symptom correlation. **To confirm, a repeat ECG at clinical-grade resolution and ruling out noise artifacts is essential, alongside advanced diagnostics such as echocardiography or continuous Holter monitoring.**

* Add extra images of normal and various abnormal ECG species.

Figure 5: Zero-shot open question performances of different LLMs and input settings for healthcare application.

Prompt:

You are an expert in finance and know the stock market information very well. We provide a time-series recording about stock price and some background information, you should provide some causal inference and impact analysis about stock price rise or fall and prospects based on the time series analysis and your knowledge of the financial field. Besides, reasoning steps can infer the current situations of time-series records. Let's begin!

The following data represents the financial stock prices of Nvidia from 2024-09-05 to 2025-01-24, which record the fluctuations in the value of stocks over time and serve as a critical tool for analyzing market trends and investment strategies. Using a time series sampled at a frequency of one trading day, we can determine patterns and detect potential anomalies in the stock market's behavior. The input recording is [107.21, 102.83, 106.47, 108.10, 116.91, 119.14, 119.10, 116.78, 115.59, 113.37, 117.87, 116.00, 116.26, 120.87, ..., 134.29, 138.31, 144.47, 149.43, 140.14, 140.11, 135.91, 133.23, 131.76,136.24, 133.57, 137.71, 140.83, 147.07, 147.22, 142.62].

Please make some statistical judgments and determine whether there are anomalies in this time series given the information above. Please provide some causal inference and impact analysis about stock price rise or fall and prospects based on the time series analysis and your knowledge of the financial field.

Ambiguous and non-specific answer

ChatGPT-o1 without other modal information:

Between 2024-09-05 and 2025-01-24, NVIDIA's stock price showed an overall upward trend, rising from around $100 to nearly $150 before settling near $140. Daily fluctuations were typically 1%-3%, with occasional larger movements of 4%-6%, typical for high-growth tech stocks. **The gains were driven by strong market expectations for AI, data center, and high-performance computing demand, along with robust GPU purchases by enterprises and cloud providers. A favorable macroeconomic environment and improved tech sector sentiment also supported the rally. Pullbacks reflected profit-taking, overvaluation concerns, and macroeconomic or geopolitical risks**. While no major anomalies were observed, uncertainties remain around supply chain bottlenecks, new competitors, and economic cycles. The stock is likely to continue oscillating with an upward trend, influenced by profitability reports, AI investment progress, and regulatory developments.

Ambiguous and non-specific answer

Deepseek-R1 without other modal information:

From September 5, 2024, to January 24, 2025, Nvidia's stock fluctuated significantly, starting at 107.21 and ending at 142.62. After an initial September decline, likely due to market uncertainty, the stock rallied in October, driven by positive earnings, AI sector growth, and favorable macroeconomic conditions. Volatility increased in November and December, oscillating between 131.60 and 148.88, **likely due to mixed macroeconomic data or competitive pressures.** A late January dip to 142.62 suggests profit-taking or external shocks, though a slight recovery highlights underlying strength. **Key drivers include Nvidia's AI dominance, product launches, and market trends,** while risks such as regulatory scrutiny and competition persist. Long-term prospects remain strong, but short-term volatility is expected, emphasizing the need for careful monitoring and risk management.

More certain answers and reasoning

ChatGPT-o1 with extra news and PDF files*:

Between September 2024 and January 2025, NVIDIA's stock price showed strong growth, driven by robust demand for AI, data centers, and high-performance computing. In Q1 FY2025, the company reported record revenue of $26 billion, up 262% year-over-year, with a 427% increase in Data Center revenue **fueled by the widespread adoption of Hopper GPUs for generative AI and large language models.** Gross margins reached 78.4% (GAAP), and net income surged 628% to $14.88 billion, reflecting strong market demand and operational efficiency. While Gaming and Professional Visualization segments experienced slight seasonal declines, the overall growth trajectory remains solid. **With Q2 FY2025 revenue projected at $28 billion, NVIDIA is well-positioned for continued success, although supply chain risks and competition warrant attention.** The stock is expected to oscillate upward, driven by AI advancements and financial performance.

* Add extra news about this company and PDF format financial reports from its official website.

Figure 6: Zero-shot open question performances of different LLMs and input settings for financial application.

Prompt:

You are an expert in the energy and electricity field. We provide an incomplete time-series recording about energy consumption and some background information, you should provide some statistical analysis based on your knowledge of the financial field and also generate full time series with missing value imputed. Besides, reasoning steps can infer the current situations of time-series records. Let's begin!

This is a London Smart Meters Dataset, which contains half-hourly energy consumption readings in kWh from 5,560 households in London, collected between November 2011 and February 2014, and categorized into 112 blocks. The frequency is one hour. The input Time Series are [0.038, 0.084, 0.038, 0.062, 0.062, 0.04, 0.086, 0.04, 0.075, 0.049, 0.087, 0.191, 'X', 0.19, 0.154, 0.162, 0.18, 0.147, 0.194, 0.153, 0.233, 0.227, 0.164, 0.139, 0.038, 0.072, 0.051, 0.041, 0.083, 0.039, 0.077, 0.046, 0.053, 0.131, 0.096, 0.139, 0.094, 0.115, 0.155, 0.148, 0.193, 'X', 0.098, 0.05, 0.049, 0.076, 0.04, 0.078, 'X', 0.058, 0.067, 0.039, 0.084, 0.04, 0.059, 0.067, 0.039, 0.084, 0.082, 0.175, 0.169, 0.157, 0.172, 0.16, 0.184, 0.16, 0.169, 0.169, 'X', 0.237, 0.181, 0.116, 0.061, 0.043, 0.079, 0.042, 0.068, 0.058, 0.049, 0.077, 0.041, 0.145, 0.096, 0.117, 0.122, 0.091, 0.184, 0.148, 0.173, 0.169, 0.071, 0.073, 0.045, 0.063, 0.056, 0.051, 'X', 0.051, ..., 0.078, 0.04, 0.078, 0.045, 0.054, 0.07, 0.039, 0.07, 0.051, 0.102, 0.166, 0.162, 0.176, 0.161, 0.177]. The missing values are represented by 'X'.

Please provide the complete time series and the rationale for generating these exact data.

ChatGPT-o1 without other modal information: No very convincing reasons

In the London Smart Meters dataset, a few readings labeled "X" were missing, and we filled those gaps by replacing each missing value with the average of its two immediate neighbors (linear interpolation). The generated time series is [XXX, ..., XXX, XXX]. This straightforward approach preserves the natural hour-to-hour continuity in residential consumption data and avoids introducing abrupt spikes that could distort analyses. While more complex models might be warranted for larger or systematic gaps, linear interpolation works effectively here, as the missing points are isolated. Keeping the series continuous and reflecting local consumption trends supports more accurate forecasting, peak-demand identification, and billing calculations in both operational and financial contexts, ultimately allowing energy suppliers and market participants to rely on a stable, reliable dataset for planning and decision-making.

Inadequate reasons and answer

Deepseek-R1 without other modal information:

The provided time series represents hourly energy consumption readings from London households, with missing values marked as 'X'. Using linear interpolation, these gaps were filled by estimating values between adjacent data points, a method well-suited for this dataset due to its temporal continuity. The generated complete time series [XXX, ..., XXX, XXX] reveals an average consumption of 0.104 kWh per hour, with a standard deviation of 0.053 kWh, indicating moderate variability. Peak consumption reaches 0.237 kWh, likely during high-activity periods, while the minimum of 0.038 kWh corresponds to low-usage times, such as late-night hours. **The data exhibits clear daily patterns, with higher consumption during mornings and evenings and lower usage at night, reflecting typical household routines.** These trends suggest strong daily seasonality, which could be further analyzed to identify anomalies, forecast demand, or inform energy efficiency initiatives. The imputed dataset provides a solid foundation for understanding consumption behavior, enabling utilities and policymakers to optimize energy management and promote sustainability. By leveraging this data, stakeholders can better address peak demand, reduce costs, and support a more efficient energy system.

More certain answers and reasoning

ChatGPT-o1 with extra news and and website links:

In the London Smart Meters dataset, several readings marked "X" were missing, and **we knew that the data were gathered during a summer period with exceptionally high temperatures and relatively consistent daily usage cycles.** In such conditions, electricity demand can spike more abruptly—often due to cooling loads like air conditioners—**so energy experts often advise using a second-order (quadratic) interpolation rather than a simple linear method. Quadratic interpolation better captures the potential for rapid increases or decreases in consumption, reflecting real-world load shapes that linear estimates might miss.** Specifically, we replaced each missing reading with a quadratic fit based on its surrounding data points. This preserves the smooth continuity of the time series while allowing for sharper changes that are common in hot summer weather. The generated time series is [XXX, ..., XXX, XXX]. As a result, subsequent tasks such as forecasting, peak-demand identification, and billing become more accurate, ensuring that the dataset reflects the true variability in summertime household energy usage.

* Add extra news of those days and some useful website links about records of local weather.

Figure 7: Zero-shot open question performances of different LLMs and input settings for electrical application.