From Time Series Analysis to Question Answering: A Survey in the LLM Era

Wei Li[1] , **Zhe Xie**[2] , **Yuxuan Liang**[3] , **Xinli Hao**[1] , **Yunyao Cheng**[4] , **Dan Pei**[2] , **Xiaofeng Meng**[1*]
[1]Renmin University of China
[2]Tsinghua University
[3]Hong Kong University of Science and Technology (Guangzhou)
[4]Aalborg University
{leeway, xinli_hao, xfmeng}@ruc.edu.cn, {xiez22, peidan}@tsinghua.edu.cn,
yuxliang@outlook.com,yunyaoc@cs.aau.dk

Abstract

Recently, Large Language Models (LLMs) have introduced a novel paradigm in Time Series Analysis (TSA), leveraging strong language capabilities to support tasks such as forecasting and anomaly detection. However, these analysis tasks cannot adequately cover temporal language tasks, such as interpretation and captioning. A fundamental gap remains between TSA and LLMs: LLMs are pre-trained to optimize natural language relevance for question answering rather than objectives specialized for TSA. To bridge this gap, TSA is evolving toward Time Series Question Answering (TSQA), shifting from expert-driven and task-specific analysis to user-driven and task-unified question answering. TSQA depends on flexible exploration rather than predefined TSA pipelines. In this survey, we first propose a taxonomy that reflects the evolution from TSA to TSQA, driven by a shift from external to internal alignment. We then organize existing literature into three alignment paradigms: Injective Alignment, Bridging Alignment, and Internal Alignment, and provide practical guidance for flexible, economical, and generalizable selection of alignment paradigms. We finally analyze datasets across domains and characteristics, identify challenges, and highlight future research directions.

1 Introduction

Time series are sequential data characterized by temporal dependencies and pattern semantics [Liang *et al.*, 2024]. Time Series Analysis (TSA) has been widely applied across multiple domains, including healthcare [Chan *et al.*, 2024], finance [Ding *et al.*, 2023], IoT [Qiu *et al.*, 2023], weather [Cheng *et al.*, 2025b], and electricity [Pan *et al.*, 2025]. Within these domains, TSA typically involves tasks such as forecasting [Jin *et al.*, 2024], anomaly detection [Liu *et al.*, 2024a], classification [Wang *et al.*, 2025b], and imputation [Wang *et al.*, 2025c]. Recently, Large Language Models (LLMs) have introduced a novel paradigm in TSA, leveraging strong natural language capabilities [Zhang *et al.*, 2024].

*Corresponding Author

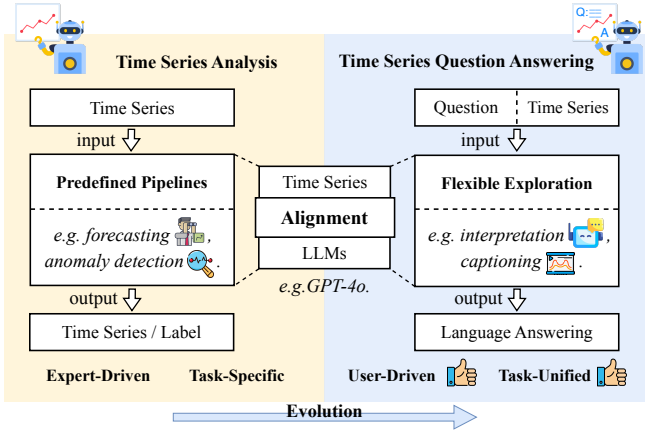


Figure 1: Comparison between Time Series Analysis and Time Series Question Answering under alignment of time series with LLMs.

However, these analysis tasks cannot adequately cover temporal language tasks. For example, in database AIOps, Time Series Question Answering (TSQA) provides comprehensive interpretations of performance metrics [Xie *et al.*, 2025]. In time series captioning tasks, TSQA summarizes temporal patterns, structures, and events [Trabelsi *et al.*, 2025]. A fundamental gap remains between TSA and LLMs: LLMs are pre-trained to optimize natural language relevance for question answering rather than objectives specialized for TSA. As shown in Figure 1, TSA is evolving toward TSQA, shifting from expert-driven and task-specific analysis to user-driven and task-unified question answering. TSA relies on predefined pipelines built on task-specific time series modeling. In contrast, TSQA depends on flexible exploration based on task-unified temporal language processing, including in-depth interpretation and captioning [Chang *et al.*, 2025a]. After aligning time series with LLMs (e.g., GPT-4o [Hurst *et al.*, 2024]), user-driven TSQA inputs time series and questions and outputs natural language answers, which are more intuitive than the numerical time series or categorical labels generated by expert-driven TSA [Zhang *et al.*, 2024]. The fundamental gap highlights alignment between time series and LLMs as a central challenge, motivating us to propose alignment paradigms that cover the evolution from TSA to TSQA.

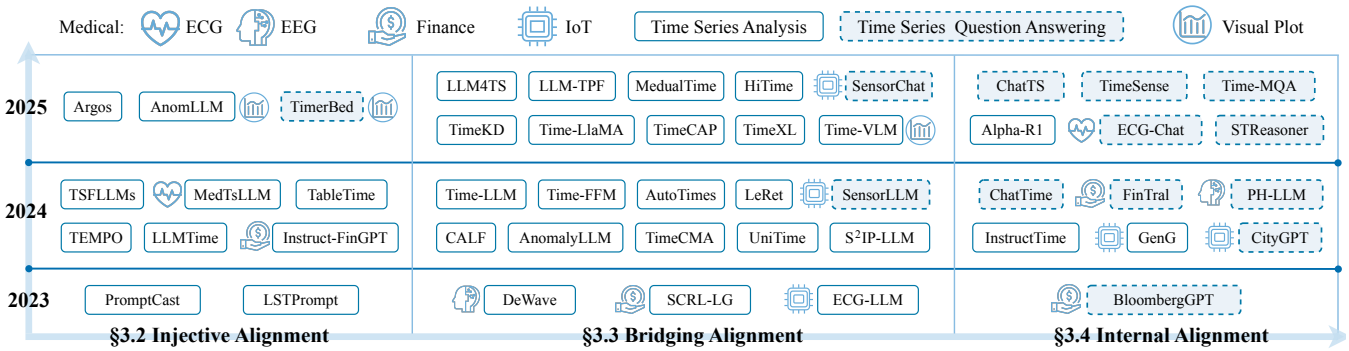


Figure 2: Taxonomy of relevant literature across three alignment paradigms. The icon to the left of each method indicates its domain, while the absence of this icon denotes a general domain. The icon on the right denotes the use of time series visual plots. The background color of each method represents task targets: lighter shades denote time series analysis (TSA), and darker shades denote time series question answering (TSQA). The TSA literature dominates external Injective and Bridging Alignment, whereas the TSQA literature dominates Internal Alignment, reflecting **the evolution from TSA to TSQA driven by a shift from external to internal alignment**.

The perspectives of representative survey literature are summarized in Table 1. Instead of considering TSA and TSQA in isolation, our survey adopts the evolution perspective from TSA to TSQA and provides an up-to-date overview. Compared with existing surveys, this evolution perspective integrates previously isolated TSA and TSQA research into a unified taxonomy of alignment paradigms in the LLM era.

As shown in Figure 2, the taxonomy consists of three alignment paradigms: (1) **Injective Alignment** injects numerical time series into textual prompts without temporal modification, using frozen LLMs. Temporal modification refers to optimizations that adapt time series to LLMs. (2) **Bridging Alignment** establishes mappings between numerical time series and prompts, introducing temporal modification while still employing frozen LLMs. (3) **Internal Alignment** leverages multiple time series representations, including numerical values, visual plots, and structured tables, combining temporal modification with internal LLM components.

The primary contributions are summarized as follows.

- We distinguish and define TSA and TSQA (Section 2).
- We propose a taxonomy of three alignment paradigms, provide practical guidance for paradigm selection, and systematically categorize the relevant literature and maintain an up-to-date GitHub repository[1] (Section 3).
- We analyze datasets from the perspectives of application domains and data characteristics (Section 4).
- We discuss challenges and future directions (Section 5).

2 Preliminaries

2.1 Definitions

A time series is represented as a numerical sequence with temporal order $\mathbf{X}_{\text{Number}} = \langle \mathbf{x}_1, \ldots, \mathbf{x}_S \rangle \in \mathbb{R}^{S \times N}$, where S denotes the sequence length and N is the number of variables. Each observation $\mathbf{x}_i$ at timestep i is an N-dimensional vector.

Textual data is defined as $\mathbf{X}_{\text{Text}} = \{s_1, s_2, \ldots, s_L\}$, where each symbol $s_i \in \mathcal{V}$ is drawn from a predefined vocabulary $\mathcal{V}$, and L denotes the length of the text sequence. Specifically, (1) **Textual time series**: This type represents numerical time series converted into discrete symbolic sequences, allowing LLMs to process time series in textual form [Liu *et al.*, 2025c]. (2) **Natural language**: This type is not derived from time series conversion, including domain knowledge, task instructions, questions, and answers.

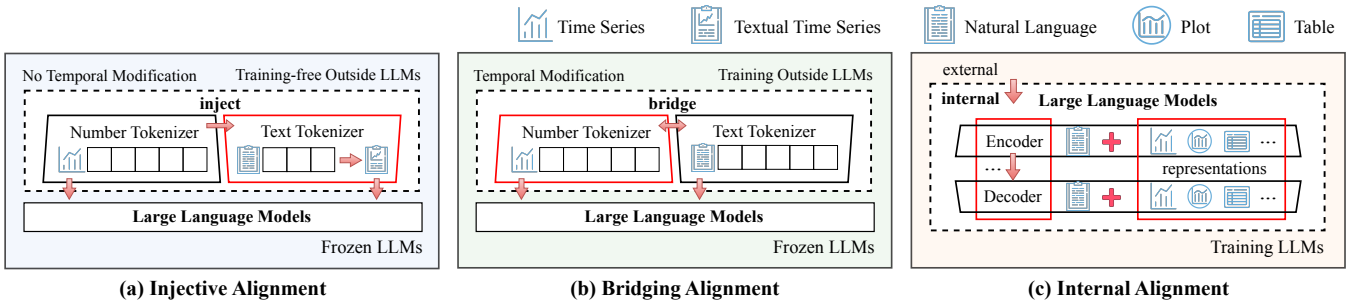
[1]https://github.com/Leeway-95/TSA-TSQA-with-LLMs

Survey Literature	Perspective
Our Survey	TSA&TSQA Evolution
[Liu *et al.*, 2025c] (IJCAI'25)	Cross-modality TSA
[Jiang *et al.*, 2025a] (KDD'25)	Multi-modality TSA
[Zhang *et al.*, 2024] (IJCAI'24)	Pipeline-based TSA
[Liang *et al.*, 2024] (KDD'24)	Foundation Model

Table 1: Comparison between our perspective and existing surveys.

2.2 Time Series Analysis with LLMs

These tasks follow a predefined processing with fixed objectives, where LLMs serve as modeling and reasoning components for TSA tasks. In addition to generating corresponding explanations, the objectives and outputs of representative analysis tasks are defined as follows. (1) **Forecasting:** Given historical observations, LLMs predict future numerical values over a predefined horizon. The output is a numerical sequence with temporal order [Tang *et al.*, 2024]. (2) **Anomaly detection:** LLMs identify deviations from normal dynamics. The outputs are binary labels (normal and abnormal) and anomaly scores [Liu *et al.*, 2024a]. (3) **Classification:** LLMs assign each time series to one category from a predefined label set. The output consists of one or more predefined class labels [Cheng *et al.*, 2025a]. (4) **Imputation:** LLMs infer missing or corrupted values using observed context. The output is a completed numerical sequence with missing entries filled [Wang *et al.*, 2025c]. Formally, given a time series $\mathbf{X}$ and contextual information $\mathbf{C}$ (e.g., natural language, plots), TSA aims to produce the task-specific output Y:



(a) Injective Alignment **(b) Bridging Alignment** **(c) Internal Alignment**

Figure 3: Comparison of three alignment paradigms. Red boxes denote focused components, and icons indicate time series representations.

$$f_p(\mathbf{X}, \mathbf{C}) \to Y$$

Where $f_p(\cdot)$ refers to a method that uses LLMs with predefined task prefixes, and p denotes the corresponding predefined task prompts, e.g., "predict the next 48 timesteps".

2.3 Time Series Question Answering

In contrast to predefined analysis tasks, TSQA formulates time series understanding as a unified question answering paradigm. This paradigm enables flexible exploration of time series while emphasizing: (1) **Deep thinking**: analyzing complex temporal semantics and dependencies to derive clear insights [Zhang *et al.*, 2025b]. (2) **Action planning**: decomposing user queries into goal-oriented steps to generate optimal answers [Chang *et al.*, 2025a]. (3) **Tool calling**: using external analytical tools or models to perform precise numerical computations, statistical tests, and other operations within the reasoning process [Ni *et al.*, 2026]. Under this paradigm, representative TSQA tasks include: (1) **In-depth interpretation**: explaining user questions through joint reasoning over temporal patterns, multivariate dependencies, and contextual information. (2) **Time series captioning**: producing natural language descriptions that summarize temporal patterns, structure, and events [Trabelsi *et al.*, 2025].

Formally, given a time series $\mathbf{X}$, additional contextual information $\mathbf{C}$, and a natural language question Q, the goal of TSQA is to produce the correct answer A:

$$f(\mathbf{X}, \mathbf{C}, Q) \to A$$

Where $f(\cdot)$ refers to a method that uses LLMs with Q, and A is a natural language answer that leverages temporal semantics from $\mathbf{X}$ and relevant $\mathbf{C}$ [Kong *et al.*, 2025].

3 Taxonomy

In this section, we propose a taxonomy with clear boundaries, formal definitions, optimization strategies, and categorize existing literature into three alignment paradigms.

3.1 Overview

The three alignment paradigms are Injective Alignment, Bridging Alignment, and Internal Alignment. As illustrated in Figure 3, the processes and emphases of each alignment paradigm are as follows: (a) Injective Alignment focuses on textual representations by injecting numerical time series into textual prompts. (b) Bridging Alignment focuses on numerical representations and builds connections between numerical time series and textual prompts. (c) Internal Alignment typically focuses on internal component modifications and multiple time series representations, including numerical values, visual plots, and structured tables. These paradigms cover the existing literature with clear boundaries, and each work can be uniquely assigned to a single category.

As shown in Figure 4, the horizontal axis indicates whether LLM parameters are trained, and the vertical axis indicates whether temporal modifications are required. Temporal modification refers to adapting time series for LLMs, including both modifications outside the LLM and adjustments to the internal LLM architecture. These two dimensions determine the three alignment paradigms. (a) Injective Alignment involves no temporal modification and adopts frozen LLMs. This design preserves the original LLM parameters. (b) Bridging Alignment introduces temporal modification while still employing frozen LLMs. This design enables joint processing of time series and textual inputs while preserving all parameters of the original LLM. (c) Internal Alignment combines temporal modification with training LLMs through parameter updating to provide native support for time series. Next, we detail the three paradigms along these boundaries.

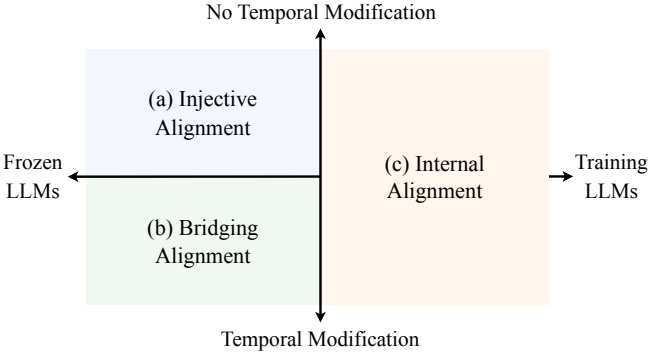


Figure 4: Boundaries of three proposed alignment paradigms.

3.2 Injective Alignment

This paradigm injects information into LLMs by directly encoding $\mathbf{X}_{\text{Number}}$ into token embeddings or appending textual time series $\mathbf{X}_{\text{Text}}$ to prompts. This process is defined as $\mathbf{Y} = f_\theta(h_\gamma(\mathbf{X}_{\text{Number}}), \mathbf{X}_{\text{Text}})$, where $h_\gamma(\cdot)$ is an injective method, and $f_\theta(\cdot)$ refers to frozen LLMs. Beyond original

Direct input strategies, two optimization strategies for $h_\gamma(\cdot)$ are In-Context Learning (ICL) and Chain of Thought (CoT).

ICL. This strategy guides LLMs with a few task demonstrations to enhance contextual understanding and reasoning. For instance, LLMAD [Liu *et al.*, 2025e] further improves ICL through diversified demonstrations and robust retrieval, finding that a small mix of positive and negative examples achieves an effective trade-off, as additional demonstrations offer limited gains. LVICL [Zhang *et al.*, 2026] compresses multiple time-series examples into a permutation-invariant context vector and injects it into each LLM layer to elicit in-context learning for improved forecasting while keeping the LLM frozen. **Benefits**: Effective in data-scarce scenarios and allows low-cost dynamic context switching. **Drawbacks**: Performance is sensitive to noisy or low-quality examples.

CoT. This strategy activates LLMs for deep thinking by decomposing complex tasks into ordered intermediate steps along explicit logical paths. For example, LSTPrompt [Liu *et al.*, 2024b] divides zero-shot forecasting of TSA into two stages: short-term reasoning for dynamic trends and long-term reasoning for periodic patterns, with periodic revisits to assumptions to enhance temporal adaptivity. TimerBed [Liu *et al.*, 2025d] applies a three-phase reasoning path for TSQA: it first identifies relevant temporal segments; then extracts numerical clues from these segments; it finally synthesizes explainable answers. **Benefits**: Converts implicit logic into explicit steps, enhancing interpretability and reliability via stepwise self-correction. **Drawbacks**: Requires predefined steps and incurs additional computational cost due to extra CoT. Practical instructions for Injective Alignment are as follows:

> This paradigm is **flexible** to operate outside LLMs. CoT and ICL are easy to implement across domains, making this paradigm suitable for domain-specific applications. Since its effectiveness depends on prompt quality, practitioners should iteratively refine the design of prompts using domain expertise and LLM feedback.

3.3 Bridging Alignment

This paradigm maps numerical time series and textual prompts, defined as $\mathbf{Y} = f_\theta(g_\phi(\mathbf{X}_{\text{Number}}, \mathbf{X}_{\text{Text}}))$, where $g_\phi(\cdot)$ is a bridging method and $f_\theta(\cdot)$ refers to frozen LLMs. Two efficient optimization strategies for $g_\phi(\cdot)$ are as follows.

Similarity Retrieval (SR). This strategy aligns numerical and textual representations via similarity. Adapters serve as a bridging space, projecting time series into an embedding space compatible with LLMs. Specifically, g_ϕ maps time series as $g_\phi(\mathbf{X}_{\text{Number}}) \rightarrow \mathbf{X}_{\text{Text}}$, with ϕ trained using a loss function $\mathcal{L}$. For example, Time-LLM [Jin *et al.*, 2024] introduces a reprogramming space that converts time series into text prototypes, bridging numeric encodings with textual embeddings. Time-CMA [Liu *et al.*, 2025b] encodes time series and text into a shared space using separate encoders, aligns embeddings via contrastive loss, and retrieves time series whose embedding best matches the query. **Benefits:** Enhances performance via flexible adapters without altering LLM parameters. **Drawbacks:** Performance depends on adapter quality and incurs computational cost during retrieval.

Low-Rank Adaptation (LoRA). This strategy is a parameter-efficient technique that bridges low-rank updates into pre-trained LLMs, enabling task-specific tuning without altering the model parameters. For instance, HiTime [Tao *et al.*, 2024] applies LoRA during generative instruction fine-tuning for time series classification, updating a small subset of weights to enable generative reasoning from numeric features. AutoTimes [Liu *et al.*, 2024e] integrates LoRA into its autoregressive forecasting framework, fine-tuning embedding and projection layers while keeping the LLM parameters frozen. **Benefits:** Parameter-efficient and task-specific adaptation. **Drawbacks:** Limited performance when the low-rank space is insufficient for complex time-series patterns. Practical instructions for Bridging Alignment are as follows:

> This paradigm is **economical** to operate outside LLMs. Similarity Retrieval and LoRA leverage aligned semantic information, making it effective for capturing time-series characteristics. Since its effectiveness depends on the quality of adapter modules, practitioners should carefully assess whether the improvements are meaningful.

3.4 Internal Alignment

This paradigm modifies internal components and parameters, defined as $\mathbf{Y} = f'_\theta(d_{\psi 1}(\mathbf{X}_{\text{Number}}), d_{\psi 2}(\mathbf{X}_{\text{Text}}), \dots)$, where $\{d_{\psi 1}(\cdot), d_{\psi 2}(\cdot), \dots\}$ are internal methods that enhance $\mathbf{X}_{\text{Number}}$ into $\mathbf{X}'_{\text{Number}}$ and $\mathbf{X}_{\text{Text}}$ into $\mathbf{X}'_{\text{Text}}$, or introduces additional representations. $f'_\theta(\cdot)$ denotes training LLMs. Two representative optimization strategies for $f'_\theta(\cdot)$ are as follows.

Re-Training (RT). This strategy aligns time series with LLMs by temporal modification and updating parameters. For example, ChatTime [Wang *et al.*, 2025a] treats time series as a foreign language. It normalizes and discretizes continuous signals into symbolic tokens, extends the LLM vocabulary, and performs continual pre-training and instruction tuning to unify temporal and textual modalities. ChatTS [Xie *et al.*, 2025] trains a native time series LLM for TSQA. It synthesizes high-quality temporal descriptions using rule-based generators and attribute selectors, followed by two-stage fine-tuning to align numerical evolution with language reasoning. **Benefits**: Enables deep internal alignment between temporal patterns and linguistic representations, and improves multi-task adaptability via parameter updates. **Drawbacks**: Requires substantial training resources and may introduce additional complexity from discretization or synthetic data.

Reinforcement Learning (RL). This strategy aligns LLM behavior with temporal objectives by optimizing decision policies using reward signals derived from sequential outcomes. For instance, TimeSense [Zhang *et al.*, 2025b] treats temporal reasoning as a self-supervised reconstruction task. It uses time-aware objectives to learn temporal order, duration, and dependencies, which implicitly ensure temporal consistency via reinforcement learning. Alpha-R1 [Jiang *et al.*, 2025c] models factor selection as a multi-step decision trajectory and directly optimizes LLM outputs via reinforcement learning. This formulation grounds language generation in long-horizon temporal feedback and aligns reasoning

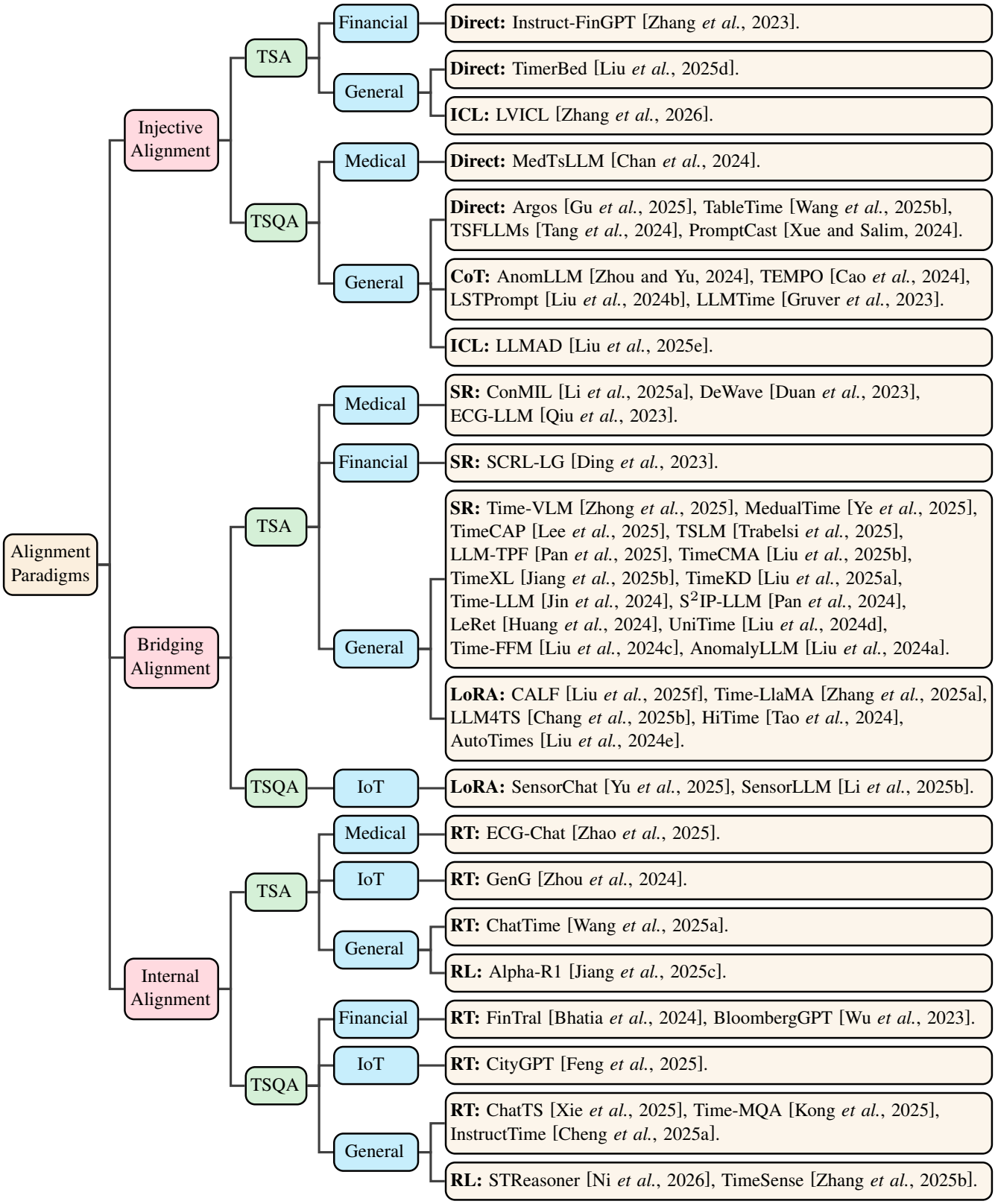


Figure 5: A comprehensive taxonomy of relevant literature, organized by three alignment paradigms, task types, domains, and optimization strategies. Notably, the category with the most literature is the Bridging Alignment, TSA task, General domain, and SR optimization strategy, where the General domain denotes literature without a specific application focus and evaluated across multiple domains.

steps with profit-driven sequential outcomes. **Benefits**: Directly optimizes long-horizon temporal objectives, adapts to dynamic environments, and provides decision interpretability through reward semantics. **Drawbacks**: Training stability depends on reward design, optimization cost is high, and performance is sensitive to data noise and quality. Practical instructions for Internal Alignment are as follows:

> This paradigm is **generalizable** for internal LLM operation. Re-Training and Reinforcement Learning provide general-purpose improvement for time series understanding and reasoning. Since its effectiveness depends on the dataset and retraining quality, practitioners should focus on data synthesis methods and evaluation metrics.

We categorize 50 relevant literature into three groups from TSA to TSQA, as shown in Figure 5. For each work, a comprehensive taxonomy is summarized across four dimensions of Alignment Paradigms, tasks, domains, and optimization strategies. Specifically, we analyze the distribution of Alignment Paradigms in Figure 6(a), including Injective Alignment 26%, Bridging Alignment 50%, and Internal Alignment 24%. We further examine the task distribution in Figure 6(b), which consists of TSA 60% and TSQA 40%. As shown in Figure 6(c), we identify the application domains covered in the literature, including Medical 10%, Finance 8%, IoT 8%, and General 74%. In addition, we focus on the optimization strategies summarized in Figure 6(d), where Similarity Retrieval (SR) accounts for the largest proportion at 36%. We present the yearly evolution trends in Figure 6(e), with TSQA surpassing TSA in growth rate from 2024 to 2025, and Internal Alignment exhibiting the most significant increase. Next, we analyze representative datasets across diverse domains, highlighting how question answering design grounds time series characteristics and drives their evolution.

4 Dataset Analysis

In this section, we introduce representative datasets organized by application domains, as shown in Table 2. Additional datasets are provided in our GitHub repository[2].

4.1 Overview

We analyze representative datasets from two perspectives: application domains and data characteristics, which respectively determine their semantics and structures. From the domain perspective, existing datasets vary in semantic dependence and regulatory constraints, but they generally require strong interpretability, contextual understanding, and integration of domain knowledge. From the characteristic perspective, existing datasets differ in temporal patterns and spatial multivariate structures. Moreover, time series can be represented in multiple forms, including numerical sequences, textual descriptions, and visual plots, each introducing different trade-offs among fidelity, interpretability, and compatibility.

[2]https://github.com/Leeway-95/TSA-TSQA-with-LLMs

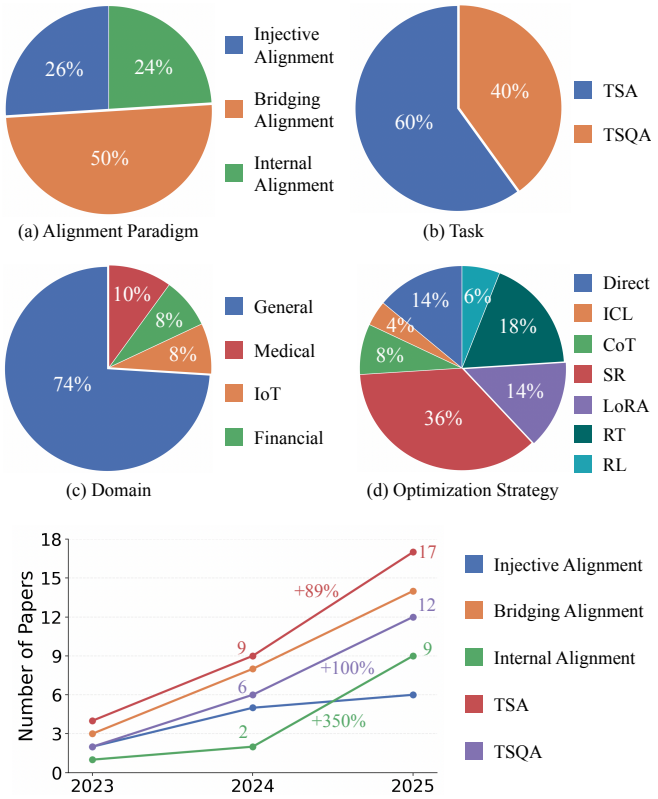


Figure 6: Taxonomy distribution and evolution across dimensions.

4.2 Domain

Time series data span a wide range of domains, covering representative medical, financial, and IoT domains.

Medical data. Medical time series include EEG, ECG, and related signals, which often exhibit strong seasonality and clear outliers. Semantic interpretation relies heavily on domain knowledge, as specific waveforms correspond to physiological states [Duan *et al.*, 2023]. Medical data are typically collected by professional equipment, but they are subject to strict privacy regulations and require transparent reasoning to support clinical decisions [Chan *et al.*, 2024].

Financial data. Financial time series are high-frequency, such as stock prices and trading volumes. They contain pronounced outliers and short-term volatility, while also exhibiting recurring trends relevant to risk assessment and market forecasting [Wu *et al.*, 2023]. Financial data usually rely on external knowledge and operate under stringent market regulations. Interpreting financial time series is often more valuable than predicting exact values [Ding *et al.*, 2023].

IoT data. IoT time series consist of multi-source sensor data that typically show seasonality and infrequent changes due to predefined operating procedures. IoT data are high-dimensional and heterogeneous, integrating multiple sensors. Data quality is influenced by sensor type, sampling rate, and hardware-induced noise [Zhou *et al.*, 2024]. Effective TSQA in this domain requires contextual fusion to ensure comprehensive and consistent interpretation [Yu *et al.*, 2025].

Dataset	Domain	Task	Size	Characteristic
ECG-QA[Oh *et al.*, 2024]	Medical	TSQA	414,348 pairs.	3 types, 12 variates, 2 modalities.
FinTexTS [Lee *et al.*, 2026]	Financial	TSA, TSQA	100M pairs.	4 variates, 2 modalities.
SensorQA [Reichman *et al.*, 2025]	IoT	TSQA	5,648 pairs.	2 types, multivariate, 5 modalities.
Time-MQA [Kong *et al.*, 2025]	General	TSQA	192,843 pairs.	6 types, multivariate, 2 modalities.
CiK [Williams *et al.*, 2025]		TSA	2,644 pairs.	5 types, univariate, 2 modalities.
TSandLanguage [Merrill *et al.*, 2024]		TSA, TSQA	230k pairs.	10 types, univariate, 2 modalities.

Table 2: Summary of datasets from TSA to TSQA, including Domain, Task, Size, and Characteristic (patterns, variables, and modalities).

4.3 Characteristic

Temporal patterns exhibit varying properties over time, including: (1) **Outliers**, representing deviant signals with anomalous semantics [Zhou and Yu, 2024]. (2) **Trend**, capturing long-term directional changes over time [Jiang *et al.*, 2025b]. (3) **Seasonality**, corresponding to recurring semantics at regular intervals [Yu *et al.*, 2025]. (4) **Volatility**, characterized by irregular fluctuations that obscure meaningful patterns and often require denoising [Pan *et al.*, 2024].

Spatial multivariates introduce temporal complexity, including: (1) **Independence**, treating variables separately to reduce complexity and avoid spurious correlations [Liu *et al.*, 2025b]. (2) **Dependence**, capturing inter-variable interactions by considering variables jointly [Pan *et al.*, 2025]. (3) **Correlation**, representing complex relational structures for coherent and context-aware reasoning [Cheng *et al.*, 2025b].

Beyond temporal patterns and spatial multivariates, multimodal representations provide an additional characteristic. (1) **Numerical representation** preserves raw time series values and offers advantages: it directly retains precise quantitative information and avoids additional transformation costs. However, numerical data also have limitations: time series are scarce compared to language, and raw values lack explicit semantic meaning [Gruver *et al.*, 2023]. Real-world signals often contain noise, missing values, and sparsity, requiring preprocessing. (2) **Textual representation** converts time series into natural language descriptions and offers advantages: text provides human-readable semantics that are intuitive and interpretable [Lee *et al.*, 2025]. This format naturally aligns with LLMs. Its main limitation lies in the alignment between the discrete nature of text and the continuous properties of time series signals [Wang *et al.*, 2025a]. (3) **Visual representation** draws time series as plots, such as line charts, and offers advantages: visual representations explicitly highlight trends, seasonality, and outliers. Visual plots can be aligned with LLMs via zero-shot visual understanding and can supplement or replace numerical or textual representations [Zhong *et al.*, 2025]. Nevertheless, visual plots introduce challenges: low resolution may cause information loss, and fine-grained temporal values are often difficult to recover precisely, limiting quantitative accuracy [Jiang *et al.*, 2025a]. These characteristics jointly form the complexity of temporal semantics and modeling, requiring LLMs to capture temporal dynamics, inter-variable dependencies, and multimodal representations for temporal understanding and reasoning.

5 Challenge and Future Direction

We further discuss the challenges and future directions under the three alignment paradigms proposed in this work.

5.1 Data

The evolution of datasets toward TSQA exposes several structural limitations. (1) **Lack of reasoning supervision** is a primary bottleneck. Most existing datasets emphasize short answers, while providing little supervision over intermediate reasoning steps or tool usage. Future datasets should explicitly incorporate reasoning traces and expected temporal behaviors to support reasoning [Divo *et al.*, 2025]. (2) **Insufficient domain diversity** constrains generalized evaluation. Time series have domain-specific variation in sequence length, granularity, noise patterns, and temporal dependencies. Constructing large-scale, multi-domain TSQA datasets with temporal diversity is critical [Chen *et al.*, 2025].

5.2 Task

TSA and TSQA are complementary rather than mutually exclusive. TSA provides effective task-specific solutions, whereas TSQA offers a unified question answering paradigm based on temporal reasoning. This evolution from TSA to TSQA also introduces fundamental challenges. (1) **Tasks shift from explicit specification to implicit extraction**. TSA assumes clearly defined tasks with specified objectives. In contrast, TSQA identifies the time series problems from user inputs that typically contain underlying temporal semantics. This shift requires models to interpret user intent into appropriate tasks [Tao *et al.*, 2024]. (2) **Lack of task tools for TSQA**. Although TSQA extends TSA toward a unified question answering paradigm, it often lacks tools explicitly designed for diverse temporal questions, such as time series slicing [Zhang *et al.*, 2025b]. Future TSQA should incorporate dedicated tools or invoke effective TSA models to enhance temporal reasoning [Kong *et al.*, 2025].

5.3 Method

Several challenges arise across optimization strategies. (1) **Limited adaptivity**. Current methods struggle with complexity and uncertainty. Future approaches should dynamically support long-horizon, multi-scale, and irregular time series under temporal reasoning [Gu *et al.*, 2025]. Incorporating multi-agent mechanisms can further enable planning,

tool use, and self-reflection, turning alignment into an adaptive process [Chang *et al.*, 2025a]. (2) **Numerical hallucination**. Current approaches struggle with numerical hallucination in TSA and TSQA. Discrete tokenization limits sensitivity to continuous values, producing plausible but incorrect reasoning. Future work should investigate how alignment paradigms mitigate this issue [Wu *et al.*, 2026].

6 Conclusion

In this survey, we categorize existing literature into three paradigms, summarize optimization strategies along the evolution from TSA to TSQA, and provide practical guidance for flexible, economical, and generalizable selection of alignment paradigms. We further analyze representative datasets, identify challenges, and highlight future research directions.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Nos. 62172423 and 62402414) and the Guangdong Basic and Applied Basic Research Foundation (No. 2025A1515011994).

References

[Bhatia *et al.*, 2024] Gagan Bhatia, El Moatez Billah Nagoudi, Hasan Cavusoglu, and Muhammad Abdul-Mageed. Fintral: A family of GPT-4 level multimodal financial large language models. In *ACL*, pages 13064–13087, 2024.

[Cao *et al.*, 2024] Defu Cao, Furong Jia, Sercan Ö. Arik, Tomas Pfister, Yixiang Zheng, Wen Ye, and Yan Liu. TEMPO: Prompt-based generative pre-trained transformer for time series forecasting. In *ICLR*, 2024.

[Chan *et al.*, 2024] Nimeesha Chan, Felix Parker, William Bennett, Tianyi Wu, Mung Yao Jia, et al. Medtsllm: Leveraging llms for multimodal medical time series analysis. *arXiv preprint arXiv:2408.07773*, 2024.

[Chang *et al.*, 2025a] Ching Chang, Yidan Shi, Defu Cao, Wei Yang, Jeehyun Hwang, Haixin Wang, Jiacheng Pang, Wei Wang, Yan Liu, Wen-Chih Peng, et al. A survey of reasoning and agentic systems in time series with large language models. *arXiv preprint arXiv:2509.11575*, 2025.

[Chang *et al.*, 2025b] Ching Chang, Wei-Yao Wang, Wen-Chih Peng, and Tien-Fu Chen. Llm4ts: Aligning pre-trained llms as data-efficient time-series forecasters. *ACM Transactions on Intelligent Systems and Technology*, 16(3):1–20, 2025.

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

[Cheng *et al.*, 2025a] Mingyue Cheng, Yiheng Chen, Qi Liu, Zhiding Liu, Yucong Luo, and Enhong Chen. Instructime: Advancing time series classification with multimodal language modeling. In *WSDM*, pages 792–800, 2025.

[Cheng *et al.*, 2025b] Yunyao Cheng, Chenjuan Guo, Kaixuan Chen, Kai Zhao, Bin Yang, Jiandong Xie, Christian S. Jensen, Feiteng Huang, and Kai Zheng. Gaussian process latent variable modeling for few-shot time series forecasting. *IEEE Trans. Knowl. Data Eng.*, 37(8):4604–4619, 2025.

[Ding *et al.*, 2023] Yujie Ding, Shuai Jia, Tianyi Ma, Bingcheng Mao, Xiuze Zhou, Liuliu Li, and Dongming Han. Integrating stock features and global information via large language models for enhanced stock return prediction. In *IJCAI*, 2023.

[Divo *et al.*, 2025] Felix Divo, Maurice Kraus, Anh Q Nguyen, Hao Xue, Imran Razzak, Flora D Salim, Kristian Kersting, and Devendra Singh Dhami. Quants: Question answering on time series. *arXiv preprint arXiv:2511.05124*, 2025.

[Duan *et al.*, 2023] Yiqun Duan, Jinzhao Zhou, Zhen Wang, Yu-Kai Wang, and Chin-Teng Lin. Dewave: Discrete EEG waves encoding for brain dynamics to text translation. In *NeurIPS*, 2023.

[Feng *et al.*, 2025] Jie Feng, Tianhui Liu, Yuwei Du, Siqi Guo, Yuming Lin, and Yong Li. Citygpt: Empowering urban spatial cognition of large language models. In *KDD*, pages 591–602, 2025.

[Gruver *et al.*, 2023] Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew Gordon Wilson. Large language models are zero-shot time series forecasters. In *NeurIPS*, 2023.

[Gu *et al.*, 2025] Yile Gu, Yifan Xiong, Jonathan Mace, Yuting Jiang, et al. Argos: Agentic time-series anomaly detection with autonomous rule generation via large language models. *arXiv preprint arXiv:2501.14170*, 2025.

[Huang *et al.*, 2024] Qihe Huang, Zhengyang Zhou, Kuo Yang, Gengyu Lin, Zhongchao Yi, and Yang Wang. Leret: Language-empowered retentive network for time series forecasting. In *IJCAI*, pages 4165–4173, 2024.

[Hurst *et al.*, 2024] Aaron Hurst, Adam Lerer, et al. Gpt-4o system card. *preprint arXiv:2410.21276*, 2024.

[Jiang *et al.*, 2025a] Yushan Jiang, Kanghui Ning, Zijie Pan, Xuyang Shen, Jingchao Ni, Wenchao Yu, Anderson Schneider, Haifeng Chen, Yuriy Nevmyvaka, and Dongjin Song. Multi-modal time series analysis: A tutorial and survey. In *KDD*, pages 6043–6053, 2025.

[Jiang *et al.*, 2025b] Yushan Jiang, Wenchao Yu, Geon Lee, Dongjin Song, Kijung Shin, Wei Cheng, Yanchi Liu, and Haifeng Chen. Timexl: Explainable multi-modal time series prediction with llm-in-the-loop. In *NeurIPS*, 2025.

[Jiang *et al.*, 2025c] Zuoyou Jiang, Li Zhao, Rui Sun, Ruohan Sun, Zhongjian Li, Jing Li, Daxin Jiang, Zuo Bai, and Cheng Hua. Alpha-r1: Alpha screening with llm reasoning via reinforcement learning. *arXiv preprint arXiv:2512.23515*, 2025.

[Jin *et al.*, 2024] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong

Wen. Time-llm: Time series forecasting by reprogramming large language models. In *ICLR*, 2024.

[Kong *et al.*, 2025] Yaxuan Kong, Yiyuan Yang, Yoontae Hwang, Wenjie Du, Stefan Zohren, Zhangyang Wang, Ming Jin, and Qingsong Wen. Time-mqa: Time series multi-task question answering with context enhancement. In *ACL*, pages 29736–29753, 2025.

[Lee *et al.*, 2025] Geon Lee, Wenchao Yu, Kijung Shin, Wei Cheng, and Haifeng Chen. Timecap: Learning to contextualize, augment, and predict time series events with large language model agents. In *AAAI*, pages 18082–18090, 2025.

[Lee *et al.*, 2026] Jaehoon Lee, Suhwan Park, Tae Yoon Lim, Seunghan Lee, Jun Seo, Dongwan Kang, Hwanil Choi, Minjae Kim, Sungdong Yoo, SoonYoung Lee, Yongjae Lee, and Wonbin Ahn. Fintexts: Financial text-paired time-series dataset via semantic-based and multi-level pairing. *preprint arXiv:2603.02702*, 2026.

[Li *et al.*, 2025a] Huayu Li, Zhengxiao He, Xiwen Chen, Ci Zhang, Stuart F. Quan, William D. S. Killgore, Shu-Fen Wung, Chen X. Chen, Geng Yuan, Jin Lu, and Ao Li. Smarter together: Combining large language models and small models for physiological signals visual inspection. *J. Heal. Informatics Res.*, 9(4):656–685, 2025.

[Li *et al.*, 2025b] Zechen Li, Shohreh Deldari, Linyao Chen, Hao Xue, and Flora D Salim. Sensorllm: Aligning large language models with motion sensors for human activity recognition. In *EMNLP*, pages 354–379, 2025.

[Liang *et al.*, 2024] Yuxuan Liang, Haomin Wen, Yuqi Nie, Yushan Jiang, Ming Jin, Dongjin Song, Shirui Pan, and Qingsong Wen. Foundation models for time series analysis: A tutorial and survey. In *KDD*, pages 6555–6565, 2024.

[Liu *et al.*, 2024a] Chen Liu, Shibo He, Qihang Zhou, Shizhong Li, and Wenchao Meng. Large language model guided knowledge distillation for time series anomaly detection. In *IJCAI*, pages 2162–2170, 2024.

[Liu *et al.*, 2024b] Haoxin Liu, Zhiyuan Zhao, Jindong Wang, Harshavardhan Kamarthi, and B. Aditya Prakash. Lstprompt: Large language models as zero-shot time series forecasters by long-short-term prompting. In *ACL*, pages 7832–7840, 2024.

[Liu *et al.*, 2024c] Qingxiang Liu, Xu Liu, Chenghao Liu, Qingsong Wen, and Yuxuan Liang. Time-ffm: Towards lm-empowered federated foundation model for time series forecasting. In *NeurIPS*, 2024.

[Liu *et al.*, 2024d] Xu Liu, Junfeng Hu, Yuan Li, Shizhe Diao, Yuxuan Liang, Bryan Hooi, and Roger Zimmermann. Unitime: A language-empowered unified model for cross-domain time series forecasting. In *WWW*, pages 4095–4106, 2024.

[Liu *et al.*, 2024e] Yong Liu, Guo Qin, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Autotimes: Autoregressive time series forecasters via large language models. In *NeurIPS*, 2024.

[Liu *et al.*, 2025a] Chenxi Liu, Hao Miao, Qianxiong Xu, Shaowen Zhou, Cheng Long, Yan Zhao, Ziyue Li, and Rui Zhao. Efficient multivariate time series forecasting via calibrated language models with privileged knowledge distillation. In *ICDE*, pages 3165–3178, 2025.

[Liu *et al.*, 2025b] Chenxi Liu, Qianxiong Xu, Hao Miao, Sun Yang, Lingzheng Zhang, Cheng Long, Ziyue Li, and Rui Zhao. Timecma: Towards llm-empowered multivariate time series forecasting via cross-modality alignment. In *AAAI*, pages 18780–18788, 2025.

[Liu *et al.*, 2025c] Chenxi Liu, Shaowen Zhou, Qianxiong Xu, Hao Miao, Cheng Long, Ziyue Li, and Rui Zhao. Towards cross-modality modeling for time series analytics: A survey in the LLM era. In *IJCAI*, pages 10564–10572, 2025.

[Liu *et al.*, 2025d] Haoxin Liu, Chenghao Liu, and B. Aditya Prakash. A picture is worth A thousand numbers: Enabling llms reason about time series via visualization. In *NAACL*, pages 7486–7518, 2025.

[Liu *et al.*, 2025e] Jun Liu, Chaoyun Zhang, Jiaxu Qian, Minghua Ma, Si Qin, Chetan Bansal, Qingwei Lin, Saravan Rajmohan, and Dongmei Zhang. Large language models can deliver accurate and interpretable time series anomaly detection. In *SIGKDD*, pages 4623–4634, 2025.

[Liu *et al.*, 2025f] Peiyuan Liu, Hang Guo, Tao Dai, Naiqi Li, Jigang Bao, Xudong Ren, Yong Jiang, and Shu-Tao Xia. CALF: aligning llms for time series forecasting via cross-modal fine-tuning. In *AAAI*, pages 18915–18923, 2025.

[Merrill *et al.*, 2024] Mike A. Merrill, Mingtian Tan, Vinayak Gupta, Thomas Hartvigsen, and Tim Althoff. Language models still struggle to zero-shot reason about time series. In *EMNLP(Findings)*, pages 3512–3533, 2024.

[Ni *et al.*, 2026] Juntong Ni, Shiyu Wang, Ming Jin, Qi He, and Wei Jin. Streasoner: Empowering llms for spatio-temporal reasoning in time series via spatial-aware reinforcement learning. *arXiv preprint arXiv:2601.03248*, 2026.

[Oh *et al.*, 2024] Jungwoo Oh, Gyubok Lee, Seongsu Bae, et al. Ecg-qa: A comprehensive question answering dataset combined with electrocardiogram. *NeurIPS*, 36, 2024.

[Pan *et al.*, 2024] Zijie Pan, Yushan Jiang, Sahil Garg, Anderson Schneider, Yuriy Nevmyvaka, and Dongjin Song. S2IP-LLM: semantic space informed prompt learning with LLM for time series forecasting. In *ICML*, 2024.

[Pan *et al.*, 2025] Qihong Pan, Haofei Tan, Guojiang Shen, Xiangjie Kong, Mengmeng Wang, and Chenyang Xu. LLM-TPF: multiscale temporal periodicity-semantic fusion llms for time series forecasting. In *IJCAI*, pages 6030–6038, 2025.

[Qiu *et al.*, 2023] Jielin Qiu, William Han, Jiacheng Zhu, Mengdi Xu, Michael A. Rosenberg, Emerson Liu, Douglas Weber, and Ding Zhao. Transfer knowledge from

natural language to electrocardiography: Can we detect cardiovascular disease through language models? In *EACL(Findings)*, pages 442–453, 2023.

[Reichman *et al.*, 2025] Benjamin Z. Reichman, Xiaofan Yu, Lanxiang Hu, Jack Truxal, Atishay Jain, Rushil Chandrupatla, Tajana Rosing, and Larry Heck. Sensorqa: A question answering benchmark for daily-life monitoring. In *SenSys*, pages 282–289, 2025.

[Tang *et al.*, 2024] Hua Tang, Chong Zhang, Mingyu Jin, Qinkai Yu, Zhenting Wang, Xiaobo Jin, Yongfeng Zhang, and Mengnan Du. Time series forecasting with llms: Understanding and enhancing model capabilities. *SIGKDD Explor.*, 26(2):109–118, 2024.

[Tao *et al.*, 2024] Xiaoyu Tao, Tingyue Pan, Mingyue Cheng, Yucong Luo, Qi Liu, and Enhong Chen. Hierarchical multimodal llms with semantic space alignment for enhanced time series classification. *ACM Transactions on Intelligent Systems and Technology*, 2024.

[Trabelsi *et al.*, 2025] Mohamed Trabelsi, Aidan Boyd, Jin Cao, and Huseyin Uzunalioglu. Time series language model for descriptive caption generation. *arXiv preprint arXiv:2501.01832*, 2025.

[Wang *et al.*, 2025a] Chengsen Wang, Qi Qi, Jingyu Wang, Haifeng Sun, Zirui Zhuang, Jinming Wu, Lei Zhang, and Jianxin Liao. Chattime: A unified multimodal time series foundation model bridging numerical and textual data. In *AAAI*, pages 12694–12702, 2025.

[Wang *et al.*, 2025b] Jiahao Wang, Mingyue Cheng, et al. Tabletime: Reformulating time series classification as training-free table understanding with large language models. In *CIKM*, pages 3009–3019, 2025.

[Wang *et al.*, 2025c] Jun Wang, Wenjie Du, Yiyuan Yang, Linglong Qian, Wei Cao, Keli Zhang, Wenjia Wang, Yuxuan Liang, and Qingsong Wen. Deep learning for multivariate time series imputation: A survey. In *IJCAI*, pages 10696–10704, 2025.

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

[Wu *et al.*, 2023] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*, 2023.

[Wu *et al.*, 2026] Xingjian Wu, Junkai Lu, Zhengyu Li, Xiangfei Qiu, Jilin Hu, Chenjuan Guo, Christian S. Jensen, and Bin Yang. Timeart: Towards agentic time series reasoning via tool-augmentation. *preprint arXiv:2601.13653*, 2026.

[Xie *et al.*, 2025] Zhe Xie, Zeyan Li, Xiao He, Longlong Xu, Xidao Wen, Tieying Zhang, Jianjun Chen, Rui Shi, and Dan Pei. Chatts: Aligning time series with llms via synthetic data for enhanced understanding and reasoning. *VLDB*, 18(8):2385–2398, 2025.

[Xue and Salim, 2024] Hao Xue and Flora D. Salim. Promptcast: A new prompt-based learning paradigm for time series forecasting. *IEEE Trans. Knowl. Data Eng.*, 36(11):6851–6864, 2024.

[Ye *et al.*, 2025] Jiexia Ye, Weiqi Zhang, Ziyue Li, Jia Li, Meng Zhao, and Fugee Tsung. Medualtime: A dual-adapter language model for medical time series-text multimodal learning. In *IJCAI*, pages 7913–7921, 2025.

[Yu *et al.*, 2025] Xiaofan Yu, Lanxiang Hu, Benjamin Z. Reichman, Dylan Chu, Rushil Chandrupatla, Xiyuan Zhang, Larry Heck, and Tajana Simunic Rosing. Sensorchat: Answering qualitative and quantitative questions during long-term multimodal sensor interactions. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 9(3):148:1–148:35, 2025.

[Zhang *et al.*, 2023] Boyu Zhang, Hongyang Yang, and Xiao-Yang Liu. Instruct-fingpt: Financial sentiment analysis by instruction tuning of general-purpose large language models. *arXiv preprint arXiv:2306.12659*, 2023.

[Zhang *et al.*, 2024] Xiyuan Zhang, Ranak Roy Chowdhury, Rajesh K. Gupta, and Jingbo Shang. Large language models for time series: A survey. In *IJCAI*, pages 8335–8343, 2024.

[Zhang *et al.*, 2025a] Juyuan Zhang, Wei Zhu, and Jiechao Gao. Adapting large language models for time series modeling via a novel parameter-efficient adaptation method. In *ACL*, 2025.

[Zhang *et al.*, 2025b] Zhirui Zhang, Changhua Pei, Tianyi Gao, Zhe Xie, Yibo Hao, Zhaoyang Yu, Longlong Xu, Tong Xiao, Jing Han, and Dan Pei. Timesense: Making large language models proficient in time-series analysis. *arXiv preprint arXiv:2511.06344*, 2025.

[Zhang *et al.*, 2026] Jianqi Zhang, Jingyao Wang, Wenwen Qiang, Fanjiang Xu, and Changwen Zheng. Enhancing large language models for time-series forecasting via vector-injected in-context learning. *preprint arXiv:2601.07903*, 2026.

[Zhao *et al.*, 2025] Yubao Zhao, Jiaju Kang, Tian Zhang, Puyu Han, and Tong Chen. Ecg-chat: A large ecg-language model for cardiac disease diagnosis. In *ICME*, pages 1–6, 2025.

[Zhong *et al.*, 2025] Siru Zhong, Weilin Ruan, Ming Jin, Huan Li, Qingsong Wen, and Yuxuan Liang. Time-vlm: Exploring multimodal vision-language models for augmented time series forecasting. In *ICML*, 2025.

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

[Zhou *et al.*, 2024] Xiaomao Zhou, Qingmin Jia, Yujiao Hu, Renchao Xie, Tao Huang, and F. Richard Yu. Geng: An llm-based generic time series data generation approach for edge intelligence via cross-domain collaboration. In *INFOCOM*, pages 1–6. IEEE, 2024.