A Survey of Reasoning and Agentic Systems in Time Series with Large Language Models

Anonymous authors
Paper under double-blind review

Abstract

Time series reasoning treats time as a first-class axis and incorporates intermediate evidence directly into the answer. This survey defines the problem and organizes the literature by reasoning topology with three families: direct reasoning in one step, linear chain reasoning with explicit intermediates, and branch-structured reasoning that explores, revises, and aggregates. The topology is crossed with the main objectives of the field, including traditional time series analysis, explanation and understanding, causal inference and decision making, and time series generation, while a compact tag set spans these axes and captures decomposition and verification, ensembling, tool use, knowledge access, multimodality, agent loops, and LLM alignment regimes. Methods and systems are reviewed across domains, showing what each topology enables and where it breaks down in faithfulness or robustness, along with curated datasets, benchmarks, and resources that support study and deployment (with an accompanying repository at `https://anonymous.4open.science/r/Time-Series-Reasoning-Survey-TMLR/`). Evaluation practices that keep evidence visible and temporally aligned are highlighted, and guidance is distilled on matching topology to uncertainty, grounding with observable artifacts, planning for shift and streaming, and treating cost and latency as design budgets. We emphasize that reasoning structures must balance capacity for grounding and self-correction against computational cost and reproducibility, while future progress will likely depend on benchmarks that tie reasoning quality to utility and on closed-loop testbeds that trade off cost and risk under shift-aware, streaming, and long-horizon settings. Taken together, these directions mark a shift from narrow accuracy toward reliability at scale, enabling systems that not only analyze but also understand, explain, and act on dynamic worlds with traceable evidence and credible outcomes.

Contents

1 Introduction

Time series data are common in everyday life, recording how variables evolve and interact over time in fields like finance, healthcare, energy, climate, transport, and manufacturing processes (Chang et al., 2025c; Ou et al., 2024; Chang et al., 2025b; 2024c; Zhao et al., 2020; Cao et al., 2024b; Niu et al., 2024; Cao et al., 2022; Li et al., 2025c). Decades of effort have made time series analysis one of the key methodologies used in monitoring, forecasting, diagnostics, and decision-making, and it has countless uses in areas such as risk modeling and patient monitoring, demand, and predictive maintenance (Chang et al., 2024a; Lin et al., 2024; Chang et al., 2024b; Lo et al., 2024; Cao et al., 2020). Existing surveys on time series have generally focused on modeling and algorithmic methods. These comprise surveys on deep learning forecasting methods (Lim & Zohren, 2021; Torres et al., 2021; Mahalakshmi et al., 2016; Liu et al., 2021; Benidis et al., 2022), architectures employing transformers (Wen et al., 2023), anomaly detection (Zamanzadeh Darban et al., 2024), classification (Ismail Fawaz et al., 2019), clustering (Liao, 2005), discovery of motifs (Torkamani & Lohweg, 2017), change-point detection (Aminikhanghahi & Cook, 2017), segmentation (Keogh et al., 2004), compression (Chiarot & Silvestri, 2023), and data augmentation (Victor & Ali, 2024; Wen et al., 2021). Respectively, these works are concerned with increasing predictive accuracy, representation, and efficiency in dealing with sequential temporal data. However, many emerging applications demand more than prediction. Domains such as personalized healthcare, adaptive risk management, and autonomous systems require models that can explain their outputs, reason about counterfactuals, and decide among alternative actions. These demands underscore that advancing time series analysis requires structured and reliable reasoning. Despite this breadth, the literature to date has not covered reasoning, explanation, or agent-based decision-making under time series. To our best knowledge, no work has been devoted to investigating how methods under time series can be used toward enabling higher-level reasoning or policy-oriented actions.

The advent of large language models (LLMs) is another turning point. Besides fitting patterns, LLMs can exhibit step-by-step reasoning traces (Ke et al., 2025; Zhang et al., 2024c; Huang & Chang, 2023; Chu et al., 2024; Zhang et al., 2025g; Yang & Thomason, 2025; Xiao et al., 2024), articulate causal hypotheses (Li et al., 2025d; Liu et al., 2025f; Kiciman et al., 2023; Zhang et al., 2022; Cao et al., 2023), and interact with external tools and environments (Shen, 2024; Ferrag et al., 2025; Yang et al., 2025a; Chen et al., 2025b). When incorporated into agentic systems, they gain the capacity for planning (Huang et al., 2024; Wei et al., 2025), reflection (Renze & Guven, 2024; Ji et al., 2023), and continual adaptation (Fujii et al., 2024; Shi et al., 2025), changing time series modeling from static prediction to interactive and explanatory processes(Ye et al., 2024). This shift opens up the space of downstream tasks: instead of just prediction or anomaly detection, models are now expected to handle causal analysis, natural language reasoning, simulating and editing temporal signals, and making policy-driven decisions.

Building on this transformation, our survey is structured on the basis of three intersecting trends that shape the future landscape. First, time series data are increasingly widespread and significant, driving practical systems that require clarity, versatility, and strong decision-making under uncertainty. Second, LLMs and multimodal LLMs have demonstrated unprecedented flexibility in reasoning and generalization, creating opportunities to recast time series problems in natural language and symbolic forms. Third, the rise of autonomous agents driven by LLMs allows models not only to analyze time series but also to act upon them—through simulation, intervention, or iterative decision loops. Motivated by these developments, we define and study time series reasoning (TSR) as the class of methods where LLMs explicitly execute structured reasoning procedures over temporally indexed data, potentially enriched by multimodal context and agentic systems. This survey presents the first systematic taxonomy of the field, organized around distinct reasoning topologies and primary objectives, and complemented by lightweight attribute tags that capture control-flow operators (such as decomposition, verification, and ensembling), actors (including tool use and agentic loops), modality and knowledge access, and alignment regimes specific to LLMs, as illustrated in Figure 1.

This survey makes three contributions. (i) We introduce the first systematic taxonomy of time series reasoning, structured along two complementary axes: reasoning topologies (execution structures) and primary objectives (task intents), and further enriched with lightweight attribute tags that capture control-flow operators, actors (tools and agent loops), modality and knowledge access, and LLM alignment regimes. (ii) We provide an integrated review that not only analyzes patterns across reasoning topologies and objectives in research papers,

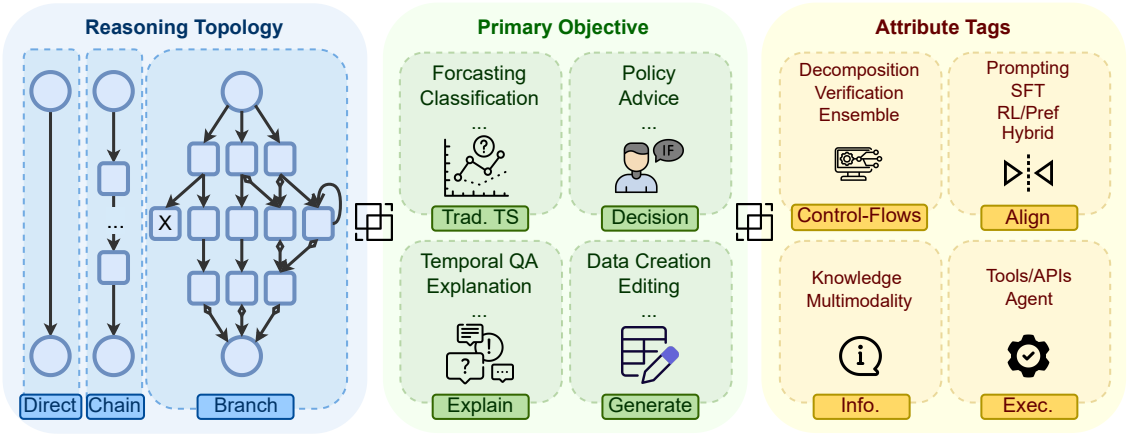

Figure 1: Framework of reasoning topologies and primary objectives, complemented by lightweight attribute tags.

but also categorizes complementary contributions such as datasets, benchmarks, surveys, tutorials, position and vision papers—highlighting how these works support and shape the development of time series reasoning. (iii) We highlight open problems in evaluation and benchmarking, multimodal fusion and alignment, retrieval and knowledge grounding, long-context reasoning, memory and efficiency, agentic control and tool use, as well as causal inference and decision support—laying out a research agenda for the next stage of time series reasoning.

The rest of the paper is organized as follows. Section 2 formalizes the notion of time series reasoning and introduces our taxonomy, decision sketches, and a systematic labeling pipeline that annotates each paper with reasoning topology, primary objectives, and attribute tags. Sections 3–5 analyze the three reasoning topologies in depth, while Section 6 surveys datasets, benchmarks, evaluation protocols, and auxiliary resources, including recent controversies and counter-evidence (§6.3) that highlight limitations and ongoing debates. Finally, Section 7 outlines open problems and future directions. Together, these sections establish a unified taxonomy, a reproducible labeling of over one hundred papers, and a synthesis of methodological trends and challenges, aiming to serve both researchers developing novel reasoning systems for time series and practitioners seeking a structured guide to the current landscape and its open questions.

2 Background and Taxonomy

2.1 What We Mean by Time Series Reasoning

Time series reasoning (TSR) refers to methods that operate over temporally indexed data while executing an explicit reasoning procedure. These methods are increasingly enabled by large language models (LLMs) and multimodal LLMs, which can articulate reasoning traces, interact with external tools, and operate as autonomous agents. In doing so, they not only strengthen traditional time series analysis tasks such as forecasting, anomaly detection, and classification, but also extend the scope of what is possible by enabling explanation, intervention, and generation of temporal dynamics. Such reasoning may take the form of single-step inference, multi-step decomposition, or branching exploration that allows both divergence and feedback across reasoning paths, reflecting an expanded view of how models can reason with time series.

In our taxonomy, TSR is defined by three complementary components: the **Reasoning Topology** (Section 2.2), which specifies the execution structure; the **Primary Objective** (Section 2.3), which clarifies the main intent of the reasoning process; and a set of **Attribute Tags** (Section 2.4), which describe auxiliary properties such as control-flow, actors, modality, and alignment. The first two levels—reasoning topology and primary objective—are *mutually exclusive*: each paper is assigned exactly one topology and exactly

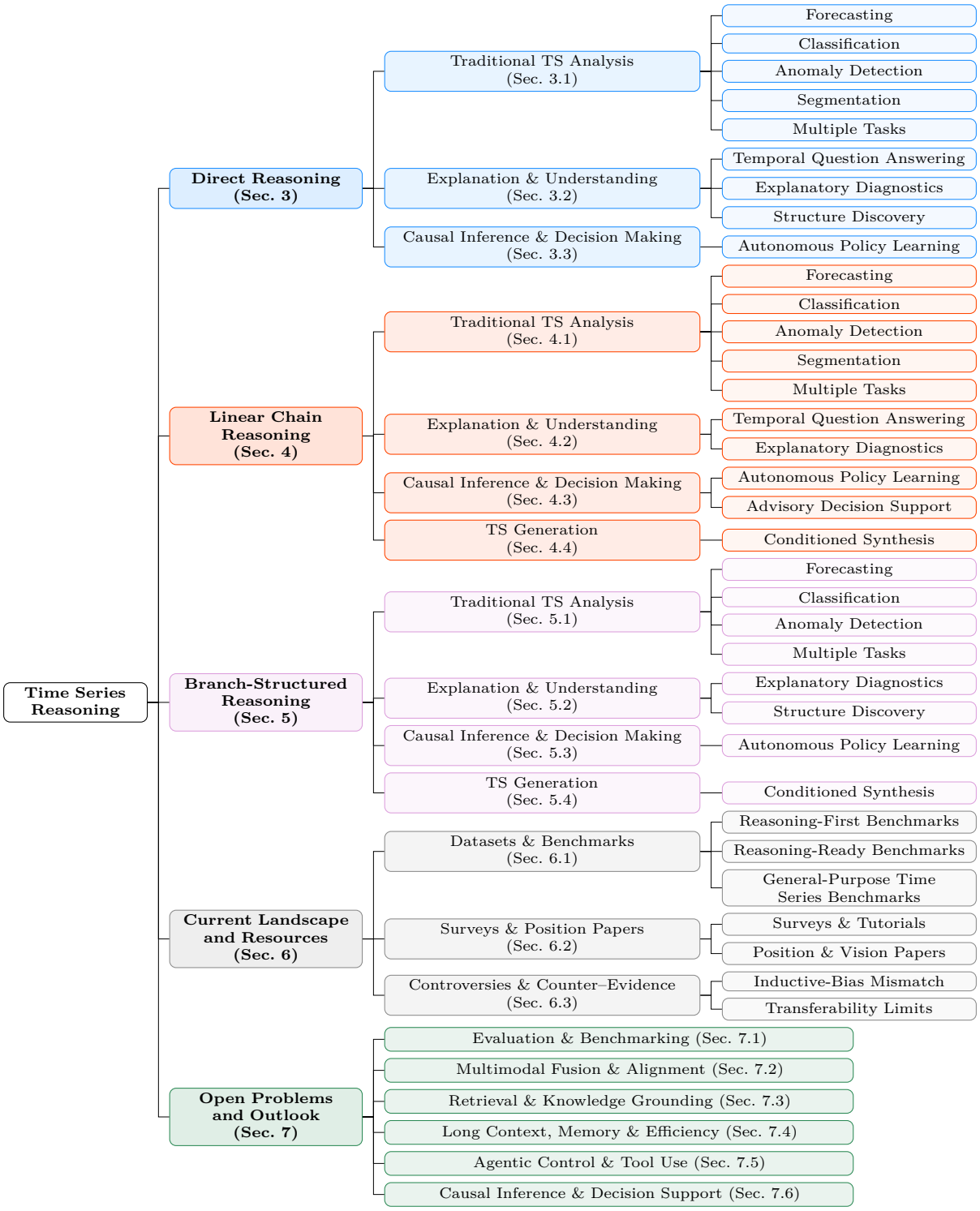

Figure 2: Taxonomy of time series reasoning literature

one objective, according to the minimal structural test and the dominant evaluation focus. In contrast, attribute tags are *non-exclusive*: a paper may carry multiple tags at once, since it can simultaneously employ decomposition, use tools, access multimodal inputs, and involve specific alignment regimes. One way to view tags is as a complete vector of attributes, where some values are explicitly marked as present and others

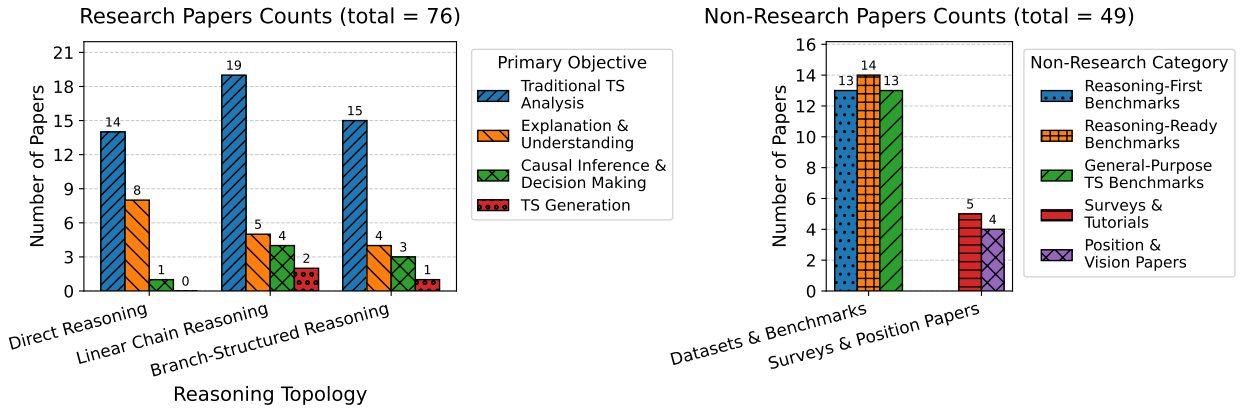

Figure 3: Number of surveyed papers: research (left) and non-research (right).

are simply inactive. Figure 2 presents the taxonomy with reasoning topologies and primary objectives, and Figure 3 complements it by showing the number of surveyed papers split into research and non-research categories. For clarity, attribute tags are not included in these figures and are discussed separately in Section 2.4. For completeness, Appendix A provides the full list of curated research papers with their assigned reasoning topologies, primary objectives, and attribute tags.

2.2 Reasoning Topology

We identify three mutually exclusive reasoning topologies, corresponding to minimal structural tests over the reasoning trace: direct reasoning, linear chain reasoning, and branch-structured reasoning (as illustrated in Figure 4). These topologies form a spectrum of increasing complexity: from direct reasoning with a single step, to linear chains of sequential steps, to branch-structured reasoning that supports in-trajectory exploration, reconnection, and feedback. This progression highlights how reasoning can evolve from simple one-shot inference to richer temporal analysis that coordinates among alternative paths.

Direct Reasoning. Direct reasoning denotes the simplest form of execution: a single-step inference or tool call without any intermediate reasoning traces. The model jumps directly from input to output, producing a forecast, classification, or anomaly label without decomposing the problem or iterating over solutions. Such execution may be implicit, with internal reasoning hidden within the model's parameters, or minimal, with only the final output exposed. Direct reasoning is commonly used as a baseline or reference point, since it maximizes efficiency and requires no orchestration overhead, but it also limits interpretability, robustness to errors, and adaptability to complex or multi-stage tasks. Despite these limitations, direct reasoning remains prevalent in practice for straightforward forecasting benchmarks, anomaly detection pipelines, or descriptive question answering when transparency and intermediate supervision are not required.

Linear Chain Reasoning. Linear chain reasoning extends beyond direct inference by introducing a sequence of reasoning steps arranged in a straight path. Each step depends on the output of the previous one, forming a logical progression such as step-by-step forecasting, causal analysis, or explanation. This sequential structure allows intermediate states to be explicitly represented, inspected, or revised in later steps, thereby offering greater interpretability and modularity than direct reasoning. Chains are especially useful when tasks naturally unfold in stages, or when human users or downstream systems benefit from observing intermediate results. However, the linear chain topology remains restricted to a single path with no branching, feedback loops, or cross-branch aggregation, which limits its flexibility in exploring multiple hypotheses or adapting dynamically during execution.

Branch-Structured Reasoning. Branch-structured reasoning represents any topology where the reasoning trace can branch into multiple paths within a single execution. Branches may arise when the model explores

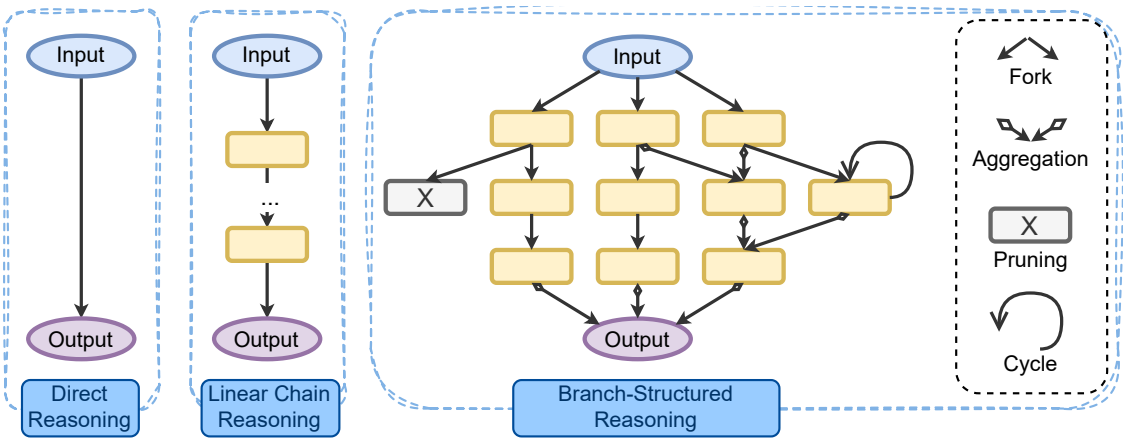

Figure 4: Three types of reasoning topologies: direct reasoning, linear chain reasoning, and branch-structured reasoning. Yellow boxes represent intermediate reasoning steps. Branch-structured reasoning additionally supports four structures: fork, aggregation, pruning, and cycle.

different hypotheses, candidate forecasts, explanations, or plans in parallel or sequentially, creating sibling nodes that diverge from a common ancestor. Branching may involve simple divergence, where alternative paths are explored independently, or more complex interactions, where later steps feed back to earlier ones or combine information from multiple branches. For example, feedback loops can revise or regenerate earlier outputs, while cross-branch operations can create new steps that depend on several existing paths. Aggregation is also part of this class: a parent node may select or rank among its children, or a new step may fuse multiple branches into a shared outcome. Compared to linear chain reasoning, branch-structured reasoning enables exploration of alternatives, adaptive revision of earlier steps, and reuse of intermediate results, but it also raises challenges such as controlling the growth of branches, handling feedback cycles, defining stopping conditions, and ensuring reproducibility. To manage combinatorial growth, branch-structured methods typically employ pruning, which eliminates low-promise branches early (for example, beam-style cutoffs or budgeted search) within the same execution trace.

2.3 Primary Objective

The primary objective of a TSR method captures the ultimate purpose of its reasoning process. While reasoning topology specifies how a model arrives at an answer, the primary objective defines the intended outcome of that process. This distinction is essential, since structurally similar reasoning strategies may pursue very different ends—for example, a chain of reasoning might be used either to forecast future values, to explain causal mechanisms, or to simulate new scenarios. We group objectives into four broad categories that span the major goals of time series reasoning: traditional time series analysis, explanation and understanding, causal inference and decision making, and time series generation. These categories provide a complementary view to reasoning topology, helping us compare methods not only by how they reason, but also by why they reason.

2.3.1 Traditional Time Series Analysis

This category covers predictive and descriptive tasks that directly model temporal dynamics. It serves as the foundation of time series reasoning, focusing on core supervised objectives such as predicting future values, assigning labels, detecting irregularities, and segmenting sequences into meaningful parts.

Forecasting. Reasoning-oriented forecasting treats prediction not only as extrapolation of past values but as an explicit reasoning process that interprets temporal patterns and conditions on context before projecting

into the future. Evaluation emphasizes point accuracy and probabilistic calibration, typically using mean absolute error, root mean squared error, and the continuous ranked probability score.

Classification. Reasoning for classification involves mapping temporal sequences to semantic categories through structured prompts, multimodal alignment, or stepwise inference, rather than treating label assignment as a black-box mapping. Evaluation focuses on robustness under imbalance and overall correctness, commonly using accuracy, F1, area under the receiver operating characteristic curve, and area under the precision–recall curve.

Anomaly Detection. Here reasoning is used to discern whether irregular points or intervals are true anomalies, often by contrasting candidate explanations, incorporating domain knowledge, or verifying suspicious patterns against context. Evaluation prioritizes correct localization and event quality, commonly using precision, recall, F1, and event-level F1, sometimes alongside detection delay.

Segmentation. Reasoning-based segmentation decomposes a sequence into meaningful sub-intervals or detects change points by combining statistical cues with interpretable decision rules, producing boundaries that reflect underlying dynamics. Evaluation emphasizes boundary accuracy and stability, for example boundary F1 and mean absolute boundary error.

Multiple Tasks. Unified reasoning frameworks tackle several objectives simultaneously, for example forecasting and classification, by reusing reasoning traces or branching workflows across tasks. Evaluation reports task-specific metrics for each included objective, for example mean squared error for forecasting and F1 for classification.

2.3.2 Explanation and Understanding

This category emphasizes reasoning that produces human-interpretable insights about temporal phenomena rather than raw predictions. It encompasses objectives such as answering scoped temporal questions, generating diagnostic narratives that clarify underlying causes, and discovering structural representations like causal tuples or symbolic rules.

Temporal Question Answering. Reasoning appears as the ability to parse a question about a time-indexed signal, retrieve relevant evidence, and articulate a direct answer grounded in temporal context. Evaluation measures answer correctness and grounding quality, commonly using question answering accuracy, exact match, and faithfulness or sufficiency scores.

Explanatory Diagnostics. These methods emphasize reasoning that connect observed outcomes to underlying causes, producing diagnostic narratives or structured explanations that clarify temporal behavior. Evaluation centers on explanation quality, commonly using human- or model-rated helpfulness, faithfulness, and coverage of salient events.

Structure Discovery. Reasoning is made explicit by generating candidate causal tuples, symbolic rules, or mechanistic abstractions and refining them into explanatory structures that capture time-series dependencies. Evaluation focuses on structure recovery quality, for example structural Hamming distance, edge precision and recall, and rule fidelity or coverage.

2.3.3 Causal Inference and Decision Making

This category focuses on reasoning about interventions and their outcomes in temporal settings. It covers autonomous policy learning, where models derive and execute action strategies directly from temporal states, as well as advisory decision support, where systems provide justified recommendations or what-if analyses to assist human decision makers.

Autonomous Policy Learning. Reasoning traces in this setting reveal how models deliberate over temporal states, weigh possible interventions, and converge on action policies without human intervention.

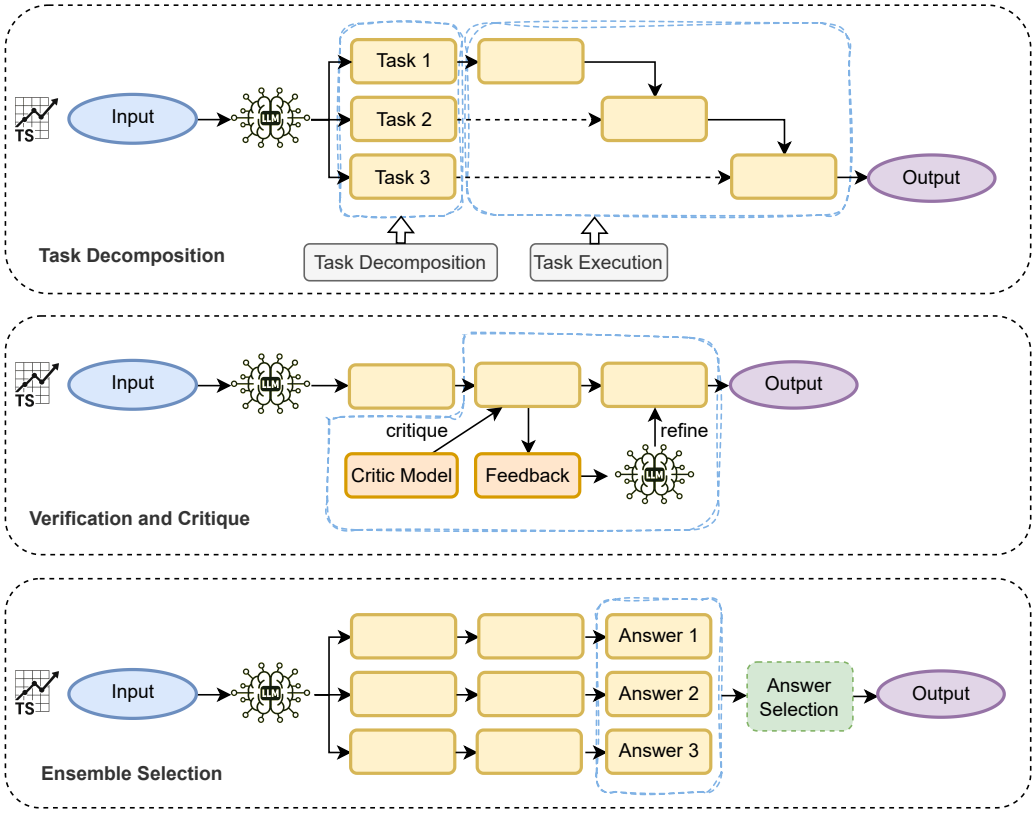

Figure 5: Control-flow operators: task decomposition, verification and critique, and ensemble selection.

Evaluation emphasizes realized performance and reliability, commonly using cumulative reward, regret, policy value, Sharpe ratio, Sortino ratio, and control-specific key performance indicators.

Advisory Decision Support. The reasoning process is used to justify and rank candidate interventions, providing humans with transparent rationales and comparative analyses rather than raw predictions alone. Evaluation measures decision quality with users and adoption in practice, commonly using outcome improvement in user studies, choice consistency, and perceived usefulness of explanations.

2.3.4 Time Series Generation

This category concerns the direct creation or modification of temporal data. It includes simulation of synthetic series and scenario-driven generation where synthetic time series follow intended patterns.

Conditioned Synthesis. Generative reasoning maps prompts or specifications into temporal dynamics, often requiring stepwise or branching inference to ensure the synthetic series follows intended trends or event patterns. Evaluation focuses on distributional fidelity, controllability, and diversity, commonly using maximum mean discrepancy and Kullback–Leibler divergence together with adherence to specified constraints.

2.4 Attribute Tags

Beyond reasoning topology and primary objective, we record lightweight, non-exclusive attribute tags to capture additional properties of each work. These tags provide finer-grained descriptors and are grouped into four categories. Most tags are *binary*, meaning they are either present or absent in a given run. Only the

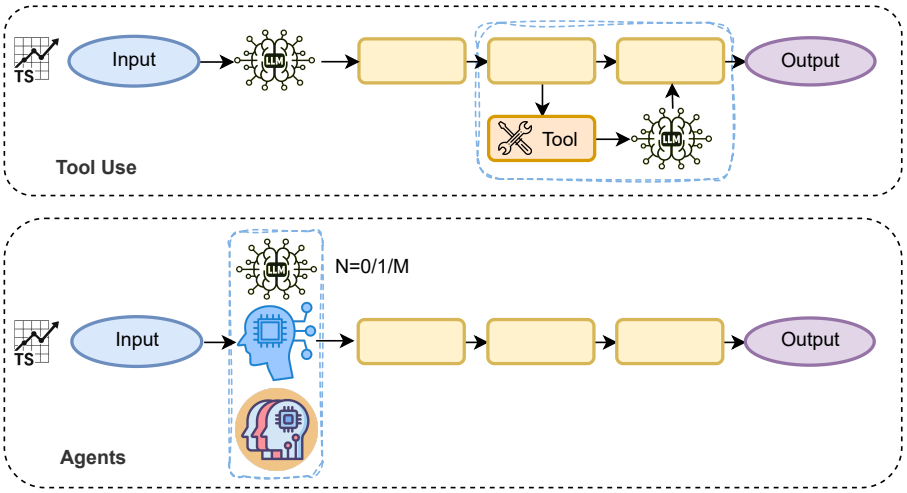

Figure 6: Execution actors: tool use, single-agent reasoning, and multi-agent reasoning.

agent tag and the LLM alignment tag are *categorical*, requiring the assignment of exactly one value from a predefined set. Importantly, tags never override the reasoning topology: they describe auxiliary behaviors or properties observed in the execution trace, while the topology is always determined directly from the structure of that trace.

2.4.1 Control-Flow Operators

Control-flow operators characterize how the reasoning process is organized at the step-to-step level. The labeling rule is simple: we assign an operator tag only when its behavior is explicitly shown in the reasoning trace or method description, not when it is merely suggested by the task interface or stated as an intention. The control-flow operators we track are task decomposition, verification and critique, and ensemble selection, as illustrated in Figure 5.

Task Decomposition. Task decomposition is present when the method explicitly enumerates subproblems, subquestions, or subplans that structure subsequent execution. The operator is evidenced by visible subgoal statements or by a planner that emits discrete substeps used downstream. Task decomposition by itself does not determine the reasoning topology. When the subgoals are executed one after another in sequence, the resulting topology is linear chain reasoning. When multiple alternatives are explored, whether in parallel, independently, or with later feedback and recombination, the resulting topology is branch-structured reasoning. This outcome is independent of verification and critique, which may or may not be present as separate operators.

Verification and Critique. Verification and critique are present when there is an explicit step that evaluates candidate outputs or intermediate reasoning through judging, checking, critiquing, or self-refinement. Silent heuristics or implicit scoring internal to a single step without an externally visible judging action do not count as verification. The operator is evidenced by a visible critic, judge, or scoring step, which may be carried out by the same model, another model, or a human. If the evaluation is performed without inducing revisions, the reasoning topology remains unchanged and can be either direct reasoning, linear chain reasoning, or branch-structured reasoning, depending on the surrounding structure. When verification leads to regeneration, edits, or revisions to earlier content, the execution trace is branch-structured reasoning, since feedback creates additional paths or reconnects to previous ones. This includes multi-round self-refinement protocols, iterative critique-and-revise loops, and agent debates where arguments trigger revisions across steps. By contrast, single-round evaluation or one-shot debates that only select among existing candidates preserve the underlying topology without adding feedback or new branches.

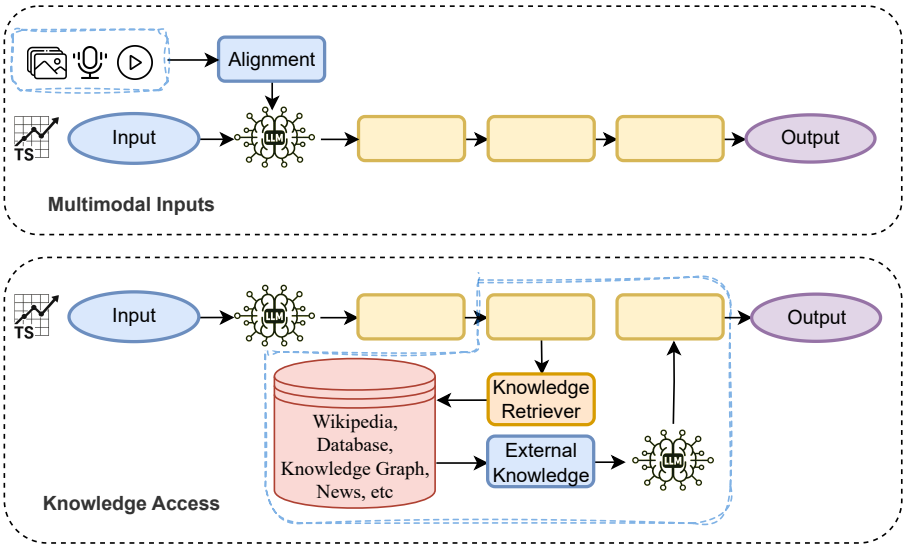

Figure 7: Information sources: multimodal inputs and external knowledge access.

Ensemble Selection. Ensemble selection, often called self-ensemble or self-consistency in the LLM literature, is present when multiple candidate reasoning traces or predictions are explicitly compared and a final outcome is chosen by a rule such as voting, ranking, or averaging. When the candidates are produced by restarting the same procedure multiple times and then resolved only at the end, the trace remains linear chain reasoning. When multiple alternatives are maintained within a single run and later resolved by selection, ranking, or fusion, the trace is branch-structured reasoning.

2.4.2 Execution Actors

Execution actors specify the entities responsible for carrying out reasoning steps during execution, as illustrated in Figure 6. They indicate whether reasoning is performed solely by the model itself, delegated to external tools, or organized through autonomous agents that act at inference time.

Tool Use. Tool use is present when the model invokes external resources such as search engines, solvers, or simulators during reasoning. The operator is evidenced by explicit calls to external systems whose outputs feed into subsequent reasoning steps. Tools are passive: they return information or computations but do not initiate new reasoning themselves.

Agents. The agent tag captures cases where autonomous agents are present at inference time. An autonomous agent is a component that, given its current state, selects the next action or message in pursuit of a goal, often powered by an LLM and sometimes using tools or memory. This tag is categorical rather than binary: it records the number of agents, with possible values *0 = no agent*, *1 = a single agent*, and *M = multiple collaborating agents*. The overall reasoning topology is determined by how the agents interact. A one-round manager–worker handoff typically corresponds to linear chain reasoning, whereas scenarios involving multiple workers that propose alternatives and are later merged, or multi-round coordination and debate with feedback, are forms of branch-structured reasoning.

2.4.3 Information Sources

Information sources capture inputs that extend beyond the raw time series itself, as illustrated in Figure 7. They cover both additional modalities, such as language or images, and external knowledge retrieved from databases, search engines, or domain resources.

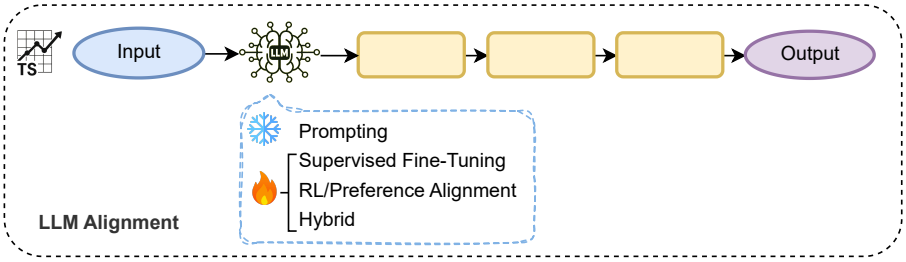

Figure 8: LLM alignment regimes: prompting, supervised fine-tuning, reinforcement/preference alignment, and hybrid approaches.

Multimodal Inputs. Multimodal inputs occur when time series are combined with other modalities such as natural language, images, audio, or structured reports. Such settings highlight scenarios where reasoning must integrate signals across different types of data rather than relying on temporal sequences alone.

Knowledge Access. Knowledge access arises when the reasoning process incorporates external information through retrieval modules, search engines, structured databases, or domain-specific resources. It is evidenced by explicit calls to knowledge sources whose retrieved content conditions or supplements the model's reasoning.

2.4.4 LLM Alignment Regimes

LLM alignment regimes specify how large language models are trained or adapted to perform reasoning on time series tasks, as shown in Figure 8. This tag is categorical: exactly one regime is assigned to each method.

Alignment. The alignment tag takes one of four regimes. *Prompting* relies on frozen models guided by instructions, few-shot examples, or chain-of-thought prompting without parameter updates. *Supervised fine-tuning* trains models on labeled temporal reasoning tasks such as instruction tuning, adapter methods, or distillation of reasoning traces. *Reinforcement or preference alignment* adapts models using feedback-based objectives, including reinforcement learning with human or AI feedback and preference optimization methods. *Hybrid approaches* combine supervised fine-tuning with reinforcement or preference alignment, for example by instruction-tuning a model before further aligning it with RLHF or direct preference optimization.

3 Direct Reasoning

Direct reasoning represents the most basic reasoning topology in the taxonomy. In this setting, a model directly maps time series inputs to outputs in a single step, without generating or exposing any intermediate reasoning trace. As such, direct reasoning can be viewed as the simplest baseline for time series reasoning: it provides efficiency and accessibility, but at the cost of limited interpretability and reduced robustness for complex tasks. Despite its simplicity, direct reasoning remains widely adopted in recent work, particularly for straightforward forecasting, anomaly detection, or descriptive question answering, and it often serves as a point of comparison for more structured reasoning topologies. The following discussion organizes direct reasoning methods according to four primary objectives, as illustrated in Figure 9.

3.1 Traditional Time Series Analysis with Direct Reasoning

Traditional time series analysis under *direct reasoning* treats the model as a one-shot mapper from temporal inputs (optionally with side context) to outputs such as forecasts, class labels, segmentation masks, or anomaly intervals. The execution topology is a single forward generation or completion without an explicit, multi-step trace. Within this topology, recent work spans zero-shot prompting, parameter-efficient adaptation, multimodal fusion, and retrieval-augmented conditioning—while retaining a single-step inference interface.

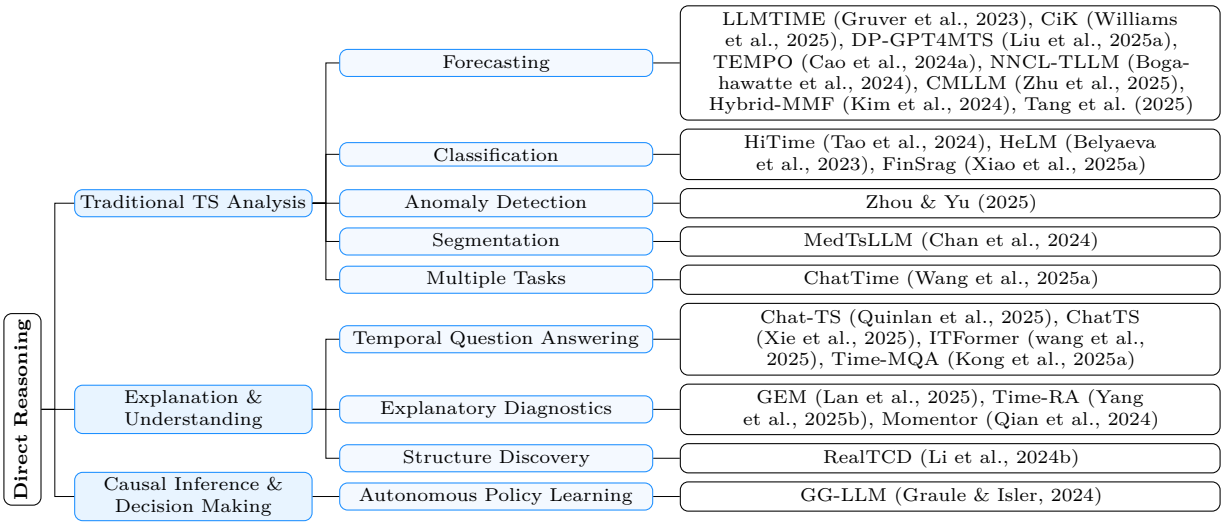

Figure 9: Taxonomy of direct reasoning approaches in time series reasoning

Forecasting. LLMTIME (Gruver et al., 2023) reframes forecasting as next-token generation over textualized numbers, sampling multiple continuations to summarize point and probabilistic predictions while analyzing calibration, tokenization, and context-length effects. CiK (Williams et al., 2025) introduces a context-aided benchmark and evaluates a direct prompt that outputs structured probabilistic forecasts in one call, showing gains when textual context is informative but exposing occasional catastrophic failures. DP-GPT4MTS (Liu et al., 2025a) conditions a largely frozen GPT-2 backbone with dual prompts, consisting of an explicit instruction and statistics prompt together with a soft textual prompt derived from timestamped text, concatenated with patched time series embeddings to decode future values directly. TEMPO (Cao et al., 2024a) attaches component-specific prompts to decomposed trend, seasonality, and residual patches and fine-tunes GPT-2 with LoRA, predicting each component in a single forward pass and then additively combining them to form the forecast. NNCL-TLLM (Bogahawatte et al., 2024) learns time-series-compatible text prototypes and forms learned prompts through nearest-neighbor selection, while retaining a one-shot inference interface. These prompts are fed with patch embeddings into a partially tuned LLM that adjusts only positional embeddings and layer norms to generate the forecasts.

CMLLM (Zhu et al., 2025) converts wind turbine supervisory control and data acquisition (SCADA) signals into text, attaches a prior-knowledge prefix, and lets a frozen LLM generate forecast tokens that are projected back to numbers. Hybrid-MMF (Kim et al., 2024) jointly forecasts future numbers and narratives by aligning numeric and textual embeddings and decoding both modalities directly, and reports a negative result that highlights fusion challenges. Tang et al. (2025) studies simple prompt strategies by injecting human background knowledge or by reprogramming numeric histories into rise and fall prose, and reports consistent error reductions while noting difficulties on multi-period series.

Classification. HiTime (Tao et al., 2024) aligns time series and textual semantics so a tuned LLM generates class labels as text, improving both accuracy and F1 on UEA datasets in a single forward pass. HeLM (Belyaeva et al., 2023) maps spirogram waveforms and clinical variables into token space and computes label likelihoods such as asthma risk without intermediate planning, achieving strong AUROC and AUPRC on UK Biobank traits. FinSrag (Xiao et al., 2025a) retrieves historical indicator segments, serializes them for prompting, and has a fine-tuned LLM directly predict stock movement as rise or fall in a single prompt call, improving both accuracy and MCC in financial forecasting.

Anomaly Detection. Zhou & Yu (2025) prompts LLMs and multimodal LLMs to return anomaly intervals from textualized sequences or plotted images in one step and finds that image inputs often outperform text, while subtle real-world anomalies remain challenging.

Segmentation. MedTsLLM (Chan et al., 2024) concatenates contextual text and signal patches (ECG, respiratory waveforms) into a frozen LLM and linearly projects output embeddings to produce segmentation masks, boundary points, or anomaly scores in one computation.

Multiple Tasks. ChatTime (Wang et al., 2025a) expands an LLM's tokenizer with discretized value symbols so a single model handles both time series forecasting and question answering, demonstrating transfer across tasks within a unified direct reasoning interface.

3.2 Explanation and Understanding with Direct Reasoning

This objective covers methods where the main product of reasoning is a natural-language answer, rationale, or causal interpretation derived from time series in a single inference step. The execution topology is direct: a model consumes temporal inputs (optionally fused with text or other modalities) and outputs explanatory text without iterative decomposition, branching, or explicit verification. Research here spans temporal QA, anomaly attribution, diagnostic reporting, and mechanism discovery.

Temporal Question Answering. Chat-TS (Quinlan et al., 2025) extends LLM vocabularies with discrete time series tokens and trains on multimodal instruction datasets to enable mixed time series text reasoning with direct answers and rationales while preserving general NLP ability. ChatTS (Xie et al., 2025) develops a multimodal LLM that integrates time series and text using synthetic QA generation and staged fine-tuning, enabling single-step explanation-oriented reasoning over trends, seasonality, anomalies, and causal queries. ITFormer (wang et al., 2025) freezes the backbone LLM and aligns temporal embeddings to the token space through a lightweight connector, enabling direct decoding of answers with strong efficiency and generalization demonstrated on new time series QA datasets. Time-MQA (Kong et al., 2025a) continually adapts LLMs on a large multi-domain QA corpus that unifies diverse time series tasks, enabling grounded and explanatory responses across forecasting, imputation, anomaly detection, classification, and open-ended reasoning.

Explanatory Diagnostics. GEM (Lan et al., 2025) aligns ECG waveforms, images, and text through frozen encoders and fine-tuned LLMs, introducing datasets and benchmarks that enable grounded diagnostic reports with clinician-style explanations. Time-RA (Yang et al., 2025b) introduces RATs40K, a large multimodal dataset for reasoning-centric anomaly detection where models generate observation–thought–action rationales in a single pass alongside detection, categorization, and explanatory reasoning refined through AI-feedback. Momentor (Qian et al., 2024) enhances video–LLMs with temporal token representations and event-sequence modeling, enabling segment-level localization and explanatory outputs in long untrimmed videos supported by a large-scale instruction dataset.

Structure Discovery. RealTCD (Li et al., 2024b) leverages an LLM to extract domain knowledge from textual system descriptions and propose candidate causal tuples that initialize a score-based causal discovery process. This meta-initialization step provides explanatory structure that guides the subsequent optimization of temporal causal graphs, while avoiding iterative reasoning during inference.

3.3 Causal Inference and Decision Making with Direct Reasoning

This objective concerns settings where the output is an action choice, policy signal, or quantified intervention effect derived from time series in a single inference step. Under direct reasoning, a model maps temporal context (and optionally auxiliary descriptions or features) to a decision-relevant score or recommendation without intermediate steps or branching.

Autonomous Policy Learning. GG-LLM (Graule & Isler, 2024) presents a framework for human-aware robot task planning where a frozen LLM, prompted once with a narration of recent human activities, scores candidate interactions whose probabilities are geometrically grounded on a semantic map. A downstream planner uses these localized scores to guide robot coverage, reducing human disturbance by about 29% in simulated apartments. The work illustrates how a single language-model output can inform temporal planning while raising questions about robustness, probability calibration, and safety in embodied settings.

3.4 Attribute Tags with Direct Reasoning

3.4.1 Control-Flow Operators with Direct Reasoning

Task Decomposition. Task decomposition is rarely adopted in direct reasoning and appears in only a small fraction of approaches. Examples include component-wise forecasting that predicts trend, seasonality, and residual components before summing them (Cao et al., 2024a), and structured output fields that separate observation, thought, and action for anomaly reasoning (Yang et al., 2025b).

Verification and Critique. Verification and critique are almost absent in direct reasoning and appear in only isolated cases. One example is the use of AI feedback to refine anomaly explanations during data construction, while inference remains single pass (Yang et al., 2025b).

Ensemble Selection. Ensemble selection is rarely employed in direct reasoning and appears in only a few approaches. One instance aggregates multiple forecast continuations by median or quantiles (Gruver et al., 2023), while another leverages outputs from a diverse model pool (GPT-4o, Gemini, DeepSeek-R1, Llama-3.3) ranked for reliability (Yang et al., 2025b).

3.4.2 Execution Actors with Direct Reasoning

Tool Use. Tool use is almost absent in direct reasoning and is demonstrated in only a single work. An example is the retrieval of historical indicator segments injected into the prompt before a single LLM decision for financial forecasting (Xiao et al., 2025a).

Agents. Agents are not employed in direct reasoning, and all pipelines operate in a non-agentic manner.

3.4.3 Information Sources with Direct Reasoning

Multimodal Inputs. The use of multimodal inputs is fairly common in direct reasoning and appears in a majority of approaches. Examples include combining time series with textual context for question answering and explanations (Quinlan et al., 2025; Xie et al., 2025; wang et al., 2025; Kong et al., 2025a), fusing health signals with text and tabular metadata for risk prediction (Belyaeva et al., 2023), and pairing signals with images such as ECG traces plus twelve-lead images (Lan et al., 2025). Other works render sequences as plots for visual anomaly prompts (Zhou & Yu, 2025), or use video as a temporal modality alongside text for fine-grained temporal understanding (Qian et al., 2024).

Knowledge Access. Knowledge access is almost absent in direct reasoning and is demonstrated in only a single work. One example is retrieval-augmented financial forecasting that conditions the LLM on retrieved historical patterns (Xiao et al., 2025a).

3.4.4 LLM Alignment Regimes with Direct Reasoning

Alignment. Both prompt-only usage and supervised tuning are widely adopted in direct reasoning. Prompt-only approaches include forecasting, context-aided forecasting, video temporal reasoning, temporal causal initialization, and zero-shot analyses (Zhu et al., 2025; Williams et al., 2025; Gruver et al., 2023; Qian et al., 2024; Li et al., 2024b; Tang et al., 2025). Instruction-tuned or adapter-based methods support multimodal reasoning, clinical interpretation, component-wise forecasting, joint numeric and text forecasting, health risk classification, and time-series question answering (Xie et al., 2025; Lan et al., 2025; Cao et al., 2024a; Kim et al., 2024; Belyaeva et al., 2023; Kong et al., 2025a). No reinforcement-only or hybrid regimes appear in this set.

4 Linear Chain Reasoning

Linear chain reasoning denotes executions that proceed through a *single, ordered sequence of steps* with no in-trajectory branching. The model may explicitly decompose a task, invoke a tool or retrieval once, and

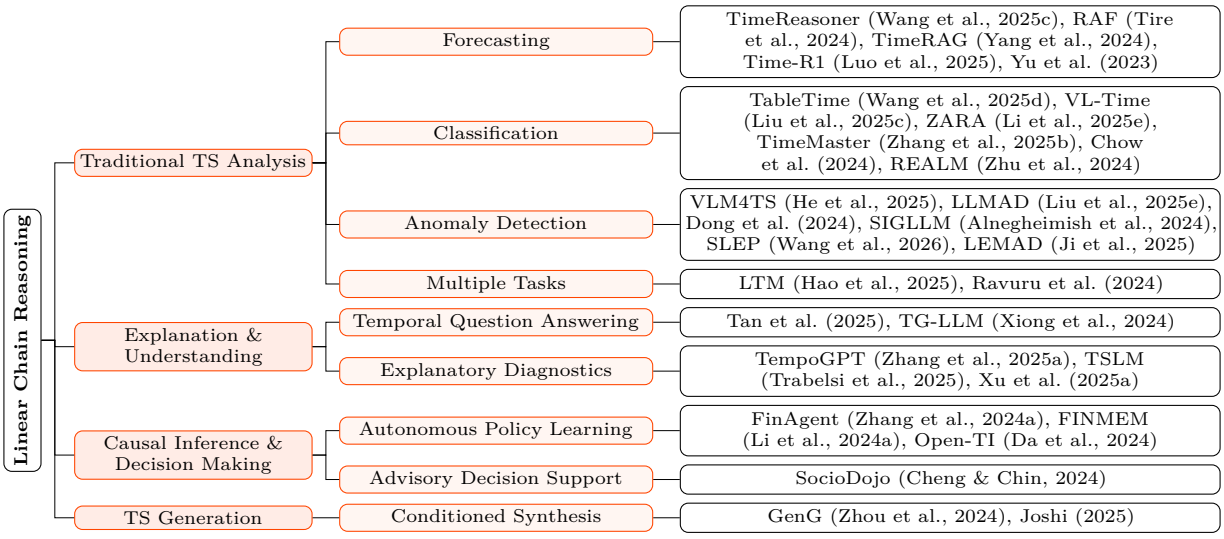

Figure 10: Taxonomy of linear chain reasoning approaches in time series reasoning

optionally perform a *one-shot* verification pass, but it does not maintain multiple concurrent hypotheses or iterate critique–revise loops. This topology preserves much of the simplicity of direct reasoning while adding mild structure that can improve grounding and numerical stability, yet still avoids the latency and complexity of branch-structured systems. The following discussion organizes linear chain methods according to four primary objectives, as illustrated in Figure 10.

4.1 Traditional Time Series Analysis with Linear Chain Reasoning

Traditional time series analysis under *linear chain reasoning* implements scripted sequences such as *analyze* $\rightarrow$ *(retrieve)* $\rightarrow$ *predict* or *detect* $\rightarrow$ *verify* $\rightarrow$ *decide*, while retaining a single-path execution.

Forecasting. TimeReasoner (Wang et al., 2025c) treats time series forecasting as deliberate reasoning, using structured prompts so that LLMs analyze patterns before generating forecasts in a fixed linear sequence. RAF (Tire et al., 2024) introduces a retrieval-augmented framework for time series foundation models that builds dataset-specific databases, retrieves the most relevant temporal segments, and integrates them into the forecasting process, showing consistent improvements across diverse benchmarks. TimeRAG (Yang et al., 2024) proposes a retrieval-based approach that slices time series into representative segments, retrieves similar histories, and reprograms them into natural-language prompts for a frozen LLM, yielding forecasting gains on the M4 benchmark without modifying model weights. Time-R1 (Luo et al., 2025) aligns the scripted chain with supervised traces followed by preference and reinforcement learning optimization while keeping inference as a fixed, linear sequence. Yu et al. (2023) integrates historical prices with company profiles and news, prompting LLMs to first summarize and contextualize signals and then output forecasts and explanations in a fixed, linear sequence, with GPT-4 few-shot outperforming financial baselines.

Classification. TableTime (Wang et al., 2025d) serializes time series into tabular prompts for training-free classification, using a fixed analyze-then-classify pass with optional self-consistency ensembles across runs and no in-trajectory branching. VL-Time (Liu et al., 2025c) renders time series as images and uses a plan-then-solve vision–language model to classify them, showing visual encoding overcomes tokenization limits of text-only LLMs. ZARA (Li et al., 2025e) performs zero-shot, classifier-free activity recognition by chaining feature-importance priors and multi-sensor retrieval into a serial LLM pipeline, with manager–worker handoffs yielding interpretable predictions. TimeMaster (Zhang et al., 2025b) trains a multimodal LLM with reinforcement learning to generate structured outputs over visualized time series data, combining reasoning, classification, and optional extension steps in a linear execution flow that boosts accuracy and context awareness. Chow et al. (2024) develops a multimodal LLM that aligns time series embeddings with a

language model's token space and fine-tunes it on chain-of-thought–augmented tasks, improving recognition and reasoning performance while retaining a single-path execution style. REALM (Zhu et al., 2024) extracts disease entities from clinical notes, matches them to a knowledge graph, and fuses their embeddings with vital-sign time series in a linear RAG pipeline for clinical risk prediction.

Anomaly Detection. VLM4TS (He et al., 2025) introduces a two-stage anomaly detection framework that first screens candidate anomalies with a pretrained vision encoder and then verifies and refines them through a VLM, producing final decisions and explanations in a single pass rather than through iterative critique–revise loops. LLMAD (Liu et al., 2025e) proposes an LLM-based framework that retrieves similar series for in-context learning, injects domain knowledge, and applies a structured reasoning process to output anomaly points with explanations, achieving interpretable results without relying on repeated revision cycles. Dong et al. (2024) evaluates large language models as explainable anomaly detectors, showing that with carefully designed prompts or lightweight fine-tuning they can identify anomalies and provide explanations, but still face instability and hallucination issues, leading to a single candidate–verification flow rather than iterative regeneration. SIGLLM (Alnegheimish et al., 2024) explores zero-shot anomaly detection by converting numeric series to text and prompting LLMs either directly or via forecasting residuals, demonstrating competitive performance against classical baselines while following a straightforward candidate–reassessment path that retains linear execution. SLEP (Wang et al., 2026) presents an LLM-based agent for anomaly detection in power system time series that leverages structured prompts, series memory, and optional reflection to provide accurate judgments and concise explanations, keeping the verification step one-shot. LEMAD (Ji et al., 2025) proposes a hierarchical multi-agent framework where specialist agents collect metrics and parse logs and a manager agent fuses information to make global anomaly decisions, enabling interpretable root-cause explanations while maintaining a single verification stage.

Multiple Tasks. LTM (Hao et al., 2025) integrates a frozen LLM with a pre-trained time series model and knowledge-graph-driven prompts, using fusion and retrieval modules in a scripted linear pipeline to improve forecasting, imputation, and anomaly detection with minimal fine-tuning. Ravuru et al. (2024) presents a hierarchical multi-agent RAG framework where a master delegates time series tasks to specialized sub-agents in a fixed sequence that retrieves dynamic prompts and improves forecasting, imputation, anomaly detection, and classification under distribution shifts, without in-run branching.

4.2 Explanation and Understanding with Linear Chain Reasoning

Explanation and understanding in the linear chain setting rely on a *single, ordered sequence* of analysis steps that culminate in an explanatory answer or narrative without maintaining concurrent alternatives or running critique–revise loops. Typical instances translate signals into structured intermediate forms (timelines, tables, or visualizations), optionally retrieve external knowledge, and then produce explanations, rationales, or summaries in one coherent trajectory.

Temporal Question Answering. Tan et al. (2025) infers natural-language event sequences that explain observed temporal segments by guiding stepwise analysis of changes and eliminating inconsistent options, producing a single answer in one path. TG-LLM (Xiong et al., 2024) translates narratives into aligned timelines and then executes deliberate reasoning over the structured representation to answer questions about order, duration, and simultaneity, while keeping verification for training rather than test-time branching.

Explanatory Diagnostics. TempoGPT (Zhang et al., 2025a) aligns temporal tokens and text in a shared space and trains the model to generate chain-of-thought rationales that culminate in conclusions such as trend analysis or fault diagnosis, executing one coherent trajectory per query. TSLM (Trabelsi et al., 2025) generates multiple local captions from a time series encoder–decoder and then consolidates them with a separate language model as a single end-of-chain summarization step, avoiding in-trajectory maintenance of alternatives. Xu et al. (2025a) orchestrates data acquisition, knowledge retrieval, analytic functions, and report writing through a serial controller for visual analytics, yielding grounded explanatory narratives and root-cause summaries without critique–revise loops at inference.

4.3 Causal Inference and Decision Making with Linear Chain Reasoning

Causal Inference and Decision Making in the linear chain setting executes a *single, ordered* observe→(retrieve)→decide pipeline, optionally with one-shot verification, to optimize a policy value over time. Typical instances ground actions with tools or memories and assess utility via returns, risk-adjusted metrics, or control rewards while avoiding in-run branching or debate.

Autonomous Policy Learning. FinAgent (Zhang et al., 2024a) sequences market intelligence, retrieval, immediate and high-level reflections, followed by a buy, sell, or hold action, achieving improved risk-adjusted returns from multimodal and tool-augmented inputs in a single path. FINMEM (Li et al., 2024a) sequences summarization, observation, retrieval, reflection, and decision within a layered memory system, extending reflection across days to adapt trading strategies while preserving a single-path execution. Open-TI (Da et al., 2024) integrates an LLM planner for configuring traffic simulations with a controller for signal actions, executing sequential thought-to-action steps to optimize throughput and travel time without branching.

Advisory Decision Support. SocioDojo (Cheng & Chin, 2024) coordinates analyst, assistant, and actuator roles to form hypotheses, retrieve evidence, and execute portfolio actions in a partially observable Markov decision process, using accept–reject as a one-shot verification step within a linear dialogue loop.

4.4 Time Series Generation with Linear Chain Reasoning

Time series generation in the linear chain setting follows a single, ordered script that first specifies targets or constraints and then executes a generator without maintaining concurrent alternatives or multi-round critique–revise loops. Typical instances use a language model to produce high-level descriptions or to guide tool configuration and data retrieval before a one-pass synthesis or consolidation step.

Conditioned Synthesis. GenG (Zhou et al., 2024) decomposes generation into a text-specification stage driven by a finetuned language model followed by conditional diffusion that synthesizes sequences under those specifications, reporting improvements in fidelity, controllability, and downstream utility while preserving a fixed two-stage path. Joshi (2025) uses publicly available language models to derive trusted-domain queries and parameter settings that are human-checked once and then applied to train GAN and VAE generators on interest-rate series, with distributional comparisons and backtesting conducted after a single linear pipeline.

4.5 Attribute Tags with Linear Chain Reasoning

4.5.1 Control-Flow Operators with Linear Chain Reasoning

Task Decomposition. Task decomposition is a prevalent feature of linear chain reasoning, reported in the majority of works. Two-stage designs recur across the literature, including planning followed by solving in visualization-guided reasoning (Liu et al., 2025c) and in table-structured classification with ordered steps (Wang et al., 2025d), localization followed by verification for anomaly detection (He et al., 2025), translation into a temporal graph followed by reasoning (Xiong et al., 2024), and description of targets followed by time series generation (Zhou et al., 2024). Generation followed by reflection within a single trajectory appears in forecasting (Wang et al., 2025c), incorporates reassessment steps for anomaly detection decisions (Liu et al., 2025e), and is implemented through immediate or extended reflections in trading agents (Zhang et al., 2024a; Li et al., 2024a). Manager-to-worker handoffs without branching within the trajectory occur in operations pipelines and traffic-control toolchains (Ji et al., 2025; Da et al., 2024).

Verification and critique. Verification and critique are less common in linear chain reasoning and appear in a minority of works. Verification and critique mechanisms span both inference- and training-time safeguards. At inference, models employ self-reflection on prior outputs in trading and forecasting (Zhang et al., 2024a; Wang et al., 2025c; Li et al., 2024a); dedicated verifier modules refine candidate predictions in visual anomaly detection, and one-shot reassessment reduces false positives (He et al., 2025; Liu et al., 2025e). During training, judging and filtering—often via reward models—are used in temporal-graph reasoning and multimodal classification (Xiong et al., 2024; Zhang et al., 2025b). Complementary oversight includes

human validation of query quality in generation frameworks (Joshi, 2025) and entity-level validation to filter hallucinations in clinical knowledge extraction (Zhu et al., 2024).

Ensemble selection. Ensemble selection is rarely used in linear chain reasoning and appears in only a few works. Self-consistency over repeated chains and aggregation of forecast samples both improve robustness, enhancing table-based classification (Wang et al., 2025d) and stabilizing zero-shot detectors (Alnegheimish et al., 2024). Generating multiple candidates followed by summarization consolidates caption candidates into a single description (Trabelsi et al., 2025).

4.5.2 Execution Actors with Linear Chain Reasoning

Tool use. Tool use is a frequent component of linear chain reasoning and features in numerous approaches. Knowledge retrieval and text tools span a range of applications, from knowledge graphs and retrieval-augmented generation for clinical prediction (Hao et al., 2025; Zhu et al., 2024), to web and news APIs for finance (Yu et al., 2023), long-term memory stores for trading (Zhang et al., 2024a), and vector databases that ground analytics pipelines (Xu et al., 2025a). Time series exemplar retrieval tools further extend these capabilities, leveraging DTW-based knowledge bases for forecasting (Yang et al., 2024), embedding retrieval for foundation forecasters (Tire et al., 2024), neighbor retrieval for classification (Wang et al., 2025d), and class-wise retrieval methods for activity recognition (Li et al., 2025e). Beyond retrieval, simulation and control tools are employed by traffic agents (Da et al., 2024). Finance and data connectors also underpin decision pipelines with retrieval and screening functions (Li et al., 2024a).

Agents. Most linear chain approaches operate without agents, reflecting this topology's dominance in the category (Liu et al., 2025c;e; Chow et al., 2024). Single-agent execution is occasionally adopted in domains such as trading and power system detection (Zhang et al., 2024a; Li et al., 2024a; Wang et al., 2026). Multi-agent coordination appears less frequently but is used for agentic RAG over time series tasks, traffic control, and zero-shot activity recognition (Ravuru et al., 2024; Da et al., 2024; Li et al., 2025e).

4.5.3 Information Sources with Linear Chain Reasoning

Multimodal inputs. The use of multimodal inputs is a frequent practice in linear chain reasoning and is present in many approaches. Incorporating text with time series signals is widespread, spanning trading that fuses prices with news (Zhang et al., 2024a), clinical prediction that blends notes with EHR series (Zhu et al., 2024), explainable stock forecasting that integrates company and macro news (Yu et al., 2023), and general time series reasoning that concatenates a dedicated series encoder with text for the LLM (Chow et al., 2024). Image or plot inputs are likewise fused with text, as in visualization-guided reasoning (Liu et al., 2025c), two-stage anomaly detection with a vision-language verifier (He et al., 2025), and structured reasoning over plotted series paired with textual prompts (Zhang et al., 2025b).

Knowledge access. Knowledge access recurs in linear chain reasoning and appears in a substantial share of approaches. Web and report retrieval actively conditions decisions in trading and portfolio studies (Zhang et al., 2024a; Yu et al., 2023; Cheng & Chin, 2024). Structured knowledge graphs steer entity-centric reasoning in both general and clinical settings (Hao et al., 2025; Zhu et al., 2024). Time series knowledge bases and prompt pools provide retrieved motifs and templates that guide downstream reasoning (Tire et al., 2024; Yang et al., 2024; Ravuru et al., 2024). Domain repositories and vector databases are used to ground both analytic workflows and anomaly-detection pipelines (Xu et al., 2025a; Liu et al., 2025e; Li et al., 2025e).

4.5.4 LLM Alignment Regimes with Linear Chain Reasoning

Alignment. Prompt-only alignment is widely reported in linear chain reasoning and appears to be the dominant regime in the majority of approaches (Zhang et al., 2024a; Liu et al., 2025c;e; Da et al., 2024). Supervised tuning with instruction-based or adapter-style methods is also common, supporting anomaly detection with synthetic supervision (Dong et al., 2024), text-guided generation (Zhou et al., 2024), temporal graph reasoning (Xiong et al., 2024), and multimodal temporal language models (Zhang et al., 2025a). Hybrid pipelines that combine supervised and reinforcement components appear in a smaller subset of works,

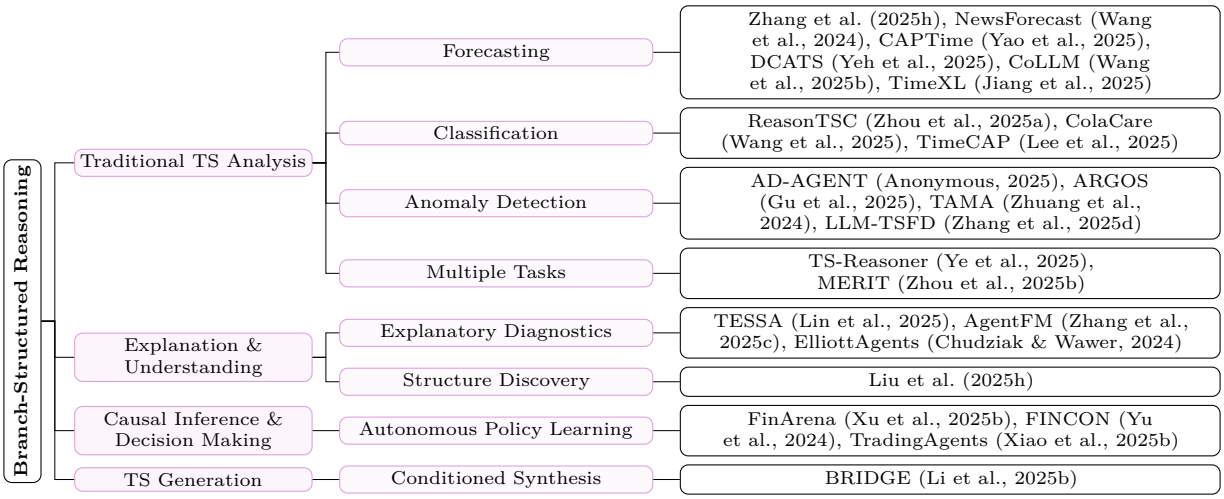

Figure 11: Taxonomy of branch-structured reasoning approaches in time series reasoning

including agentic RAG over time series tasks, event inference from win probability series, slow-thinking forecasting with reinforced LLMs, and structured multimodal reasoning (Ravuru et al., 2024; Tan et al., 2025; Luo et al., 2025; Zhang et al., 2025b). No works in this set rely solely on reinforcement or preference alignment without supervision.

5 Branch-Structured Reasoning

Branch-structured reasoning captures executions that *fork, revise, and fuse* intermediate hypotheses within a single run. A pipeline qualifies as branch-structured whenever it explores alternatives in parallel or sequentially, runs multi-round critique–revise loops, or aggregates across concurrent candidates; cross-branch fusion and debate-driven updates are canonical. Compared to linear chains, branching expands search and self-correction capacity but raises challenges in cost control, stability, and reproducibility. We organize branch-structured systems by primary objectives, as illustrated in Figure 11.

5.1 Traditional Time Series Analysis with Branch-Structured Reasoning

Forecasting. Zhang et al. (2025h) frames news-driven forecasting as a competitive multi-agent process where parallel hypotheses are iteratively pruned and refined through self-reflection, with surviving agents' predictions aggregated into improved forecasts. NewsForecast (Wang et al., 2024) iterates reflection on errors against missed contextual factors and updates selection logic before regeneration, forming critique–revise loops that adjust earlier choices. CAPTime (Yao et al., 2025) routes among probabilistic experts with token-level fusion, enabling mixture-style decoding that reconciles concurrent generative paths and improves multimodal time series forecasting with a frozen LLM backbone. DCATS (Yeh et al., 2025) employs proposal–evaluate–refine cycles where an LLM agent iteratively selects and validates sub-datasets, improving data quality and forecasting accuracy across diverse backbone models. CoLLM (Wang et al., 2025b) routes predictions between small and large models using confidence scores, invoking stronger solvers only when needed and fusing uncertain outputs, achieving efficient and accurate remaining-life prediction. TimeXL (Jiang et al., 2025) couples a prototype-based encoder with prediction, reflection, and refinement agents that iteratively critique and revise forecasts, branching through feedback loops before fusing outputs into accurate, explainable time series predictions.

Classification. ReasonTSC (Zhou et al., 2025a) conducts structured multi-turn reasoning that explicitly backtracks to explore alternatives before a fused decision, yielding internal branches that are reconciled. ColaCare (Wang et al., 2025) elicits divergent agent reviews with retrieved evidence and reconciles them through iterative debate and synthesis, generating fused clinical reports that improve predictions of mortality

and readmission. TimeCAP (Lee et al., 2025) combines a contextualization–prediction branch driven by frozen LLMs with a multimodal encoder branch, fusing their outputs after independent reasoning to improve event prediction through cross-branch aggregation.

Anomaly Detection. AD-AGENT (Anonymous, 2025) decomposes anomaly detection into specialized agents coordinated by memory, iterating code and logic through generator–reviewer loops that revise upstream steps based on critiques and conditional retrieval. ARGOS (Gu et al., 2025) iteratively generates, repairs, and reviews detector rules with multiple agents, then fuses selected candidates with a base detector to yield more accurate and efficient anomaly detection. TAMA (Zhuang et al., 2024) converts series to images, analyzes across stages with a self-reflection pass that revises prior outputs, and applies sliding windows with pointwise voting for in-trajectory aggregation. LLM-TSFD (Zhang et al., 2025d) leverages human-in-the-loop critique and knowledge-based corrections to iteratively refine pipelines, tools, and diagnostic explanations for industrial time series fault detection.

Multiple Tasks. TS-Reasoner (Ye et al., 2025) decomposes time series tasks into workflows of specialized operators and refines them with execution feedback, maintaining multiple candidate branches that are adaptively revised and fused for improved forecasting, anomaly detection, and causal discovery. MERIT (Zhou et al., 2025b) replaces hand-crafted augmentations with a multi-agent system that generates multiple candidate sequence views in parallel, verifies them, and selects the most suitable variants, yielding universal representations that enhance classification, forecasting, imputation, and anomaly detection across diverse datasets.

5.2 Explanation and Understanding with Branch-Structured Reasoning

Explanation and understanding in the branch-structured setting maintain multiple alternatives and allow critique–revise loops within a single run to reach an interpretable conclusion. Typical instances coordinate parallel roles or hypotheses and reconcile them through verification or fusion to produce diagnoses, symbolic rules, or explanatory annotations.

Explanatory Diagnostics. TESSA (Lin et al., 2025) coordinates a general annotator, a domain-specific annotator, and a reviewer that critiques and feeds back revisions, producing cross-domain explanatory annotations for time series through multi-agent branching and reconciliation. AgentFM (Zhang et al., 2025c) orchestrates system, data, and task agents over metrics, logs, and traces, aggregates divergent findings via a meta-agent, and delivers failure diagnoses and mitigation rationales by fusing parallel role-specific analyses. ElliottAgents (Chudziak & Wawer, 2024) explores stock time series through multiple candidate Elliott Wave patterns that are verified by deep reinforcement learning–based backtesting and synthesized into explanatory wave-structured reports, with agents branching across detection, validation, and reporting before convergence.

Structure Discovery. Liu et al. (2025h) iteratively generates, evaluates, and refines symbolic structures for time series dynamics through propose–verify–refine loops, producing interpretable equations and causal rules validated by quantitative metrics and rubric-based checks.

5.3 Causal Inference and Decision Making with Branch-Structured Reasoning

Causal Inference and Decision Making in the branch-structured setting maintain multiple alternatives, run critique–revise loops, and fuse candidate plans to select actions over time. Typical pipelines coordinate specialized roles, retrieve external knowledge, and reconcile disagreements via debate or confidence-aware fusion, with evaluation by policy metrics such as cumulative/annual return, Sharpe ratio, or drawdown.

Autonomous Policy Learning. FinArena (Xu et al., 2025b) coordinates agents for time series, news, and statements, aggregates their outputs with user risk preferences, and produces personalized trading actions, combining parallel analysis, adaptive retrieval, and iterative reasoning to improve financial decision making. FINCON (Yu et al., 2024) coordinates analyst and manager agents with dual-level risk control that critiques trajectories and updates beliefs, producing portfolio policies optimized for returns, Sharpe

ratio, and drawdown in partially observable financial markets. TradingAgents (Xiao et al., 2025b) conducts multi-round debates between bullish and bearish researchers with a facilitator selection and a risk team's adjustments, integrating tool-augmented retrieval and producing day-by-day trading decisions evaluated by backtest returns and risk.

5.4 Time Series Generation with Branch-Structured Reasoning

Time series generation in the branch-structured setting maintains multiple alternatives, runs critique–revise loops, and fuses candidates to produce controllable synthetic sequences. Typical instances coordinate parallel roles and reconcile textual or symbolic specifications before conditioning a generator in a single execution.

Conditioned Synthesis. BRIDGE (Li et al., 2025b) maintains parallel agent teams that iteratively propose, critique, and refine textual descriptions for target series, retrieve external evidence, and reconcile alternatives via consensus to produce conditioning inputs. The refined descriptions are then fused with learned temporal prototypes and a frozen text encoder to condition a diffusion generator, yielding controllable synthesis evaluated by fidelity and adherence metrics across diverse domains. This branching procedure improves semantic controllability over single-path baselines, albeit at the cost of additional compute for iterative refinement.

5.5 Attribute Tags with Branch-Structured Reasoning

5.5.1 Control-Flow Operators with Branch-Structured Reasoning

Task Decomposition. Task decomposition is almost universal in branch-structured reasoning, appearing in most systems. Role-specialized stages, where distinct agents or modules own complementary subgoals, are standard (Xu et al., 2025b; Yu et al., 2024; Lin et al., 2025; Zhang et al., 2025c). Propose–repair–review loops that enumerate candidates, then iteratively fix and reassess, are also frequent (Gu et al., 2025). Hierarchical orchestration with a manager that integrates or routes among workers appears in several systems (Xu et al., 2025b; Zhang et al., 2025c). Multi-round discussions where branches debate and a coordinator reconciles outcomes further illustrate decomposition with feedback (Xiao et al., 2025b).

Verification and Critique. Verification and critique are prominent elements of branch-structured reasoning, reported in most literature. Self-reflection that checks predictions or intermediate text and then triggers revisions is explicit in closed-loop designs (Jiang et al., 2025). Code-level review and repair, where candidate detection rules are debugged and refined through explicit repair and review agents, occurs in rule-generation pipelines (Gu et al., 2025). Reviewer modules or human-in-the-loop checks that send feedback upstream are used to refine annotations and decisions (Lin et al., 2025; Xu et al., 2025b). Debate-style critique, where opposing branches argue and a facilitator selects or adjusts the plan, is another recurring pattern (Xiao et al., 2025b).

Ensemble Selection. Within branch-structured reasoning, ensemble selection is relatively uncommon, described in a smaller subset of the literature. Top-k selection with later fusion of alternatives is used in rule-based anomaly detection (Gu et al., 2025). Late fusion that combines encoder outputs with LLM predictions within the same run is used in multimodal forecasting pipelines (Jiang et al., 2025; Lee et al., 2025).

5.5.2 Execution Actors with Branch-Structured Reasoning

Tool Use. Tool use emerges as a prominent element of branch-structured reasoning and is present in many published approaches. Systems call external retrieval over the web and market providers (Xu et al., 2025b; Xiao et al., 2025b), optimization or analytics components such as portfolio solvers and risk calculators (Yu et al., 2024), and code execution or indicator calculators inside the loop (Xiao et al., 2025b).

Agents. Multi-agent execution with coordinated specialists is the dominant pattern in branch-structured reasoning, appearing in the majority of works (Xu et al., 2025b; Yu et al., 2024; Zhang et al., 2025c; Lin

et al., 2025). Single-agent branches with iterative self-refinement are occasionally reported (Yeh et al., 2025), while some approaches operate without explicit agent components (Zhuang et al., 2024).

5.5.3 Information Sources with Branch-Structured Reasoning

Multimodal Inputs. The use of multimodal inputs is a common practice in branch-structured reasoning, reported in many works. Time series combined with text is common in forecasting and analysis frameworks (Jiang et al., 2025; Lee et al., 2025; Xu et al., 2025b; Lin et al., 2025). Some works also incorporate audio transcripts alongside market time series and text (Yu et al., 2024). Visual inputs such as time series plots, when treated as an image modality for MLLMs, are used in symbolic reasoning pipelines (Liu et al., 2025h).

Knowledge Access. Knowledge access is fairly common in branch-structured reasoning and appears in a majority of reported approaches. Adaptive web search and provider APIs supply external evidence that conditions downstream reasoning (Xu et al., 2025b; Xiao et al., 2025b). Other pipelines retrieve domain documents or stored memories to ground decisions (Yu et al., 2024; Zhou et al., 2025b), and retrieve in-context examples for augmentation (Lee et al., 2025).

5.5.4 LLM Alignment Regimes with Branch-Structured Reasoning

Alignment. Prompt-only usage with frozen backbones remains the dominant regime in this bucket (Gu et al., 2025; Lin et al., 2025; Jiang et al., 2025; Xu et al., 2025b; Lee et al., 2025; Zhang et al., 2025c; Xiao et al., 2025b). Supervised fine-tuning also appears in a few task-specific systems (Zhang et al., 2025h; Wang et al., 2025b; 2024), but we find no literatures in branch-structured reasoning that rely solely on reinforcement or preference alignment, nor any hybrids that combine supervised and preference alignment.

6 Current Landscape and Resources

This section surveys key resources for time series reasoning, focusing on datasets and benchmarks that vary in how directly they test reasoning, on surveys and position papers that synthesize progress and outline research agendas, and on recent studies that critically examine model performance and generalization. Together, these components provide a comprehensive view of both the resources that enable research and the evidence that shapes current understanding. This organization is summarized in Figure 12.

6.1 Datasets and Benchmarks

We group datasets and benchmarks by how directly they test time series reasoning. First, reasoning-first benchmarks define tasks and splits that explicitly require skills such as feature understanding, compositional generalization, temporal question answering, intervention reasoning, or agentic planning. In contrast, reasoning-ready benchmarks were not built primarily for reasoning but naturally support it through aligned side information, chronological or event-aligned protocols, or other structures that can be prompted into reasoning tasks. Finally, general-purpose time series benchmarks are standard collections for forecasting, detection, imputation, and related tasks that do not target reasoning by default, yet serve as solid references and can be adapted with minimal modification.

6.1.1 Reasoning-First Benchmarks.

Fons et al. (2024); Potosnak et al. (2024; 2025) introduce synthetic evaluations that test feature understanding and compositional generalization through controlled templates and held-out compositions, while ReC4TS (Liu et al., 2025d) develops a complementary evaluation targeting reasoning strategies and test-time sampling in time series forecasting. MTBench (Chen et al., 2025a) and TSQA (Kong et al., 2025a) frame temporal question-answering over time series: MTBench emphasizes cross-modal QA that links textual reports with series to test semantic trend understanding, indicator prediction, and correlation-based questions, while TSQA frames a broad set of time-series tasks as natural-language QA with prompts and rationale generation to elicit stepwise temporal reasoning. PUB (Pawelec et al., 2024) introduces a synthetic plot-understanding evaluation where generated charts, including time series plots with anomalies and degradations, are paired

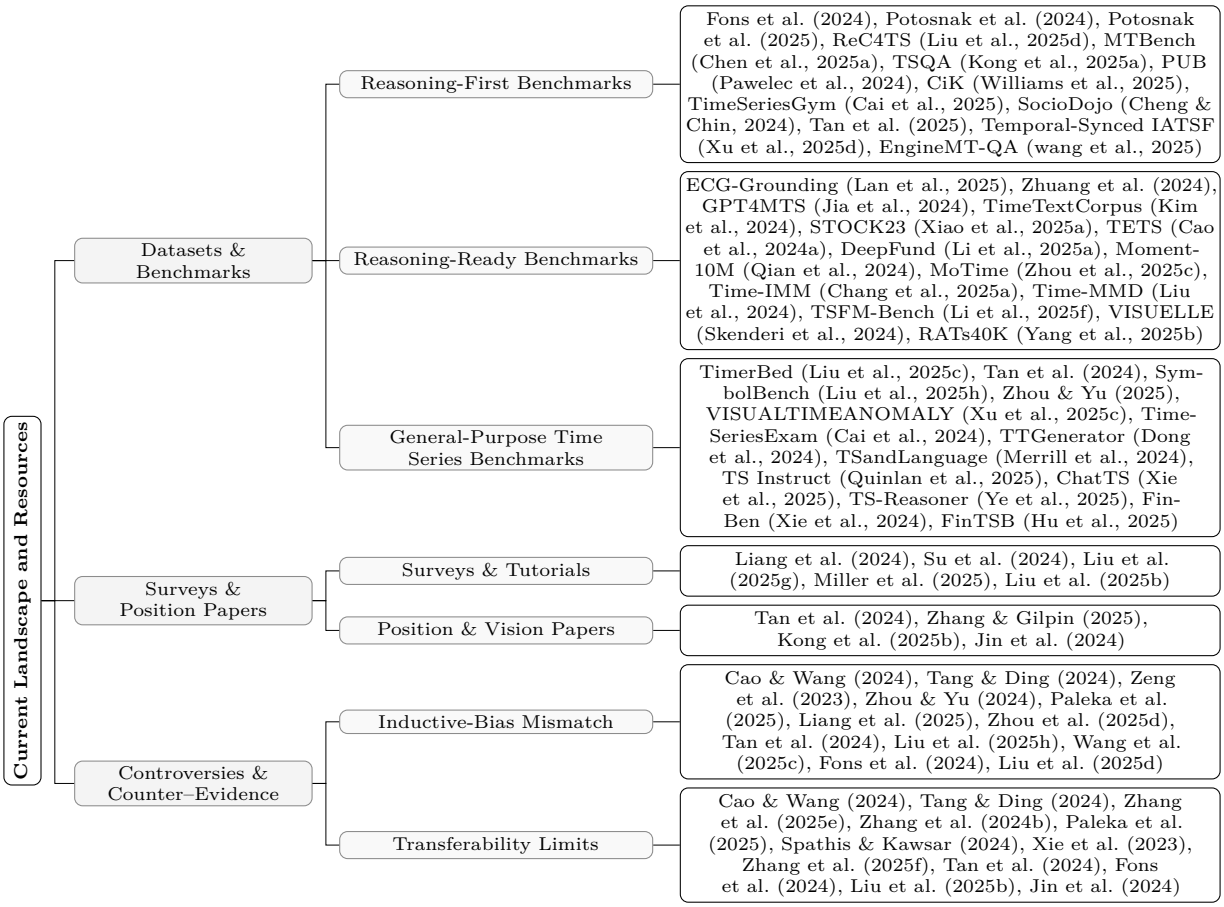

Figure 12: Taxonomy of current landscape and resources in time series reasoning

with JSON-structured questions to test visual and temporal reasoning. CiK (Williams et al., 2025) introduces context-dependent forecasting tasks that pair time series histories with textual cues and evaluates models by comparing performance with context present versus withheld, using a context-weighted CRPS-style scoring rule and prompting baselines for comparison. TimeSeriesGym (Cai et al., 2025) and SocioDojo (Cheng & Chin, 2024) construct episodic, reproducible environments for evaluating agentic decision making over time series, covering planning, tool use, iterative refinement, and seeded runs for systematic comparison. Tan et al. (2025), Temporal-Synced IATSF (Xu et al., 2025d), and EngineMT-QA (wang et al., 2025) align time-indexed series with textual descriptions of events or interventions and use release-aligned or event-aligned protocols, so models must infer timing, effect windows, and procedure-aware dynamics.

6.1.2 Reasoning-Ready Benchmarks.

ECG-Grounding (Lan et al., 2025) and Zhuang et al. (2024) pair clinical signals with contextual records and use case- or patient-level evaluation to support diagnostic and temporal reasoning under clinician-like constraints. GPT4MTS (Jia et al., 2024), TimeTextCorpus (Kim et al., 2024), STOCK23 (Xiao et al., 2025a), and TETS (Cao et al., 2024a) couple time series with news, descriptions, or structured text for context-aware forecasting and use chronology-preserving evaluations with ablations that toggle auxiliary context to isolate reasoning over external information, while DeepFund (Li et al., 2025a) provides a related live evaluation of model workflows on streaming market data but does not primarily target text-ablation protocols. Moment-10M (Qian et al., 2024) provides temporal localization with aligned segment boundaries that enable fine-grained reasoning over time, while MoTime (Zhou et al., 2025c) supplies multimodal, description-aligned time series with standardized splits and protocols for alignment-based forecasting and cold-start evaluation. Time-IMM (Chang et al., 2025a), Time-MMD (Liu et al., 2024), and TSFM-Bench (Li

et al., 2025f) assemble large, multi-domain suites with unified loaders, canonical splits, and baseline code to facilitate controlled, reproducible reasoning studies at scale. VISUELLE (Skenderi et al., 2024) targets cold-start demand forecasting by combining product metadata and images with exogenous time series signals, and uses release-based splits to enforce evaluation under information asymmetry. RATs40K (Yang et al., 2025b) provides anomaly-focused multimodal resources built with critique-and-revise label construction and protocols that evaluate detection, localization, categorization, and explanatory reasoning across disturbance regimes and data sources.

6.1.3 General-Purpose Time Series Benchmarks.

TimerBed (Liu et al., 2025c) and Tan et al. (2024) provide stratified benchmarks and cross-dataset evaluations that use standardized chronological partitions, serving as reference points for comparing forecasting and LLM-based methods across domains. SymbolBench (Liu et al., 2025h), Zhou & Yu (2025), VISUALTIMEANOMALY (Xu et al., 2025c), and TimeSeriesExam (Cai et al., 2024) supply controllable, reproducible suites with fixed seeds, perturbation controls, and procedurally generated scenario catalogs that enable exact reproducibility for forecasting, anomaly, symbolic, and reasoning evaluations. TTGenerator (Dong et al., 2024) curates anomaly and change-point resources with event and tolerance-window protocols that emphasize temporal localization, while TSandLanguage (Merrill et al., 2024) provides a broader reasoning-focused suite including etiological reasoning, question answering, and context-aided forecasting using large-scale multiple-choice evaluation rather than change-point localization. TS Instruct (Quinlan et al., 2025) and ChatTS (Xie et al., 2025) provide multimodal classification and segmentation resources for sensor and biomedical streams and use subject-wise or group-wise evaluation splits when applicable to assess generalization across entities. TS-Reasoner (Ye et al., 2025) proposes an LLM-driven agentic framework that combines imputation and forecasting under prescribed masking patterns and rolling chronological splits to evaluate reconstruction and next-step prediction under missingness, using execution-feedback self-refinement and specialized operators to improve numeric fidelity. FinBen (Xie et al., 2024) aggregates finance time series with leakage-aware backtesting protocols oriented toward realistic temporal deployment. FinTSB (Hu et al., 2025) assembles financial forecasting testbeds with fixed temporal ranges, leakage-aware splits, and realistic backtesting suitable for standardized comparisons.

6.2 Surveys and Position Papers

The literature in this area includes both surveys and tutorials as well as position and vision papers. Surveys and tutorials synthesize progress by mapping methods, datasets, and open challenges in time series reasoning, offering structured overviews that guide researchers and practitioners. Position and vision papers, on the other hand, put forward arguments for or against specific approaches, outline perspectives on emerging opportunities, and set research agendas for the field.

6.2.1 Surveys and Tutorials.

Liang et al. (2024); Su et al. (2024) map the landscape of foundation models and LLM applications for time series, organizing architectures, pretraining and adaptation regimes, datasets, and task coverage. They identify gaps in robustness, multimodality, and evaluation practice, motivating evaluation that prioritizes reasoning and robustness rather than incremental accuracy gains. Liu et al. (2025g) surveys synthetic data for time series and organizes the literature by generation methods and lifecycle use in pretraining, finetuning, and evaluation. It explains why particular design choices make synthetic corpora useful for probing feature understanding, compositional behavior, and temporal consistency, and it highlights limitations and future directions. Miller et al. (2025); Liu et al. (2025b) provide tutorials that unify common pipelines across forecasting, detection, classification, and analytics. They curate representative methods and reading paths and discuss pitfalls such as data leakage and misaligned splits that can obscure or mimic reasoning signals.

6.2.2 Position and Vision Papers.

Tan et al. (2024); Zhang & Gilpin (2025) question default uses of large language models for time series by presenting extensive replications and simple context-based baselines. They argue that progress should

target settings where structured reasoning and external context matter and call for protocols that make those requirements explicit. Kong et al. (2025b); Jin et al. (2024) propose research agendas that emphasize compositional generalization, causal grounding, multimodal integration, and transparent evaluation. They advocate benchmarks and tools that reveal when a model is reasoning, when it is copying context, and when it fails under distribution shift.

6.3 Controversies and Counter–Evidence

A complementary line of research reports cases where Transformer and large language model methods for time series need careful interpretation. Two recurring themes appear across replication studies, ablation analyses, and position papers. The goal is to improve evaluation practices and clarify the limits of current approaches without changing the overall focus of the survey.

6.3.1 Inductive-Bias Mismatch.

A consistent observation is that generic Transformer and large language model backbones often lack the inductive structure that many time series tasks require, including clear seasonality, multiple repeating patterns, strong local correlations, and shifts between regimes Cao & Wang (2024); Liang et al. (2025); Zhou et al. (2025d); Liu et al. (2025h). Across forecasting and related tasks, several studies report results similar to tuned classical or specialized time series methods, and in many cases those baselines perform better, especially when horizon-wise evaluation is used with careful control of splits and data leakage Tan et al. (2024); Zeng et al. (2023); Zhou & Yu (2024); Wang et al. (2025c). The gap is most visible for long-horizon prediction and for series with strong frequency patterns Zeng et al. (2023); Wang et al. (2025c); Liu et al. (2025d); Zhou et al. (2025d). These results suggest that some reported improvements linked to reasoning or longer context are actually affected by missing time series priors when baselines and reporting methods are not carefully designed Tan et al. (2024); Tang & Ding (2024); Fons et al. (2024); Paleka et al. (2025). In practice, horizon-based reporting and strong baseline selection are often needed to interpret gains reliably compared to established time series methods Tan et al. (2024); Tang & Ding (2024); Fons et al. (2024); Paleka et al. (2025). When gains are observed, they often appear alongside components that include decomposition, frequency awareness, state-space modeling, or convolutional structures Cao & Wang (2024).

6.3.2 Transferability Limits.

A second line of evidence links mixed or negative results to a mismatch between pretraining priors and downstream time series dynamics, as well as to weak transfer across datasets and domains Tan et al. (2024); Tang & Ding (2024); Fons et al. (2024); Paleka et al. (2025). Pretraining on text and code does not by itself provide models with strong numeric accuracy, periodic structure, or stability over time Zhang et al. (2025e); Spathis & Kawsar (2024); Zhang et al. (2025f); Liu et al. (2025b). Reported gains can depend on dataset characteristics, prompt design, or evaluation choices that do not hold under domain shift or cross-benchmark testing Fons et al. (2024); Liu et al. (2025b); Jin et al. (2024); Xie et al. (2023). Replications document rank reversals across datasets, sensitivity to prompt wording, shot order, and random seed, and performance drops under shift or cold start settings Cao & Wang (2024); Fons et al. (2024); Zhang et al. (2024b); Liu et al. (2025b). These findings indicate that adaptation, such as small parameter-efficient finetuning or numerically aware output heads, and drift-aware protocols, are prerequisites for assessing generalization Tan et al. (2024); Zhang et al. (2025e); Xie et al. (2023); Zhang et al. (2025f). Together they motivate reporting that includes leave-one-dataset-out in addition to averages, versioned splits with fixed seeds, explicit leakage checks, and calibration or coverage evaluations when claims involve transfer, cross-domain utility, or robustness Paleka et al. (2025).

7 Open Problems and Outlook

We highlight six themes that recur across direct, chain, and branch topologies, reflecting open problems that are repeatedly emphasized in author-stated outlooks and pointing to areas where further progress is needed.

7.1 Evaluation and Benchmarking.

Many works call for standardized, stress-tested evaluation rather than small curated sets, which often mean narrow domains with few series or short horizons, handpicked or templated prompts, pre-segmented windows that hide boundary effects, single-source cohorts, and limited resampling or shift tests (Liang et al., 2024; Su et al., 2024). To move beyond these limitations, stronger practice should include versioned splits with release-aligned protocols, fixed seeds, and reporting templates for reproducibility, as well as human studies where domain judgment matters (Tan et al., 2024; Fons et al., 2024; Lan et al., 2025).

Because reasoning quality directly shapes utility, evaluations must also audit faithfulness. This includes metrics that test sufficiency and necessity of cited evidence, agreement between explanations and underlying data, and consistency with numeric outputs (Dong et al., 2024; Yao et al., 2025; Yeh et al., 2025; Wang et al., 2024; Li et al., 2025e). Post-hoc grading and LLM-as-judge approaches, while useful, are not sufficient unless paired with groundable signals, audit trails, or counterevidence checks (Liu et al., 2025h; Quinlan et al., 2025).

Looking ahead, evaluation suites should connect intermediate reasoning to user value, through measures such as decision impact or policy loss, rather than relying only on proxy scores (Hu et al., 2025; Xie et al., 2024). While forecasting and causal tasks already benefit from multiple datasets and benchmarks, related areas such as time series editing—repair, counterfactual editing, and constrained rewriting—remain underrepresented. More explicit task definitions, faithful scoring protocols, and stress-tested datasets will be key to advancing this frontier (Liu et al., 2025c; Zhou et al., 2025a; Fons et al., 2024).

7.2 Multimodal Fusion and Alignment.

Improving the alignment between time series, text, and images or videos remains a central challenge (Quinlan et al., 2025; Williams et al., 2025; Liang et al., 2024). Alignment spans three levels: instance pairing across modalities for the same example, temporal correspondence that ties tokens or pixels to the correct time indices, and semantic grounding that links claims to the measured signal (Liang et al., 2024). Fine-grained temporal localization requires identifying the exact time point or interval in the series that a word, phrase, or visual element refers to, and verifying that the predicted span matches the ground truth at the correct scale (Yeh et al., 2025; Xie et al., 2024). Further progress calls for stronger connectors and cross-modal objectives that explicitly link time indices to tokens or pixels. Promising directions include contrastive training with hard near misses (Liu et al., 2024), segment-level alignment losses (Yeh et al., 2025), pointer-style time-span decoders (Qian et al., 2024), and learned similarity beyond DTW (Yang et al., 2024).

Modality imbalance is common because the text channel often provides more tokens and denser labels, while the time series offers weaker supervision and more noise. This imbalance can cause models to overfit to text and overlook subtle temporal changes (Zhou et al., 2025c; Williams et al., 2025; Skenderi et al., 2024). Mitigation strategies include balanced sampling and loss reweighting across modalities, modality dropout and learned gating, per-modality normalization, and distillation that transfers signal from the stronger to the weaker modality (Zhou et al., 2025c; Williams et al., 2025; Belyaeva et al., 2023; Liu et al., 2025a).

Temporal synchronization across modalities is another frequent failure mode. Differences in sampling rates, logging delays, and clock drift can introduce misalignment that accumulates over long horizons (Chen et al., 2025a; Liu et al., 2024). Practical remedies include timestamp normalization, learned shift predictors, cross-correlation to estimate lag, multi-scale encoders that align coarsely before refining, and anchoring on events shared across modalities (Xie et al., 2025; Liu et al., 2025c).

To address plotted-series bias and style overfitting, evaluations should incorporate render-swap controls and provide access to raw signals in addition to plots (Dong et al., 2024; Liu et al., 2025c). Outlook work should prioritize broader datasets and libraries with synchronized multimodal pairs and multilingual text, enabling more faithful evaluation of alignment under longer horizons and richer modalities (Liu et al., 2024; Liang et al., 2024).

7.3 Retrieval and Knowledge Grounding.

Grounding answers in external sources such as tables, knowledge bases, logs, and domain corpora is widely requested across tasks (Zhu et al., 2025; Zhang et al., 2025d; Yang et al., 2024). This line of work seeks to reduce hallucinations and improve domain specificity while maintaining efficiency (Hao et al., 2025; Gu et al., 2025; Zhang et al., 2025d). A central design choice is whether to retrieve at inference time or to pre-encode knowledge during training or adaptation. The decision depends on factors such as update frequency, domain shift, and memory or latency budgets (Tire et al., 2024; Bogahawatte et al., 2024; Liu et al., 2025h).

Time-aware retrieval is particularly promising. This involves building segment-level indexes over subsequence representations, aligning events and entities to timestamps, and retrieving windows that capture motifs or regimes rather than entire series (Tire et al., 2024; Yang et al., 2024). Tool-augmented retrieval provides further benefits by issuing structured queries such as SQL over tables, log filters, or simulator calls, and then feeding results back into the reasoning step with explicit citations (Zhu et al., 2025; Zhang et al., 2025d; Xu et al., 2025b).

When multiple candidates are retrieved, robustness improves through learned re-ranking with cross-modal checks and late fusion via evidence-weighted voting, rather than relying on a single source (Yao et al., 2025; Tire et al., 2024). Calibration also benefits from evidence-linked decoding, which constrains claims to cited spans, down-weights unsupported tokens, and abstains when retrieved evidence is weak or conflicting (Hao et al., 2025). For long data streams, these methods can be complemented by streaming retrieval with rolling caches and periodic re-indexing to maintain relevance without exceeding context limits (Tire et al., 2024; Yang et al., 2024).

The outlook is retrieval pipelines that are more time-aware, fault-tolerant, and auditable, supported by benchmarks that stress-test robustness under shift, incompleteness, and cost constraints (Hu et al., 2025).

7.4 Long Context, Memory, and Efficiency.

Scaling to longer histories with manageable latency and memory remains a recurring challenge (Tan et al., 2024; Zhou & Yu, 2025; Zhou et al., 2025a; Yu et al., 2024). Compression strategies include multi-resolution encoders, learnable downsampling of sequences or KV-caches, segment pooling or sketching, and value-aware sparsification that preserves extremes and change points (Chan et al., 2024; Zhou et al., 2025a; Yu et al., 2023). Streaming inference combines sliding windows with warm-start states and truncated backpropagation, with state handover across windows to avoid recomputing long prefixes (Kong et al., 2025a; Wang et al., 2025b). Stateful memory augments the model with episodic or event-indexed slots and explicit write/read policies so that long-range dependencies persist beyond the context window (Li et al., 2024a; Kong et al., 2025a).

Lightweight adaptation methods such as LoRA or IA3 adapters, prefix prompts, and linear probes for numeric channels help retain throughput while capturing domain specifics (Tao et al., 2024; Zhou et al., 2025a). Practical deployment further depends on compute-aware training and decoding, including mixed precision and quantization, block-sparse or chunked attention, early-exit or confidence-based halting, and speculative decoding to reduce latency (Hu et al., 2025; He et al., 2025; Chan et al., 2024; Li et al., 2025f).

Open problems include mitigating recency bias from limited windows, preserving temporal semantics under compression, and balancing efficiency with rare-event retention. Future progress will depend on benchmarks that jointly measure accuracy, latency, memory, and energy, alongside protocols that capture streaming and drift-aware conditions (Liu et al., 2025g; Li et al., 2025f; Hu et al., 2025; Yu et al., 2024).

7.5 Agentic Control and Tool Use.

Many works envision systems that perceive streams, plan, call tools or simulators, and then verify and act, moving from passive prediction to closed-loop control (Zhang et al., 2024a; Yeh et al., 2025; Da et al., 2024; Cai et al., 2025). A central open problem is action selection under uncertainty and delayed feedback, where sparse rewards and partial observability make credit assignment brittle. Promising directions include

uncertainty-aware planners, policy regularization, and counterfactual rollouts before committing to actions (Yu et al., 2024; Li et al., 2024a; Da et al., 2024).

Termination and rollback are often under-specified in deployed settings. Agents require explicit stop criteria, safe fallbacks, and recovery policies that bound risk during reversible and irreversible operations (Zhang et al., 2025c; Anonymous, 2025). Tool integration can be brittle: while domain solvers often improve correctness, they may increase latency and risk version drift. Recent approaches explore cost-aware tool selection, caching or batching of calls, and learned simulators that are periodically recalibrated to ground-truth solvers (Anonymous, 2025; Xiao et al., 2025b; Da et al., 2024).

Interface robustness is another recurring pain point, since schema or unit changes, simulator API evolution, and sampling inconsistencies often destabilize pipelines. Outlook work should add contract tests, unit and scale checks, and fault injection to agent evaluations (Anonymous, 2025; Da et al., 2024). Verification layers remain under-formalized: beyond heuristic checkers, agents need principled critics that test invariants, cross-validate with independent tools, and trigger abstention or rollback when evidence conflicts (Ye et al., 2025; Liu et al., 2025h; Jiang et al., 2025).

Evaluation protocols have not yet kept pace with real-world practice. The community needs closed-loop benchmarks with standard tool APIs, rate limits, explicit costs, and safety budgets, reporting regret, constraint violations, and tool spend alongside accuracy (Anonymous, 2025; Cai et al., 2025; Da et al., 2024; Hu et al., 2025). Looking forward, human-in-the-loop governance will be critical, including audited logs, reproducible decision traces, and handoff protocols when confidence or safety falls below thresholds (Yeh et al., 2025).

7.6 Causal Inference and Decision Support.

Bridging descriptive explanations to causal conclusions, such as counterfactuals, treatment effects, and policies, remains a central goal for time series reasoning (Tan et al., 2024; Xu et al., 2025d; Liu et al., 2025c). Key open problems include identification under time-varying confounding, latent common causes, and feedback loops that arise in interactive or controlled settings (Xu et al., 2025d; Li et al., 2024b).

Benchmarks and simulators with known data-generating processes are needed, including longitudinal interventions, dynamic policies, and realistic constraints. Semi-synthetic designs that splice real series with scripted interventions can help establish ground truth (Da et al., 2024; Xu et al., 2025d; Liang et al., 2024). Methodologically, counterfactual forecasting and dynamic treatment learning should model interventions explicitly and test invariances implied by causal structure (Xu et al., 2025d; Li et al., 2024b).

Evaluations should link rationales to causal evidence through sufficiency and necessity checks, counterfactual consistency, and refutation tests such as placebos or sensitivity analyses (Liu et al., 2025h). Principled off-policy evaluation is required before deployment, combining importance-sampling and model-based estimators with calibrated uncertainty to report policy value, regret, and constraint violations (Yu et al., 2024; Cheng & Chin, 2024; Zhang et al., 2024a).

Heterogeneous effects and fairness under domain shift remain underexplored. Future work should report subgroup treatment effects with coverage guarantees and specify safety budgets for interventions (Belyaeva et al., 2023; Liu et al., 2024; Xu et al., 2025b). A practical outlook is end-to-end pipelines that couple causal objectives with closed-loop evaluation, where policies are audited, costs are explicit, and rollback rules are triggered when uncertainty or shift exceeds thresholds (Da et al., 2024; Anonymous, 2025; Xu et al., 2025b).

8 Conclusion

Time series reasoning treats time as a first-class axis and integrates intermediate evidence into the answer itself. We organize the field by reasoning topology, distinguishing three main families: direct reasoning in a single step, linear chain reasoning with explicit intermediate steps, and branch-structured reasoning that explores, revises, and aggregates. Alongside topology, we consider the main objectives of the literature—traditional time series analysis, explanation and understanding, causal inference and decision making, and time series generation. Common techniques such as decomposition and verification, ensembling, tool use, knowledge

access, multimodality, agent loops, and LLM alignment regimes cut across these perspectives, offering a compact way to describe methods and their strengths and weaknesses.

Several themes emerge for design and evaluation. The choice of reasoning structure is central: moving from direct to chain to branch increases capacity for grounding, search, and self-correction, but also raises computational cost, variance, and reproducibility challenges. Evidence must remain visible and tightly linked to data, retrieved context, and tool outputs, with strict temporal alignment to keep narratives faithful to signals. Agents and tools can extend analysis into action, but they require clear stop rules, rollback plans, and cost-aware strategies. Evaluation should mirror deployment through shift-aware protocols, long horizons, and streaming settings, with checks that test whether rationales truly reflect the data. Cost and latency should be treated as design budgets, and lightweight adaptation often provides the right balance when domain specificity is needed.

Looking ahead, the field should pursue benchmarks that tie reasoning quality to utility, closed-loop testbeds that balance cost and risk, and streaming evaluations that capture long-horizon challenges. No single topology will dominate, as domains vary in constraints, costs, and tolerance for risk. What matters is deliberate structural choice, alignment with primary objectives, and evaluation that keeps evidence and faithfulness at the center. By advancing along these lines, time series reasoning can move from narrow accuracy toward broad reliability, enabling systems that not only analyze but also understand, explain, and act on dynamic worlds.

References

Sarah Alnegheimish, Linh Nguyen, Laure Berti-Equille, and Kalyan Veeramachaneni. Large language models can be zero-shot anomaly detectors for time series?, 2024. URL https://arxiv.org/abs/2405.14755.

Samaneh Aminikhanghahi and Diane J Cook. A survey of methods for time series change point detection. *Knowledge and information systems*, 51(2):339–367, 2017.

Anonymous. AD-AGENT: A multi-agent framework for end-to-end anomaly detection. In *Submitted to ACL Rolling Review - July 2025*, 2025. URL https://openreview.net/forum?id=2vUbjvUYdE. under review.

Anastasiya Belyaeva, Justin Cosentino, Farhad Hormozdiari, Krish Eswaran, Shravya Shetty, Greg Corrado, Andrew Carroll, Cory Y. McLean, and Nicholas A. Furlotte. Multimodal llms for health grounded in individual-specific data. In *Machine Learning for Multimodal Healthcare Data: First International Workshop, ML4MHD 2023, Honolulu, Hawaii, USA, July 29, 2023, Proceedings*, pp. 86–102, Berlin, Heidelberg, 2023. Springer-Verlag. ISBN 978-3-031-47678-5. doi: 10.1007/978-3-031-47679-2_7. URL https://doi.org/10.1007/978-3-031-47679-2_7.

Konstantinos Benidis, Syama Sundar Rangapuram, Valentin Flunkert, Yuyang Wang, Danielle Maddix, Caner Turkmen, Jan Gasthaus, Michael Bohlke-Schneider, David Salinas, Lorenzo Stella, et al. Deep learning for time series forecasting: Tutorial and literature survey. *ACM Computing Surveys*, 55(6):1–36, 2022.

Jayanie Bogahawatte, Sachith Seneviratne, Maneesha Perera, and Saman Halgamuge. Rethinking time series forecasting with llms via nearest neighbor contrastive learning, 2024. URL https://arxiv.org/abs/2412.04806.

Yifu Cai, Arjun Choudhry, Mononito Goswami, and Artur Dubrawski. Timeseriesexam: A time series understanding exam. In *NeurIPS Workshop on Time Series in the Age of Large Models*, 2024. URL https://openreview.net/forum?id=oXRX7ABgLz.

Yifu Cai, Xinyu Li, Mononito Goswami, Michał Wiliński, Gus Welter, and Artur Dubrawski. Timeseriesgym: A scalable benchmark for (time series) machine learning engineering agents, 2025. URL https://arxiv.org/abs/2505.13291.

Defu Cao, Yujing Wang, Juanyong Duan, Ce Zhang, Xia Zhu, Congrui Huang, Yunhai Tong, Bixiong Xu, Jing Bai, Jie Tong, et al. Spectral temporal graph neural network for multivariate time-series forecasting. *Advances in neural information processing systems*, 33:17766–17778, 2020.

Defu Cao, Yousef El-Laham, Loc Trinh, Svitlana Vyetrenko, and Yan Liu. A synthetic limit order book dataset for benchmarking forecasting algorithms under distributional shift. In *NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2022.

Defu Cao, James Enouen, Yujing Wang, Xiangchen Song, Chuizheng Meng, Hao Niu, and Yan Liu. Estimating treatment effects from irregular time series observations with hidden confounders. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 6897–6905, 2023.

Defu Cao, Furong Jia, Sercan O Arik, Tomas Pfister, Yixiang Zheng, Wen Ye, and Yan Liu. TEMPO: Prompt-based generative pre-trained transformer for time series forecasting. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=YH5w12OUuU.

Defu Cao, Wen Ye, Yizhou Zhang, and Yan Liu. Timedit: General-purpose diffusion transformers for time series foundation model. *arXiv preprint arXiv:2409.02322*, 2024b.

Rui Cao and Qiao Wang. An evaluation of standard statistical models and llms on time series forecasting. In *2024 IEEE International Conference on Future Machine Learning and Data Science (FMLDS)*, pp. 533–538. IEEE, 2024.

Nimeesha Chan, Felix Parker, William Bennett, Tianyi Wu, Mung Yao Jia, James Fackler, and Kimia Ghobadi. Medtsllm: Leveraging llms for multimodal medical time series analysis, 2024. URL https://arxiv.org/abs/2408.07773.

Ching Chang, Chiao-Tung Chan, Wei-Yao Wang, Wen-Chih Peng, and Tien-Fu Chen. Timedrl: Disentangled representation learning for multivariate time-series. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pp. 625–638, 2024a. doi: 10.1109/ICDE60146.2024.00054.

Ching Chang, Chan Chiao-Tung, Wei-Yao Wang, Wen-Chih Peng, and Tien-Fu Chen. Self-supervised learning of disentangled representations for multivariate time-series. In *NeurIPS 2024 Workshop: Self-Supervised Learning-Theory and Practice*, 2024b.

Ching Chang, Wei-Yao Wang, Wen-Chih Peng, Tien-Fu Chen, and Sagar Samtani. Align and fine-tune: Enhancing LLMs for time-series forecasting. In *NeurIPS Workshop on Time Series in the Age of Large Models*, 2024c. URL https://openreview.net/forum?id=AaRCmJieG4.

Ching Chang, Jeehyun Hwang, Yidan Shi, Haixin Wang, Wen-Chih Peng, Tien-Fu Chen, and Wei Wang. Time-imm: A dataset and benchmark for irregular multimodal multivariate time series, 2025a. URL https://arxiv.org/abs/2506.10412.

Ching Chang, Ming-Chih Lo, Wen-Chih Peng, and Tien-Fu Chen. Prompttss: A prompting-based approach for interactive multi-granularity time series segmentation. *arXiv preprint arXiv:2506.11170*, 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 Trans. Intell. Syst. Technol.*, 16(3), April 2025c. ISSN 2157-6904. doi: 10.1145/3719207. URL https://doi.org/10.1145/3719207.

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, 2025a. URL https://arxiv.org/abs/2503.16858.

Yiqun Chen, Qi Liu, Yi Zhang, Weiwei Sun, Xinyu Ma, Wei Yang, Daiting Shi, Jiaxin Mao, and Dawei Yin. Tourrank: Utilizing large language models for documents ranking with a tournament-inspired strategy. In *Proceedings of the ACM on Web Conference 2025*, pp. 1638–1652, 2025b.

Junyan Cheng and Peter Chin. Sociodojo: Building lifelong analytical agents with real-world text and time series. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=s9z0HzWJJp.

Giacomo Chiarot and Claudio Silvestri. Time series compression survey. *ACM Computing Surveys*, 55(10): 1–32, 2023.

Winnie Chow, Lauren E. Gardiner, Haraldur T Hallgrimsson, Maxwell A Xu, and Shirley You Ren. Towards time-series reasoning with LLMs. In *NeurIPS Workshop on Time Series in the Age of Large Models*, 2024. URL https://openreview.net/forum?id=W0UTR3LLwj.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. Navigate through enigmatic labyrinth a survey of chain of thought reasoning: Advances, frontiers and future. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1173–1203, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.65. URL https://aclanthology.org/2024.acl-long.65/.

Jaroslaw A. Chudziak and Michal Wawer. ElliottAgents: A natural language-driven multi-agent system for stock market analysis and prediction. In Nathaniel Oco, Shirley N. Dita, Ariane Macalinga Borlongan, and Jong-Bok Kim (eds.), *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pp. 961–970, Tokyo, Japan, December 2024. Tokyo University of Foreign Studies. URL https://aclanthology.org/2024.paclic-1.91/.

Longchao Da, Kuanru Liou, Tiejin Chen, Xuesong Zhou, Xiangyong Luo, Yezhou Yang, and Hua Wei. Open-ti: Open traffic intelligence with augmented language model. *International Journal of Machine Learning and Cybernetics*, 15(10):4761–4786, 2024.

Manqing Dong, Hao Huang, and Longbing Cao. Can llms serve as time series anomaly detectors?, 2024. URL https://arxiv.org/abs/2408.03475.

Mohamed Amine Ferrag, Norbert Tihanyi, and Merouane Debbah. From llm reasoning to autonomous ai agents: A comprehensive review. *arXiv preprint arXiv:2504.19678*, 2025.

Elizabeth Fons, Rachneet Kaur, Soham Palande, Zhen Zeng, Tucker Balch, Manuela Veloso, and Svitlana Vyetrenko. Evaluating large language models on time series feature understanding: A comprehensive taxonomy and benchmark. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 21598–21634, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1204. URL https://aclanthology.org/2024.emnlp-main.1204/.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. Continual pre-training for cross-lingual LLM adaptation: Enhancing japanese language capabilities. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=TQdd1VhWbe.

Moritz A. Graule and Volkan Isler. Gg-llm: Geometrically grounding large language models for zero-shot human activity forecasting in human-aware task planning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 568–574, 2024. doi: 10.1109/ICRA57147.2024.10611090.

Nate Gruver, Marc Anton Finzi, Shikai Qiu, and Andrew Gordon Wilson. Large language models are zero-shot time series forecasters. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=md68e8iZK1.

Yile Gu, Yifan Xiong, Jonathan Mace, Yuting Jiang, Yigong Hu, Baris Kasikci, and Peng Cheng. Argos: Agentic time-series anomaly detection with autonomous rule generation via large language models, 2025. URL https://arxiv.org/abs/2501.14170.

Shule Hao, Junpeng Bao, and Chuncheng Lu. A time series multitask framework integrating a large language model, pre-trained time series model, and knowledge graph, 2025. URL https://arxiv.org/abs/2503.07682.

Zelin He, Sarah Alnegheimish, and Matthew Reimherr. Harnessing vision-language models for time series anomaly detection, 2025. URL https://arxiv.org/abs/2506.06836.

Yifan Hu, Yuante Li, Peiyuan Liu, Yuxia Zhu, Naiqi Li, Tao Dai, Shu tao Xia, Dawei Cheng, and Changjun Jiang. Fintsb: A comprehensive and practical benchmark for financial time series forecasting, 2025. URL https://arxiv.org/abs/2502.18834.

Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 1049–1065, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.67. URL https://aclanthology.org/2023.findings-acl.67/.

Xu Huang, Weiwen Liu, Xiaolong Chen, Xingmei Wang, Hao Wang, Defu Lian, Yasheng Wang, Ruiming Tang, and Enhong Chen. Understanding the planning of llm agents: A survey. *arXiv preprint arXiv:2402.02716*, 2024.

Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Deep learning for time series classification: a review. *Data mining and knowledge discovery*, 33(4):917–963, 2019.

Xin Ji, Le Zhang, Wenya Zhang, Fang Peng, Yifan Mao, Xingchuang Liao, and Kui Zhang. Lemad: Llm-empowered multi-agent system for anomaly detection in power grid services. *Electronics*, 14(15), 2025. ISSN 2079-9292. doi: 10.3390/electronics14153008. URL https://www.mdpi.com/2079-9292/14/15/3008.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. Towards mitigating llm hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 1827–1843, 2023.

Furong Jia, Kevin Wang, Yixiang Zheng, Defu Cao, and Yan Liu. Gpt4mts: Prompt-based large language model for multimodal time-series forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(21):23343–23351, Mar. 2024. doi: 10.1609/aaai.v38i21.30383. URL https://ojs.aaai.org/index.php/AAAI/article/view/30383.

Yushan Jiang, Wenchao Yu, Geon Lee, Dongjin Song, Kijung Shin, Wei Cheng, Yanchi Liu, and Haifeng Chen. Explainable multi-modal time series prediction with llm-in-the-loop, 2025. URL https://arxiv.org/abs/2503.01013.

Ming Jin, Yifan Zhang, Wei Chen, Kexin Zhang, Yuxuan Liang, Bin Yang, Jindong Wang, Shirui Pan, and Qingsong Wen. Position: what can large language models tell us about time series analysis. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Satyadhar Joshi. Using gen ai agents with gae and vae to enhance resilience of us markets. *Available at SSRN 5123068*, 2025.

Zixuan Ke, Fangkai Jiao, Yifei Ming, Xuan-Phi Nguyen, Austin Xu, Do Xuan Long, Minzhi Li, Chengwei Qin, PeiFeng Wang, silvio savarese, Caiming Xiong, and Shafiq Joty. A survey of frontiers in LLM reasoning: Inference scaling, learning to reason, and agentic systems. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=S1sZZ25InC. Survey Certification.

Eamonn Keogh, Selina Chu, David Hart, and Michael Pazzani. Segmenting time series: A survey and novel approach. *Data mining in time series databases*, 57(2004):1–22, 2004.

Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research*, 2023.

Kai Kim, Howard Tsai, Rajat Sen, Abhimanyu Das, Zihao Zhou, Abhishek Tanpure, Mathew Luo, and Rose Yu. Multi-modal forecaster: Jointly predicting time series and textual data, 2024. URL https://arxiv.org/abs/2411.06735.

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 Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the*

63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 29736–29753, Vienna, Austria, July 2025a. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1437. URL https://aclanthology.org/2025.acl-long.1437/.

Yaxuan Kong, Yiyuan Yang, Shiyu Wang, Chenghao Liu, Yuxuan Liang, Ming Jin, Stefan Zohren, Dan Pei, Yan Liu, and Qingsong Wen. Position: Empowering time series reasoning with multimodal llms, 2025b. URL https://arxiv.org/abs/2502.01477.

Xiang Lan, Feng Wu, Kai He, Qinghao Zhao, Shenda Hong, and Mengling Feng. Gem: Empowering mllm for grounded ecg understanding with time series and images, 2025. URL https://arxiv.org/abs/2503.06073.

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 *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025. ISBN 978-1-57735-897-8. doi: 10.1609/aaai.v39i17.33989. URL https://doi.org/10.1609/aaai.v39i17.33989.

Changlun Li, Yao Shi, Chen Wang, Qiqi Duan, Runke Ruan, Weijie Huang, Haonan Long, Lijun Huang, Yuyu Luo, and Nan Tang. Time travel is cheating: Going live with deepfund for real-time fund investment benchmarking, 2025a. URL https://arxiv.org/abs/2505.11065.

Hao Li, Yu-Hao Huang, Chang Xu, Viktor Schlegel, Renhe Jiang, Riza Batista-Navarro, Goran Nenadic, and Jiang Bian. BRIDGE: Bootstrapping text to control time-series generation via multi-agent iterative optimization and diffusion modeling. In *Forty-second International Conference on Machine Learning*, 2025b. URL https://openreview.net/forum?id=uRD6wkqulN.

Haohang Li, Yangyang Yu, Zhi Chen, Yuechen Jiang, Yang Li, Denghui Zhang, Rong Liu, Jordan W. Suchow, and Khaldoun Khashanah. Finmem: A performance-enhanced LLM trading agent with layered memory and character design. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*, 2024a. URL https://openreview.net/forum?id=sstfVOwbiG.

Peiwen Li, Xin Wang, Zeyang Zhang, Yuan Meng, Fang Shen, Yue Li, Jialong Wang, Yang Li, and Wenwu Zhu. Realtcd: Temporal causal discovery from interventional data with large language model. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, pp. 4669–4677, New York, NY, USA, 2024b. Association for Computing Machinery. ISBN 9798400704369. doi: 10.1145/3627673.3680042. URL https://doi.org/10.1145/3627673.3680042.

Shixuan Li, Wei Yang, Peiyu Zhang, Xiongye Xiao, Defu Cao, Yuehan Qin, Xiaole Zhang, Yue Zhao, and Paul Bogdan. Climatellm: Efficient weather forecasting via frequency-aware large language models. *arXiv preprint arXiv:2502.11059*, 2025c.

Xin Li, Zhuo Cai, Shoujin Wang, Kun Yu, and Fang Chen. A survey on enhancing causal reasoning ability of large language models. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 399–416. Springer, 2025d.

Zechen Li, Baiyu Chen, Hao Xue, and Flora D. Salim. Zara: Zero-shot motion time-series analysis via knowledge and retrieval driven llm agents, 2025e. URL https://arxiv.org/abs/2508.04038.

Zhe Li, Xiangfei Qiu, Peng Chen, Yihang Wang, Hanyin Cheng, Yang Shu, Jilin Hu, Chenjuan Guo, Aoying Zhou, Christian S. Jensen, and Bin Yang. Tsfm-bench: A comprehensive and unified benchmark of foundation models for time series forecasting. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, pp. 5595–5606, New York, NY, USA, 2025f. Association for Computing Machinery. ISBN 9798400714542. doi: 10.1145/3711896.3737442. URL https://doi.org/10.1145/3711896.3737442.

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 *Proceedings of the 30th ACM*

SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '24, pp. 6555–6565, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704901. doi: 10.1145/3637528.3671451. URL https://doi.org/10.1145/3637528.3671451.

Zida Liang, Jiayi Zhu, and Weiqiang Sun. Why attention fails: The degeneration of transformers into mlps in time series forecasting. *arXiv preprint arXiv:2509.20942*, 2025.

T Warren Liao. Clustering of time series data—a survey. *Pattern recognition*, 38(11):1857–1874, 2005.

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

Cheng-Ming Lin, Ching Chang, Wei-Yao Wang, Kuang-Da Wang, and Wen-Chih Peng. Root cause analysis in microservice using neural granger causal discovery. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(1):206–213, Mar. 2024. doi: 10.1609/aaai.v38i1.27772. URL https://ojs.aaai.org/index.php/AAAI/article/view/27772.

Minhua Lin, Zhengzhang Chen, Yanchi Liu, Xujiang Zhao, Zongyu Wu, Junxiang Wang, Xiang Zhang, Suhang Wang, and Haifeng Chen. Decoding time series with llms: A multi-agent framework for cross-domain annotation, 2025. URL https://arxiv.org/abs/2410.17462.

Chanjuan Liu, Shengzhi Wang, and Enqiang Zhu. Dp-gpt4mts: Dual-prompt large language model for textual-numerical time series forecasting, 2025a. URL https://arxiv.org/abs/2508.04239.

Chenxi Liu, Hao Miao, Cheng Long, Yan Zhao, Ziyue Li, and Panos Kalnis. Llms meet cross-modal time series analytics: Overview and directions, 2025b. URL https://arxiv.org/abs/2507.10620.

Haoxin Liu, Shangqing Xu, Zhiyuan Zhao, Lingkai Kong, Harshavardhan Kamarthi, Aditya B. Sasanur, Megha Sharma, Jiaming Cui, Qingsong Wen, Chao Zhang, and B. Aditya Prakash. Time-MMD: Multi-domain multimodal dataset for time series analysis. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=fuD0h4R1IL.

Haoxin Liu, Chenghao Liu, and B. Aditya Prakash. A picture is worth a thousand numbers: Enabling LLMs reason about time series via visualization. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7486–7518, Albuquerque, New Mexico, April 2025c. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.383. URL https://aclanthology.org/2025.naacl-long.383/.

Haoxin Liu, Zhiyuan Zhao, Shiduo Li, and B. Aditya Prakash. Evaluating system 1 vs. 2 reasoning approaches for zero-shot time series forecasting: A benchmark and insights, 2025d. URL https://arxiv.org/abs/2503.01895.

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 *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, pp. 4623–4634, New York, NY, USA, 2025e. Association for Computing Machinery. ISBN 9798400714542. doi: 10.1145/3711896.3737239. URL https://doi.org/10.1145/3711896.3737239.

Xiaoyu Liu, Paiheng Xu, Junda Wu, Jiaxin Yuan, Yifan Yang, Yuhang Zhou, Fuxiao Liu, Tianrui Guan, Haoliang Wang, Tong Yu, et al. Large language models and causal inference in collaboration: A comprehensive survey. *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 7668–7684, 2025f.

Xu Liu, Taha Aksu, Juncheng Liu, Qingsong Wen, Yuxuan Liang, Caiming Xiong, Silvio Savarese, Doyen Sahoo, Junnan Li, and Chenghao Liu. Empowering time series analysis with synthetic data: A survey and outlook in the era of foundation models, 2025g. URL https://arxiv.org/abs/2503.11411.

Zewen Liu, Juntong Ni, Xianfeng Tang, Max S. Y. Lau, Wenpeng Yin, and Wei Jin. Can large language models adequately perform symbolic reasoning over time series?, 2025h. URL https://arxiv.org/abs/2508.03963.

Zhenyu Liu, Zhengtong Zhu, Jing Gao, and Cheng Xu. Forecast methods for time series data: A survey. *Ieee Access*, 9:91896–91912, 2021.

Ming-Chih Lo, Ching Chang, and Wen-Chih Peng. Text2freq: Learning series patterns from text via frequency domain. In *NeurIPS Workshop on Time Series in the Age of Large Models*, 2024. URL https://openreview.net/forum?id=Pi6sA1MSSr.

Yucong Luo, Yitong Zhou, Mingyue Cheng, Jiahao Wang, Daoyu Wang, Tingyue Pan, and Jintao Zhang. Time series forecasting as reasoning: A slow-thinking approach with reinforced llms, 2025. URL https://arxiv.org/abs/2506.10630.

Ganapathy Mahalakshmi, S Sridevi, and Shyamsundar Rajaram. A survey on forecasting of time series data. In *2016 international conference on computing technologies and intelligent data engineering (ICCTIDE'16)*, pp. 1–8. IEEE, 2016.

Mike A Merrill, Mingtian Tan, Vinayak Gupta, Thomas Hartvigsen, and Tim Althoff. Language models still struggle to zero-shot reason about time series. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 3512–3533, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.201. URL https://aclanthology.org/2024.findings-emnlp.201/.

Tymoteusz Miller, Irmina Durlik, Ewelina Kostecka, Polina Kozlovska, Adrianna Łobodzińska, Sylwia Sokołowska, and Agnieszka Nowy. Integrating artificial intelligence agents with the internet of things for enhanced environmental monitoring: Applications in water quality and climate data. *Electronics*, 14(4), 2025. ISSN 2079-9292. doi: 10.3390/electronics14040696. URL https://www.mdpi.com/2079-9292/14/4/696.

Hao Niu, Guillaume Habault, Defu Cao, Yizhou Zhang, Roberto Legaspi, Huy Quang Ung, James Enouen, Shinya Wada, Chihiro Ono, Atsunori Minamikawa, et al. Mixture of projection experts for multivariate long-term time series forecasting. In *2024 International Conference on Machine Learning and Applications (ICMLA)*, pp. 1798–1803. IEEE, 2024.

Ting-Yun Ou, Ching Chang, and Wen-Chih Peng. Coke: Causal discovery with chronological order and expert knowledge in high proportion of missing manufacturing data. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, pp. 4803–4810, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704369. doi: 10.1145/3627673.3680083. URL https://doi.org/10.1145/3627673.3680083.

Daniel Paleka, Shashwat Goel, Jonas Geiping, and Florian Tramèr. Pitfalls in evaluating language model forecasters. *arXiv preprint arXiv:2506.00723*, 2025.

Aneta Pawelec, Victoria Sara Wesołowska, Zuzanna Bączek, and Piotr Sankowski. Pub: Plot understanding benchmark and dataset for evaluating large language models on synthetic visual data interpretation, 2024. URL https://arxiv.org/abs/2409.02617.

Willa Potosnak, Cristian Ignacio Challu, Mononito Goswami, Michał Wiliński, Nina Żukowska, and Artur Dubrawski. Implicit reasoning in deep time series forecasting. In *NeurIPS Workshop on Time Series in the Age of Large Models*, 2024. URL https://openreview.net/forum?id=PFZiJwX9j5.

Willa Potosnak, Cristian Challu, Mononito Goswami, Kin G. Olivares, Michał Wiliński, Nina Żukowska, and Artur Dubrawski. Investigating compositional reasoning in time series foundation models, 2025. URL https://arxiv.org/abs/2502.06037.

Long Qian, Juncheng Li, Yu Wu, Yaobo Ye, Hao Fei, Tat-Seng Chua, Yueting Zhuang, and Siliang Tang. Momentor: advancing video large language model with fine-grained temporal reasoning. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Paul Quinlan, Qingguo Li, and Xiaodan Zhu. Chat-ts: Enhancing multi-modal reasoning over time-series and natural language data, 2025. URL https://arxiv.org/abs/2503.10883.

Chidaksh Ravuru, Sagar Srinivas Sakhinana, and Venkataramana Runkana. Agentic retrieval-augmented generation for time series analysis, 2024. URL https://arxiv.org/abs/2408.14484.

Matthew Renze and Erhan Guven. Self-reflection in llm agents: Effects on problem-solving performance. *arXiv preprint arXiv:2405.06682*, 2024.

Zhuocheng Shen. Llm with tools: A survey. *arXiv preprint arXiv:2409.18807*, 2024.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. Continual learning of large language models: A comprehensive survey. *ACM Comput. Surv.*, May 2025. ISSN 0360-0300. doi: 10.1145/3735633. URL https://doi.org/10.1145/3735633. Just Accepted.

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

Dimitris Spathis and Fahim Kawsar. The first step is the hardest: Pitfalls of representing and tokenizing temporal data for large language models. *Journal of the American Medical Informatics Association*, 31(9): 2151–2158, 2024.

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

Mingtian Tan, Mike A Merrill, Vinayak Gupta, Tim Althoff, and Thomas Hartvigsen. Are language models actually useful for time series forecasting? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=DV15UbHCY1.

Mingtian Tan, Mike A. Merrill, Zack Gottesman, Tim Althoff, David Evans, and Tom Hartvigsen. Inferring events from time series using language models, 2025. URL https://arxiv.org/abs/2503.14190.

Francis Tang and Ying Ding. Are large language models useful for time series data analysis? *arXiv preprint arXiv:2412.12219*, 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. Newsl.*, 26(2):109–118, January 2025. ISSN 1931-0145. doi: 10.1145/3715073.3715083. URL https://doi.org/10.1145/3715073.3715083.

Xiaoyu Tao, Tingyue Pan, Mingyue Cheng, and Yucong Luo. Hierarchical multimodal llms with semantic space alignment for enhanced time series classification, 2024. URL https://arxiv.org/abs/2410.18686.

Kutay Tire, Ege Onur Taga, Muhammed Emrullah Ildiz, and Samet Oymak. Retrieval augmented time series forecasting, 2024. URL https://arxiv.org/abs/2411.08249.

Sahar Torkamani and Volker Lohweg. Survey on time series motif discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 7(2):e1199, 2017.

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

Mohamed Trabelsi, Aidan Boyd, Jin Cao, and Huseyin Uzunalioglu. Time series language model for descriptive caption generation, 2025. URL https://arxiv.org/abs/2501.01832.

Alexander Okhuese Victor and Muhammad Intizar Ali. Enhancing time series data predictions: A survey of augmentation techniques and model performances. In *Proceedings of the 2024 Australasian Computer Science Week*, ACSW '24, pp. 1–13, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400717307. doi: 10.1145/3641142.3641143. URL https://doi.org/10.1145/3641142.3641143.

Bingrui Wang, Yuan Zhou, Leijiao Ge, and Sun-Yuan Kung. Large-model-based smart agent for time series anomaly detection in power systems. *Expert Systems with Applications*, 296:128917, 2026. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2025.128917. URL https://www.sciencedirect.com/science/article/pii/S0957417425025345.

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 *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025a. ISBN 978-1-57735-897-8. doi: 10.1609/aaai.v39i12.33384. URL https://doi.org/10.1609/aaai.v39i12.33384.

Haiteng Wang, Lei Ren, Tuo Zhao, and Lu Jiao. Collm: Industrial large-small model collaboration with fuzzy decision-making agent and self-reflection. *IEEE Transactions on Fuzzy Systems*, pp. 1–11, 2025b. doi: 10.1109/TFUZZ.2025.3594229.

Jiahao Wang, Mingyue Cheng, and Qi Liu. Can slow-thinking llms reason over time? empirical studies in time series forecasting, 2025c. URL https://arxiv.org/abs/2505.24511.

Jiahao Wang, Mingyue Cheng, Qingyang Mao, Yitong Zhou, Feiyang Xu, and Xin Li. Tabletime: Reformulating time series classification as training-free table understanding with large language models, 2025d. URL https://arxiv.org/abs/2411.15737.

Xinlei Wang, Maike Feng, Jing Qiu, Jinjin Gu, and Junhua Zhao. From news to forecast: Integrating event analysis in LLM-based time series forecasting with reflection. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=tj8nsfxi5r.

Yilin wang, Peixuan Lei, Jie Song, Yuzhe Hao, Tao Chen, Yuxuan Zhang, LEI JIA, Yuanxiang Li, and zhongyu wei. ITFormer: Bridging time series and natural language for multi-modal QA with large-scale multitask dataset. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=GByP03itA.

Zixiang Wang, Yinghao Zhu, Huiya Zhao, Xiaochen Zheng, Dehao Sui, Tianlong Wang, Wen Tang, Yasha Wang, Ewen Harrison, Chengwei Pan, Junyi Gao, and Liantao Ma. Colacare: Enhancing electronic health record modeling through large language model-driven multi-agent collaboration. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, pp. 2250–2261, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400712746. doi: 10.1145/3696410.3714877. URL https://doi.org/10.1145/3696410.3714877.

Hui Wei, Zihao Zhang, Shenghua He, Tian Xia, Shijia Pan, and Fei Liu. Plangenllms: A modern survey of llm planning capabilities. *arXiv preprint arXiv:2502.11221*, 2025.

Qingsong Wen, Liang Sun, Fan Yang, Xiaomin Song, Jingkun Gao, Xue Wang, and Huan Xu. Time series data augmentation for deep learning: A survey. In Zhi-Hua Zhou (ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pp. 4653–4660. International Joint Conferences on Artificial Intelligence Organization, 8 2021. doi: 10.24963/ijcai.2021/631. URL https://doi.org/10.24963/ijcai.2021/631. Survey Track.

Qingsong Wen, Tian Zhou, Chaoli Zhang, Weiqi Chen, Ziqing Ma, Junchi Yan, and Liang Sun. Transformers in time series: a survey. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI '23, 2023. ISBN 978-1-956792-03-4. doi: 10.24963/ijcai.2023/759. URL https://doi.org/10.24963/ijcai.2023/759.

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, 2025. URL https://arxiv.org/abs/2410.18959.

Mengxi Xiao, Zihao Jiang, Lingfei Qian, Zhengyu Chen, Yueru He, Yijing Xu, Yuecheng Jiang, Dong Li, Ruey-Ling Weng, Min Peng, Jimin Huang, Sophia Ananiadou, and Qianqian Xie. Retrieval-augmented large language models for financial time series forecasting, 2025a. URL https://arxiv.org/abs/2502.05878.

Xiongye Xiao, Gengshuo Liu, Gaurav Gupta, Defu Cao, Shixuan Li, Yaxing Li, Tianqing Fang, Mingxi Cheng, and Paul Bogdan. Neuro-inspired information-theoretic hierarchical perception for multimodal learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Z9AZsU1Tju.

Yijia Xiao, Edward Sun, Di Luo, and Wei Wang. Tradingagents: Multi-agents LLM financial trading framework. In *The First MARW: Multi-Agent AI in the Real World Workshop at AAAI 2025*, 2025b. URL https://openreview.net/forum?id=4QPrXwMQt1.

Qianqian Xie, Weiguang Han, Yanzhao Lai, Min Peng, and Jimin Huang. The wall street neophyte: A zero-shot analysis of chatgpt over multimodal stock movement prediction challenges. *arXiv preprint arXiv:2304.05351*, 2023.

Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, Yijing Xu, Haoqiang Kang, Ziyan Kuang, Chenhan Yuan, Kailai Yang, Zheheng Luo, Tianlin Zhang, Zhiwei Liu, GUOJUN XIONG, Zhiyang Deng, Yuechen Jiang, Zhiyuan Yao, Haohang Li, Yangyang Yu, Gang Hu, Huang Jiajia, Xiao-Yang Liu, Alejandro Lopez-Lira, Benyou Wang, Yanzhao Lai, Hao Wang, Min Peng, Sophia Ananiadou, and Jimin Huang. Finben: An holistic financial benchmark for large language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=loDHZstVP6.

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. *Proc. VLDB Endow.*, 18(8):2385–2398, 2025. URL https://www.vldb.org/pvldb/vol18/p2385-xie.pdf.

Siheng Xiong, Ali Payani, Ramana Kompella, and Faramarz Fekri. Large language models can learn temporal reasoning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10452–10470, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.563. URL https://aclanthology.org/2024.acl-long.563/.

Chao Xu, Qi Zhang, Baiyan Li, Anmin Wang, and Jingsong Bao. Visual analysis of time series data for multi-agent systems driven by large language models. In *Proceedings of the 3rd International Conference on Signal Processing, Computer Networks and Communications*, SPCNC '24, pp. 427–431, New York, NY, USA, 2025a. Association for Computing Machinery. ISBN 9798400710834. doi: 10.1145/3712335.3712410. URL https://doi.org/10.1145/3712335.3712410.

Congluo Xu, Zhaobin Liu, and Ziyang Li. Finarena: A human-agent collaboration framework for financial market analysis and forecasting, 2025b. URL https://arxiv.org/abs/2503.02692.

Xiongxiao Xu, Haoran Wang, Yueqing Liang, Philip S. Yu, Yue Zhao, and Kai Shu. Can multimodal llms perform time series anomaly detection?, 2025c. URL https://arxiv.org/abs/2502.17812.

Zhijian Xu, Hao Wang, and Qiang Xu. Intervention-aware forecasting: Breaking historical limits from a system perspective, 2025d. URL https://arxiv.org/abs/2405.13522.

Silin Yang, Dong Wang, Haoqi Zheng, and Ruochun Jin. Timerag: Boosting llm time series forecasting via retrieval-augmented generation, 2024. URL https://arxiv.org/abs/2412.16643.

Wei Yang and Jesse Thomason. Learning to deliberate: Meta-policy collaboration for agentic llms with multi-agent reinforcement learning. *arXiv preprint arXiv:2509.03817*, 2025.

Wei Yang, Defu Cao, and Yan Liu. Foundation models for demand forecasting via dual-strategy ensembling. In *KDD 2025 Workshop on AI for Supply Chain: Today and Future*, 2025a.

Yiyuan Yang, Zichuan Liu, Lei Song, Kai Ying, Zhiguang Wang, Tom Bamford, Svitlana Vyetrenko, Jiang Bian, and Qingsong Wen. Time-ra: Towards time series reasoning for anomaly with llm feedback, 2025b. URL https://arxiv.org/abs/2507.15066.

Yueyang Yao, Jiajun Li, Xingyuan Dai, MengMeng Zhang, Xiaoyan Gong, Fei-Yue Wang, and Yisheng Lv. Context-aware probabilistic modeling with llm for multimodal time series forecasting, 2025. URL https://arxiv.org/abs/2505.10774.

Wen Ye, Yizhou Zhang, Wei Yang, Lumingyuan Tang, Defu Cao, Jie Cai, and Yan Liu. Beyond forecasting: Compositional time series reasoning for end-to-end task execution. *CoRR*, abs/2410.04047, 2024. URL https://doi.org/10.48550/arXiv.2410.04047.

Wen Ye, Wei Yang, Defu Cao, Yizhou Zhang, Lumingyuan Tang, Jie Cai, and Yan Liu. Domain-oriented time series inference agents for reasoning and automated analysis, 2025. URL https://arxiv.org/abs/2410.04047.

Chin-Chia Michael Yeh, Vivian Lai, Uday Singh Saini, Xiran Fan, Yujie Fan, Junpeng Wang, Xin Dai, and Yan Zheng. Empowering time series forecasting with llm-agents, 2025. URL https://arxiv.org/abs/2508.04231.

Xinli Yu, Zheng Chen, Yuan Ling, Shujing Dong, Zongyi Liu, and Yanbin Lu. Temporal data meets llm – explainable financial time series forecasting, 2023. URL https://arxiv.org/abs/2306.11025.

Yangyang Yu, Zhiyuan Yao, Haohang Li, Zhiyang Deng, Yuechen Jiang, Yupeng Cao, Zhi Chen, Jordan W. Suchow, Zhenyu Cui, Rong Liu, Zhaozhuo Xu, Denghui Zhang, Koduvayur Subbalakshmi, GUOJUN XIONG, Yueru He, Jimin Huang, Dong Li, and Qianqian Xie. Fincon: A synthesized LLM multi-agent system with conceptual verbal reinforcement for enhanced financial decision making. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=dG1HwKMYbC.

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

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

Haochuan Zhang, Chunhua Yang, Jie Han, Liyang Qin, and Xiaoli Wang. Tempogpt: Enhancing time series reasoning via quantizing embedding, 2025a. URL https://arxiv.org/abs/2501.07335.

Junru Zhang, Lang Feng, Xu Guo, Yuhan Wu, Yabo Dong, and Duanqing Xu. Timemaster: Training time-series multimodal llms to reason via reinforcement learning, 2025b. URL https://arxiv.org/abs/2506.13705.

Lingzhe Zhang, Yunpeng Zhai, Tong Jia, Xiaosong Huang, Chiming Duan, and Ying Li. Agentfm: Role-aware failure management for distributed databases with llm-driven multi-agents. In *Proceedings of the 33rd ACM International Conference on the Foundations of Software Engineering*, FSE Companion '25, pp. 525–529, New York, NY, USA, 2025c. Association for Computing Machinery. ISBN 9798400712760. doi: 10.1145/3696630.3728492. URL https://doi.org/10.1145/3696630.3728492.

Qi Zhang, Chao Xu, Jie Li, Yicheng Sun, Jinsong Bao, and Dan Zhang. Llm-tsfd: An industrial time series human-in-the-loop fault diagnosis method based on a large language model. *Expert Syst. Appl.*, 264(C), March 2025d. ISSN 0957-4174. doi: 10.1016/j.eswa.2024.125861. URL https://doi.org/10.1016/j.eswa.2024.125861.

Wentao Zhang, Lingxuan Zhao, Haochong Xia, Shuo Sun, Jiaze Sun, Molei Qin, Xinyi Li, Yuqing Zhao, Yilei Zhao, Xinyu Cai, Longtao Zheng, Xinrun Wang, and Bo An. A multimodal foundation agent for financial trading: Tool-augmented, diversified, and generalist. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, pp. 4314–4325, New York, NY, USA, 2024a. Association for Computing Machinery. ISBN 9798400704901. doi: 10.1145/3637528.3671801. URL https://doi.org/10.1145/3637528.3671801.

Xinyu Zhang, Shanshan Feng, and Xutao Li. From text to time? rethinking the effectiveness of the large language model for time series forecasting. *arXiv preprint arXiv:2504.08818*, 2025e.

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

Xiyuan Zhang, Boran Han, Haoyang Fang, Abdul Fatir Ansari, Shuai Zhang, Danielle C. Maddix, Cuixiong Hu, Andrew Gordon Wilson, Michael W. Mahoney, Hao Wang, Yan Liu, Huzefa Rangwala, George Karypis, and Bernie Wang. When does multimodality lead to better time series forecasting?, 2025f. URL https://arxiv.org/abs/2506.21611.

Yadong Zhang, Shaoguang Mao, Tao Ge, Xun Wang, Yan Xia, Wenshan Wu, Ting Song, Man Lan, and Furu Wei. LLM as a mastermind: A survey of strategic reasoning with large language models. In *First Conference on Language Modeling*, 2024c. URL https://openreview.net/forum?id=iMqJsQ4evS.

Yizhou Zhang, Defu Cao, and Yan Liu. Counterfactual neural temporal point process for estimating causal influence of misinformation on social media. *Advances in Neural Information Processing Systems*, 35: 10643–10655, 2022.

Yizhou Zhang, Defu Cao, Lun Du, Qiang Fu, and Yan Liu. When splitting makes stronger: A theoretical and empirical analysis of divide-and-conquer prompting in LLMs. In *Second Conference on Language Modeling*, 2025g. URL https://openreview.net/forum?id=rAR7iPI8Kh.

Yuanzhao Zhang and William Gilpin. Context parroting: A simple but tough-to-beat baseline for foundation models in scientific machine learning, 2025. URL https://arxiv.org/abs/2505.11349.

Yuxuan Zhang, Yangyang Feng, Daifeng Li, Kexin Zhang, Junlan Chen, and Bowen Deng. Can competition enhance the proficiency of agents powered by large language models in the realm of news-driven time series forecasting?, 2025h. URL https://arxiv.org/abs/2504.10210.

Hang Zhao, Yujing Wang, Juanyong Duan, Congrui Huang, Defu Cao, Yunhai Tong, Bixiong Xu, Jing Bai, Jie Tong, and Qi Zhang. Multivariate time-series anomaly detection via graph attention network. In *2020 IEEE international conference on data mining (ICDM)*, pp. 841–850. IEEE, 2020.

Jiahui Zhou, Dan Li, Lin Li, Zhuomin Chen, Shunyu Wu, Haozheng Ye, Jian Lou, and Costas J. Spanos. Enhancing llm reasoning for time series classification by tailored thinking and fused decision, 2025a. URL https://arxiv.org/abs/2506.00807.

Shu Zhou, Yunyang Xuan, Yuxuan Ao, Xin Wang, Tao Fan, and Hao Wang. MERIT: Multi-agent collaboration for unsupervised time series representation learning. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 24011–24028, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1231. URL https://aclanthology.org/2025.findings-acl.1231/.

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 *IEEE INFOCOM 2024 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 1–6, 2024. doi: 10.1109/INFOCOMWKSHPS61880.2024.10620716.

Xin Zhou, Weiqing Wang, Francisco J. Baldán, Wray Buntine, and Christoph Bergmeir. Motime: A dataset suite for multimodal time series forecasting, 2025c. URL https://arxiv.org/abs/2505.15072.

Yufa Zhou, Yixiao Wang, Surbhi Goel, and Anru R Zhang. Why do transformers fail to forecast time series in-context? *arXiv preprint arXiv:2510.09776*, 2025d.

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

Zihao Zhou and Rose Yu. Can LLMs understand time series anomalies? In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=LGafQ1g2D2.

Guopeng Zhu, Weiqing Jia, Zhitai Xing, Ling Xiang, Aijun Hu, and Rujiang Hao. Cmllm: A novel cross-modal large language model for wind power forecasting. *Energy Conversion and Management*, 330:119673, 2025. ISSN 0196-8904. doi: https://doi.org/10.1016/j.enconman.2025.119673. URL https://www.sciencedirect.com/science/article/pii/S0196890425001967.

Yinghao Zhu, Changyu Ren, Shiyun Xie, Shukai Liu, Hangyuan Ji, Zixiang Wang, Tao Sun, Long He, Zhoujun Li, Xi Zhu, and Chengwei Pan. Realm: Rag-driven enhancement of multimodal electronic health records analysis via large language models, 2024. URL https://arxiv.org/abs/2402.07016.

Jiaxin Zhuang, Leon Yan, Zhenwei Zhang, Ruiqi Wang, Jiawei Zhang, and Yuantao Gu. See it, think it, sorted: Large multimodal models are few-shot time series anomaly analyzers, 2024. URL https://arxiv.org/abs/2411.02465.

A Full Taxonomy Assignments

To keep the tables concise, we use abbreviations for reasoning topologies, objectives, tasks, and attribute tags. Table 1 lists these abbreviations. The complete taxonomy assignments of all curated **research papers** are then provided separately for each reasoning topology: Table 2 (direct reasoning), Table 3 (linear chain reasoning), and Table 4 (branch-structured reasoning). Each table reports the corresponding papers together with their primary objectives, specific tasks, and attribute tags. In addition, Table 5 reports curated **non-research papers**, which correspond to the resources surveyed in Section 6 (*Current Landscape and Resources*). Since attribute tags are not applicable, these works are presented only with their type and outlook. Together, these tables serve as a comprehensive reference for the taxonomy developed in the main text.

Table 1: Abbreviations and value definitions for table headers, primary objectives, task values, and attribute tags used in the curated research paper taxonomy tables (Table 2, Table 3, and Table 4). Non-research papers (Table 5) are listed without attribute tags.

Full Name	**Abbreviation / Values**
General Table Headers	
Primary Objective	Prim. Obj.
Attribute Tag Headers (8 total)	
Task Decomposition	T-Dec
Verification and Critique	T-Ver
Ensemble Selection	T-Ens
Tool Use	T-Tool
Knowledge Access	T-Know
Multimodal Inputs	T-Multi
Agents	T-Agent
LLM Alignment	T-Align
Primary Objective Values	
Traditional Time Series Analysis	Trad. TS Anal.
Explanation and Understanding	Expl. & Und.
Causal Inference and Decision Making	Causal Inf.
Time Series Generation	TS Gen.
Task Values	
Forecasting	Forc.
Classification	Class.
Anomaly Detection	Anom. Det.
Segmentation	Segm.
Multiple Tasks	Mult. Tasks
Temporal Question Answering	Temp. QA
Explanatory Diagnostics	Expl. Diagn.
Structure Discovery	Struct. Disc.
Autonomous Policy Learning	Auto. Policy
Advisory Decision Support	Adv. Dec. Supp.
Conditioned Synthesis	Cond. Synth.
Attribute Tag Values	
T-Dec, T-Ver, T-Ens, T-Tool, T-Know, T-Multi	✓= present, empty = absent
T-Agent	0 = no agent, 1 = single agent, M = multiple agents
T-Align	P = Prompting, S = Supervised fine-tuning R = Reinforcement/preference alignment, H = Hybrid

Table 2: Full taxonomy assignments of curated research papers with direct reasoning topology only, including primary objectives, tasks, and attribute tags.

Method	Primary Obj.	Task	T-Dec	T-Ver	T-Ens	T-Tool	T-Know	T-Multi	T-Agent	T-Align
LLMTIME (Gruver et al., 2023)	Trad. TS Anal.	Forc.			✓				0	P
CiK (Williams et al., 2025)	Trad. TS Anal.	Forc.						✓	0	P
DP-GPT4MTS (Liu et al., 2025a)	Trad. TS Anal.	Forc.						✓	0	S
TEMPO (Cao et al., 2024a)	Trad. TS Anal.	Forc.	✓					✓	0	S
NNCL-TLLM (Bogahawatte et al., 2024)	Trad. TS Anal.	Forc.							0	S
CMLLM (Zhu et al., 2025)	Trad. TS Anal.	Forc.							0	P
Hybrid-MMF (Kim et al., 2024)	Trad. TS Anal.	Forc.						✓	0	S
Tang et al. (2025)	Trad. TS Anal.	Forc.							0	P
HiTime (Tao et al., 2024)	Trad. TS Anal.	Class.						✓	0	S
HeLM (Belyaeva et al., 2023)	Trad. TS Anal.	Class.						✓	0	S
FinSrag (Xiao et al., 2025a)	Trad. TS Anal.	Class.				✓	✓		0	S
Zhou & Yu (2025)	Trad. TS Anal.	Anom. Det.	✓					✓	0	P
MedTsLLM (Chan et al., 2024)	Trad. TS Anal.	Segm.						✓	0	P
ChatTime (Wang et al., 2025a)	Trad. TS Anal.	Mult. Tasks						✓	0	S
Chat-TS (Quinlan et al., 2025)	Expl. & Und.	Temp. QA						✓	0	S
ChatTS (Xie et al., 2025)	Expl. & Und.	Temp. QA						✓	0	S
ITFormer (wang et al., 2025)	Expl. & Und.	Temp. QA						✓	0	P
Time-MQA (Kong et al., 2025a)	Expl. & Und.	Temp. QA						✓	0	S
GEM (Lan et al., 2025)	Expl. & Und.	Expl. Diagn.						✓	0	S
Time-RA (Yang et al., 2025b)	Expl. & Und.	Expl. Diagn.	✓	✓	✓			✓	0	S
Momentor (Qian et al., 2024)	Expl. & Und.	Expl. Diagn.						✓	0	P
RealTCD (Li et al., 2024b)	Expl. & Und.	Struct. Disc.							0	P
GG-LLM (Graule & Isler, 2024)	Causal Inf.	Auto. Policy							0	P

Table 3: Full taxonomy assignments of curated research papers with linear chain reasoning topology only, including primary objectives, tasks, and attribute tags.

Method	Primary Obj.	Task	T-Dec	T-Ver	T-Ens	T-Tool	T-Know	T-Multi	T-Agent	T-Align
TimeReasoner (Wang et al., 2025c)	Trad. TS Anal.	Forc.	✓	✓				✓	0	P
RAF (Tire et al., 2024)	Trad. TS Anal.	Forc.				✓	✓		0	S
TimeRAG (Yang et al., 2024)	Trad. TS Anal.	Forc.				✓	✓		0	P
Time-R1 (Luo et al., 2025)	Trad. TS Anal.	Forc.	✓						0	H
Yu et al. (2023)	Trad. TS Anal.	Forc.	✓			✓	✓	✓	0	S
TableTime (Wang et al., 2025d)	Trad. TS Anal.	Class.	✓		✓	✓			0	P
VL-Time (Liu et al., 2025c)	Trad. TS Anal.	Class.	✓					✓	0	P
ZARA (Li et al., 2025e)	Trad. TS Anal.	Class.	✓			✓	✓		M	P
TimeMaster (Zhang et al., 2025b)	Trad. TS Anal.	Class.	✓	✓		✓		✓	0	H
Chow et al. (2024)	Trad. TS Anal.	Class.	✓					✓	0	S
REALM (Zhu et al., 2024)	Trad. TS Anal.	Class.	✓	✓		✓	✓	✓	0	P
VLM4TS (He et al., 2025)	Trad. TS Anal.	Anom. Det.	✓	✓				✓	0	P
LLMAD (Liu et al., 2025e)	Trad. TS Anal.	Anom. Det.	✓	✓		✓	✓		0	P
Dong et al. (2024)	Trad. TS Anal.	Anom. Det.	✓						0	S
SIGLLM (Alnegheimish et al., 2024)	Trad. TS Anal.	Anom. Det.			✓				0	P
SLEP (Wang et al., 2026)	Trad. TS Anal.	Anom. Det.		✓					1	P
LEMAD (Ji et al., 2025)	Trad. TS Anal.	Anom. Det.	✓					✓	M	P
LTM (Hao et al., 2025)	Trad. TS Anal.	Mult. Tasks				✓	✓	✓	0	P
Ravuru et al. (2024)	Trad. TS Anal.	Mult. Tasks				✓	✓		M	H
Tan et al. (2025)	Expl. & Und.	Temp. QA	✓					✓	0	H
TG-LLM (Xiong et al., 2024)	Expl. & Und.	Temp. QA	✓	✓					0	S
TempoGPT (Zhang et al., 2025a)	Expl. & Und.	Expl. Diagn.	✓					✓	0	S
TSLM (Trabelsi et al., 2025)	Expl. & Und.	Expl. Diagn.			✓	✓		✓	0	S
Xu et al. (2025a)	Expl. & Und.	Expl. Diagn.	✓			✓	✓	✓	M	P
FinAgent (Zhang et al., 2024a)	Causal Inf.	Auto. Policy	✓	✓		✓	✓	✓	1	P
FINMEM (Li et al., 2024a)	Causal Inf.	Auto. Policy	✓	✓		✓	✓	✓	1	P
Open-TI (Da et al., 2024)	Causal Inf.	Auto. Policy	✓			✓			M	P
SocioDojo (Cheng & Chin, 2024)	Causal Inf.	Adv. Dec. Supp.	✓	✓		✓	✓	✓	M	P
GenG (Zhou et al., 2024)	TS Gen.	Cond. Synth.	✓					✓	0	S
Joshi (2025)	TS Gen.	Cond. Synth.		✓		✓	✓		0	P

Table 4: Full taxonomy assignments of curated research papers with branch-structured reasoning topology only, including primary objectives, tasks, and attribute tags.

Method	Primary Obj.	Task	T-Dec	T-Ver	T-Ens	T-Tool	T-Know	T-Multi	T-Agent	T-Align
Zhang et al. (2025h)	Trad. TS Anal.	Forc.	✓	✓	✓	✓	✓	✓	M	S
NewsForecast (Wang et al., 2024)	Trad. TS Anal.	Forc.	✓	✓			✓	✓	M	S
CAPTime (Yao et al., 2025)	Trad. TS Anal.	Forc.			✓			✓	0	P
DCATS (Yeh et al., 2025)	Trad. TS Anal.	Forc.	✓	✓		✓	✓		1	P
CoLLM (Wang et al., 2025b)	Trad. TS Anal.	Forc.		✓	✓				0	S
TimeXL (Jiang et al., 2025)	Trad. TS Anal.	Forc.	✓	✓	✓			✓	M	P
ReasonTSC (Zhou et al., 2025a)	Trad. TS Anal.	Class.	✓	✓		✓			0	P
ColaCare (Wang et al., 2025)	Trad. TS Anal.	Class.	✓	✓		✓	✓	✓	M	P
TimeCAP (Lee et al., 2025)	Trad. TS Anal.	Class.	✓		✓	✓	✓	✓	M	P
AD-AGENT (Anonymous, 2025)	Trad. TS Anal.	Anom. Det.	✓	✓		✓	✓		M	P
ARGOS (Gu et al., 2025)	Trad. TS Anal.	Anom. Det.	✓	✓	✓				0	P
TAMA (Zhuang et al., 2024)	Trad. TS Anal.	Anom. Det.	✓	✓	✓			✓	0	P
LLM-TSFD (Zhang et al., 2025d)	Trad. TS Anal.	Anom. Det.	✓	✓		✓	✓		1	P
TS-Reasoner (Ye et al., 2025)	Trad. TS Anal.	Mult. Tasks	✓	✓		✓	✓		1	P
MERIT (Zhou et al., 2025b)	Trad. TS Anal.	Mult. Tasks	✓	✓		✓	✓		M	P
TESSA (Lin et al., 2025)	Expl. & Und.	Expl. Diagn.	✓	✓				✓	M	P
AgentFM (Zhang et al., 2025c)	Expl. & Und.	Expl. Diagn.	✓			✓	✓	✓	M	P
ElliottAgents (Chudziak & Wawer, 2024)	Expl. & Und.	Expl. Diagn.	✓	✓		✓	✓		M	P
Liu et al. (2025h)	Expl. & Und.	Struct. Disc.	✓	✓	✓	✓		✓	0	P
FinArena (Xu et al., 2025b)	Causal Inf.	Auto. Policy	✓	✓		✓	✓	✓	M	P
FINCON (Yu et al., 2024)	Causal Inf.	Auto. Policy	✓	✓		✓	✓	✓	M	P
TradingAgents (Xiao et al., 2025b)	Causal Inf.	Auto. Policy	✓	✓		✓	✓	✓	M	P
BRIDGE (Li et al., 2025b)	TS Gen.	Cond. Synth.	✓	✓		✓	✓	✓	M	P

Table 5: Curated non-research papers relevant to time series reasoning, including datasets, benchmarks, surveys, tutorials, and position/vision papers. These correspond to the resources discussed in Section 6 (Current Landscape and Resources).

Paper	Type
Fons et al. (2024)	Reasoning-First Benchmarks
Potosnak et al. (2024)	Reasoning-First Benchmarks
Potosnak et al. (2025)	Reasoning-First Benchmarks
ReC4TS (Liu et al., 2025d)	Reasoning-First Benchmarks
MTBench (Chen et al., 2025a)	Reasoning-First Benchmarks
TSQA (Kong et al., 2025a)	Reasoning-First Benchmarks
PUB (Pawelec et al., 2024)	Reasoning-First Benchmarks
CiK (Williams et al., 2025)	Reasoning-First Benchmarks
TimeSeriesGym (Cai et al., 2025)	Reasoning-First Benchmarks
SocioDojo (Cheng & Chin, 2024)	Reasoning-First Benchmarks
Tan et al. (2025)	Reasoning-First Benchmarks
Temporal-Synced IATSF (Xu et al., 2025d)	Reasoning-First Benchmarks
EngineMT-QA (wang et al., 2025)	Reasoning-First Benchmarks
ECG-Grounding (Lan et al., 2025)	Reasoning-Ready Benchmarks
Zhuang et al. (2024)	Reasoning-Ready Benchmarks
GPT4MTS (Jia et al., 2024)	Reasoning-Ready Benchmarks
TimeTextCorpus (Kim et al., 2024)	Reasoning-Ready Benchmarks
STOCK23 (Xiao et al., 2025a)	Reasoning-Ready Benchmarks
TETS (Cao et al., 2024a)	Reasoning-Ready Benchmarks
DeepFund (Li et al., 2025a)	Reasoning-Ready Benchmarks
Moment-10M (Qian et al., 2024)	Reasoning-Ready Benchmarks
MoTime (Zhou et al., 2025c)	Reasoning-Ready Benchmarks
Time-IMM (Chang et al., 2025a)	Reasoning-Ready Benchmarks
Time-MMD (Liu et al., 2024)	Reasoning-Ready Benchmarks
TSFM-Bench (Li et al., 2025f)	Reasoning-Ready Benchmarks
VISUELLE (Skenderi et al., 2024)	Reasoning-Ready Benchmarks
RATs40K (Yang et al., 2025b)	Reasoning-Ready Benchmarks
TimerBed (Liu et al., 2025c)	General-Purpose Time Series Benchmarks
Tan et al. (2024)	General-Purpose Time Series Benchmarks
SymbolBench (Liu et al., 2025h)	General-Purpose Time Series Benchmarks
Zhou & Yu (2025)	General-Purpose Time Series Benchmarks
VISUALTIMEANOMALY (Xu et al., 2025c)	General-Purpose Time Series Benchmarks
TimeSeriesExam (Cai et al., 2024)	General-Purpose Time Series Benchmarks
TTGenerator (Dong et al., 2024)	General-Purpose Time Series Benchmarks
TSandLanguage (Merrill et al., 2024)	General-Purpose Time Series Benchmarks
TS Instruct (Quinlan et al., 2025)	General-Purpose Time Series Benchmarks
ChatTS (Xie et al., 2025)	General-Purpose Time Series Benchmarks
TS-Reasoner (Ye et al., 2025)	General-Purpose Time Series Benchmarks
FinBen (Xie et al., 2024)	General-Purpose Time Series Benchmarks
FinTSB (Hu et al., 2025)	General-Purpose Time Series Benchmarks
Liang et al. (2024)	Surveys and Tutorials
Su et al. (2024)	Surveys and Tutorials
Liu et al. (2025g)	Surveys and Tutorials
Miller et al. (2025)	Surveys and Tutorials
Liu et al. (2025b)	Surveys and Tutorials
Tan et al. (2024)	Position and Vision Papers
Zhang & Gilpin (2025)	Position and Vision Papers
Kong et al. (2025b)	Position and Vision Papers
Jin et al. (2024)	Position and Vision Papers