Computers and Artificial Intelligence; Vol.3, No.6; 2026
E-ISSN: 3006-3728 P-ISSN: 3006-371X
DOI: https://doi.org/10.70267/cai.v3n6.0120
Published by: Zeus Press

Controversies in Zero-Shot Forecasting with Time Series Foundation Models and a Three-Layer Framework for Trustworthy Evaluation

Huiyun Yang*

School of Information Science and Technology (School of Artificial Intelligence), Beijing Forestry University, Beijing 100083, China

**Corresponding author: Huiyun Yang.*

Abstract

Time series foundation models (TSFMs) aim to forecast new datasets without target-specific parameter updates. Assessing this capability requires scrutiny of how evaluation evidence is generated and interpreted. We present a semi-systematic review of 80 core papers on the evaluation of TSFMs and large language model (LLM)-based forecasting methods. The review comprises 68 papers from the initial search and 12 from a targeted supplementary search within the same January 2022 to August 2026 publication window. We organize the evidence around benchmark integrity, ranking robustness, and domain transferability. In the reviewed literature, empirical audits of overlap between TSFM pretraining corpora and test benchmarks remain limited. Protocol choices, including normalization, forecast horizons, information availability, and metric aggregation, can alter model comparisons. Domain studies report both competitive performance and limitations, depending on task formulation, the relevance of pretraining data, sampling granularity, and data quality. These findings warrant conditional interpretations of zero-shot results rather than general claims of success or failure. We propose a five-part checklist for trustworthy evaluation and identify research priorities in corpus auditing, statistical comparison, domain-specific calibration, broader empirical coverage, and model selection. Reliable deployment requires evidence relevant to the intended task and operating conditions.

Keywords

time series foundation models, zero-shot forecasting, benchmark evaluation, data leakage, evaluation protocols

1. Introduction

Time series forecasting informs decisions in power dispatch, medical monitoring, financial risk management, and cloud operations. Conventional approaches typically fit a model to each dataset or group of related series. Since 2022, researchers have increasingly explored forecasting with foundation models, including Chronos [1], TimesFM [2], Moirai [3], MOMENT [4], and Lag-Llama [5]. These models explore whether pretraining on heterogeneous time series can reduce the need for target-specific training, although their downstream adaptation requirements differ. In parallel, LLM-based methods adapt language representations for numerical forecasting. Time-LLM [6], for example, freezes the language-model backbone while training reprogramming and output components. A frozen backbone alone therefore does not make the

© 2026 The Author(s). Authors retain copyright in their articles.

complete forecasting system zero-shot. Claims of transfer without target-specific updates must be assessed against each study's training and evaluation protocol.

However, the evidence underpinning current evaluation practices is not yet robust. Zero-shot capability is assessed through protocols and metrics, so distortions in measurement can distort the resulting conclusions. General benchmarks report competitive performance in many settings, but other studies raise three concerns about the interpretation of these results. First, incomplete disclosure of pretraining corpora makes overlap with test benchmarks difficult to assess. Second, model rankings may reverse when protocol choices, such as normalization, change [7]. Third, favorable benchmark results may not hold in independent evaluations. Zhang et al. [8] found that a simple model could match mainstream TSFMs, while domain evaluations report outcomes that vary by task and application. Tan et al. [10] further showed that removing the LLM component from widely used LLM-based forecasting methods need not reduce performance and can improve it. Together, these findings motivate the central question of this review: do current evaluation practices provide credible measures of TSFMs' zero-shot forecasting capability?

Existing surveys have not directly resolved this question. Liang et al. [11] review the development of TSFM methods; Jin et al. [12] survey the methods and applications of large models and pretrained foundation models; and Wang et al. [13] review deep time series models more broadly and introduce a comparative benchmark. These surveys primarily organize model developments and reported performance. This review instead emphasizes the credibility and comparability of the underlying evaluation evidence. Chinese-language surveys similarly cover large models for time series forecasting in decision intelligence [14] and advances in time series and spatio-temporal data analysis [15], with limited critical discussion of evaluation credibility. This gap complicates deployment decisions in high-stakes settings.

We conduct a semi-systematic review of evaluation evidence, examining whether pretraining corpora overlap with test benchmarks, whether protocol choices change model rankings, and whether conclusions vary systematically across benchmarks and domains. We then assess how independent domain evaluations qualify favorable results from general benchmarks and derive practical standards for trustworthy evaluation.

This review makes three contributions. First, we develop a three-layer framework covering benchmark integrity, ranking robustness, and domain transferability, and explain how concerns at one layer affect the others (Section 2.5). Second, we synthesize positive and negative evidence across benchmarks and domains. We identify three conditions associated with changes in comparative conclusions: benchmark selection, trade-offs between model scale and efficiency, and baseline choice and strength (Section 5.4). We also examine three recurring conditions affecting domain transfer: distribution shift, data granularity, and data quality (Section 6.5). Third, we propose a minimum-standards checklist, organized into five groups, to guide benchmark design and model reporting (Section 7).

2. Review Methodology and Research Design

2.1 Review Type and Research Questions

We adopt the semi-systematic review design described by Snyder [16] to identify recurring evaluation controversies across research communities with different terminology, tasks, and protocols. Structured searches and explicit inclusion criteria make study selection transparent, while narrative synthesis and thematic analysis support comparisons across heterogeneous evidence. Systematic reviews can also use qualitative synthesis; the absence of a quantitative meta-analysis is therefore not, by itself, a reason to choose a different design. We use a semi-systematic approach to map themes and boundary conditions across a broad, evolving literature rather than to answer a narrowly defined question about effects exhaustively. Unlike an unstructured narrative review, this approach makes search and selection decisions explicit. We do not claim exhaustive coverage, and we conditionally include influential preprints (Section 2.3). Section 2.5 discusses the resulting limitations.

The central question in Section 1 is addressed through three related questions: do pretraining corpora overlap with test benchmarks; can protocol choices change model rankings; and do general benchmark conclusions transfer to specific application domains? These questions correspond to Sections 4, 5, and 6, respectively, and guide both the inclusion criteria and the analytical framework.

© 2026 The Author(s). Authors retain copyright in their articles.

2.2 Search Strategy and Data Sources

The search comprises identification, screening, and inclusion stages; Section 2.4 reports the number of papers at each stage. Search terms are organized into five thematic groups (Table 1). Group A combines terms for time series foundation models with evaluation terminology. Group B extends this search to LLM-based forecasting and ablation studies. Group C targets leakage, reproducibility, weak baselines, and critical evaluations. Group D uses corresponding Chinese subject terms in CNKI. Group E uses title searches and forward citation tracking from eight seed papers: Chronos [1], TimesFM [2], Moirai [3], MOMENT [4], Lag-Llama [5], Time-LLM [6], GIFT-Eval [17], and the ablation study by Tan et al. [10]. Citation tracking is the main source of domain-specific evidence, which is often published outside the venues used for general benchmarking research.

Table 1: Search terms by thematic group

Group	Theme	Search terms
A	TSFM evaluation	"time series foundation model" / "foundation model for time series" / "large time series model" / "pre-trained time series model", combined with benchmark*, evaluat*, "zero-shot", "empirical study", reprodu*
B	Evaluation of LLMs for time series forecasting	"large language model" / LLM / GPT / LLaMA, combined with "time series" and forecast* / predict*, then combined with benchmark*, evaluat*, "zero-shot", ablation
C	Controversies and credibility	"time series", combined with "data leakage" / contamination / reproducib* / "simple baseline" / "actually useful" / "pitfalls" / critique, then combined with foundation / "large language model" / "pre-trained"
D	Chinese-language search (CNKI)	(time series) × (foundation model / large model / large language model / pre-trained model) × (evaluation / assessment / benchmark / zero-shot / empirical)
E	Citation tracking	Forward citation tracking from eight seed papers (listed in the main text)

We searched Google Scholar, Web of Science, arXiv, IEEE Xplore, ACM Digital Library, and CNKI. Together, these sources cover interdisciplinary research, gray literature, emerging work awaiting peer review, peer-reviewed publications, and Chinese-language literature. The publication window extends from January 2022 to August 2026, with the initial search completed on August 15, 2026. A targeted supplementary search conducted on October 21, 2026 retained the same publication cutoff and identified additional studies relevant to the existing analytical framework.

2.3 Inclusion and Exclusion Criteria

Eligible core papers must meet all four criteria: (1) publication within the search window; (2) a focus on TSFMs or LLM-based time series models, with a central contribution to forecasting evaluation through benchmark construction, comparative evaluation, independent replication, critical ablation, or domain evaluation; (3) identifiable datasets, metrics, baselines, and protocols; and (4) preferably, peer-reviewed publication. Under criterion (4), arXiv preprints are eligible if they have at least 30 citations or constitute key benchmarks or critical studies without a published version. We exclude method papers whose evaluation serves only to demonstrate the proposed model, retaining them as background rather than core evidence. Other exclusions are engineering application reports without an evaluation contribution, short papers and abstracts, studies limited to non-forecasting tasks such as classification or anomaly detection, and duplicate versions of the same study. In the supplementary search, preprints [78,84] qualified as key benchmarks and [85,86] as critical evaluations under criterion (4); their inclusion does not depend on an assumed citation count.

2.4 Screening Process and Composition of the Literature

The initial search across six databases, supplemented by citation tracking, yielded 1,300 records. After deduplication, 900 records underwent title and abstract screening, and 68 core papers were included following full-text assessment. This initial set comprised 9 benchmark papers, 17 independent evaluations and replication studies, 10 critical studies and negative-result papers, 20 domain application evaluations, 5 evaluation-methodology and position papers, and 7 cross-domain methodological references and Chinese-language papers. Grouped by evaluation purpose and modeling approach, these papers fall into seven categories: general benchmark construction, debates on LLM-based forecasting, comparative evaluation and replication, domain-

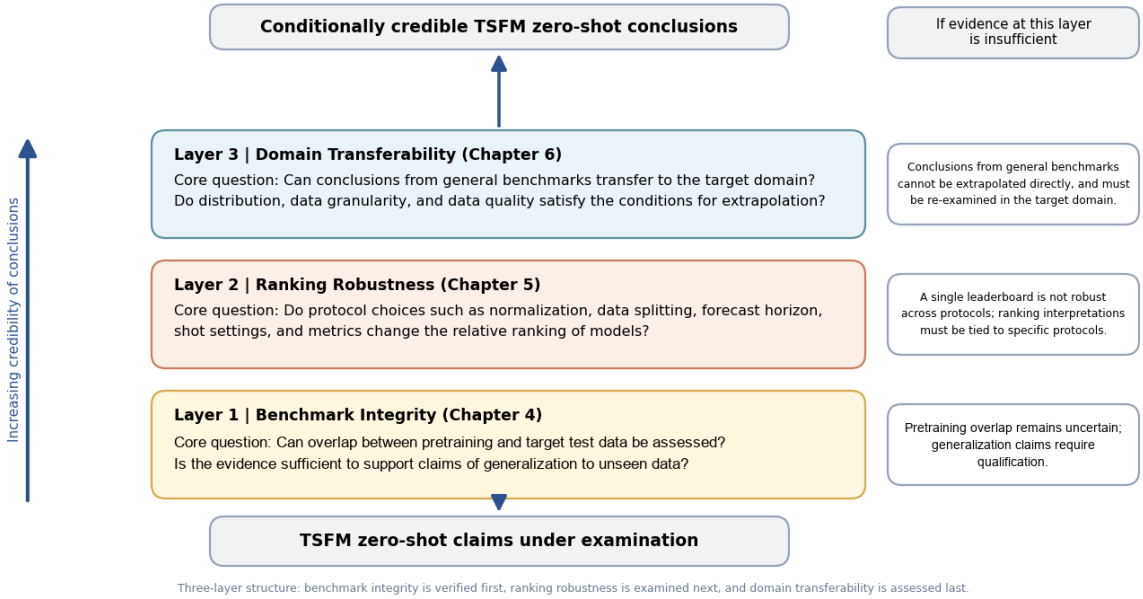
© 2026 The Author(s). Authors retain copyright in their articles.

specific evaluation, protocols and metrics, leakage and credibility, and Chinese-language research. These categories inform the organization of Sections 3–6.

2.5 Analysis Method, Screening Reliability, and Methodological Limitations

Following the qualitative synthesis approach used in semi-systematic reviews [16], we extract each paper's research focus, protocol settings, findings, and principal evaluation concerns. We then use thematic analysis to identify recurring controversies. Our analysis of benchmark integrity draws on the eight-category data leakage taxonomy proposed by Kapoor and Narayanan [18]. The framework has three layers (Figure 1). Benchmark integrity concerns whether pretraining corpora overlap with test benchmarks. Ranking robustness concerns whether normalization, data splitting, forecast horizons, and metrics affect model rankings. Domain transferability concerns whether general benchmark conclusions extend to specific applications. These layers correspond to Sections 4, 5, and 6, respectively.

Figure 1: A three-layer framework for trustworthy evaluation of zero-shot forecasting with TSFMs



For the initial search, the author reported applying the criteria in Section 2.3, reviewing each candidate twice, and archiving the screening decisions. The supplementary search checked source records, relevance to the existing sections, and duplication against the reference list. Several limitations remain. Screening by a single author may introduce selection bias despite repeated review. The semi-systematic design does not require exhaustive coverage, so relevant studies, particularly non-English publications, may have been omitted. The August 15, 2026 publication cutoff also applies to the supplementary search; later studies are outside the scope of this review. Preprints and workshop papers are identified separately and should be interpreted in light of their review status. Section 8 considers these limitations when defining the scope of the conclusions.

3. Current Evaluation Landscape

3.1 Zero-Shot Forecasting and Pretraining Exposure

Zero-shot forecasting generally means predicting a target dataset without task-specific parameter updates on that dataset. Historical observations may still be provided as input context; using them in this way does not itself constitute training. The zero-shot, few-shot, and full-shot settings in GIFT-Eval [17] and TSFM-Bench [19] distinguish levels of target-specific adaptation. A separate issue is whether target test observations, or closely related information, appeared in the pretraining corpus. Excluding such exposure strengthens evidence of generalization, but its absence is not a defining requirement of every valid zero-shot setting. A system may make no target-specific updates even when its pretraining exposure is uncertain. Studies should therefore report both the adaptation setting and the available evidence on pretraining overlap. Limited corpus transparency can constrain this assessment [20], motivating the discussion of benchmark integrity in Section 4.

© 2026 The Author(s). Authors retain copyright in their articles.

3.2 Existing Evaluation Methods and Benchmarks

Benchmarks differ in data provenance, data splits, and openness (Table 2). GIFT-Eval, introduced by Aksu et al. [17], is a foundational contribution. It compares pretrained foundation models and traditional deep learning baselines across 23 datasets under a common protocol, with results that favor TSFMs overall. TSFM-Bench, introduced by Li, Qiu, et al. [19], covers both LLM-based models and models pretrained on time series. Its zero-shot, few-shot, and full-shot comparisons indicate that TSFM advantages depend on data characteristics and that gains in data efficiency vary across settings. The independent comparison by Montet et al. [21] similarly shows that adaptation settings and efficiency metrics affect conclusions. Although both major benchmarks use standardized protocols internally, differences in dataset composition, split boundaries, and result availability limit comparisons between them [20]. These limitations have increased interest in auditing benchmark design. Qiao et al. [22] review narrow task coverage and protocol weaknesses and propose a task-centric framework for next-generation benchmarks. Controlled analysis is still needed to distinguish differences in model capability from the effects of evaluation choices. Complementary benchmarks broaden the scope: ProbTS [77] evaluates point and distributional forecasts across horizons, while fev-bench [78] includes tasks with covariates and quantifies uncertainty in aggregate model comparisons. These designs underscore the need to define the forecasting task before interpreting leaderboard results.

Table 2: Comparison of general benchmarks and evaluation frameworks

Benchmark / study	Coverage	Datasets and task sources	Evaluation settings and data splits	Protocol and availability
GIFT-Eval (Aksu et al.)	Statistical models, deep learning models, and TSFMs (17 baselines)	23 datasets with training and test splits (15 univariate, 8 multivariate), 144,000 time series, and 177 million observations; 7 domains and 10 frequencies; an additional pretraining set of approximately 23 billion observations without test-set overlap	Separate training and test sets; the final 10% of each dataset used for testing with non-overlapping rolling evaluation; short, medium, and long forecast horizons defined by frequency and domain	Standardized evaluation components and metrics; publicly available code, data, and leaderboard
TSFM-Bench / FoundTS (Li, Qiu et al.)	LLM-based TSFMs, models pretrained on time series, and dataset-specific baselines (14 TSFMs)	21 multivariate datasets from 10 domains, including electricity, traffic, and energy, annotated with 7 categories of statistical features	Zero-shot, few-shot, and full-shot settings; standardized splitting, loading, normalization, and few-shot sampling; proportional splits (e.g., 7:1:2 or 6:2:2)	Standardized experimental protocol; publicly available KDD paper and code repository
Montet, Pasquier, and Wolf	Foundation models from providers including Amazon, Salesforce, and Google, alongside statistical and deep learning baselines (16 models)	8 academic and industrial datasets: Energy, ETTh1, ETTm1, ExchangeRate, Weather, ZurichElectricity, HEIA1h, and MeteoSwiss; applications include energy, electricity, exchange rates, weather, and buildings	Zero-shot, few-shot (0%, 33%, 67%), and full-shot (100%) settings; 80%/20% training/test split; context length of 512; forecast horizons of 24/48/96/192; sliding-window test evaluation	Built on the onTime framework; publicly available paper; most academic datasets accessible through Darts; MeteoSwiss data available on request from the authors; no public leaderboard
It's TIME (Qiao et al.)	12 representative TSFMs (TimesFM, Moirai, Chronos,	50 new datasets and 98 forecasting tasks across 8 domains and multiple	Strict zero-shot setting; rolling evaluation for each	Human-in-the-loop benchmark

© 2026 The Author(s). Authors retain copyright in their articles.

Benchmark / study	Coverage	Datasets and task sources	Evaluation settings and data splits	Protocol and availability
	Toto, Sundial, VisionTS++, Kairos, TiRex, etc.)	frequencies; task configurations reflect practical needs and differences in predictability	task; MASE and CRPS aggregated across tasks after normalization against Seasonal Naive	construction emphasizing leakage prevention; publicly available code and an interactive leaderboard with multiple aggregation levels

3.3 Credibility of Evaluation Evidence

The benchmarks and tools reviewed in Section 3.2 do not yet resolve all concerns about the credibility of model comparisons. Three issues merit particular attention.

First, protocol choices can alter conclusions. Controlled normalization experiments show that changing the normalization strategy can change zero-shot errors by nearly an order of magnitude [7]. Prompt design can likewise affect conclusions about the usefulness of LLM-based forecasting [23], while forecast horizons and adaptation settings can change a model's relative performance [19]. Data-splitting and metric-related pitfalls are also well documented [24,25]. Protocol sensitivity and statistical uncertainty therefore require explicit assessment; a single leaderboard is insufficient. Some newer benchmarks quantify comparison uncertainty [78], but rankings remain specific to their task definitions. Section 5 examines this evidence in detail.

Second, evaluation often covers a limited set of dimensions. General benchmarks focus largely on forecast accuracy, although conclusions can change when other dimensions are considered. A systematic calibration study found that low point-forecast errors do not guarantee reliable interval coverage [26]. Incorporating energy consumption can change model rankings [27], and tests on consumer hardware show that the accuracy advantages of large models may not offset their deployment costs [28]. Robustness tests report sharp performance declines as noise increases [29]. Studies of learned representations also find substantial redundancy across TSFMs, suggesting that existing benchmarks may not adequately distinguish their capabilities [30]. Transferability metrics [31] and analyses of observability [32] help explain how aggregate metrics can mask structural failures. Such blind spots can arise when aggregate accuracy takes priority over calibration, efficiency, and robustness. Controlled diagnostics provide complementary evidence: SimpleTimeBench [80] tests basic temporal patterns and covariate use, while adversarial evaluations [86] examine the effects of small input perturbations. These tests probe different aspects of robustness and are not interchangeable.

Third, the capabilities to be evaluated remain contested. Early evaluations of LLM-based forecasting assessed complete systems, whereas recent studies isolate specific capabilities. Comparative studies find substantial differences among adaptation strategies [33], and feature-understanding tests reveal weaker recognition of trends and cycles than numerical forecasting claims might suggest [34]. Evaluation now also includes multimodal time series question answering [35], event-driven agent tasks [36], and financial numerical forecasting [37]. Some benchmarks distinguish fast from slow reasoning and test whether additional inference-time computation improves forecasting [38]. As tasks and evaluation methods evolve, the capabilities attributed to LLM-based forecasting require continued empirical scrutiny.

4. Benchmark Integrity

4.1 Transparency of Pretraining Corpora

Data leakage is a longstanding concern in forecasting evaluation. Hewamalage et al. [24] document normalization leakage, data-splitting errors, and metric misuse, all of which can be addressed through appropriate experimental design. TSFMs introduce an additional challenge upstream, during pretraining. Their large, often opaque corpora make it difficult to determine whether test observations have already been seen.

© 2026 The Author(s). Authors retain copyright in their articles.

As discussed in Section 3.1, the absence of target-specific updates and the absence of pretraining overlap are distinct properties. Limited corpus transparency constrains claims about generalization even when the operational zero-shot setting is clear. Meyer et al. [20] identify interacting problems involving leakage, protocol design, and benchmark comparability. Lu's review of financial TSFMs [39] highlights related concerns across pretraining corpora, architectures, and benchmarks, including opaque data provenance and insufficient attention to risk-aware evaluation.

4.2 Forms of Contamination and Audit Coverage

We examine the coverage of existing audits by contamination type. For direct overlap between pretraining corpora and test benchmarks, the reviewed literature offers limited empirical evidence; assessments often rely on the scope and quality of model-provider disclosures. Most relevant audit methods have been developed in natural language processing. Balloccu et al. [40] propose methods for detecting contamination and identifying problematic evaluation practices in closed-source LLMs. Dong et al. [41] introduce a diagnostic framework for distinguishing generalization from memorization. These approaches could help determine whether time series benchmark performance reflects generalization or prior exposure, but their transfer remains incomplete. Related time series research includes label-free assessment of pretraining data quality by Wen et al. [42] and a survey of synthetic pretraining data by Liu, Aksu, et al. [43]. These studies support corpus analysis but do not establish the absence of benchmark overlap. REAL-V-TSFM [83] adds evidence from time series extracted from real-world videos. Such data can test generalization beyond conventional benchmark distributions, but performance on new data is not itself an audit of pretraining overlap.

At the protocol level, normalization leakage and data-splitting errors are well documented [24]. The leakage taxonomy developed by Kapoor and Narayanan [18] for machine learning-based science also provides a framework applicable to TSFMs. At the reporting level, weak baselines and selective reporting can inflate the apparent benefits of new methods. In an adjacent field, McGreivy and Hakim [44] found that 79% of papers claiming to outperform standard numerical methods with machine learning used weak baselines. A similar mechanism may affect TSFM evaluation, although dedicated evidence remains limited.

5. Ranking Robustness

5.1 Evidence of Protocol Sensitivity

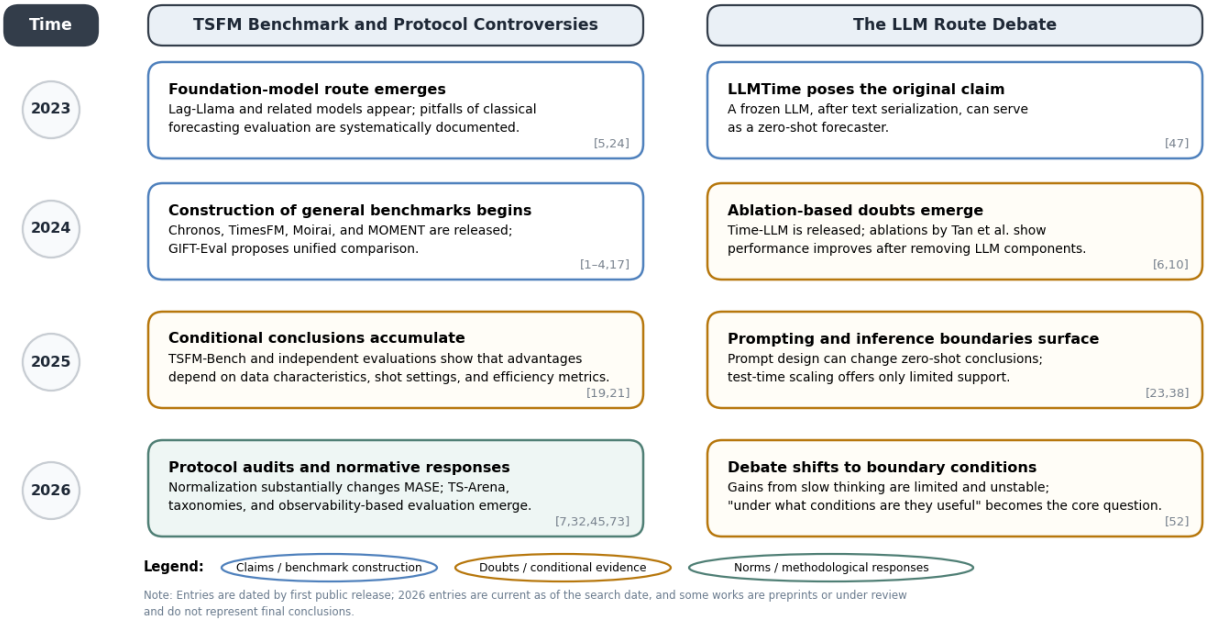
Multiple studies provide direct evidence of protocol sensitivity. Figure 2 traces these concerns across two research trajectories. Ahmed et al. [7] compare six normalization methods across four TSFMs with different architectures. Changing normalization reduced zero-shot MASE from 9.38 without normalization to 1.02 with RevIN, with an improvement of about 44% relative to the other normalization methods. The effect varied with the scale sensitivity of each model's training objective. Rankings under one protocol therefore should not be assumed to hold under another: measured differences can reflect both model capability and evaluation choices. Data splitting, forecast horizons, and adaptation settings also affect comparisons [19,24]. In LLM-based forecasting, Ashok et al. [23] show that the difference between naive and optimized prompts can change conclusions about effectiveness. O'Donnell et al. [81] provide a complementary controlled example: changing the temporal grid used to represent scientific observations affects architectures differently. Resampling should therefore also be included in protocol sensitivity analyses.

5.2 Contrasting Responses

Researchers have responded to protocol sensitivity in two contrasting ways. One response calls for more differentiated evaluation. Saqur et al. [45] criticize the pursuit of state-of-the-art rankings and advocate comparisons within data categories, arguing that a single leaderboard can obscure differences among time series tasks. A second response questions the benefits attributed to the models themselves. Large-scale ablations by Chakraborty et al. [46] suggest that the advantages of pretrained large models over smaller alternatives have been overstated. When gains from scale depend on the dataset and comparator, model size alone cannot justify ranking claims. Despite their different emphases, both responses question whether current protocols adequately support the stronger claims drawn from their results.

© 2026 The Author(s). Authors retain copyright in their articles.

Figure 2: Two research trajectories in debates on TSFM zero-shot evaluation (2023–2026)



5.3 The Debate over the Usefulness of LLMs

The usefulness of LLMs has generated a sustained debate in time series forecasting. LLMTime, introduced by Gruver et al. [47], demonstrated zero-shot forecasting by converting time series to text for a frozen LLM. Tan et al. [10] directly tested the LLM component's contribution by removing it from widely used methods such as Time-LLM; performance did not decline and sometimes improved. Merrill, Tan, et al. [48] provide complementary evidence of limited symbolic reasoning over time series in the tested settings, consistent with the feature-understanding results of Fons et al. [34] discussed in Section 3.3. The Time-LLM team's position paper [49] counters that these ablations apply to the tested methods and cannot be generalized to all LLM-based approaches. Other independent studies support a cautious interpretation. Cao and Wang [50] find no systematic advantage over classical statistical models under a common evaluation setting, while Hassani and Silva [51] argue that LLMs should serve as forecasting comparators rather than default solutions.

The debate has consequently shifted toward identifying when LLMs provide benefits. Liu et al. [38] report test-time scaling in time series forecasting, suggesting potential gains from additional inference-time computation. However, Wang et al. [52] find that extended reasoning yields limited and inconsistent improvements in time series reasoning tasks. Together with evidence of prompt sensitivity [23], these results support the conditional interpretation developed in Section 5.4. Context is Key [79] examines a distinct setting in which textual information is needed for accurate forecasts. Its results support assessing access to relevant information separately from the contribution of the language-model architecture.

5.4 Conditions Underlying Ranking Reversals

Evidence across benchmarks and settings identifies several conditions under which comparative conclusions change (Table 3). The first is benchmark selection. The favorable results in GIFT-Eval [17] and the conditional findings in TSFM-Bench [19] need not conflict: they arise from different protocols and dataset compositions. The readiness evaluation [8] further shows that apparent advantages depend on the criterion used to assess practical usability. The second condition is the trade-off between scale and efficiency. Ablation studies [46] and energy evaluations [27,28] show that large models' accuracy advantages may not preserve their rankings once computational costs are considered. The third is baseline choice and strength. In a large replication study predating TSFMs, Shao et al. [53] found that successive state-of-the-art claims in multivariate forecasting often reflected marginal gains. Zhu, Carpentier, and Verbeke [54] illustrate the importance of simple baselines in anomaly detection, although their results are not direct forecasting evidence. Domain and distribution also matter. Kamarthi and Prakash [55] show that pretraining gains depend strongly on similarity between source and target domains, providing a basis for the domain analysis in Section 6. Many current

© 2026 The Author(s). Authors retain copyright in their articles.

debates thus revisit established forecasting-evaluation problems, while pretraining contamination and divergent protocols make those problems harder to detect and assess.

Table 3: Conditions underlying ranking reversals

Boundary condition	Representative evidence	Mechanism of reversal	Impact on ranking interpretation	Recommendation
Benchmark and criterion selection	GIFT-Eval reports favorable results overall; TSFM-Bench emphasizes conditional gains; Zhang et al. find that PCA+Linear can compete with TSFMs in a readiness evaluation	Differences in dataset composition, task criteria, and protocols; questions about whether zero-shot operation is necessary for the task	A model's rank can change with the benchmark, evaluation criterion, or protocol	Specify the benchmark version, dataset composition, task criteria, and protocol for each ranking
Scale–efficiency trade-off	Chakraborty et al. find inconsistent benefits from model scale across datasets; Guibert et al. report divergence between accuracy and energy efficiency; Simeone compares accuracy, calibration, and robustness on consumer hardware	Parameter count, architecture, data characteristics, and hardware constraints jointly affect the balance of costs and benefits	Model size alone does not justify a higher ranking	Report accuracy, parameter count, latency, energy consumption, and target hardware jointly
Baseline history and strength	Shao et al. find inconsistent conclusions across studies and effects of data heterogeneity on model selection; Zhu et al. find no significant TSFM advantage over one-line baselines in anomaly detection; Saqur et al. note that strongly periodic benchmarks favor simple methods	Weak baselines, aggregate metrics, and benchmark selection can make marginal gains appear to represent major advances	A new state-of-the-art result does not necessarily indicate a substantial gain in capability	Include strong baselines, effect sizes, significance tests, and results for individual datasets
Domain and distribution match	Kamarthi and Prakash emphasize data heterogeneity in cross-domain pretraining; Zhang et al.'s OOD analysis favors diverse pretraining sources over scaling a single source; the evaluations in Section 6 identify task-dependent strengths and limitations	Differences between target and pretraining data in distribution, sampling granularity, or quality	General leaderboard rankings cannot be directly extended to specific domains	Stratify results by domain, distribution shift, granularity, and data quality

6. Domain Transferability

Table 4 summarizes the domain-specific studies discussed in this section. We examine their findings by application domain and then identify conditions that may limit transferability.

Table 4: Summary of domain-specific evaluation evidence

Domain / application	Representative study	Main findings	Scope and limitations
Energy and power: load	[56] Benchmarking Time Series Foundation Models for Short-Term Household Electricity Load Forecasting	TSFMs are broadly comparable to Transformers trained from scratch; models including TimesFM perform better with longer input contexts, while Lag-Llama and Moirai do not consistently outperform Seasonal Average (conditional)	Limited to household load forecasting; seasonal baselines remain stronger on some metrics and longer horizons

© 2026 The Author(s). Authors retain copyright in their articles.

Domain / application	Representative study	Main findings	Scope and limitations
Energy and power: load	[57] Comparative Analysis of Zero-Shot Capability of Time Series Foundation Models in Short-Term Load Prediction	Chronos, TimesFM, and TimeGPT can outperform traditional methods such as Gaussian process regression and support vector regression (conditional)	Limited application settings and dataset scale; no system-level validation across multiple conditions
Energy and power: electricity price	[58] Assessing Time Series Foundation Models for Probabilistic Electricity Price Forecasting	Moirai shows zero-shot potential but underperforms domain-specific neural networks; fine-tuning improves calibration (conditional)	Limited use of exogenous variables; calibration and point-accuracy findings do not fully align
Energy and power: electricity price	[59] Benchmarking Pre-Trained Time Series Models for Electricity Price Forecasting	Chronos-Bolt and Time-MoE are the strongest TSFMs tested, but none significantly outperforms the double-seasonal MSTL baseline (negative)	No significant advantage over the statistical baseline; conclusions depend on the protocol and metrics
Energy and power: wind and deployment	[60] Zero-Shot Forecasting of Volatile Wind Power Against Data Missing with Large Language Models	Proposes attentive residual prompt tuning to improve LLM forecasting under missing-data conditions (methodological reference)	Method-development study that informs the analysis of missing data; not an independent cross-domain evaluation
Energy and power: wind and deployment	[28] Time Series Foundation Models for Energy Load Forecasting on Consumer Hardware: A Multi-Dimensional Zero-Shot Benchmark	Leading TSFMs reduce MASE by about 47% relative to Seasonal Naive, but differ substantially in calibration, extreme-event performance, and inference latency (conditional)	Deployment suitability depends on the balance of accuracy, calibration, and cost
Energy and power: wind and deployment	[61] A Survey of Large-Language-Model-Based Forecasting Techniques for Power Systems (review)	Reviews challenges involving data quality, privacy, security, and computational resources (methodological reference)	Review evidence without a common benchmark experiment
Healthcare and biology	[62] Evaluating Zero-Shot Foundation Models for Time Series Forecasting in Clinical Settings	Task-specific models are more accurate; zero-shot models show some generalization in external validation and may suit settings where retraining is impossible (conditional)	Findings depend on the electronic health record data distribution and task settings
Healthcare and biology	[63] Zero-Shot Forecasting of ECG Time Series Data Using Generative Foundation Models	Some generative models outperform LSTM in specific settings, but Chronos and Lag-Llama require substantially longer inference times (conditional)	Chronos and Lag-Llama require more than 600 times the LSTM inference time in this benchmark; the ratio is specific to these models and settings
Healthcare and biology	[64] Multivariate Influenza Forecasting Using Time-Series Foundation Models	Chronos and TimesFM are competitive with SARIMA at a one-week horizon; Chronos performs better at longer horizons in the surveillance data evaluated (conditional)	Evidence is limited to Hong Kong influenza surveillance; advantages depend on model, horizon, and comparator
Cloud computing	[65] Performance of Zero-Shot Time Series Foundation Models on Cloud Data	Several widely used TSFMs consistently underperform simple linear baselines and produce erratic, near-random forecasts (negative)	Cloud-data characteristics and mismatch with general pretraining corpora limit transfer
Cloud computing	[66] Pushing the Limits of Pre-Training for Time Series Forecasting in the CloudOps Domain	Models pretrained within the domain provide a strong zero-shot baseline, reducing error by about 27% on the largest dataset (conditional)	Gains depend mainly on pretraining within the domain and do not establish general transfer across domains

© 2026 The Author(s). Authors retain copyright in their articles.

Domain / application	Representative study	Main findings	Scope and limitations
Finance and economics	[67] Generalisation Bounds of Zero-Shot Economic Forecasting Using Time Series Foundation Models	Models can match or outperform classical methods during stable periods, but performance deteriorates during rapid shocks (conditional)	Risk of overlap between pretraining data and the test period; conclusions may be overstated
Finance and economics	[39] Time Series Foundation Models in Finance: Pretraining Corpora, Architectures, Financial Benchmarks, and Risk-Aware Evaluation	Reviews pretraining corpora, architectures, financial benchmarks, and risk-aware evaluation dimensions (methodological reference)	Review evidence rather than a single benchmark result
Finance and economics	[68] An LLM-Fused Framework for Financial Multivariate Time Series Forecasting and Its Experimental Evaluation	The framework outperforms most deep models, but removing the LLM layer yields comparable or better results; text-prompt fine-tuning reduces performance (conditional)	The benefit of the LLM component is not robustly established
Retail demand	[69] Critical Evaluation of Time Series Foundation Models in Demand Forecasting	TimesFM performs best overall across temporal granularities, followed closely by deep learning and conventional machine-learning models; foundation models provide strong baselines (conditional)	Findings are limited to the evaluated retail demand data and temporal granularities
Natural environment	[70] Zero-Shot Forecasting of Streamflow Using Time Series Foundation Models: Are We There Yet?	The best TSFM matches domain-specific LSTMs in univariate zero-shot forecasting but performs substantially worse in multivariate forecasting; TTM improves significantly with fine-tuning (conditional)	Performance remains weaker on multivariate and extreme-value tasks; results require stratification by distribution shift
Natural environment	[71] Evaluating the Generalizability of Foundation Models for Extreme Environments	Zero-shot forecasts improve only slightly on persistence; LoRA fine-tuning does not outperform the recurrent baseline; BiLSTM performs best overall (negative)	Zero-shot gains do not hold in the tested extreme environments and spatial extrapolation settings
Meteorology (methodological reference)	[72] Assessment of AI Large-Model Forecasts of the 2024 Meiyu Rainfall in the Middle and Lower Reaches of the Yangtze River	Circulation patterns are reliable at a 10-day lead; FourCastNet and Fuxi show significant precipitation forecast skill at 11–15 days; EC-S2S precipitation forecasts outperform the large models (methodological reference)	Specialized meteorological models rather than general TSFMs; retrospective assessment of a single event

6.1 Energy and Power

Energy studies illustrate the importance of interpreting results within specific forecasting settings. For household load forecasting, Meyer et al. [56] report performance comparable to models trained from scratch, with TimesFM performing better when given longer input contexts. Lin et al. [57] find that several TSFMs outperform the traditional comparators tested in their load-prediction setting. Electricity-price studies highlight further constraints: Marchesi et al. [58] examine probabilistic accuracy and calibration, while Hornek et al. [59] find that the evaluated TSFMs do not significantly outperform the double-seasonal MSTL baseline. Simeone [28] assesses accuracy, calibration, robustness, and inference cost on consumer hardware. Together, these studies support evaluating deployment requirements jointly rather than relying on point accuracy alone. Duan et al. [60] provide methodological context through a prompt-tuning approach for volatile wind power and missing observations. Missing data can remove informative recent history and alter model inputs, motivating targeted sensitivity tests. This method-development study is not treated as independent evidence of

© 2026 The Author(s). Authors retain copyright in their articles.

general zero-shot failure. The Chinese-language power-system review [61] provides further context on deployment challenges.

6.2 Healthcare and Biology

Healthcare findings are mixed and task-dependent. Pucher et al. [62] find that task-specific models are more accurate in their clinical use cases, while zero-shot models generalize usefully to external datasets and may suit settings where retraining is impractical. In ECG forecasting, Ali et al. [63] report that generative models can outperform traditional methods in some settings, but computational costs vary substantially. In that benchmark, Chronos and Lag-Llama required more than 600 times as much inference time as LSTM. Chu et al. [64] report competitive influenza forecasts from Chronos and TimesFM, with Chronos performing better at longer horizons. Zhang et al. [9] also find that TimeGPT produces competitive forecasts of daily emergency-department arrivals against statistical and machine-learning baselines in a hospital dataset. These studies support assessing accuracy, transferability, and computational requirements for each clinical task; they do not establish uniformly negative findings across healthcare.

6.3 Cloud Computing and Industry

Studies of cloud operations distinguish transfer across domains from pretraining within the target domain. Toner et al. [65] find that simple baselines outperform several general-purpose TSFMs on cloud monitoring data. Woo et al. [66], however, obtain a strong zero-shot baseline after pretraining on CloudOps data, with lower error on their largest dataset. Because these studies use different pretraining conditions, they should not be interpreted as repeated negative findings. Together, they motivate testing whether learned temporal patterns are relevant to the target operational data.

6.4 Finance, Economics, and Other Domains

In finance and economics, model comparisons depend on the task and evaluation objective. Jetwiriyanon et al. [67] examine the limits of zero-shot economic forecasting, while Lu [39] reviews financial data provenance and risk-aware evaluation. The Chinese-language study [68] evaluates a financial forecasting framework incorporating an LLM, but its results should not be generalized to all TSFMs. Rahimikia et al. [85] distinguish off-the-shelf transfer from finance-specific pretraining, and FinVerse [84] separates point-forecast accuracy from cross-sectional ranking and portfolio evaluation. Jiang et al. [87] compare foundation models with conventional models for carbon-price forecasting across horizons and temporal granularities, reinforcing the need to specify both when interpreting performance.

For retail demand, Puvvada and Chaudhuri [69] find that TimesFM provides a strong forecasting baseline across the temporal granularities tested.

For streamflow, Sun and Sun [70] report competitive univariate zero-shot forecasts but weaker multivariate results. These findings indicate task-specific limitations rather than a general lack of useful transfer. The extreme-environment evaluation [71] concerns different conditions and should be interpreted within that setting. Li et al. [88] also report favorable zero-shot traffic-flow results, with trade-offs between accuracy and inference time as context length and model size increase. The meteorological assessment [72] is retained as a methodological reference because it evaluates specialized weather models rather than general TSFMs.

6.5 Cross-Domain Boundary Conditions

The reviewed studies identify three recurring conditions to examine before applying benchmark conclusions to a new domain. The first is the match between distributions and tasks. Pretraining can yield transferable patterns, but their usefulness depends on the target process and the information available at prediction time. Differences between general-purpose and domain-pretrained CloudOps models [65,66], and between univariate and multivariate streamflow results [70], show why distribution shift should be evaluated rather than assumed to cause failure. Syed et al. [82] further support comparisons within application domains instead of reliance on pooled rankings alone.

The second condition concerns data granularity and temporal representation. Changes in sampling frequency or aggregation can remove patterns used by a model. The household-load evaluation [56] and

© 2026 The Author(s). Authors retain copyright in their articles.

controlled temporal-grid analysis by O'Donnell et al. [81] therefore motivate sensitivity checks under the intended sampling conditions.

The third condition is data quality. Missing values reduce the available historical information, and imputation can alter temporal dependence. Duan et al. [60] address these challenges in wind-power forecasting, providing a methodological reference for their possible mechanisms. Noise tests [29] and adversarial evaluations [86] offer distinct forms of robustness evidence. None of these studies establishes a universal failure rate.

Overall, the available evidence does not justify directly extending benchmark rankings to specific applications. Performance should be reassessed for the target distribution, task, sampling granularity, and data quality, taking account of both positive and negative findings.

7. Minimum Standards and Methodological Outlook

The three-layer analysis leads to a practical question: what minimum evidence is needed to support a credible, appropriately qualified zero-shot claim? This section proposes a checklist based on the preceding evaluation concerns, discusses methodological developments, and identifies open research problems.

7.1 Minimum-Standards Checklist

We propose a checklist for trustworthy TSFM evaluation organized into five groups (Table 5). The first concerns leakage disclosure and verifiability. Reports must describe pretraining corpus composition and decontamination procedures in a form that permits independent verification [20]; studies addressing highly contested claims should preferably be preregistered [73]. The second concerns complete protocol reporting. Benchmark developers and researchers must disclose normalization, split boundaries, forecast horizons, adaptation settings, and metric definitions, and provide a sensitivity analysis of at least one protocol factor. The third concerns statistical reliability. Ranking differences require appropriate statistical comparisons and confidence intervals; marginal differences should not support state-of-the-art claims [45,53,78]. The fourth concerns multidimensional reporting. Evaluations must report calibration [25,26] and efficiency [27] alongside accuracy, with interval coverage reported separately for probabilistic forecasts [58]. The fifth concerns detailed failure analysis. Aggregate metrics must be supplemented with analyses of structural failures [32] and explicit consideration of weak baselines [44]. The dimensions–protocols–biases framework in the LLM evaluation survey [74] addresses related issues and can inform further refinement. The checklist defines a feasible minimum for research papers, technical reports, and benchmark projects.

Table 5: Minimum standards for trustworthy zero-shot evaluation of TSFMs

Group	No.	Minimum standard	Rationale / risk
1. Leakage disclosure and verifiability	1	Disclose pretraining corpus composition, including sources, time coverage, and filtering rules	Opaque corpora make the absence of pretraining exposure difficult to verify
1. Leakage disclosure and verifiability	2	Describe decontamination procedures, including checks for overlap with test benchmarks	Pretraining corpora may contain test observations
1. Leakage disclosure and verifiability	3	Provide reproducible audit trails that permit independent verification	Provider declarations alone do not establish high-confidence conclusions
1. Leakage disclosure and verifiability	4	Prefer preregistered studies when assessing highly contested claims	Preregistration limits tuning informed by prior access to test data
2. Complete protocol reporting	1	Disclose the normalization method and whether it introduces information leakage	Normalization can change zero-shot MASE by nearly an order of magnitude
2. Complete protocol reporting	2	Report split boundaries and how temporal adjacency between training and test data is handled	Random splitting and inappropriate boundary handling inflate measured performance
2. Complete protocol reporting	3	Report forecast horizons, adaptation settings, and metric definitions	Relative rankings vary across zero-shot, few-shot, and full-shot settings
2. Complete protocol reporting	4	Provide a sensitivity analysis of at least one protocol factor	Normalization and prompt design can both change conclusions

© 2026 The Author(s). Authors retain copyright in their articles.

Group	No.	Minimum standard	Rationale / risk
3. Statistical reliability	1	Support ranking differences with significance tests	Ranking instability has been observed repeatedly
3. Statistical reliability	2	Report confidence intervals or equivalent uncertainty measures	Noise should not be interpreted as a difference in capability
3. Statistical reliability	3	Do not base state-of-the-art claims on marginal differences	Replication studies show that successive state-of-the-art results often reflect marginal gains
3. Statistical reliability	4	Disclose protocol incompatibilities in cross-benchmark comparisons	GIFT-Eval and TSFM-Bench use incompatible protocols
4. Multidimensional reporting	1	Report calibration in addition to accuracy	Accuracy and calibration can diverge significantly
4. Multidimensional reporting	2	Report computational efficiency and energy consumption	Energy metrics can substantially change model rankings
4. Multidimensional reporting	3	Report interval coverage separately for probabilistic forecasts	Unstable coverage has been reported in electricity-price forecasting
4. Multidimensional reporting	4	Report deployment costs on the intended hardware or under consumer-hardware constraints	Small models lead in energy efficiency on consumer hardware
5. Fine-grained failure analysis	1	Supplement aggregate metrics with detailed analyses of structural failures	Aggregate metrics systematically mask structural failures
5. Fine-grained failure analysis	2	Explicitly discuss risks from weak baselines and selective reporting	Weak baselines and selective reporting systematically overstate the benefits of new methods
5. Fine-grained failure analysis	3	Stratify results by data category or domain conditions	A single leaderboard may obscure differences among time series tasks
5. Fine-grained failure analysis	4	Report stress or sensitivity tests for distribution shift, granularity mismatch, and data quality problems	Aggregate benchmark scores may conceal sensitivity to domain, granularity, and data quality

The checklist has three defining features. First, it follows the three-layer framework in Section 2: the first two groups address benchmark credibility, the next two address ranking robustness, and the final group addresses transferability and accountability. Second, it establishes minimum requirements rather than replacing domain-specific standards. High-risk applications, including healthcare and energy, should apply stricter criteria. Third, its value depends on continued use and refinement. Journals, conferences, and benchmark organizations can incorporate it into submission or acceptance requirements, while practitioners can adapt it for assessment before deployment.

7.2 Methodological Outlook

Evaluation methodology is receiving increasing attention. Saqur et al. [45] advocate comparisons within data categories, Cohen et al. [32] emphasize structured diagnostics, and Qiao et al. [22] organize benchmarking around tasks. Supplementary studies extend these directions: fev-bench [78] quantifies uncertainty in aggregate comparisons, while Syed et al. [82] call for explicit domain-level evaluation. Model-selection approaches, including SwiftTS [75] and evaluation-free selection [76], aim to reduce the cost of choosing among models, although their recommendations still need validation under the intended operating conditions. Future work should combine transparent data provenance, explicit adaptation settings, and diagnostic evaluation so that efficiency gains do not obscure uncertainty about deployment performance.

7.3 Open Problems

This synthesis identifies five research priorities. First, concerns about corpus contamination should be translated into reproducible audits of overlap between individual models' pretraining data and benchmark observations. Relevant starting points include corpus-analysis methods [42] and contamination diagnostics developed in NLP [40,41].

Second, statistical comparisons need more consistent treatment of dependence among forecast errors, uncertainty in aggregate rankings, and multiple comparisons. Tools such as the bootstrap intervals in fev-bench

© 2026 The Author(s). Authors retain copyright in their articles.

[78] offer useful starting points; the remaining challenge is to apply them appropriately and consistently across evaluation settings.

Third, calibration should be assessed within application domains and under distribution shift, building on general calibration studies [26] and domain-specific evaluations such as electricity-price forecasting [58].

Fourth, more independent evaluations in the Chinese-language literature would complement existing surveys and application studies [14,15,61,68].

Fifth, model-selection research should link observable data characteristics to validated selection criteria [24,75,76].

8. Conclusion

This semi-systematic review examines 80 core papers on TSFM evaluation published within the January 2022 to August 2026 window, including 12 studies identified through a targeted supplementary search. It organizes the evidence around benchmark integrity, ranking robustness, and domain transferability. The synthesis identifies conditions that affect the interpretation of zero-shot forecasting results; it does not estimate the proportion of successful or unsuccessful studies across the field.

Three findings follow. First, the absence of target-specific parameter updates should be distinguished from the absence of pretraining exposure to target test data. Reporting both strengthens the interpretation of zero-shot results. Second, normalization, forecast horizons, contextual information, temporal representation, and metric aggregation can affect model comparisons, making sensitivity analysis and uncertainty reporting essential. Third, domain studies reveal both competitive performance and limitations. A model's usefulness depends on the application, data characteristics, adaptation setting, and computational requirements. The proposed minimum-standards checklist translates these findings into reporting and validation practices. It should be used alongside domain-specific requirements and refined as evaluation methods develop.

The framework and checklist are intended to help researchers, reviewers, and practitioners distinguish demonstrated forecasting performance from claims that exceed the available evidence. They support reproducible comparisons and conclusions whose scope is justified by the evaluation.

References

[1] Ansari A F, Stella L, Turkmen C, et al. Chronos: Learning the language of time series. Transactions on Machine Learning Research, 2024.

[2] Das A, Kong W, Sen R, Zhou Y. A decoder-only foundation model for time-series forecasting. In: Proceedings of the 41st International Conference on Machine Learning (ICML). PMLR, 2024.

[3] Woo G, Liu C, Kumar A, Xiong C, Savarese S, Sahoo D. Unified training of universal time series forecasting transformers. In: Proceedings of the 41st International Conference on Machine Learning (ICML). PMLR, 2024: 53140-53164.

[4] Goswami M, Szafer K, Choudhry A, et al. MOMENT: A family of open time-series foundation models. In: Proceedings of the 41st International Conference on Machine Learning (ICML). PMLR, 2024.

[5] Rasul K, Ashok A, Williams A R, et al. Lag-Llama: Towards foundation models for probabilistic time series forecasting. arXiv preprint arXiv:2310.08278, 2023.

[6] Jin M, Wang S, Ma L, et al. Time-LLM: Time series forecasting by reprogramming large language models. In: Proceedings of the 12th International Conference on Learning Representations (ICLR), 2024.

[7] Ahmed I, Krompaß D, Feng C, et al. A comparative study on how data normalization affects zero-shot generalization in time series foundation models. In: Proceedings of ICASSP. IEEE, 2026.

[8] Zhang Y, Zeng Q, Zhang Y, Xu Z, Zheng M, Gao C, Jiang M, Zheng Z. Are time series foundation models ready for zero-shot forecasting? In: 1st ICML Workshop on Foundation Models for Structured Data, 2025 (OpenReview: kQDzqIkXLO).

© 2026 The Author(s). Authors retain copyright in their articles.

[9] Zhang X, Sridharan S, Zahrin N H B, Venkataraman N, Goh S H. Benchmarking a time-series foundation model (TimeGPT) for real-world forecasting applications. Machine Learning with Applications, 2026, 23: 100801. https://doi.org/10.1016/j.mlwa.2025.100801

[10] Tan M, Merrill M, Gupta V, Hartvigsen T, Althoff T. Are language models actually useful for time series forecasting? In: Advances in Neural Information Processing Systems 37 (NeurIPS), 2024: 60162-60191.

[11] Liang Y, Wen H, Nie Y, et al. Foundation models for time series analysis: A tutorial and survey. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2024: 6555-6565.

[12] Jin M, Kong Y, Liang Y, et al. Large models for time series and spatio-temporal data: A survey and outlook. ACM Computing Surveys, 2026.

[13] Wang Y, Wu H, Dong J, Liu Y, Long M, Wang J. Deep time series models: A comprehensive survey and benchmark. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2026, 48: 11642-11662.

[14] Shao Z, Yu C, et al. Large language models for time series forecasting in decision intelligence. Journal of Command and Control, 2025.

[15] Luo Y, Sun J, Tao L. Research progress on large models for time series and spatiotemporal data analysis. Science & Technology Review, 2025, 43(18): 48-56.

[16] Snyder H. Literature review as a research methodology: An overview and guidelines. Journal of Business Research, 2019, 104: 333-339.

[17]Aksu T, Woo G, Liu J, et al. GIFT-Eval: A benchmark for general time series forecasting model evaluation. arXiv preprint arXiv:2410.10393, 2024.

[18] Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns, 2023, 4(9): 100804.

[19] Li Z, Qiu X, Chen P, et al. TSFM-Bench: A comprehensive and unified benchmark of foundation models for time series forecasting. In: Proceedings of ACM SIGKDD, 2025.

[20] Meyer M, Kaltenpoth S, Zalipski K, Müller O. Time series foundation models: Benchmarking challenges and requirements. arXiv:2510.13654, 2025. https://doi.org/10.48550/arXiv.2510.13654

[21] Montet C, et al. Benchmarking foundation models for time series forecasting: Zero-shot, few-shot, and full-shot. MDPI CS&Math Forum, 2025, 11(1): 32.

[22] Qiao Z, et al. It's TIME: Towards the next generation of time series forecasting benchmarks. arXiv preprint arXiv:2602.12147, 2026.

[23] Ashok A, et al. Beyond naive prompting: Strategies for improved zero-shot context-aided forecasting. OpenReview, 2025.

[24] Hewamalage H, Ackermann K, Bergmeir C. Forecast evaluation for data scientists: Common pitfalls and best practices. Data Mining and Knowledge Discovery, 2023, 37: 1688-1742.

[25]Adachi N, Fujisawa K, Osborne M A. Fixing the pitfalls of probabilistic time-series forecasting evaluation. arXiv preprint arXiv:2503.06079, 2025.

[26] Adler D, et al. Beyond accuracy: Are time series foundation models well-calibrated? In: Proceedings of ICLR, 2026.

[27] Guibert Q, et al. Benchmarking time series foundation models on their accuracy and energy. In: Swiss AI Days, PMLR 309, 2026: 46-55.

[28] Simeone A. Time series foundation models for energy load forecasting on consumer hardware: A multi-dimensional zero-shot benchmark. arXiv preprint arXiv:2602.10848, 2026.

[29] Gupta S. Evaluating time series foundation models on noisy periodic time series. arXiv preprint arXiv:2501.00889, 2025.

© 2026 The Author(s). Authors retain copyright in their articles.

[30] Bao Y, et al. Universal redundancies in time series foundation models. arXiv preprint arXiv:2602.01605, 2026.

[31] Zhang W, Chen X, Li X, Chen K, Guan W, Nie L. Unified transferability metrics for time series foundation models. In: Advances in Neural Information Processing Systems 38 (NeurIPS), 2025.

[32] Cohen B, Khwaja E, Doubli Y, Lemaachi S, et al. This time is different: An observability perspective on time series foundation models. arXiv preprint arXiv:2505.14766, 2025.

[33] Chuang Y N, et al. LTSM-Bundle: A toolbox and benchmark on large language models for time series. In: Proceedings of ACM SIGKDD, 2025.

[34] Fons E, et al. Evaluating large language models on time series feature understanding: A comprehensive taxonomy and benchmark. In: Proceedings of EMNLP, 2024.

[35] Chen J, et al. MTBench: A multimodal time series benchmark for temporal reasoning and question answering. arXiv preprint arXiv:2503.16858, 2025.

[36] Weng Z, et al. TemporalBench: Evaluating LLM-based agents on contextual and event-informed time series tasks. In: Proceedings of ACM SIGKDD, 2026.

[37] Xie Q, et al. FinBen: A holistic financial benchmark for large language models. In: NeurIPS Datasets and Benchmarks, 2024.

[38] Liu H, Zhao Z, Wang J, Kamarthi H, Prakash B A. Evaluating System 1 vs. 2 reasoning approaches for zero-shot time series forecasting: A benchmark and insights. arXiv preprint arXiv:2503.01895, 2025.

[39] Lu J. Time series foundation models in finance: Pretraining corpora, architectures, financial benchmarks, and risk-aware evaluation. In: Proceedings of the 2025 2nd International Conference on Digital Economy and Computer Science. ACM, 2025: 144-151.

[40] Balloccu S, Schmidtova P, Lango M, Dusek O. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In: Proceedings of EACL, 2024: 67-93.

[41] Dong Y, et al. Generalization or memorization: Data contamination and trustworthy evaluation. In: Findings of ACL, 2024.

[42] Wen S, Feofanov V, Zhang Y. Measuring pre-training data quality without labels for time series foundation models. arXiv preprint arXiv:2412.06368, 2024.

[43] Liu Z, Aksu T, et al. Empowering time series analysis with synthetic data: A survey and outlook in the era of foundation models. arXiv preprint arXiv:2503.11411, 2025.

[44] McGreivy N, Hakim A. Weak baselines and reporting biases lead to overoptimism in machine learning for fluid-related partial differential equations. Nature Machine Intelligence, 2024, 6(10): 1256-1269.

[45] Saqur R, et al. Seeking SOTA: Time-series forecasting must adopt taxonomy-specific evaluation. arXiv preprint arXiv:2603.15506, 2026.

[46] Chakraborty S, Delibasoglu I, Heintz F. Scaling transformers for time series forecasting: Do pretrained large models outperform small-scale alternatives? Artificial Intelligence Review, 2026, 59(2).

[47] Gruver N, Finzi M, Qiu S, Wilson A G. Large language models are zero-shot time series forecasters. In: Advances in Neural Information Processing Systems 36 (NeurIPS), 2023.

[48] Merrill M A, Tan M, Gupta V, Hartvigsen T, Althoff T. Language models still struggle to zero-shot reason about time series. In: Findings of EMNLP, 2024.

[49] Jin M, et al. Position: What can large language models tell us about time series analysis? arXiv preprint arXiv:2402.02713, 2024.

[50] Cao D, Wang Y. An evaluation of standard statistical models and LLMs on time series forecasting. In: Proceedings of IEEE FMLDS, 2024.

[51] Hassani H, Silva E S. Large language models as benchmarks in forecasting practice. International Journal of Applied Forecasting, 2024.

© 2026 The Author(s). Authors retain copyright in their articles.

[52] Wang J, Cheng M, Wang D, Tao X, Liu Q, Chen E. Can slow-thinking LLMs reason over time? Empirical studies in time series forecasting. arXiv preprint arXiv:2505.24511, 2025.

[53] Shao Z, et al. Exploring progress in multivariate time series forecasting: Comprehensive benchmarking and heterogeneity analysis. IEEE Transactions on Knowledge and Data Engineering, 2025, 37(1): 291-305.

[54] Zhu L, Carpentier M, Verbeke W. When foundation models are one-liners: Limitations and future directions for time series anomaly detection. In: Proceedings of ICLR, 2026.

[55] Kamarthi H, Prakash B A. Large pre-trained time series models for cross-domain time series analysis tasks. In: Advances in Neural Information Processing Systems 37 (NeurIPS), 2024: 56190-56214.

[56] Meyer M, et al. Benchmarking time series foundation models for short-term household electricity load forecasting. IEEE Access, 2025.

[57] Lin N, Yun D, Xia W, Palensky P, Vergara P P. Comparative analysis of zero-shot capability of time series foundation models in short-term load prediction. In: 2025 IEEE Kiel PowerTech. IEEE, 2025: 1-6.

[58] Marchesi G, Ballarino A, Brusaferri A. Assessing time series foundation models for probabilistic electricity price forecasting. Energies, 2025, 18(23): 6269.

[59] Hornek T, Sartipi A, Tchappi I, Fridgen G. Benchmarking pre-trained time series models for electricity price forecasting. In: 16th International Conference on the European Energy Market (EEM). IEEE, 2025: 1-7.

[60] Duan Z, Bian C, Yang S. Zero-shot forecasting of volatile wind power against data missing with large language model through attentive residual prompt tuning. Renewable Energy, 2026, 257: 124808. https://doi.org/10.1016/j.renene.2025.124808

[61] Zhang Z, Liu D, et al. A review of power system forecasting techniques based on large language models. Power Generation Technology, 2025, 46(3): 438-453.

[62] Pucher G, Dada A, Nensa F, Schuler M, Reinhardt C, Kleesiek J, Sauer C. Evaluating zero-shot foundation models for time series forecasting in clinical settings: A simulation study with electronic health records. Studies in Health Technology and Informatics, 2025, 329: 820-824.

[63] Ali M, Moore P W, Barkouki T, Brattain L J. Zero-shot forecasting for ECG time series data using generative foundation models. In: 2024 IEEE 20th International Conference on Body Sensor Networks (BSN). IEEE, 2024: 1–4. https://doi.org/10.1109/BSN63547.2024.10780646

[64] Chu R Y K, Szeto K C, Chung P H. Multivariate influenza forecasting using time-series foundation models: A comparative study with classical statistical models using 11 years of surveillance data in Hong Kong. Scientific Reports, 2026. https://doi.org/10.1038/s41598-026-64335-9

[65] Toner W, Lee T L, Joosen A, Singh R, Asenov M. Performance of zero-shot time series foundation models on cloud data. In: I Can't Believe It's Not Better: Challenges in Applied Deep Learning at ICLR 2025 Workshops. PMLR, 2025, 296: 1–12. https://proceedings.mlr.press/v296/toner25a.html

[66] Woo G, et al. Pushing the limits of pre-training for time series forecasting in the CloudOps domain. arXiv preprint arXiv:2310.05063, 2023.

[67] Jetwiriyanon J, Susnjak T, Ranathunga S. Generalisation bounds of zero-shot economic forecasting using time series foundation models. Machine Learning and Knowledge Extraction (MAKE), 2025, 7(4): 135.

[68] Xiao W Y L, Hou C, Han S. A financial multivariate time series forecasting framework integrating large language models: Experimental evaluation. Journal of Beijing Normal University (Natural Science), 2025, 61(6): 797-804.

[69] Puvvada S, Chaudhuri A. Critical evaluation of time series foundation models in demand forecasting. In: TSFALM Workshop (OpenReview: TS42sRKINd), 2024.

[70] Sun A Y, Sun A A. Zero-shot forecasting of streamflow using time series foundation models: Are we there yet? Machine Learning: Earth, 2026, 2: 010501. https://doi.org/10.1088/3049-4753/ae4982

© 2026 The Author(s). Authors retain copyright in their articles.

[71] Huang Y, et al. Evaluating the generalizability of foundation models for extreme environmental events: Case study of California wildfire PM2.5. arXiv preprint arXiv:2607.07951, 2026.

[72] Cao X, Liang P, Huang W, Zhang Z, Gong Y. Forecast evaluation of large artificial intelligence models for the 2024 Meiyu season in the middle and lower reaches of the Yangtze River. Journal of Tropical Meteorology, 2024, 40(6): 1045-1062.

[73] Meyer M, et al. TS-Arena: A live forecast pre-registration platform for leakage-free evaluation. In: Proceedings of ACM SIGKDD, 2026.

[74]Luo W, Wang H. A review of large language model evaluation. Journal of Chinese Information Processing, 2024, 38(1): 1-23.

[75] Zhang T, Ouyang B, Shu Y, Chen X, Guo C, Yang B. SwiftTS: A swift selection framework for time series pre-trained models via multi-task meta-learning. In: Proceedings of ICLR, 2026.

[76] Abdallah Z, et al. Evaluation-free time series forecasting model selection via meta-learning. ACM Transactions on Knowledge Discovery from Data, 2025, 19(3): 1-41.

[77] Zhang J, Wen X, Zhang Z, Zheng S, Li J, Bian J. ProbTS: Benchmarking point and distributional forecasting across diverse prediction horizons. Advances in Neural Information Processing Systems 37, Datasets and Benchmarks Track, 2024. https://doi.org/10.52202/079017-1523

[78] Shchur O, Ansari A F, Turkmen C, et al. fev-bench: A realistic benchmark for time series forecasting. arXiv:2509.26468, 2025. https://doi.org/10.48550/arXiv.2509.26468

[79] Williams A R, Ashok A, Marcotte É, et al. Context is Key: A benchmark for forecasting with essential textual information. In: Proceedings of ICML. PMLR, 2025, 267: 66887–66944. https://proceedings.mlr.press/v267/williams25a.html

[80] Ghoroghchian N, Zhang H, Han S, Labach A, Stein G. Foundations without fundamentals: Zero-shot blind spots in time series FMs. In: 2nd ICML Workshop on Foundation Models for Structured Data, 2026. https://openreview.net/forum?id=iIRdd86Xkr

[81] O'Donnell L, Kolbeinsson A, Kolbeinsson B, Deisenroth M P. Towards benchmarking time series foundation models on native scientific data. In: 2nd ICML Workshop on Foundation Models for Structured Data, 2026. https://openreview.net/forum?id=NtU2gLl4hN

[82] Syed M A B, Ahamed M Y, Wasi A T. Position: Time-series foundation models require explicit domain-level benchmarks. In: Proceedings of ICML. PMLR, 2026, 306: 172221–172247. https://proceedings.mlr.press/v306/syed26a.html

[83] Li L, Sleem L, Wang Y, Xu Y, Gentile N, State R. Uncovering zero-shot generalization gaps in time-series foundation models using real-world videos. In: AI4TS Workshop at AAAI, 2026. arXiv:2509.26347. https://doi.org/10.48550/arXiv.2509.26347

[84] Lee J, Seo J, Lee S, et al. FinVerse: Financial time-series benchmark. arXiv:2608.03259v1, 2026. https://arxiv.org/abs/2608.03259v1

[85] Rahimikia E, Ni H, Wang W. Re(Visiting) time series foundation models in finance. arXiv:2511.18578, 2025. https://doi.org/10.48550/arXiv.2511.18578

[86] Zhang J, Zhang Z, Zheng S, Wen X, Li J, Bian J. Are time-series foundation models deployment-ready? A systematic study of adversarial robustness across domains. arXiv:2505.19397, 2025. https://doi.org/10.48550/arXiv.2505.19397

[87] Jiang L L, Cheong W S, Ng J S L. Multi-step ahead carbon credit price forecasting using time series foundation models. Applied Soft Computing, 2026, 198: 115290. https://doi.org/10.1016/j.asoc.2026.115290

[88] Li Y, Zhao Q, Wang M. Zero-shot traffic flow prediction with foundation models: A comparison with deep learning approaches. Scientific Reports, 2026, 16: 20472. https://doi.org/10.1038/s41598-026-50464-8

© 2026 The Author(s). Authors retain copyright in their articles.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Copyrights

Copyright for this article is retained by the author (s), with first publication rights granted to the journal. This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/).

© 2026 The Author(s). Authors retain copyright in their articles.