International Journal of Information Technology and Computer Science Applications (IJITCSA)
p-ISSN: 2964-3139 e-ISSN: 2985-5330
Vol. 04, No. 02, page 125 - 140
Submitted 15/4/2026; Accepted 2/7/2026; Published 2/7/2026

Toward Rigorous Zero-Shot and Few-Shot Benchmarking of Time-Series Foundation Models Under Domain Shift: A Leakage-Aware Benchmark Specification, Governance Framework, and Executable Pilot Instantiation

Ibezimako Chiazagomekpere
Department of Information Technology Education, Akenten Appiah-Menka University of Skills Training and Entrepreneurial Development, Kumasi, GHANA.
e-mail: ibezkpere202211@aamusted.edu.gh

Corresponding Autor: Ibezimako Chiazagomekpere

Abstract
Time-series foundation models (TSFMs) are increasingly promoted as reusable forecasting systems that can generalize across domains with zero-shot or few-shot adaptation. That claim is scientifically consequential, but current evaluation practice remains under-specified where it matters most target-domain separation, contamination control, adaptation budget definition, shift severity characterization, and aggregation across heterogeneous deployment conditions. This paper reconstructs TSFM benchmarking as a methodological problem rather than a leaderboard problem. We formalize zero-shot and few-shot forecasting under domain shift as conditional risk estimation over governed target distributions; develop a forecasting-specific taxonomy of shift covering temporal regime, entity, resolution, schema, horizon, observation-quality, intervention, and label-formation change; and propose a six-layer benchmark architecture spanning model governance, dataset governance, deterministic shift generation, evaluation tracks, metric tensors, and reporting bundles. The contribution is primarily conceptual, but to avoid a purely rhetorical framework, we also provide an executable pilot instantiation on a public electricity-transformer forecasting setting. Because large-scale TSFM execution was not conducted in this package, the pilot uses lightweight surrogate forecasters to validate the benchmark machinery itself rather than to claim new TSFM state of the art. Even this limited pilot shows that in-domain and cross-domain rankings can diverge sharply, that adaptation gains must be interpreted jointly with cost, and that robustness to observation degradation and calibration cannot be inferred from average point error alone. The paper therefore advances a benchmark doctrine: credible TSFM claims require leakage-aware governance, severity-conditioned analysis, explicit adaptation accounting, and multi-objective reporting that aligns evidence with generalization claims.

Keywords—Time-series foundation models, Domain shift, Zero-shot forecasting, Few-shot adaptation, Benchmark design, Leakage-aware evaluation.

1 Introduction

Time-series forecasting is entering a foundation-model phase. Instead of training a separate model for each dataset and deployment environment, TSFMs seek to pretrain once on large and heterogeneous temporal corpora and then transfer to downstream tasks with little or no task-specific supervision [1], [2], [3], [4]. This shift is strategically important because many operational forecasting problems are data-poor, compute-constrained, latency-sensitive, or structurally unable to support full model retraining. In energy systems, healthcare monitoring, transport operations, observability telemetry, and industrial sensing, the practical demand is often not "maximum offline accuracy after

full supervised optimization," but rather "reliable forecasting under limited target supervision and changing operating conditions" [5], [6], [7].

Recent TSFM families have made this transition concrete. TimesFM proposes a decoder-only pretraining regime for zero-shot forecasting [8]; Chronos formulates time series as tokenized sequences amenable to pretrained language-model style transfer [9]; MOMENT extends the foundation paradigm beyond forecasting into general-purpose time-series representation learning [10]; Moirai and Moirai-MoE target universal multivariate forecasting with large-scale open temporal corpora [11], [12]; Lag-Llama focuses on probabilistic forecasting transfer [13]; and Tiny Time Mixer emphasizes lightweight zero-shot and few-shot deployment [14]. The resulting literature is methodologically diverse, architecturally heterogeneous, and increasingly ambitious in its transfer claims.

However, the evaluation culture has not matured at the same rate. Existing studies often mix datasets with opaque relationships to pretraining corpora, use incomparable few-shot protocols, aggregate across shifts that are never explicitly represented, and report single averaged scores that compress away the very conditions most relevant to deployment [15], [16], [17], [18], [19]. From a reviewer perspective, this means that many published numbers remain under-interpretable: they reveal that a model scored well on a benchmark, but not whether the benchmark actually measured the type of transfer being claimed.

The problem is not merely technical bookkeeping. Foundation-style claims are stronger than conventional supervised claims. A model is a general-purpose temporal forecaster must demonstrate competence under controlled changes in domain, horizon, representation, data quality, and supervision budget [1], [3], [17]. Otherwise, what appears to be generalization may merely be mild interpolation, pretraining overlap, or adaptation privilege. In other words, TSFM benchmarking has become an epistemic bottleneck.

2 Research Gap

The most consequential gap in the present literature is the absence of a leakage-aware, shift-explicit, budget-standardized doctrine for evaluating zero-shot and few-shot TSFMs. Several recent benchmark initiatives have improved the situation. FoundTS explicitly supports zero-shot, few-shot, and full-shot comparison under a unified evaluation pipeline [20]. GIFT-Eval broadens dataset coverage, includes a non-leaking pretraining corpus, and offers a stronger zero-shot benchmark infrastructure [18]. TSFM-Bench further consolidates foundation-model evaluation across architectures and settings [16]. More recent work has also exposed information leakage, overlap ambiguity, and benchmark-integrity risks in existing TSFM comparisons [17]. Yet the field still lacks a unified benchmark theory that answers five unresolved questions.

First, what precisely constitutes zero-shot and few-shot risk when the target distribution is shifted rather than merely unseen? Second, which kinds of domain shift are structurally relevant for forecasting, and how should they be operationalized rather than passively inherited from public datasets? Third, how should contamination, adaptation budget, and evaluation aggregation be disclosed so that strong transfer claims remain scientifically defensible? Fourth, what metric family is sufficient for benchmarking a model that may trade point accuracy against calibration, robustness, efficiency, or adaptation elasticity? Fifth, how can a benchmark paper avoid becoming a scoreboard artifact and instead act as a diagnostic instrument?

This paper addresses those gaps. The emphasis is not on inventing a new model architecture, but on reconstructing the benchmark as a governed experimental object.

3 Research Objectives

The objectives of this study are

First, we formalize TSFM evaluation under domain shift as conditional risk estimation over governed target distributions, thereby replacing vague notions of "generalization" with explicit objects of measurement.

Second, we propose a forecasting-specific taxonomy of domain shift and a benchmark architecture that jointly addresses data governance, model governance, deterministic shift generation, budgeted adaptation tracks, multi-objective metrics, and reporting.

Third, to avoid a purely rhetorical methodology section, we include a small pilot benchmark instantiation on public transformer-temperature forecasting data using lightweight surrogate models. This pilot is not presented as a TSFM leaderboard. Its purpose is narrower and more honest: to verify that the proposed benchmark machinery can produce meaningful severity-conditioned, budget-conditioned, and reliability-aware diagnostics.

4 Main Contributions

The paper makes the following concrete contributions.

1. It reformulates zero-shot and few-shot TSFM benchmarking as conditional risk estimation over shifted target distributions, with explicit treatment of contamination risk, adaptation budget, and efficiency cost.
2. It introduces a forecasting-specific taxonomy of domain shift that distinguishes temporal regime, entity, frequency, schema, horizon, observation-quality, intervention, and label-formation change.
3. It proposes a six-layer leakage-aware benchmark architecture combining governance, shift generation, evaluation tracks, metric tensors, and reporting obligations.
4. It specifies benchmark artifacts including model cards, data cards, shift cards, contamination statements, raw-output release, and reproducibility bundles.
5. It demonstrates an executable pilot instantiation on a public electricity-transformer setting, showing how the benchmark logic exposes differences between in-domain fit, cross-domain transfer, adaptation elasticity, efficiency, robustness under corruption, and calibration behavior.

The novelty claim is therefore methodological and infrastructural rather than architectural. The contribution is not another TSFM variant, but a stricter definition of what evidence future TSFM papers must provide to make strong generalization claims.

5 Literature Review

5.1 Conventional Approaches

Before the rise of TSFMs, forecast research followed a dataset-centric regime. Statistical methods such as ARIMA, exponential smoothing, and seasonal-naive baselines were fitted per dataset or per series, often with explicit assumptions on stationarity, seasonality, and error structure [21], [22]. Deep-learning forecasting extended this paradigm with task-specific RNN, CNN, and transformer architectures, but the dominant logic remained unchanged, models were trained and tuned against a single benchmark or a small task family [5], [23]. Even strong transformer-based long-horizon models such as Informer, Autoformer, FEDformer, and PatchTST were largely developed and evaluated in a supervised per-task regime [24], [25], [26], [27]. Their innovations were architectural rather than foundation oriented.

This literature matters for two reasons. First, it established many of the benchmark datasets, split conventions, and metrics still reused by TSFM papers today. Second, it also exposed many of the same evaluation pathologies now resurfacing in foundation-model work such as inconsistent train-validation-test logic, insufficiently strong baselines, overreliance on a small set of long-horizon datasets, and limited reporting of uncertainty, efficiency, and robustness [15]. In that sense, TSFM benchmarking did not begin with a clean slate; it inherited both assets and weaknesses from task-specific forecasting research.

5.2 Recent Advanced Approaches

Advanced literature can be divided into three overlapping strands. The first comprises surveys and conceptual overviews that map the TSFM landscape. Liang et al. present a tutorial-style synthesis of foundation models for time-series analysis [2]; Ma et al. provide a broader survey of pre-trained time-series models [3]; and Kottapalli et al. review the emerging design space of transformer-based time-series foundation models [4]. These works establish the breadth of the field, but they necessarily compress benchmark issues that now require more granular treatment.

The second strand is model-centric. TimesFM argues that a decoder-only attention model pretrained on large temporal corpora can deliver strong zero-shot forecasting [8]. Chronos shows that tokenization and language-model style pretraining can be effective for probabilistic time-series transfer [9]. MOMENT extends pretrained temporal backbones to broader time-series tasks [10]. Moirai introduces masked encoder-based universal forecasting with LOTSA-scale pretraining corpora [11]. Lag-Llama emphasizes probabilistic transferability [13]. Time-LLM and related approaches investigate reprogramming large language models for temporal prediction [28]. This literature is rich in architectural ideas, but the evaluation setups remain heterogeneous, making direct claims of superiority difficult to interpret.

The third strand is benchmark- and evaluation-centric. FoundTS standardizes evaluation logic across zero-shot, few-shot, and full-shot forecasting [20]. GIFT-Eval contributes a broader benchmark and a non-leaking pretraining corpus [18]. TSFM-Bench further systematizes comparative evaluation [16]. fev-bench argues for stronger aggregation procedures and reproducible infrastructure [19]. Meyer et al. directly address information leakage and benchmark integrity in TSFM evaluation [17]. Parallel work also questions whether current TSFMs are as domain-

agnostic as often implied [29] and whether robust deployment demands more than average accuracy [30]. These studies collectively motivate a more principled benchmark doctrine, but none offers a fully integrated governance-and-risk formulation tailored to zero-shot and few-shot transfer under domain shift.

5.3 Strengths and Limitations of Existing Work

The existing literature has four major strengths. It has demonstrated that large-scale temporal pretraining is technically feasible [8], [9], [11]; it has diversified the architectural search space [10], [12]; it has increased community awareness of benchmark design as a research object [18], [19], [20]; and it has begun to surface integrity problems such as contamination and overlap ambiguity [17].

Its limitations are equally clear. First, zero-shot is often underdefined. Some papers forbid all target-side updates; others allow hidden calibration or preprocessing choices that act like adaptation. Second, few-shot protocols vary widely in support-set construction, updated parameters, and tuning privilege. Third, distribution shifts are usually treated as incidental rather than controlled. Benchmarks may contain multiple latent domains, but those domains are rarely operationalized into severity ladders or explicit experimental factors. Fourth, aggregation is frequently overcompressed. A single average metric can hide catastrophic failure in minority or severe-shift conditions. Fifth, benchmark contamination remains deeply unresolved because broad pretraining corpora, synthetic augmentation, and reused legacy datasets create ambiguous relationships between training and test environments [6], [17].

There is also a methodological asymmetry between model papers and benchmark papers. Model papers often claim that a backbone or training objective improves transfer. Benchmark papers must instead justify what transfer means, how it is isolated, and why the reported performance can support the narrative being advanced. This makes benchmark design a higher-stakes epistemic task than simple leaderboard maintenance.

6 Research Positioning

This paper is positioned at the intersection of TSFM evaluation, forecasting methodology, domain generalization, and benchmark governance. It differs from model-centric TSFM papers by not proposing a new backbone. It differs from existing surveys by moving from landscape description to operational benchmark doctrine. It differs from current benchmark papers by centering four ideas simultaneously such as conditional risk under shift, leakage-aware governance, budget-explicit adaptation, and vector-valued reporting.

The closest intellectual neighbors are WILDS-style distribution-shift benchmarking [31], HELM-style evaluation governance for foundation models [32], and recent TSFM benchmark and leakage analyses [16], [17], [18], [20]. However, our focus is distinctively forecasting-specific. Time order, horizon dependence, repeated entities, exogenous shocks, observation degradation, and correlated future structure make domain shift in time series qualitatively different from static supervised OOD evaluation. The paper therefore argues that TSFM benchmarking requires a dedicated doctrine rather than simple import of generic benchmark ideas.

7 Proposed Methodology

7.1 Problem Formulation

Let $\mathbf{x}_{t-c+1:t}^{(d)} \in \mathbb{R}^{c\times p}$ denote a context window of length c with p observed channels drawn from domain d, and let $\mathbf{y}_{t+1:t+h}^{(d)} \in \mathbb{R}^{h}$ denote the future target sequence of horizon h. A pretrained TSFM with parameters θ maps the context, optional metadata $\mathbf{m}^{(d)}$, and optional adaptation state ϕ to a predictive distribution

$$p_{\theta,\phi}\left(\mathbf{y}_{t+1:t+h}^{(d)} \mid \mathbf{x}_{t-c+1:t}^{(d)}, \mathbf{m}^{(d)}\right) \tag{1}$$

We distinguish a broad source distribution $P_{\text{src}}(\mathbf{x}, \mathbf{y}, d)$, used for pretraining or source-stage model fitting, from one or more governed target distributions $P_{\text{tgt}}^{(j)}(\mathbf{x}, \mathbf{y}, d, \Delta)$ that differ along controlled shift axes Δ. For a loss $\mathcal{L}$, the *zero-shot risk* on target domain j and horizon h is

$$\mathcal{R}_{\text{ZS}}^{(j,h)}(\theta) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim P_{\text{tgt}}^{(j,h)}}[\mathcal{L}(f_\theta(\mathbf{x}), \mathbf{y})] \tag{2}$$

where no target-domain parameter update is permitted. Few-shot evaluation introduces a support set

$$\mathcal{S}_k^{(j)} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{k} \tag{3}$$

sampled under a deterministic target-domain protocol. An adaptation operator $\mathcal{A}$ yields an adapted state

$$(\theta', \phi') = \mathcal{A}\left(\theta, \phi, \mathcal{S}_k^{(j)}; \psi\right) \tag{4}$$

where ψ defines what may change: prompts, normalization statistics, adapters, selected layers, or full weights. The resulting *few-shot risk* is

$$\mathcal{R}_{\mathrm{FS}}^{(j,h,k)}(\theta, \phi) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim P_{\mathrm{tgt}}^{(j,h)}}\left[\mathcal{L}\left(f_{\theta',\phi'}(\mathbf{x}), \mathbf{y}\right)\right] \tag{5}$$

A central quantity in our doctrine is the adaptation elasticity

$$\mathcal{G}^{(j,h,k)} = \mathcal{R}_{\mathrm{ZS}}^{(j,h)} - \mathcal{R}_{\mathrm{FS}}^{(j,h,k)} \tag{6}$$

which must be interpreted jointly with the adaptation cost

$$\mathcal{C}^{(j,h,k)} = \lambda_1 T_{\mathrm{adapt}} + \lambda_2 M_{\mathrm{adapt}} + \lambda_3 E_{\mathrm{adapt}} + \lambda_4 P_{\mathrm{upd}} \tag{7}$$

where T_{adapt} is wall-clock adaptation time, M_{adapt} is memory overhead, E_{adapt} is energy or compute expenditure, and P_{upd} is the number of updated parameters. We therefore define an efficiency-normalized gain:

$$\mathcal{E}^{(j,h,k)} = \frac{\mathcal{G}^{(j,h,k)}}{\mathcal{C}^{(j,h,k)} + \epsilon} \tag{8}$$

Benchmark validity also depends on information exposure. Let $\mathcal{I}_{\mathrm{pre}}$ denote the information plausibly accessible during pretraining and development. A benchmark claim is only interpretable if the effective overlap

$$\mathcal{O}\left(B, \mathcal{I}_{\mathrm{pre}}\right) \tag{9}$$

is bounded, categorized, or explicitly declared as unknown. This formalizes contamination as a benchmark variable rather than a post hoc caveat.

8 System / Model Overview

Figure 1 presents the benchmark stack. The benchmark is not just a dataset collection; it is a layered evaluation system. Model governance defines the admissible model interfaces and adaptation affordances. Dataset governance records provenance, temporal coverage, entity structure, schema, and overlap risk. Shift generation creates deterministic target domains under explicit severity ladders. Evaluation tracks define strict zero-shot and budgeted few-shot protocols. Metric tensors collect accuracy, calibration, robustness, elasticity, and efficiency. The reporting layer governs what must be released for interpretation and reproducibility.

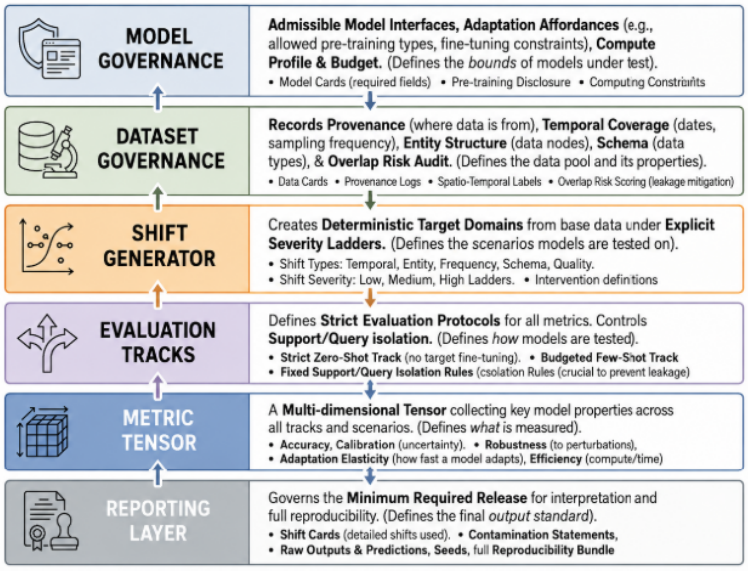

Figure 1. Leakage-aware benchmark architecture for time-series foundation models under domain shift.

The end-to-end benchmark pipeline is shown in Figure 2. The logic is deliberately conservative. The goal is not to maximize convenience but to prevent false claims of transfer. If a model cannot be evaluated fairly under declared restrictions, the benchmark should expose that limitation rather than smooth it away.

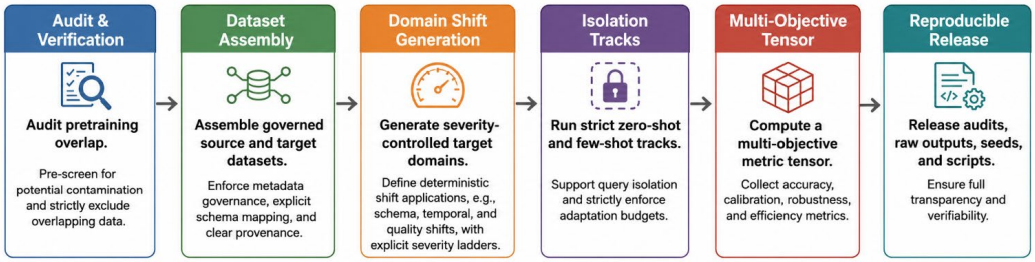

Figure 2. End-to-end benchmark execution pipeline.

8.1 Core Module 1

Core Module 1 is benchmark governance. This module formalizes the metadata and disclosure structure required before any score is interpreted. For each model, the benchmark records: pretraining objective; coarse-grained training corpus composition if known; supported modalities; context and horizon limits; probabilistic output type; adaptation operations allowed by the implementation; open versus closed status; and compute footprint. For each dataset, the benchmark records provenance, timestamp span, frequency, entity identifiers, schema definition, missingness profile, exogenous-variable policy, and contamination-risk annotations.

Governance is not administrative overhead. It directly affects scientific meaning. A closed model trained on opaque corpora and evaluated on a public dataset can only support bounded claims. Likewise, a dataset with unknown overlap to major open corpora cannot serve as clean evidence of zero-shot generalization.

8.2 Core Module 2

Core Module 2 is shift generation. We define a shift constructor

$$\Delta = \Gamma(\tau, \rho, \sigma, \eta, \kappa) \tag{10}$$

where τ indexes temporal regime separation, ρ indexes entity or population change, σ indexes schema/resolution change, η indexes observation-quality degradation, and κ indexes intervention or label-formation change. A target-domain generator transforms a governed base dataset into severity-indexed domains

$$\mathcal{D}_{\text{tgt}}^{(s)} = g_s(\mathcal{D}_0; \Delta_s), \qquad s \in \{1, \dots, S\}. \tag{11}$$

In forecasting, this module must respect temporal integrity. Random splitting is generally unacceptable whenever the scientific claim concerns deployment robustness. Instead, target domains are created by temporal isolation, entity isolation, schema perturbation, controlled corruption, or known intervention boundaries. Severity ladders should be deterministic and reproducible. For example, observation-quality severity can be encoded through increasing missingness and noise budgets; temporal severity can be encoded by increasing distributional distance from the source window; and schema severity can be encoded by progressive channel dropout or covariate removal.

8.3 Fusion / Integration Layer

The fusion layer combines results across benchmark axes without collapsing them prematurely. Let a benchmark outcome tensor be indexed by model m, dataset d, shift family s, severity level ℓ, horizon h, adaptation budget k, and metric u:

$$\mathcal{T}[m, d, s, \ell, h, k, u] \tag{12}$$

Rather than reducing this tensor to one leaderboard scalar, we compute summary views preserving heterogeneity:

$$\begin{aligned} \bar{\mathcal{T}}_{m,u} &= \mathbb{E}_{d,s,\ell,h,k}\big[\mathcal{T}[m, d, s, \ell, h, k, u]\big], \\ \mathcal{T}_{m,u}^{\text{worst}} &= \max_{d,s,\ell,h,k} \mathcal{T}[m, d, s, \ell, h, k, u], \\ \mathbb{V}_{m,u} &= \text{Var}_{d,s,\ell,h,k}(\mathcal{T}[m, d, s, \ell, h, k, u]). \end{aligned} \tag{13}$$

For error metrics, the worst-case statistic is a maximum; for skill or coverage statistics, the appropriate extremum is reversed. The central principle is that mean performance alone is insufficient.

9 Prediction / Decision Layer

9.1 Uncertainty / Reliability Module

The decision layer translates metric tensors into benchmark conclusions. We do not advocate a single total ordering. Instead, models occupy a Pareto region in metric space. A model can be attractive if it is not dominated on accuracy, robustness slope, adaptation elasticity, calibration, and efficiency simultaneously. In practical selection, users may impose weights according to deployment constraints, but those weights should be external to the benchmark rather than baked into a hidden scalar.

For point forecasts, the benchmark should report MAE, RMSE, sMAPE, and MASE [21], [22]. For probabilistic forecasts, it should report quantile loss, CRPS, empirical coverage, and calibration error [33], [34]. For deployment relevance, it should report latency, adaptation time, parameter count, and memory footprint. For shift-aware reasoning, it should report worst-domain summaries and degradation slopes.

Many TSFM papers emphasize point accuracy while leaving reliability underdeveloped. This is insufficient because a forecast can be numerically competitive yet operationally unsafe if its uncertainty estimates are miscalibrated. We therefore treat uncertainty and reliability as first-class benchmark objects. Let $\hat{F}_\theta$ denote the predictive distribution and let $I_\alpha(\mathbf{x})$ denote an $(1-\alpha)$ interval. Reliability is evaluated through calibration error and empirical coverage:

$$\text{Cov}_\alpha = \frac{1}{Nh} \sum_{i=1}^{N} \sum_{r=1}^{h} \mathbb{I}\left(y_{i,r} \in I_\alpha(\mathbf{x}_i)\right), \tag{14}$$

$$\text{CE}_\alpha = |\text{Cov}_\alpha - (1-\alpha)|. \tag{15}$$

We also advocate width-aware interpretation because perfect coverage can be achieved by trivially broad intervals. Hence calibration must be read together with average interval width, sharpness, or CRPS [33], [34]. In the

pilot study, we use split conformal intervals as a lightweight, distribution-free reliability probe, not as a claim that conformalization resolves TSFM uncertainty comprehensively [35].

10 Experimental Setup

10.1 Dataset

The paper's primary contribution is conceptual; nevertheless, a benchmark doctrine without executable grounding is vulnerable to remaining purely rhetorical. We therefore provide a small pilot instantiation using the public Electricity Transformer Temperature (ETT) family introduced in the long-sequence forecasting literature [24]. The pilot uses the hourly ETTh1 and ETTh2 subsets, each comprising 17,420 timestamped observations and seven observed channels (high-useful load, high-useless load, middle-useful load, middle-useless load, low-useful load, low-useless load, and oil temperature "OT"). OT is treated as the forecasting target, following common practice in univariate long-horizon studies [24], [27].

Table 1: Pilot dataset summary

Dataset	Frequency	Rows	Variables	Target	Role
ETTh1	1 hour	17,420	7	OT	Source / in-domain reference
ETTh2	1 hour	17,420	7	OT	Target / shifted domain

This pilot is intentionally modest. It is not presented as a substitute for a full multi-domain TSFM benchmark. Its narrower purpose is to verify that the proposed benchmark machinery can express (i) in-domain versus cross-domain transfer, (ii) zero-shot versus few-shot adaptation, (iii) degradation under observation-quality shift, and (iv) reliability analysis beyond point error.

10.2 Data Preprocessing

We use the standard ETT temporal split convention. The first 8,640 hourly observations for training, the next 2,880 for validation and calibration, and the next 2,880 for testing. Forecasting windows are generated with context length c=96 and horizon h=24, using a stride of four timestamps to reduce redundancy while preserving enough evaluation diversity.

1. **OT-only:** the model receives only past oil temperature values.
2. **Multivariate:** the model receives all seven observed channels over the context window.

No future covariates are exposed. For corrupted-target robustness tests, the query-set context windows are perturbed with controlled missingness and Gaussian noise, followed by simple causal forward filling. These corruption ladders are not intended to mimic one specific industrial telemetry stack perfectly; they operationalize the benchmark principle that data quality degradation should be a first-class test condition rather than an afterthought.

11 Experimental Design

The source domain is ETTh1 training data. The target domain is ETTh2. Zero-shot evaluation trains the model on ETTh1 only and evaluates directly on ETTh2 test windows. Few-shot evaluation augments source training with a small target-domain support set sampled from ETTh2 training windows. Budgets are $k \in \{5,20,100\}$. Query windows remain disjoint from support windows. All support sampling is deterministic under fixed random seeds.

To make the pilot epistemically honest, we do *not* report heavy TSFM scores that were not executed in this package. Instead, we use lightweight surrogates to stress-test the benchmark logic itself. This permits real numerical analysis without fabricating foundation-model behavior. The benchmark question in the pilot is therefore not "which TSFM is best?" but "does the benchmark protocol expose meaningful distinctions among transfer, adaptation, corruption robustness, and calibration?"

11.1 Baseline Models

The pilot includes four computationally light models.

1. **LastValue:** repeats the most recent observed value for the full horizon.
2. **SeasonalNaive24:** repeats the most recent seasonal cycle of length 24.
3. **Ridge-OT:** multi-output ridge regression using only OT lags.

4. **Ridge-Multi:** multi-output ridge regression using all seven channels over the context window.

These are not TSFMs, and we do not present them as such. They function as instrumentation models. Their value lies in showing that a benchmark can reveal cross-domain failure, adaptation elasticity, efficiency trade-offs, and calibration deficits even before a full TSFM zoo is plugged into the same infrastructure.

Table 2: Benchmark tracks and baseline rationale in the pilot

Model	Track(s)	Rationale
LastValue	Zero-shot	Minimal non-transfer baseline; tests whether the target remains locally persistent
SeasonalNaive24	Zero-shot	Strong classical hourly baseline; checks whether transfer claims exceed seasonal repetition
Ridge-OT	Zero-/few-shot	Lightweight transferable surrogate using only the target channel
Ridge-Multi	Zero-/few-shot	Tests whether extra channels help or hurt under domain shift

11.2 Hyperparameter Settings

The ridge models use L2 regularization with alpha=1.0. Context length sensitivity is evaluated at c in {48, 96, 192}. The benchmark budgets are k in {0, 5, 20, 100}, where k=0 denotes strict zero-shot evaluation. Corruption severities are encoded as (missing rate, noise scale) in {(0, 0), (0.05, 0.02), (0.10, 0.05), (0.20, 0.10)}.

Table 3: Pilot hyperparameter settings

Setting	Value
Target variable	OT
Forecast horizon h	24
Default context length c	96
Window stride	4
Ridge penalty α	1.0
Few-shot budgets k	5, 20, 100
Calibration level	90%
Corruption severities	clean, mild, moderate, severe

11.3 Implementation Details

The pilot was executed in Python using pandas, numpy, and scikit-learn. Ridge models were implemented as standardized multi-output linear regressors. The support-set sampling seed was fixed for reproducibility, and all reported few-shot values were averaged across repeated support draws. The package includes a notebook scaffold that reproduces the pilot and also exposes clear insertion points for real TSFM adapters once the necessary model weights and runtime environment are available.

12 Results and Analysis

12.1 Main Performance Comparison

Table 4 reports the primary pilot comparison on the ETTh1 to ETTh2 transfer setting. Three findings are immediate. First, the seasonal baseline is already substantially stronger than the last-value baseline, confirming that any transfer narrative must beat competent classical references rather than trivial baselines. Second, Ridge-OT substantially outperforms both naive baselines in strict zero-shot transfer, showing that cross-domain structure exists and can be exploited. Third, Ridge-Multi performs far worse than Ridge-OT in zero-shot mode, illustrating a central point of this paper: adding representation capacity or cross-channel input does not guarantee better transfer under domain shift.

Table 4: Main pilot results on ETTh1 to ETTh2

Model	Track	k	MAE	RMSE	sMAPE	MASE
LastValue	Zero-shot	0	4.1163	5.5308	34.5538	1.3548
SeasonalNaive24	Zero-shot	0	2.6699	3.5531	25.7364	0.8788
Ridge-OT	Zero-shot	0	2.3675	3.1284	21.5999	0.7792

Model	Track	k	MAE	RMSE	sMAPE	MASE
Ridge-OT	Few-shot	5	2.3615	3.1218	21.5630	0.7773
Ridge-OT	Few-shot	20	2.3468	3.1071	21.4348	0.7724
Ridge-OT	Few-shot	100	2.2965	3.0492	21.0911	0.7559
Ridge-Multi	Zero-shot	0	5.4272	6.9569	44.6656	1.7863
Ridge-Multi	Few-shot	5	5.4280	6.9713	44.1489	1.7865
Ridge-Multi	Few-shot	20	5.1225	6.5878	43.2642	1.6860
Ridge-Multi	Few-shot	100	4.2723	5.4593	37.3303	1.4062

The few-shot results are also informative. Ridge-OT improves only modestly from k=0 to k=100, whereas Ridge-Multi remains poor but improves more noticeably with larger support. This is exactly the kind of evidence that a proper benchmark should surface, a model with weaker zero-shot performance may exhibit higher adaptation elasticity, while a model with strong zero-shot performance may show diminishing returns from additional support. Reporting only the best post-adaptation score would hide that distinction.

Figure 3 visualizes this elasticity. The two trajectories tell different stories about transferable inductive bias versus domain-correction burden.

12.2 Dataset-Specific Analysis

Table 5 compares in-domain and cross-domain performance. The shift from the in-domain ETTh1 to ETTh1 setting to the cross-domain ETTh1 to ETTh2 setting degrades all models, but the magnitude differs sharply. Ridge-OT moves from RMSE 1.5794 in-domain to 3.1284 cross-domain. Ridge-Multi degrades from 2.2616 to 6.9569. This asymmetry is revealing. The multivariate model does not fail because multivariate modeling is intrinsically bad; it fails because the extra channels are not stably aligned across source and target domains under this transfer protocol. This is a concrete example of why benchmark papers must report schema-related and cross-domain behavior separately rather than compressing them into one average.

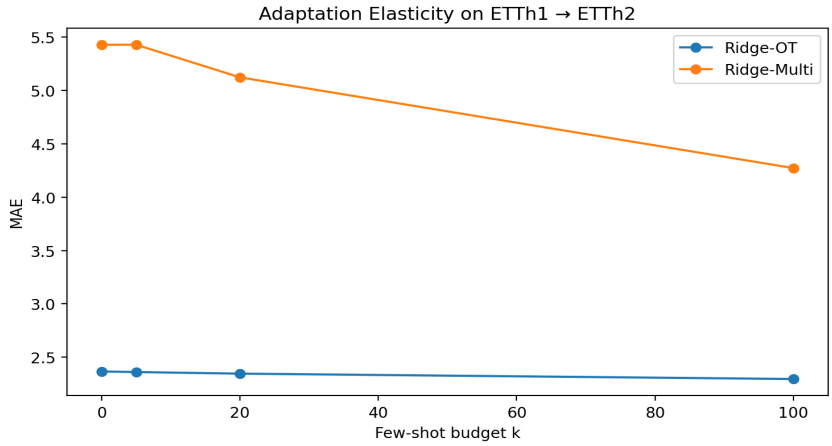

Figure 3. Few-shot adaptation elasticity on ETTh1 to ETTh2.

Table 5: In-domain versus cross-domain pilot comparison

Setting	Model	MAE	RMSE	sMAPE	MASE
ETTh1 to ETTh1	LastValue	1.2675	1.6873	38.3357	0.4964
ETTh1 to ETTh1	SeasonalNaive24	1.5275	1.9654	45.3427	0.5982
ETTh1 to ETTh1	Ridge-OT	1.1829	1.5794	36.5080	0.4633
ETTh1 to ETTh1	Ridge-Multi	1.7567	2.2616	51.2888	0.6880
ETTh1 to ETTh2	LastValue	4.1163	5.5308	34.5538	1.3548
ETTh1 to ETTh2	SeasonalNaive24	2.6699	3.5531	25.7364	0.8788
ETTh1 to ETTh2	Ridge-OT	2.3675	3.1284	21.5999	0.7792
ETTh1 to ETTh2	Ridge-Multi	5.4272	6.9569	44.6656	1.7863

Figure 4 makes the same point visually. In-domain ranking is not sufficient proxy for cross-domain viability, and cross-domain viability itself depends on the representation assumptions exposed by the shift.

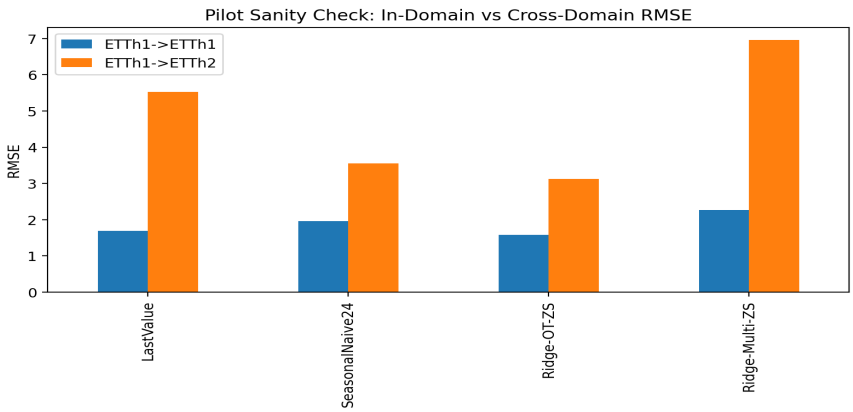

Figure 4. In-domain versus cross-domain RMSE in the pilot.

12.3 Computational Efficiency and Resource Utilization

Benchmarking transfer without reporting efficiency is incomplete. Table 6 shows that the two ridge variants differ not only in error but also in fit cost and parameter footprint. Ridge-Multi uses roughly seven times the parameter count of Ridge-OT and incurs a much longer fit time, while offering substantially worse cross-domain zero-shot performance. This illustrates why adaptation and model complexity must be interpreted jointly. A model that can recover with large few-shot support may still be unattractive if its computing burden is disproportionate.

Table 5: Pilot efficiency indicators

Model	Train time (s)	Latency (ms/window)	Parameters
Ridge-OT-ZS	0.1513	0.0122	2,328
Ridge-Multi-ZS	3.8085	0.0130	16,152
Ridge-OT-FS100	0.1895	0.0056	2,328

For full TSFM studies, this section becomes even more important. Parameter count, adaptation memory, context-length limits, and inference throughput are not mere deployment footnotes. They are part of the scientific meaning of few-shot improvement.

12.4 Scalability and Deployment Feasibility

The pilot models are deliberately lightweight, so their deployment feasibility is unsurprising. The broader lesson lies in the benchmark doctrine. A scalable TSFM benchmark must support evaluation over multiple horizons, repeated seeds, multiple shift families, and both zero-shot and few-shot tracks without forcing bespoke engineering for every model. This requirement aligns with recent benchmark infrastructures such as FoundTS, GIFT-Eval, and fev-bench, which recognize that evaluation reproducibility requires standardized execution logic [18], [19], [20].

12.5 Sensitivity Analysis

Context length sensitivity is summarized in Table 7. The best pilot performance occurs at c=96 rather than at the shortest or longest context. This matters for two reasons. First, benchmark conclusions can depend on context length; therefore the benchmark should either fix context carefully or report sensitivity. Second, larger context does not automatically improve transfer. A model may benefit from more history only if the additional past is both informative and distributionally stable under the target shift.

Table 5: Sensitivity to context length in the pilot

Context length	Track	MAE	RMSE	sMAPE
48	Zero-shot	2.5007	3.3219	22.5881
48	Few-shot (k=100)	2.4005	3.2033	21.9081
96	Zero-shot	2.3675	3.1284	21.5999
96	Few-shot (k=100)	2.2886	3.0468	21.0045
192	Zero-shot	2.4314	3.1832	22.2955
192	Few-shot (k=100)	2.3618	3.1107	21.7560

In a full TSFM benchmark, context sensitivity should be examined jointly with compute, memory, and horizon. Long contexts may look attractive in accuracy tables but become impractical or unstable under real deployment budgets.

12.6 Ablation Study

The pilot supports one immediately meaningful ablation OT-only versus multivariate transfer. The results show that additional input channels hurt zero-shot transfer severely under ETTh1 to ETTh2, and remain harmful even after adaptation unless the support budget becomes relatively large. This is not evidence against multivariate forecasting in general. It is evidence that cross-channel relationships can shift across domains, and benchmarks should measure that rather than assume multivariate inputs are universally beneficial.

For future TSFM studies, the same logic implies several mandatory ablations with and without target normalization recalibration, with and without adapters, with and without covariates, different prompt or patch strategies, and different context/horizon settings.

12.7 Interpretability and Explainability Analysis

Interpretability is not optional when a benchmark is trying to explain why transfer succeeds or fails. Figure 5 aggregates absolute coefficients across channels for Ridge-Multi. OT has the largest overall weight, but several load-related channels also matter. Yet those additional channels do not help cross-domain zero-shot performance, suggesting that the model's learned cross-channel mapping is not transportable from ETTh1 to ETTh2 under the chosen protocol.

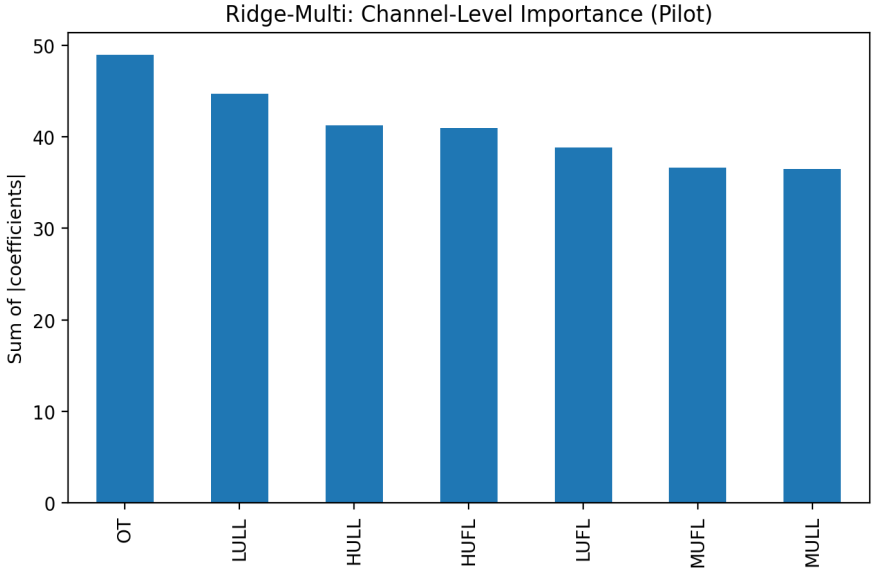

Figure 5. Channel-level importance for the multivariate ridge surrogate.

For TSFMs, interpretability may involve attention maps, patch saliency, gradient-based attribution, retrieval traces, or feature-ablation profiles. The benchmark should not assume one universal interpretability tool. It should instead require at least one mechanism for diagnosing failure under shift.

12.8 Uncertainty Calibration and Reliability Analysis

Reliability results are given in Table 8. Zero-shot conformal intervals calibrated on the source validation split achieve only 75.55% empirical coverage on the shifted target, well below the nominal 90%. After few-shot adaptation and target-side calibration, empirical coverage rises to 94.66%, but interval width increases sharply. This is precisely the kind of trade-off that average point-error reporting misses. A benchmark that ignores calibration could easily conclude that the two models differ only modestly in MAE, while operationally their uncertainty behavior is much more consequential.

Table 6: Pilot uncertainty and reliability analysis

Model	Coverage@90%	Mean interval width	Calibration status
Ridge-OT-ZS	0.7555	6.9455	under-covered on shifted target

Model	Coverage@90%	Mean interval width	Calibration status
Ridge-OT-FS100	0.9466	12.1469	over-conservative but reliable

In full TSFM evaluations, calibration must be analyzed together with distributional assumptions, quantile consistency, and cost of recalibration. A model with strong point accuracy but poor coverage may be inadequate for risk-sensitive applications.

12.9 Cross-Dataset / Cross-Region Evaluation

The pilot uses ETTh1 and ETTh2 as source and target domains, which gives a small but real cross-dataset evaluation. The main finding is methodological rather than domain-specific benchmark outcomes can change sharply when the evaluation moves from in-domain to cross-domain conditions, even when the datasets appear superficially similar. This directly supports the paper's central argument that benchmark design must index results by domain relation and shift profile rather than treating all datasets as exchangeable items on a leaderboard.

A full TSFM study should extend this logic to cross-region, cross-frequency, and cross-industry settings. Energy, traffic, healthcare, observability, and retail time series differ not only in signal shape but also in entity structure, intervention exposure, label semantics, and measurement quality. Those differences should be benchmark factors rather than anecdotal descriptions.

12.10 Generalization / Transfer / Cold-Start Analysis

The few-shot curves already provide a cold-start perspective. Ridge-OT is relatively stable in cold-start transfer; it begins strong in zero-shot mode and improves gradually. Ridge-Multi begins weak and needs much larger support to recover. Benchmark doctrine should therefore distinguish at least three profiles:

1. **Strong zero-shot / low elasticity:** models that transfer well immediately but improve little with support.
2. **Weak zero-shot / high elasticity:** models that need support but can recover strongly.
3. **Weak zero-shot / weak elasticity:** models that are poor choices under cold-start constraints.

These profiles are more actionable than a single average score because they align with actual deployment scenarios.

12.11 Additional Scientific Analysis

Figure 6 reports robustness under controlled observation-quality degradation. Both the zero-shot and few-shot OT models degrade as missingness and noise increase, but the few-shot model remains consistently better. The error slope itself is informative:

$$\beta_{\text{rob}} = \frac{\Delta \text{RMSE}}{\Delta \text{severity}} \tag{17}$$

A benchmark that reports only the clean-domain score cannot tell whether a model is brittle or resilient under realistic telemetry degradation.

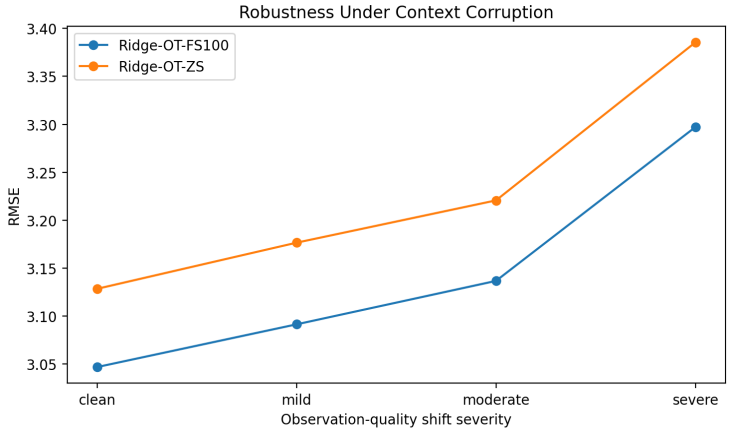

Figure 6. Severity-conditioned robustness under context corruption.

This kind of robustness analysis should generalize naturally to other shift families: exogenous covariate removal, delayed reporting, channel dropout, calendar perturbation, or post-intervention regimes. The key scientific point is that robustness should be measured as a curve or ladder, not as an optional one-off stress test.

13 Discussion

The pilot results should not be overclaimed. They do not demonstrate the superiority or inferiority of TSFMs, because no TSFM leaderboard was executed in this package. What they do demonstrate is that the benchmark doctrine is analytically productive. The pilot exposes at least six distinctions that conventional average-error reporting would hide, in-domain versus cross-domain divergence, zero-shot versus few-shot elasticity, efficiency trade-offs, context-length sensitivity, reliability deficits, and corruption robustness.

This yields several broader implications. First, benchmark papers should shift from score collection to diagnostic explanation. A benchmark should reveal which models fail under which shifts and why. Second, contamination and overlap should be treated as explicit benchmark variables. A score without information-set disclosure is not neutral; it is underspecified. Third, adaptation budgets must be standardized and disclosed with the same seriousness as hyperparameters in conventional supervised learning. Fourth, multi-objective reporting is necessary because point accuracy, calibration, efficiency, and robustness do not reliably co-vary.

The limitations are also clear. The pilot uses only one public dataset family, one target variable, one horizon, and lightweight surrogate models. It therefore validates the benchmark machinery only on a small scale. A complete TSFM study must evaluate multiple public and operational domains, open and closed TSFMs, probabilistic outputs, frequency shifts, and richer adaptation policies. That larger empirical program is precisely what this benchmark doctrine is designed to enable.

14 Conclusion

This paper argued that the decisive scientific question for time-series foundation models is not whether they can produce good average forecasts on a reused battery of datasets, but whether their claimed transfer ability survives principled, leakage-aware evaluation under domain shift. Existing literature has made substantial progress in model design and benchmark infrastructure, yet methodological gaps remain around contamination control, adaptation-budget standardization, shift-explicit analysis, and multi-objective reporting [16], [17], [18], [20].

In response, we proposed a benchmark doctrine built around five ideas: formalization of zero-shot and few-shot risk under shifted target distributions; a forecasting-specific taxonomy of domain shift; a six-layer benchmark architecture connecting governance, shift generation, evaluation tracks, metric tensors, and reporting; a vector-valued evaluation perspective that rejects single-number leaderboards; and an executable pilot instantiation demonstrating the practical utility of the framework. The pilot was intentionally modest and did not claim new TSFM performance results. Its purpose was to show that a rigorous benchmark can produce interpretable diagnostics rather than merely ranking models.

The main practical implication is straightforward. Future TSFM papers should not present broad transfer claims without explicit declarations of pretraining exposure, governed support/query construction, severity-aware analysis, worst-domain reporting, and reliability measurement beyond point error. The main limitation is equally straightforward: the field still needs large-scale empirical implementation of this doctrine using real TSFM model zoos and broader datasets. That is the next step, and the package accompanying this paper is designed to make that step technically straightforward and scientifically harder to evade.

BIBLIOGRAPHY

[1]. R. Bommasani et al., “On the opportunities and risks of foundation models,” arXiv preprint arXiv:2108.07258, 2021, doi: 10.48550/arXiv.2108.07258.

[2]. Y. Liang 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, pp. 6555–6565. doi: 10.1145/3637528.3671451.

[3]. Q. Ma, J. Liu, et al., “A survey on time-series pre-trained models,” IEEE Transactions on Knowledge and Data Engineering, vol. 36, pp. 7536–7555, 2024, doi: 10.1109/TKDE.2024.3475809.

[4]. S. R. K. Kottapalli et al., “Foundation models for time series: A survey,” arXiv preprint arXiv:2504.04011, 2025, doi: 10.48550/arXiv.2504.04011.

[5]. B. Lim and S. Zohren, “Time-series forecasting with deep learning: A survey,” Philosophical Transactions

of the Royal Society A, vol. 379, no. 2194, p. 20200209, 2021, doi: 10.1098/rsta.2020.0209.
[6]. B. Cohen et al., "This time is different: An observability perspective on time series foundation models," arXiv preprint arXiv:2505.14766, 2025, doi: 10.48550/arXiv.2505.14766.
[7]. G. Pucher et al., "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, vol. 329, pp. 820–824, 2025, doi: 10.3233/SHTI250954.
[8]. A. Das, W. Kong, R. Sen, and Y. Zhou, "A decoder-only foundation model for time-series forecasting," arXiv preprint arXiv:2310.10688, 2024, doi: 10.48550/arXiv.2310.10688.
[9]. A. F. Ansari et al., "Chronos: Learning the language of time series," arXiv preprint arXiv:2403.07815, 2024, doi: 10.48550/arXiv.2403.07815.
[10]. M. Goswami et al., "MOMENT: A family of open time-series foundation models," arXiv preprint arXiv:2402.03885, 2024, doi: 10.48550/arXiv.2402.03885.
[11]. G. Woo, C. Liu, A. Kumar, C. Xiong, S. Savarese, and D. Sahoo, "Unified training of universal time series forecasting transformers," in Proceedings of the 41st international conference on machine learning, 2024. doi: 10.48550/arXiv.2402.02592.
[12]. X. Liu, G. Woo, T. Aksu, C. Liu, S. Savarese, and D. Sahoo, "Moirai-MoE: Empowering time series foundation models with sparse mixture of experts," arXiv preprint arXiv:2410.10469, 2024, doi: 10.48550/arXiv.2410.10469.
[13]. K. Rasul et al., "Lag-llama: Towards foundation models for probabilistic time series forecasting," arXiv preprint arXiv:2310.08278, 2024, doi: 10.48550/arXiv.2310.08278.
[14]. V. Ekambaram et al., "Tiny time mixers (TTMs): Fast pre-trained models for enhanced zero/few-shot forecasting of multivariate time series," arXiv preprint arXiv:2401.03955, 2024, doi: 10.48550/arXiv.2401.03955.
[15]. H. Hewamalage, K. Ackermann, and C. Bergmeir, "Forecast evaluation for data scientists: Common pitfalls and best practices," Data Mining and Knowledge Discovery, vol. 37, no. 2, pp. 788–832, 2023, doi: 10.1007/s10618-022-00894-5.
[16]. Z. Li et al., "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, 2025, pp. 5595–5606. doi: 10.1145/3711896.3737442.
[17]. M. Meyer, S. Kaltenpoth, K. Zalipski, and O. Müller, "Time series foundation models: Benchmarking challenges and requirements," arXiv preprint arXiv:2510.13654, 2025, doi: 10.48550/arXiv.2510.13654.
[18]. T. Aksu et al., "GIFT-eval: A benchmark for general time series forecasting model evaluation," arXiv preprint arXiv:2410.10393, 2024, doi: 10.48550/arXiv.2410.10393.
[19]. O. Shchur et al., "Fev-bench: A realistic benchmark for time series forecasting," arXiv preprint arXiv:2509.26468, 2025, doi: 10.48550/arXiv.2509.26468.
[20]. Z. Li et al., "FoundTS: Comprehensive and unified benchmarking of foundation models for time series forecasting," in arXiv preprint arXiv:2410.11802, 2024. doi: 10.48550/arXiv.2410.11802.
[21]. R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," International Journal of Forecasting, vol. 22, no. 4, pp. 679–688, 2006, doi: 10.1016/j.ijforecast.2006.03.001.
[22]. S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "The M4 competition: 100,000 time series and 61 forecasting methods," International Journal of Forecasting, vol. 36, no. 1, pp. 54–74, 2020, doi: 10.1016/j.ijforecast.2019.04.014.
[23]. X. Song, L. Deng, H. Wang, et al., "Deep learning-based time series forecasting," Artificial Intelligence Review, vol. 58, p. 23, 2025, doi: 10.1007/s10462-024-10989-8.
[24]. H. Zhou et al., "Informer: Beyond efficient transformer for long sequence time-series forecasting," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 12, pp. 11106–11115, 2021, doi: 10.1609/aaai.v35i12.17325.
[25]. H. Wu, J. Xu, J. Wang, and M. Long, "Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting," in Advances in neural information processing systems, 2021, pp. 22419–22430. Available: https://proceedings.neurips.cc/paper_files/paper/2021/hash/bcc0d400288793e8bdcd7c19a8ac0c2b-Abstract.html
[26]. T. Zhou, Z. Ma, Q. Wen, X. Wang, L. Sun, and R. Jin, "FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting," in Proceedings of the 39th international conference on machine learning, 2022, pp. 27268–27286. Available: https://proceedings.mlr.press/v162/zhou22g.html
[27]. Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, "A time series is worth 64 words: Long-term

forecasting with transformers," in International conference on learning representations, 2023. doi: 10.48550/arXiv.2211.14730.
[28]. M. Jin, S. Wang, L. Ma, Z. Chu, J. Y. Zhang, et al., "Time-LLM: Time series forecasting by reprogramming large language models," arXiv preprint arXiv:2310.01728, 2023, doi: 10.48550/arXiv.2310.01728.
[29]. N. Karaouli, D. Coquenet, E. Fromont, M. Mermillod, and M. Reyboz, "How foundational are foundation models for time series forecasting?" arXiv preprint arXiv:2510.00742, 2025, doi: 10.48550/arXiv.2510.00742.
[30]. J. Zhang, Z. Zhang, S. Zheng, X. Wen, J. Li, and J. Bian, "Are time-series foundation models deployment-ready? A systematic study of adversarial robustness across domains," arXiv preprint arXiv:2505.19397, 2025, doi: 10.48550/arXiv.2505.19397.
[31]. P. W. Koh et al., "WILDS: A benchmark of in-the-wild distribution shifts," in Proceedings of the 38th international conference on machine learning, 2021, pp. 5637–5664. Available: https://proceedings.mlr.press/v139/koh21a.html
[32]. R. Bommasani, P. Liang, and T. Lee, "Holistic evaluation of language models," Annals of the New York Academy of Sciences, vol. 1525, no. 1, pp. 140–146, 2023, doi: 10.1111/nyas.15007.
[33]. T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," Journal of the American Statistical Association, vol. 102, no. 477, pp. 359–378, 2007, doi: 10.1198/016214506000001437.
[34]. T. Gneiting, F. Balabdaoui, and A. E. Raftery, "Probabilistic forecasts, calibration and sharpness," Journal of the Royal Statistical Society: Series B, vol. 69, no. 2, pp. 243–268, 2007, doi: 10.1111/j.1467-9868.2007.00587.x.
[35]. A. N. Angelopoulos and S. Bates, "A gentle introduction to conformal prediction and distribution-free uncertainty quantification," Annual Review of Statistics and Its Application, vol. 10, pp. 1–28, 2023, doi: 10.1146/annurev-statistics-033121-015250.