# ARE UNIVERSAL TIME SERIES FOUNDATION MODELS BASED ON A CATEGORY ERROR?

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Time series foundation models (TSFMs) achieve strong forecasts across diverse domains with little or no task-specific training. Yet what makes this transfer work, and why can adding data or longer histories sometimes fail to help? We investigate a possible *category error*: treating a shared numerical format as sufficient grounds for shared predictive structure and information. Time series are a container for observations from different processes; the same history can imply different futures when their generating rules remain unresolved. Under squared loss, averaging those rules suppresses conflicting components while retaining common structure, whereas informative context can make the rules distinguishable. We develop this concern into four questions about task symmetries, rule information, source mixing, and lightweight reuse of forecasting ability. Controlled interventions verify predicted costs of invalid sharing and show when additional information restores suppressed components. Small-scale pretraining establishes that the value of auxiliary data depends on its source and allocation. An exploratory distillation study finds that a tiny, history-conditioned operator model acquires useful predictive ability from few TSFM examples, with better performance when its training includes the target source. These findings give concrete grounds for reconsidering the assumptions behind universal forecasting and motivate models, losses, training procedures, and data construction that account for what can be shared and what must be distinguished.

## 1 INTRODUCTION

The success of broadly pretrained vision models (Dosovitskiy et al., 2021) motivates a compelling program for time series: aggregate observations across applications, train a shared predictor, and transfer its structure with little additional supervision. Time series foundation models (TSFMs) now forecast across sampling frequencies, scales, dimensions, and applications (Das et al., 2024; Ansari et al., 2025; Liu et al., 2026; 2025; Kottapalli et al., 2025). Their success raises a deeper question: what is shared when these observations are all called *time series*?

A time-indexed array is a data container; it specifies neither the generating mechanism, the variables determining its continuation, nor the meaning of an error. A spike can mark a promotion or a mechanical fault; a zero trace can indicate sensor disconnection or store closure. Longer histories or covariates may reveal these distinctions, or they may be absent from the input. Inferring a common, sufficiently observed forecasting problem from this representation is the *category error* we examine. Language and vision transfer rests on task regularities; seasonality, smoothness, and shared dynamics likewise support time-series transfer where those regularities hold. Shared architectures can represent heterogeneous collections as expressively as local ones (Montero-Manso & Hyndman, 2021), and modern TSFMs incorporate covariates and related series (Ansari et al., 2025). Such models can infer regimes or preserve source-specific behavior, provided the input and learning procedure make the distinctions usable. Parameter sharing, distributional compatibility, and predictive sufficiency are different properties.

The concern about *generic filters* follows from a particular information regime. When indistinguishable histories have opposing futures, squared-loss prediction averages the unresolved rules, suppressing their disputed component while retaining what they share. Source cues and historical evidence can change this balance. This motivates an exploratory question about reusable predic-



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

tion structure: how much forecasting ability can a tiny, history-conditioned combination of simple operators acquire from a pretrained TSFM? Training on a small set of teacher forecasts makes the required learned scale and sample budget part of the question.

Reliability matters in open, changing systems (Ferdaus et al., 2026), where useful sharing depends on both observations and the prediction task. We investigate the category-error concern through four questions: **(A)** which transformations preserve the prediction task, **(B)** how unresolved rules affect forecasts and what information distinguishes them, **(C)** when additional training sources help a fixed target, and **(D)** how much forecasting ability can be transferred into a lightweight adaptive operator model. A and B isolate mechanisms; C tests conditions for useful sharing in natural data; D explores the complexity and transferability of reusable prediction structure. These questions give the category-error concern testable consequences, without presupposing an unconditional answer to the title.

Our contributions are:

- We provide a task-relative framework for assessing universality claims, distinguishing numerical form from predictive structure and sufficient information, and quantifying the cost of invalid sharing.
- We connect rule identifiability to the forecast components that survive sharing, with quantitative predictions for cue reliability and historical information; mixture-risk analysis explains how auxiliary-data value depends on the target and learning regime.
- We establish a practical distinction between the value of additional training sources and the complexity needed to reuse forecasting ability. Target-specific sharing gains and lightweight teacher-trained operators expose these two dimensions of universality, informing source allocation and compact reuse of learned structure.

## 2 Related work

### 2.1 Time series foundation models

TSFM research follows native-pretraining and language-model-adaptation routes (Kottapalli et al., 2025; Ye et al., 2027). Native systems span MOMENT's general representations, Moirai's unified forecasting, TimesFM's decoder-based forecasts (Goswami et al., 2024; Woo et al., 2024; Das et al., 2024), and compact pretrained mixers (Ekambaram et al., 2024). Chronos models quantized values, whereas Chronos-2 uses continuous patches and multivariate, covariate, and cross-series conditioning (Ansari et al., 2024; 2025); Moirai 2.0 and Sundial develop autoregressive and generative forecasting (Liu et al., 2026; 2025). LLMTime and Time-LLM adapt large language models (LLMs) through numerical serialization or temporal reprogramming (Gruver et al., 2023; Jin et al., 2024), while AutoTimes and UniTime support autoregressive and cross-domain forecasting (Liu et al., 2024b;a). Patch-vocabulary, spectral-space, and design-bias analyses examine transferable representations and forecasting behavior (Xie et al., 2025; Ma et al., 2026; Yu et al., 2026). Conditioning interfaces and domain instructions can expose distinctions absent from history-only inputs; our analysis asks whether the available distinctions supply sufficient predictive information.

### 2.2 Architectures and transferable structure in time series

Forecasting architectures incorporate sparse attention, decomposition, frequency modeling, and cross-variable interaction (Zhou et al., 2021; Wu et al., 2021; Zhou et al., 2022; Zhang & Yan, 2023), while strong linear baselines motivate scrutiny of architectural complexity and realized behavior (Zeng et al., 2023). The Time-series Dense Encoder (TiDE) and TimeMixer exploit dense and multiscale mixing (Das et al., 2023; Wang et al., 2024a); PatchTST uses patching and channel-independent sharing to improve longer-history forecasting (Nie et al., 2023). Shared structure also underlies group-equivariant and geometric learning (Cohen & Welling, 2016; Bronstein et al., 2021), and augmentation can reduce estimation variance under valid invariance (Chen et al., 2020). Our task-relative analysis examines which transformations preserve forecasting structure, while distillation tests transferable behavior through a restricted student (Hinton et al., 2015; Frosst & Hinton, 2017). A shared operator dictionary makes the represented behavior explicit.

### 2.3 DISTRIBUTION SHIFT AND PREDICTIVE INFORMATION IN TIME SERIES

Reversible instance normalization (RevIN) addresses instance-level location and scale changes, while Non-stationary Transformers retain distributional information in attention (Kim et al., 2022; Liu et al., 2022). Cross-source heterogeneity and training dynamics affect learned representations (Feng et al., 2026; Zhai et al., 2023; Syed et al., 2026); UniTime addresses domain identification and unequal convergence rates (Liu et al., 2024a). Reliability also depends on sensitivity to adversarial perturbations (Zhang et al., 2025a). Global forecasting can nevertheless match local expressiveness and benefit from heterogeneous data (Montero-Manso & Hyndman, 2021). Transfer theory considers source risk, distributional discrepancy, and shared predictor quality (Ben-David et al., 2010); invariant representations can fail through changed conditional structure or lost predictive information (Zhao et al., 2019; Johansson et al., 2019). Our mixture analysis separates these issues from finite learning, using classical information and forecast-decision principles (Blackwell, 1951; Gneiting, 2011; Gneiting & Raftery, 2007; Perez-Diaz et al., 2025). Cross-domain pretraining also produces source-dependent gains and negative transfer (Zhang et al., 2025b). Transferability estimators predict adaptation performance from model–dataset relationships (Yao et al., 2025). We examine source composition and allocation under matched exposure controls. External context can change the prediction problem (Potapczynski et al., 2026): intervention-aware forecasting and Context is Key supply information beyond numerical histories (Xu et al., 2025; Williams et al., 2025; Alipour et al., 2026), as can news-conditioned forecasts (Wang et al., 2024b). Our interventions distinguish the value of predictive context from its use by a learner.

## 3 FROM A SHARED CONTAINER TO A TASK-RELATIVE MODALITY

A time series specifies an indexing convention and a value space; a forecasting task also specifies a continuation law, information, and a loss. The *category error* is to infer shared task-preserving structure or sufficient information from that container alone. Question A makes this distinction relative to a task family.

**Definition 1** (Conditional-mean task symmetry)**.** *Let $X \in \mathcal{X}$ be a length-$L$ history and $Y \in \mathbb{R}^H$ its $H$-step future. Use expectation $\mathbb{E}$, probability $\Pr$, and Euclidean norm $\|\cdot\|$. Fix a group $G$ (identity $e$) acting on $\mathcal{X}$ and output representation $\rho$; $g_{\#}P$ is the law of $gX$ for $X \sim P$, and a.s. means almost surely. For a domain–task pair $\theta = (P_\theta, m_\theta)$, let $P_\theta$ denote the history law and $m_\theta(x) = \mathbb{E}_\theta[Y \mid X = x]$ its squared-loss prediction target. For a task family $S$, define*

$$\mathrm{Stab}(\theta) = \{g \in G : g_{\#}P_\theta = P_\theta,\ m_\theta(gx) = \rho(g)m_\theta(x)\ P_\theta\text{-a.s.}\}, \qquad G_S = \bigcap_{\theta \in S} \mathrm{Stab}(\theta). \quad (1)$$

*The family has shared conditional-mean task structure for this action when $G_S$ contains useful nonidentity transformations. This definition concerns the conditional mean, rather than invariance of the full conditional distribution.*

Both conditions matter: preserving history frequencies need not preserve continuation rules, and relating forecasting rules need not preserve the history law. A family can share an architecture or numerical representation without sharing every useful symmetry. The stabilizer specifies what an imposed invariance or equivariance must preserve.

**Proposition 1** (The shared-symmetry spectrum)**.** *For task families $S \subseteq T$, $G_T \subseteq G_S$. If every $g \neq e$ is excluded by at least one task stabilizer, then $G_S = \{e\}$. A faithful action and an unconstrained family containing every point-supported input law suffice for this separation.*

A finite related corpus may retain useful shared structure (Montero-Manso & Hyndman, 2021); a smaller shared group instead imposes fewer invariance constraints and can enlarge the admissible predictor class. To quantify the cost of imposing an invalid symmetry, augment each observation with equal probability by $(X, Y) \mapsto (gX, \rho(g)Y)$ and evaluate predictor $f$ under the original joint law $P$ (history marginal $P^X$) with risk $\mathcal{R}_P(f) = \mathbb{E}_P\|Y - f(X)\|^2/H$.

**Proposition 2** (Risk cost of an invalid shared symmetry)**.** *Let $g^2 = e$, let $\rho(g)$ be orthogonal, and assume $g_{\#}P^X = P^X$ and $\mathbb{E}\|Y\|^2 < \infty$. For $m(x) = \mathbb{E}[Y \mid X = x]$, define the task defect*

$$\delta_g = H^{-1}\mathbb{E}\|m(gX) - \rho(g)m(X)\|^2. \quad (2)$$

*The equal-augmentation Bayes predictor is* $m_g(x) = \frac{1}{2}[m(x) + \rho(g)m(gx)]$*, and its risk on the original task satisfies*

$$\mathcal{R}_P(m_g) - \mathcal{R}_P(m) = \delta_g/4. \tag{3}$$

A task-preserving transformation has zero defect; Section 5.2 tests zero- and positive-defect tasks under the same intervention. The proof and geometric extension are in Appendix A.1.

# 4 Information, mixture learning, and transferable prediction structure

## 4.1 Source information and Bayes conflict

Question B asks what happens when the available observations do not distinguish different continuation rules. Assume $Y$ has finite second moment; let $D$ denote the source or regime. Write $m_X = \mathbb{E}[Y \mid X]$, $m_{X,D} = \mathbb{E}[Y \mid X, D]$, and $\mathcal{R}(f) = \mathbb{E}\|Y - f\|^2/H$.

**Proposition 3** (Predictive value of source information)**.**

$$\mathcal{R}(m_X) - \mathcal{R}(m_{X,D}) = H^{-1}\mathbb{E}\|m_{X,D} - m_X\|^2 \equiv \Delta_D \geq 0. \tag{4}$$

*The gap is positive exactly when source identity changes the conditional mean on a set of positive probability.*

For finitely many sources indexed by $d, e$, write $p_d(x) = \Pr(D = d \mid X = x)$ and $m_d(x) = \mathbb{E}[Y \mid X = x, D = d]$. The gap has the pairwise form

$$\Delta_D = \frac{1}{2H}\mathbb{E}_X\left[\sum_{d,e} p_d(X)p_e(X)\|m_d(X) - m_e(X)\|^2\right]. \tag{5}$$

Each disagreement is weighted by the product of the two rules' posterior probabilities for the observed history. Sources with disjoint observable support contribute no conflict, even when their dynamics differ. Splitting one source into identical copies divides its posterior mass and leaves the gap unchanged. Conversely, increasing the mass of an opposing rule can increase conflict without introducing a new domain. Dataset count therefore cannot quantify this conflict; the relevant quantities are posterior overlap and continuation disagreement. The controlled source-code experiments vary whether the forecaster can resolve that overlap.

For contextual code $Z$, write $m_{X,Z} = \mathbb{E}[Y \mid X, Z]$. If $Z$ reveals only part of $D$, then, under $\mathbb{E}[Y \mid X, D, Z] = m_{X,D}$,

$$\Delta_D = H^{-1}\mathbb{E}\|m_{X,Z} - m_X\|^2 + H^{-1}\mathbb{E}\|m_{X,D} - m_{X,Z}\|^2. \tag{6}$$

The terms measure information recovered by $Z$ and unresolved after it. Codes help by separating predictive means; a conditionally independent random code recovers none of the gap, and informative codes require a learner that can use them. For deterministic finite codes, the residual is posterior variance of source means within code cells: extra values help only when they separate continuation means; domain names are neither necessary nor sufficient. Preprocessing incurs the same cost when it merges histories with different predictive means. Normalization can aid finite learning while discarding predictive level or scale; retaining those statistics changes the information set. Appendix A.2 gives both forms.

General context can also predict information beyond the source. With $m_{X,D,Z} = \mathbb{E}[Y \mid X, D, Z]$, let $V_Z = \mathcal{R}(m_X) - \mathcal{R}(m_{X,Z})$, $J = \mathcal{R}(m_{X,D}) - \mathcal{R}(m_{X,D,Z})$, and $Q = \mathcal{R}(m_{X,Z}) - \mathcal{R}(m_{X,D,Z})$. The nested risks satisfy

$$V_Z = \Delta_D + J - Q. \tag{7}$$

Here $J \geq 0$ measures context information beyond source identity, and $Q \geq 0$ the source information still missing. Under sufficiency, $J = 0$ and the recovered gap is $\Delta_D - Q$. External events or covariates may instead yield $J > 0$: their value depends on continuation information, not dataset identification. This separates source-code interventions from richer information interfaces.

### 4.2 CANCELLATION, SURVIVING STRUCTURE, AND HISTORY

Question B also predicts which forecast components survive unresolved rules and what historical evidence can recover them. For opposing means $m_{X,D} = c(X) + Br(X)$, let $c, r$ be common and disputed components, $B \in \{-1, 1\}$ the rule sign, $p(X) = \Pr(B = 1 \mid X)$, and $\Delta(X)$ the conditional information gap. Then

$$m_X = c(X) + (2p(X) - 1)r(X), \qquad \Delta(X) = 4p(X)(1 - p(X))\|r(X)\|^2/H. \tag{8}$$

Balanced ambiguity cancels the opposing component while retaining the common component. For a balanced binary rule and an otherwise uninformative history, a symmetric sign cue $Z$ with flip probability $\eta$ gives $\mathbb{E}[B \mid X, Z] = (1-2\eta)Z$. The disputed direction therefore has relative amplitude $1 - 2\eta$, yielding a quantitative prediction at intermediate cue reliability. For a finite group $K$ with orthogonal representation $\rho$, a uniform unresolved orbit of a reference vector $a(X)$ similarly gives $m_X = c(X) + \Pi_K a(X)$, where $\Pi_K = |K|^{-1} \sum_{g \in K} \rho(g)$ projects onto the invariant subspace. The surviving component need not be low frequency, and a nonuniform posterior average generally is not a projection. This averaging of unresolved output rules differs from intersecting task stabilizers. Appendix A.2 gives the group formula and these qualifications.

Longer nested histories cannot increase optimal squared risk, but source-label value need not decrease: new observations can make a previously irrelevant label predictive. In the opposing-rule model, historical evidence helps when it changes the posterior $p(X)$ away from balanced ambiguity; extra samples that leave this posterior unchanged do not restore the disputed component. Thus two equally long prefixes can have different predictive value. Section 5.3 fixes the recent suffix and future while changing prefix informativeness, using a structure-assisted learner to test whether the added rule evidence can be recovered. The intervention separates useful historical information from additional context length alone.

### 4.3 MIXTURE PREDICTION AND FINITE LEARNING

Question C asks when adding sources helps a fixed target; its answer also depends on how well a shared predictor is learned. For source laws $P_d$ and weights $\pi_d$, define mixture $P_\pi = \sum_d \pi_d P_d$ with source posterior $p_\pi(d \mid x)$. Its Bayes rule is $m_\pi(x) = \sum_d p_\pi(d \mid x) m_d(x)$. Fix target A with $\pi_A > 0$, and write $\|v\|_A^2 = H^{-1}\mathbb{E}_A\|v(X)\|^2$. For target A and auxiliary B, let $w_B(x) = p_\pi(B \mid x)$. The population mismatch is

$$\|m_\pi - m_A\|_A^2 = H^{-1}\mathbb{E}_A\left[w_B(X)^2\|m_B(X) - m_A(X)\|^2\right]. \tag{9}$$

The mixture-wide conflict and the target mismatch weight disagreements differently. A rare target can contribute little to the former yet incur a large mismatch when the auxiliary source dominates the posterior on its histories. Consequently, low pooled risk can coexist with damage to an individual target. This distinction makes per-target comparisons necessary: averaging over sources before measuring transfer can conceal the cases that motivate conditioning or a different training allocation.

Conflict requires overlap in observable histories and different conditional means on that overlap. Fitted models additionally incur approximation, estimation, and optimization effects. For fitted predictor $\widehat{f}$, write $b_A = m_\pi - m_A$ and $\langle u, v \rangle_A = H^{-1}\mathbb{E}_A[u(X)^\top v(X)]$. Target-risk orthogonality gives the exact decomposition

$$\mathcal{R}_A(\widehat{f}) - \mathcal{R}_A(m_A) = \|b_A\|_A^2 + \|\widehat{f} - m_\pi\|_A^2 + 2\langle b_A, \widehat{f} - m_\pi \rangle_A. \tag{10}$$

The interaction prevents a general split into two independent nonnegative penalties. More auxiliary observations can improve estimation of a shared rule while shifting its population optimum away from the target rule. A change in training allocation affects both processes, so the sign of observed transfer is a property of the fitted comparison, not a direct measurement of $\|b_A\|_A^2$.

Let $a_{\mathcal{F}}$ be the mixture approximation excess risk of predictor class $\mathcal{F}$, $\gamma$ its uniform empirical-to-population risk deviation, and $\varepsilon_{\text{opt}}$ the empirical suboptimality of $\widehat{f} \in \mathcal{F}$.

**Proposition 4** (A conditional mixture-learning bound)**.** *Under history-law domination $P_A^X \le C P_\pi^X$ for a constant $C > 0$,*

$$\mathcal{R}_A(\widehat{f}) - \mathcal{R}_A(m_A) \le \left(\|b_A\|_A + \sqrt{C(a_{\mathcal{F}} + 2\gamma + \varepsilon_{\text{opt}})}\right)^2. \tag{11}$$

Because the mixture includes the target, $C = 1/\pi_A$ suffices. Auxiliary data can reduce estimation error when the continuation rules agree, whereas conflicting sources can introduce population mismatch even with exact learning. Finite training can nevertheless improve target performance if its learning gains outweigh that mismatch and their interaction. Changing source weights also changes posterior overlap and target coverage. These distinct effects motivate both fixed-target-exposure and fixed-total-budget comparisons in the natural experiment. Appendix A.3 gives the precise error definitions and proof; the bound does not assume that adding sources improves either term.

### 4.4 From surviving structure to lightweight prediction

Question D extends Question B's surviving-component analysis: how much forecasting ability can a small adaptive repertoire acquire from a TSFM? This study combines $K$ forecasting operators $F_1, \ldots, F_K$ for level, trend, seasonality, and autoregression. A network $g_\theta$ with parameters $\theta$ produces weights $w_\theta$ and student forecast $D_\theta$:

$$D_\theta(x) = \sum_{k=1}^{K} w_{\theta,k}(x) F_k(x), \qquad w_\theta(x) = \operatorname{softmax} g_\theta(x). \tag{12}$$

Forecasts lie in the operator outputs' convex hull; learning changes their weights. Teacher predictions supervise $g_\theta$ from few examples without retraining the teacher. Appendix A.4 describes the representation. We measure future-prediction risk at specified learning and inference budgets. Same-source training tests within-task prediction; source holdout tests transfer. Their comparison measures dependence on target-source exposure, giving the filter interpretation an operational meaning: the forecasting ability explicit operators can carry and the resources required.

## 5 Experiments

### 5.1 Experimental design

We test Questions A–B through controlled interventions, C through small-scale pretraining, and D through pretrained-model behavior (Jander et al., 2026). The operator probe uses Chronos-2, Moirai 2.0, TimesFM-2.5, Sundial, TiRex-2, and Toto-2.0-313m (Ansari et al., 2025; Liu et al., 2026; Das et al., 2024; Liu et al., 2025; Podest et al., 2026; Khwaja et al., 2026). Natural training uses five GIFT-Eval sources: Electricity, Electricity Transformer Temperature (ETT1 and ETT2), Jena weather, and LOOP Seattle (Aksu et al., 2024); a fixed-recipe coverage panel expands this comparison to 12 sources (Appendix D.4). Natural comparisons use disjoint training, validation, and test intervals with three initialization seeds. We report window-normalized mean squared error (NMSE). Natural training and the operator probe use $L = 512$, $H = 48$; the latter trains lightweight, input-dependent operator mixtures from the teachers' native median forecasts. Natural uncertainty uses paired entity or temporal-block resampling after seed averaging, with limited independent blocks for ETT and Jena. Appendix B defines splits and scoring, and Appendices C–E give the experiment-specific protocols.

### 5.2 Question A: does a shared transformation preserve the task?

**Symmetry intervention.** We hold the numerical interface and input law fixed while varying whether the sign action $(z, y) \mapsto (-z, -y)$ preserves the continuation rule. For standard Gaussian state $z$, waveform $u$ with $\|u\|^2/H = 1$, and even-component weight $\alpha$, let $m_\alpha(z) = [\sqrt{1-\alpha}\, z + \sqrt{\alpha}(z^2 - 1)/\sqrt{2}]u$. The sign action preserves the odd component and cancels the even component; all six tasks have unit signal energy and noise variance .01. Equation 3 predicts excess risk $\alpha$. Across $\alpha \in \{0, .1, .25, .5, .75, 1\}$, original-task NMSE stays within .01002–.01007, while augmented NMSE rises to 1.00495. The measured conditional-mean risk increments track the finite-panel prediction $\delta_g/4$, with maximum absolute deviation .0041 across seed means (Figure 1). A polynomial reference gives the same trend. Appendix C.1 specifies the paired protocol.

### 5.3 Question B: what information resolves conflicting rules?

**Rule visibility.** A noisy trajectory reaches a boundary and either transmits or reflects its velocity. The rules have identical history distributions and opposing continuations. With the same small pre-

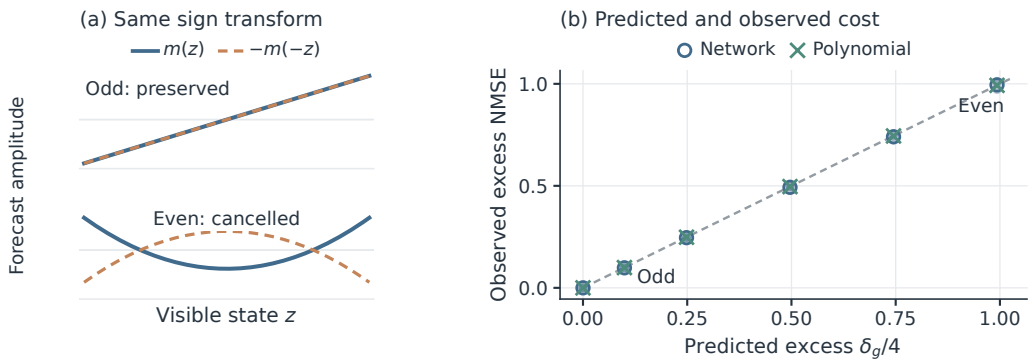


Figure 1: **Question A: shared transformations need not preserve the prediction task.** (a) Sign transformation preserves the odd rule and cancels the even rule. (b) Excess conditional-mean NMSE follows defect cost $\delta_g/4$ across six mixtures. Small markers are three seeds; large markers their mean. The dashed line denotes equality.

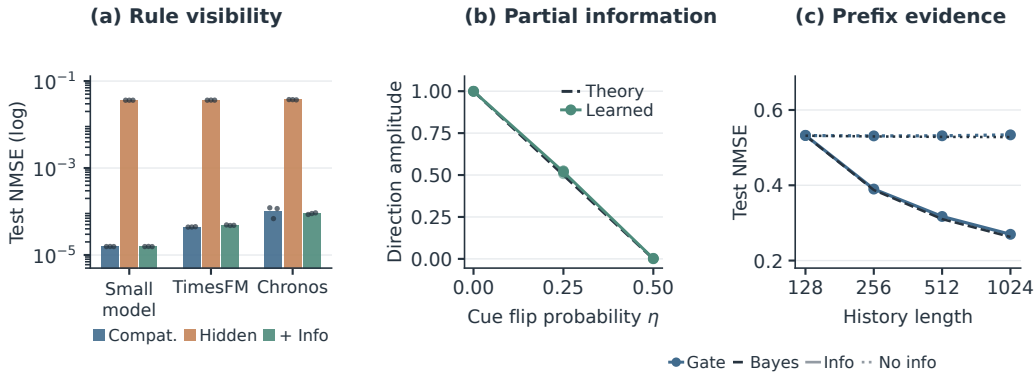


Figure 2: **Question B: rule visibility controls averaging and recoverability.** (a) Compatible, hidden, and informative conditions; comparisons are within each model, with dots for three seeds. (b) Cue reliability controls amplitude. (c) Informative versus equally long uninformative prefixes on 2,048 paired trajectories, comparing the structure-assisted learner with conditional Bayes risk.

dictor, compatible pooling attains NMSE $1.566 \times 10^{-5}$, versus .036279 under hidden conflict. Informative source identifiers restore the compatible risk (Figure 2(a)); independent random identifiers do not (Appendix C). Single-source and compatible risks lie within 1.02 times their visible-information least-squares references. As a historical cue's error probability $\eta$ increases, the learned direction follows Equation 8's attenuation $1-2\eta$ (panel b). The decision also depends on the loss: at $\eta = .25$, mean absolute error (MAE) training restores near-unit amplitude rather than the mean squared error (MSE)-trained value .521, improving absolute error while worsening NMSE. Mean, median, and distributional forecasts therefore need different interpretations (Gneiting & Raftery, 2007; Scheuerer & Hamill, 2015; Ferro, 2014). The same intervention extends to pretrained TimesFM and Chronos: hidden-rule NMSE is .03640 and .03728, whereas informative conditions reduce it to $4.83 \times 10^{-5}$ and $8.95 \times 10^{-5}$. Figure 2(a) compares the three conditions within each model, linking the information mechanism to TSFM adaptation under their respective training objectives.

**Informative history.** The prefix intervention fixes the recent suffix and future, adding informative events or equally long uninformative observations. A structure-assisted gate reduces hidden-conflict NMSE from .5327 at $L = 128$ to .2697 at $L = 1024$, near the conditional-risk reference .2628; uninformative-prefix risk stays near .53 (panel c). The paired information interaction is $-.2648$, with a 95% trajectory-bootstrap interval excluding zero. The gate uses event and frequency priors from the generator. Appendices C and C.3 report the generic Transformer comparison, full intervals, and sensitivity to those priors.

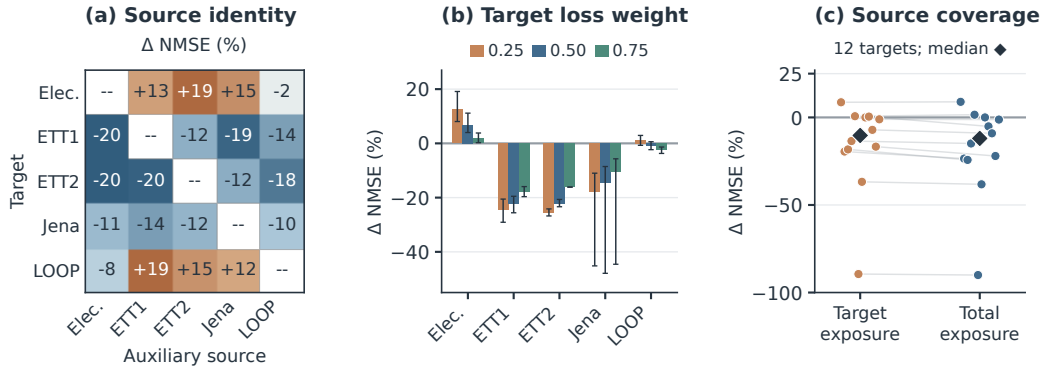


Figure 3: **Question C: auxiliary-data value depends on the target and training allocation.** NMSE percentage changes ($\Delta$) below zero favor mixing over target-only training. (a) Auxiliary-source identity and (b) target loss weight, with total exposure fixed; whiskers in (b) are paired 95% entity/block intervals. (c) Twelve targets under fixed target or total exposure: connected points represent the same target and diamonds mark source medians. All effects use three-seed mean risks.

### 5.4 Question C: When do additional sources help a target?

**Controlled training comparisons.** We train the same patch Transformer from random initialization on target A alone, A plus one auxiliary, or A plus the other four sources. All arms use 400 updates and a fixed endpoint; mixed batches contain equal numbers of target and auxiliary windows. Figure 3(a,b) compares with target-only training at the same total examples and updates, so negative NMSE changes favor transfer. A second control holds target exposure fixed to test the value of additional auxiliary examples. Appendix D retains both five-source exposure comparisons and validation-based source selection.

**Source identity.** On Electricity, ETT2 raises NMSE by $19.47\%$, whereas LOOP lowers it by $1.92\%$, with paired entity-bootstrap intervals $[10.60, 31.85]\%$ and $[-4.17, -.35]\%$ (Figure 3(a)). The reverse pair improves ETT2 by $20.28\%$, so a source pair has no single transfer value. This directionality accords with Equation 10, which evaluates both mismatch and learning effects on the target distribution. The Electricity contrast persists at 800 updates, with changes $+24.90\%$ and $-3.40\%$; validation checkpoint selection preserves both directions (Appendix D.2).

**Loss allocation.** Figure 3(b) holds sampled counts fixed and varies the target's share of the loss. Increasing this weight from .25 to .75 reduces Electricity's excess NMSE from $12.49\%$ to $1.84\%$, but reduces ETT2's improvement from $25.57\%$ to $16.13\%$. LOOP changes from $+1.19\%$ to $-2.48\%$. Emphasizing the target can therefore protect its fitted solution or weaken useful auxiliary learning. Because these fits use the same observations, this intervention separates training allocation from simply adding data, as anticipated by Equation 10.

**Breadth of useful sharing.** Figure 3(c) extends the same recipe to twelve sources spanning five domains and three sampling frequencies. Mixing improves NMSE on 9/12 targets under fixed total exposure (median $-11.97\%$) and 8/12 under fixed target exposure ($-10.29\%$). The gains include Solar and M Dense, while Electricity retains a penalty of $8.91\%$ under fixed total exposure, with paired interval $[4.98, 14.67]\%$. Useful broad transfer can coexist with target-specific penalties. Appendix D.4 reports all targets and intervals; temporal-range and natural pretrained-adaptation comparisons remain in Appendix D.

### 5.5 Question D: Can lightweight operators acquire forecasting ability?

**A small learned repertoire.** A 4,590-parameter network learns history-dependent weights over 14 level, trend, seasonal, and autoregressive operators from a frozen TSFM's forecasts. Each student uses 256 or 1,024 training examples and 128 validation examples. Same-source training uses earlier target-dataset observations; source-family holdout excludes that dataset and related variants from

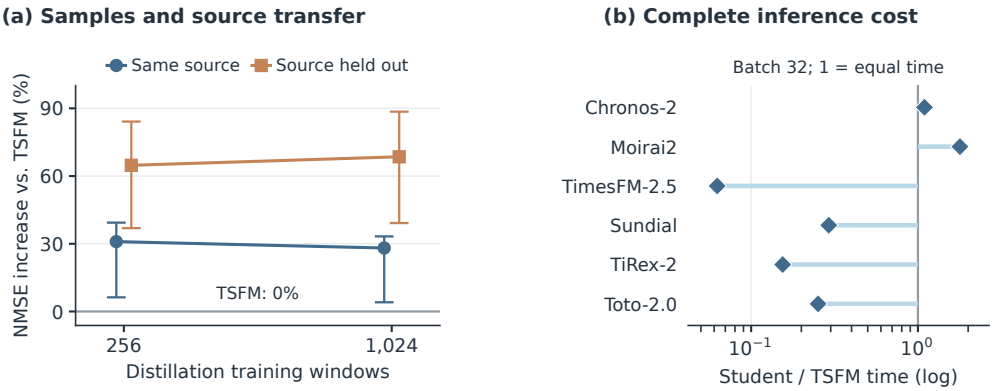


Figure 4: **Question D: exploring forecasting ability with a lightweight teacher-trained repertoire.** (a) NMSE excess at two sample budgets, pooling six teachers within each of five sources; whiskers are paired 95% entity intervals. Zero denotes teacher-level risk. (b) Complete student/teacher batch-32 inference-time ratios; below one is faster. All 14 operators run online.

training and validation. Both protocols share five test sources, six teachers, sample budgets, and three seeds (Appendix E).

**Prediction from few teacher examples.** At 1,024 training windows, same-source student NMSE is $28.12\%$ above the teachers; source-family-held-out NMSE is $68.54\%$ above them (Figure 4(a)). Within each source, we average the six teacher risks and corresponding three-seed student risks, compute the student's percentage excess, then average the five percentages. Appendix E reports teacher-specific differences. At 256 windows the gaps are $30.96\%$ and $64.77\%$. Small target-source samples already support useful forecasts; exposure to that source matters more here than the fourfold difference in sample count.

**Learning scale and deployment cost.** Complete student inference takes 18.7–26.9 milliseconds (ms) per batch of 32 histories, including all operators. It is faster than four teachers, with student/teacher time ratios .063–.292 (Figure 4(b)). The timings quantify deployment cost alongside learning scale.

**What is being reused?** Limited teacher supervision lets the network select among explicit forecasting operations; source-family holdout tests their reuse without target examples. Together with Questions A–C, this gives the filter interpretation a concrete scope: simple shared operations carry useful forecasting ability, while information and training distribution determine where they apply.

## 6 Discussion and Conclusion

**Discussion.** Controlled interventions quantify how task structure and rule information affect risk; natural pretraining tests source choice and allocation; teacher-trained operators reveal the reach of a small adaptive repertoire. The last question concerns the prediction gap achievable at a specified learning scale and deployment cost. These branches connect the category-error concern to mechanisms of invalid sharing, practical source allocation, and reusable prediction structure. Our evidence comes from small-scale training on 12 sources and behavioral analysis of six TSFMs; broader pretraining interventions would test how these effects develop with scale. The resulting choices concern source composition, usable context, and learning objectives, beyond the shared numerical format.

**Conclusion.** Universal time-series forecasting warrants a task-relative account of what transfers. A common numerical container guarantees neither shared predictive structure nor sufficient information; the tested consequences give this concern empirical substance. We encourage models and information interfaces that distinguish rules, losses aligned with forecast decisions, and training and data construction that preserve useful sharing as mixtures and predictors scale.

### Artificial intelligence (AI) use statement

In this work, we used generative AI tools to aid or polish writing and for retrieval and discovery. We have not used generative AI tools for other tasks with required disclosure. We have reviewed all AI-assisted work and checked LLM-generated results manually. We take responsibility for the final content of this work, including text produced with the aid of generative AI.

### Reproducibility statement

Experimental code is provided in the supplementary material.

## References

Taha Aksu, Gerald Woo, Juncheng Liu, Xu Liu, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. Gift-eval: A benchmark for general time series forecasting model evaluation. *arXiv preprint arXiv:2410.10393*, 2024.

Elham Alipour, xiaoli zhang, and Boyang Tom Jin. Impact-driven event covariates for time-series foundation models. In *2nd ICML Workshop on Foundation Models for Structured Data*, 2026.

Abdul Fatir Ansari, Lorenzo Stella, Ali Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Bernie Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. Expert Certification.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, Mononito Goswami, Shubham Kapoor, Danielle C. Maddix, Pablo Guerron, Tony Hu, Junming Yin, Nick Erickson, Prateek Mutalik Desai, Hao Wang, Huzefa Rangwala, George Karypis, Yuyang Wang, and Michael Bohlke-Schneider. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2010. doi: 10.1007/s10994-009-5152-4.

David Blackwell. *Comparison of Experiments*, pp. 93–102. University of California Press, December 1951. ISBN 9780520411586. doi: 10.1525/9780520411586-009.

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

Shuxiao Chen, Edgar Dobriban, and Jane H. Lee. A group-theoretic framework for data augmentation. *Journal of Machine Learning Research*, 21(245):1–71, 2020.

Panayiotis Christou, Shichu Chen, Xupeng Chen, and Parijat Dube. Test time learning for time series forecasting. *arXiv preprint arXiv:2409.14012*, 2024.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2990–2999. PMLR, 2016.

Abhimanyu Das, Weihao Kong, Andrew Leach, Shaan K Mathur, Rajat Sen, and Rose Yu. Long-term forecasting with tiDE: Time-series dense encoder. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Forty-first International Conference on Machine Learning*, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2021.

Vijay Ekambaram, Arindam Jati, Pankaj Dayama, Sumanta Mukherjee, Nam H. Nguyen, Wesley M. Gifford, Chandra Reddy, and Jayant Kalagnanam. Tiny Time Mixers (TTMs): Fast pre-trained models for enhanced zero/few-shot forecasting of multivariate time series. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2359.

Kun Feng, Shaocheng Lan, Yuchen Fang, Wenchao He, Sihan Lu, Shuqi Gu, Lintao Ma, Xingyu Lu, and Kan Ren. Kairos: Toward adaptive and parameter-efficient time series foundation models. *arXiv preprint arXiv:2509.25826*, 2026.

Md Meftahul Ferdaus, Tanmoy Dam, Md Rasel Sarkar, Moslem Uddin, and Sreenatha G Anavatti. Foundation models for clean energy forecasting: A comprehensive review. *Renewable and Sustainable Energy Reviews*, 226:116452, 2026.

C. A. T. Ferro. Fair scores for ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 140(683):1917–1923, 2014. doi: https://doi.org/10.1002/qj.2270.

Nicholas Frosst and Geoffrey Hinton. Distilling a neural network into a soft decision tree. *arXiv preprint arXiv:1711.09784*, 2017.

Tilmann Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011. doi: 10.1198/jasa.2011.r10138.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: A family of open time-series foundation models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 16115–16152. PMLR, 21–27 Jul 2024.

Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew G Wilson. Large language models are zero-shot time series forecasters. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 19622–19635. Curran Associates, Inc., 2023. doi: 10.52202/075280-0861.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Mathis Jander, Wouter van Heeswijk, and Martijn Mes. Causal analysis for time series foundation models. *arXiv preprint arXiv:2608.24303*, 2026.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. Time-llm: Time series forecasting by reprogramming large language models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 23857–23880, 2024.

Fredrik D. Johansson, David Sontag, and Rajesh Ranganath. Support and invertibility in domain-invariant representations. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 527–536. PMLR, 2019.

Emaad Khwaja, Chris Lettieri, Gerald Woo, Eden Belouadah, Marc Cenac, Guillaume Jarry, Enguerrand Paquin, Xunyi Zhao, Viktoriya Zhukov, Othmane Abou-Amal, Chenghao Liu, Ameet Talwalkar, and David Asker. Toto 2.0: Time series forecasting enters the scaling era. *arXiv preprint arXiv:2605.20119*, 2026.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International Conference on Learning Representations*, 2022.

Siva Rama Krishna Kottapalli, Karthik Hubli, Sandeep Chandrashekhara, Garima Jain, Sunayana Hubli, Gayathri Botla, and Ramesh Doddaiah. Foundation models for time series: A survey. *arXiv preprint arXiv:2504.04011*, 2025.

Chenghao Liu, Taha Aksu, Juncheng Liu, Xu Liu, Hanshu Yan, Quang Pham, Silvio Savarese, Doyen Sahoo, Caiming Xiong, and Junnan Li. Moirai 2.0: When less is more for time series forecasting. *arXiv preprint arXiv:2511.11698*, 2026.

Xu Liu, Junfeng Hu, Yuan Li, Shizhe Diao, Yuxuan Liang, Bryan Hooi, and Roger Zimmermann. Unitime: A language-empowered unified model for cross-domain time series forecasting. In *Proceedings of the ACM Web Conference 2024*, pp. 4095–4106, 2024a.

Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: Exploring the stationarity in time series forecasting. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 9881–9893. Curran Associates, Inc., 2022. doi: 10.52202/068431-0718.

Yong Liu, Guo Qin, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Autotimes: Autoregressive time series forecasters via large language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 122154–122184. Curran Associates, Inc., 2024b. doi: 10.52202/079017-3882.

Yong Liu, Guo Qin, Zhiyuan Shi, Zhi Chen, Caiyin Yang, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Sundial: A family of highly capable time series foundation models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 39295–39317, 2025.

Xiangkai Ma, Xiaobin Hong, Wenzhong Li, and Sanglu Lu. A wave is worth 100 words: Investigating cross-domain transferability in time series. *ACM Transactions on Knowledge Discovery from Data*, 20(3):1–35, 2026. doi: 10.1145/3788674.

Pablo Montero-Manso and Rob J. Hyndman. Principles and algorithms for forecasting groups of time series: Locality and globality. *International Journal of Forecasting*, 37(4):1632–1653, 2021. doi: 10.1016/j.ijforecast.2021.03.004.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

Alvaro Perez-Diaz, James C. Loach, Danielle E. Toutoungi, and Lee Middleton. Foundation model forecasts: Form and function. *arXiv preprint arXiv:2510.19345*, 2025.

Patrick Podest, Marco Pichler, Elias Bürger, Levente Zólyomi, Bernhard Voggenberger, Wilhelm Berghammer, Daniel Klotz, Sebastian Böck, Günter Klambauer, and Sepp Hochreiter. TiRex-2: Generalizing TiRex to multivariate data and streaming. *arXiv preprint arXiv:2607.01204*, 2026.

Andres Potapczynski, Ravi Kiran Selvam, Tatiana Konstantinova, Malcolm Wolff, Kin G. Olivares, Ruijun Ma, Michael W. Mahoney, Andrew Gordon Wilson, Boris N. Oreshkin, and Dmitry Efimov. The arrow of time: What tabular foundation models miss in time series forecasting. In *ICLR Workshop on Time Series in the Age of Large Models*, 2026.

Michael Scheuerer and Thomas M. Hamill. Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review*, 143(4):1321–1334, 2015. doi: 10.1175/MWR-D-14-00269.1.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9229–9248. PMLR, 13–18 Jul 2020.

Md Asif Bin Syed, Md Younus Ahamed, and Azmine Toushik Wasi. Position: Time-series foundation models require explicit domain-level benchmarks. In *Forty-third International Conference on Machine Learning Position Paper Track*, 2026.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Zhang, and JUN ZHOU. Timemixer: Decomposable multiscale mixing for time series forecasting. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 38626–38652, 2024a.

Xinlei Wang, Maike Feng, Jing Qiu, Jinjin Gu, and Junhua Zhao. From news to forecast: Integrating event analysis in llm-based time series forecasting with reflection. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 58118–58153. Curran Associates, Inc., 2024b. doi: 10.52202/079017-1853.

Andrew Robert Williams, Arjun Ashok, Étienne Marcotte, Valentina Zantedeschi, Jithendaraa Subramanian, Roland Riachi, James Requeima, Alexandre Lacoste, Irina Rish, Nicolas Chapados, and Alexandre Drouin. Context is key: A benchmark for forecasting with essential textual information. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 66887–66944, 2025.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53140–53164. PMLR, 21–27 Jul 2024.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 22419–22430. Curran Associates, Inc., 2021.

Yi Xie, Yun Xiong, Zejian Shi, Hao Niu, and Zhengfu Liu. The language of time: a language model perspective on time-series foundation models. *arXiv preprint arXiv:2507.00078*, 2025.

Zhijian Xu, Hao Wang, and Qiang Xu. Intervention-aware forecasting: Breaking historical limits from a system perspective. *arXiv preprint arXiv:2405.13522*, 2025.

Qingren Yao, Ming Jin, Chengqi Zhang, Chao-Han Huck Yang, Jun Qi, and Shirui Pan. Estimating time series foundation model transferability via in-context learning. *arXiv preprint arXiv:2509.23695*, 2025.

Jiexia Ye, Yongzi Yu, Weiqi Zhang, Le Wang, Jia Li, and Fugee Tsung. Empowering time series analysis with foundation models: A comprehensive survey. *Information Fusion*, 137:104601, 2027. doi: 10.1016/j.inffus.2026.104601.

Annan Yu, Danielle Maddix, Boran Han, Xiyuan Zhang, Abdul Fatir Ansari, Oleksandr Shchur, Christos Faloutsos, Andrew Gordon Wilson, Michael W Mahoney, and Bernie Wang. Understanding the implicit biases of design choices for time series foundation models. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 71998–72049, 2026.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9):11121–11128, June 2023. ISSN 2159-5399. doi: 10.1609/aaai.v37i9.26317.

Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Joshua M. Susskind. Stabilizing transformer training by preventing attention entropy collapse. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 40770–40803. PMLR, 23–29 Jul 2023.

Jiawen Zhang, Zhenwei Zhang, Shun Zheng, Xumeng Wen, Jia Li, and Jiang Bian. Are time-series foundation models deployment-ready? a systematic study of adversarial robustness across domains. *arXiv preprint arXiv:2505.19397*, 2025a.

Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *The eleventh international conference on learning representations*, 2023.

Zhenwei Zhang, Jiawen Zhang, Shun Zheng, Yuantao Gu, and Jiang Bian. Does cross-domain pre-training truly help time-series foundation models? In *ICLR 2025 Workshop on Foundation Models in the Wild*, 2025b.

Han Zhao, Remi Tachet des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7523–7532. PMLR, 2019.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):11106–11115, May 2021. ISSN 2159-5399. doi: 10.1609/aaai.v35i12.17325.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 27268–27286. PMLR, 17–23 Jul 2022.

# A PROOFS FOR QUESTIONS A–D

Functions are identified up to almost-sure equivalence, with finite norms and risks. The identities use a complete future vector under Euclidean squared loss in one common coordinate system; masks, normalization denominators, and nonquadratic scores can define different risks.

## A.1 TASK SYMMETRIES (QUESTION A)

The task stabilizer in Equation 1 is a subgroup: the identity satisfies both requirements, and if $g, h \in G$ preserve $P_\theta$, then $(gh)_\# P_\theta = g_\#(h_\# P_\theta) = P_\theta$. Since $h$ preserves the measure, it also preserves null sets, and $m_\theta(ghx) = \rho(g)m_\theta(hx) = \rho(gh)m_\theta(x)$ almost surely. The inverse relation follows by changing variables with the measure-preserving map $g$. Thus intersections of these stabilizers are also subgroups. For $S \subseteq T$, every element satisfying all the constraints indexed by $T$ satisfies those indexed by $S$, establishing $G_T \subseteq G_S$.

If the family is separating, no nonidentity element can belong to the intersection. For the unconstrained-family condition, faithfulness ensures that each $g \neq e$ moves at least one $x$. The point-supported law (Dirac measure) $\delta_x$ is then sent to $\delta_{gx} \neq \delta_x$. It excludes $g$ irrespective of the output rule. Without faithfulness, elements acting identically everywhere cannot be excluded this way.

For Proposition 2, distribution preservation makes the posterior weights of the original and transformed tasks equal to $1/2$ at almost every history. Since $g^{-1} = g$ and $\rho(g)^{-1} = \rho(g)$, the transformed conditional mean is $\rho(g)m(gx)$. Their average is $m_g$. Conditional-expectation orthogonality on the original task gives

$$\mathcal{R}_P(m_g) - \mathcal{R}_P(m) = \frac{1}{4H}\mathbb{E}\|\rho(g)m(gX) - m(X)\|^2 = \delta_g/4,$$

where the last equality uses orthogonality. No conditional-variance equality between the two tasks is needed. Under the assumed history-law invariance, $\delta_g = 0$ is exactly the conditional-mean equivariance requirement in the task stabilizer.

For a geometric extension, additionally assume the relevant shared groups are closed Lie subgroups and act smoothly, freely, and properly on the same $d$-dimensional manifold. The quotient-manifold theorem gives $\dim(\mathcal{X}/G_S) = d - \dim G_S$. Inclusion of Lie subgroups implies $\dim G_T \leq \dim G_S$, hence $\dim(\mathcal{X}/G_T) \geq \dim(\mathcal{X}/G_S)$. General group actions can have varying stabilizers and singular quotient spaces; the formula is not asserted for them. For an equivariant output rule, reconstructing the output may require orbit-position information in addition to a quotient coordinate.

## A.2 SOURCE INFORMATION AND SURVIVING STRUCTURE (QUESTION B)

For nested sigma-algebras $\mathcal{F} \subseteq \mathcal{G}$, set $m_\mathcal{F} = \mathbb{E}[Y \mid \mathcal{F}]$ and $m_\mathcal{G} = \mathbb{E}[Y \mid \mathcal{G}]$. Expand

$$Y - m_\mathcal{F} = (Y - m_\mathcal{G}) + (m_\mathcal{G} - m_\mathcal{F}).$$

The cross term vanishes by conditioning on $\mathcal{G}$, proving Equation 4; the risk gap is zero exactly when the conditional means agree almost surely. For a finite source set, condition on $X = x$ and use $\sum_d p_d(x)m_d(x) = m_X(x)$. Expanding the pairwise squared distances gives

$$\begin{aligned}\frac{1}{2}\sum_{d,e} p_d(x)p_e(x)\|m_d(x) - m_e(x)\|^2 &= \sum_d p_d(x)\|m_d(x)\|^2 - \|m_X(x)\|^2 \\ &= \sum_d p_d(x)\|m_d(x) - m_X(x)\|^2.\end{aligned}$$

Taking expectation and dividing by $H$ proves Equation 5. Conditional means on source-null sets may be assigned arbitrary values because their posterior weights vanish. For metadata, the assumption $\mathbb{E}[Y \mid X, D, Z] = m_{X,D}$ ensures $m_{X,Z} = \mathbb{E}[m_{X,D} \mid X, Z]$. Decompose

$$m_{X,D} - m_X = (m_{X,D} - m_{X,Z}) + (m_{X,Z} - m_X).$$

The first term has conditional mean zero given $(X, Z)$, so orthogonality proves Equation 6. Without the sufficiency assumption, subtracting the nested risks defining $V_Z$, $J$, and $Q$ gives Equation 7.

Orthogonality ensures $J, Q \geq 0$; sufficiency sets $J = 0$, whereas context that predicts additional future information can have $J > 0$.

The same projection identity quantifies preprocessing. If raw history $W$ is replaced by $X = \phi(W)$, its population cost in a common output coordinate system is

$$H^{-1}\mathbb{E}\|\mathbb{E}[Y \mid W] - \mathbb{E}[Y \mid X]\|^2.$$

Retaining location and scale statistics after normalization changes this information set.

For finitely many sources and a deterministic code $z = \phi(x, d)$, write $p_d(x) = \Pr(D = d \mid X = x)$ and $C_z(x) = \{d : \phi(x, d) = z\}$. For a nonempty positive-mass cell,

$$\bar{m}_z(x) = \frac{\sum_{d \in C_z(x)} p_d(x) m_d(x)}{\sum_{d \in C_z(x)} p_d(x)}.$$

The unresolved conditional gap is

$$\frac{1}{H} \sum_z \sum_{d \in C_z(x)} p_d(x) \|m_d(x) - \bar{m}_z(x)\|^2.$$

It is zero exactly when all positive-probability source means within each cell coincide. If there are more distinct positive-probability means than code values, no such deterministic code can remove the entire gap at that history. For random metadata with $Z \perp D \mid X$, the posterior source weights remain unchanged, giving $m_{X,Z} = m_X$ under the same conditional-mean sufficiency assumption.

The sign formula follows from $\mathbb{E}[B \mid X] = 2p(X) - 1$ and $\mathrm{Var}(B \mid X) = 4p(X)(1 - p(X))$. For a finite orthogonal group $K$, let $D \mid X$ be uniform on $K$ and $m_{X,D} = c(X) + \rho(D)a(X)$. Then

$$m_X = c(X) + \Pi_K a(X), \qquad \Pi_K = |K|^{-1} \sum_{g \in K} \rho(g), \tag{13}$$

and multiplication permutes group elements, so $\Pi_K^2 = \Pi_K$ and $\rho(g)\Pi_K = \Pi_K$. Orthogonality and the inverse permutation imply $\Pi_K^\top = \Pi_K$. Its range is exactly the common fixed subspace. Writing $I$ for the identity matrix, each $\rho(g)$ preserves the norm of the orthogonal complement, so the mean squared distance of $\rho(D)a$ from $\Pi_K a$ equals $\|(I - \Pi_K)a\|^2$. Thus the conditional information gap is $\|(I - \Pi_K)a(X)\|^2/H$.

Uniformity must hold conditional on $X$: with sign-group weights $3/4$ and $1/4$, the average is $I/2$, not an idempotent projection. A projection also need not be low-pass: for any unit alternating-sign vector $v$, $\{I, 2vv^\top - I\}$ is an orthogonal group averaging to $vv^\top$, which preserves that high-frequency direction. Common structure $c(X)$ remains in either case. For any fixed linear $Q$, conditional orthogonality gives

$$\mathbb{E}\|Qm_{X,D}\|^2 - \mathbb{E}\|Qm_X\|^2 = \mathbb{E}\|Q(m_{X,D} - m_X)\|^2.$$

Temporal demeaning and first differences select different structures; this identity concerns conditional means, not arbitrary learned predictors or medians.

Nested histories yield nonincreasing Bayes squared risk by the same projection proof. For a counterexample to monotonicity of the label gap, let $D, Z$ be independent fair signs, let $Y = DZ$, let the short history be empty, and let the longer history reveal $Z$. Knowing $D$ adds no predictive value to the short history but removes all unit risk with the longer history. Thus longer histories need not reduce the source-label gap. Renormalizing histories while discarding their statistics can also violate information nesting; common scoring coordinates alone do not restore the lost input information.

### A.3 Mixture prediction and finite learning (Question C)

For $v : \mathcal{X} \to \mathbb{R}^H$, define $\|v\|_\pi^2 = H^{-1}\mathbb{E}_\pi\|v(X)\|^2$ and $\|v\|_A^2 = H^{-1}\mathbb{E}_A\|v(X)\|^2$. Write $\widehat{\mathcal{R}}_\pi$ for empirical mixture risk. Let

$$a_{\mathcal{F}} = \inf_{f \in \mathcal{F}} \{\mathcal{R}_\pi(f) - \mathcal{R}_\pi(m_\pi)\}, \qquad \gamma = \sup_{f \in \mathcal{F}} |\widehat{\mathcal{R}}_\pi(f) - \mathcal{R}_\pi(f)|.$$

The fitted predictor satisfies $\widehat{f} \in \mathcal{F}$ and $\widehat{\mathcal{R}}_\pi(\widehat{f}) \leq \inf_{f \in \mathcal{F}} \widehat{\mathcal{R}}_\pi(f) + \varepsilon_{\mathrm{opt}}$.

Disintegrating $P_\pi = \sum_d \pi_d P_d$ gives $m_\pi(x) = \sum_d p_\pi(d \mid x) m_d(x)$ almost surely; versions on source-null sets do not affect positive-posterior contributions. Target-risk orthogonality gives $\mathcal{R}_A(f) - \mathcal{R}_A(m_A) = \|f - m_A\|_A^2$. Splitting this distance at $m_\pi$ gives Equation 10, where $b_A = m_\pi - m_A$.

This decomposition can also be averaged over the training randomness. Let $\mathbb{E}_{\text{train}}$ average over training randomness and $\bar{f} = \mathbb{E}_{\text{train}}[\widehat{f}]$; assume the required integrability, and set $V_A = \mathbb{E}_{\text{train}}\|\widehat{f} - \bar{f}\|_A^2$. Then

$$\begin{aligned}\mathbb{E}_{\text{train}}[\mathcal{R}_A(\widehat{f}) - \mathcal{R}_A(m_A)] = \|m_\pi - m_A\|_A^2 + \|\bar{f} - m_\pi\|_A^2 + V_A \\ + 2\langle m_\pi - m_A, \bar{f} - m_\pi\rangle_A.\end{aligned}$$

The interaction disappears under unbiasedness for $m_\pi$ on the target support, which general neural training does not guarantee.

For Proposition 4, the assumed empirical guarantee and uniform deviation imply

$$\begin{aligned}\mathcal{R}_\pi(\widehat{f}) &\le \widehat{\mathcal{R}}_\pi(\widehat{f}) + \gamma \\ &\le \inf_{f \in \mathcal{F}} \widehat{\mathcal{R}}_\pi(f) + \varepsilon_{\text{opt}} + \gamma \\ &\le \inf_{f \in \mathcal{F}} \mathcal{R}_\pi(f) + \varepsilon_{\text{opt}} + 2\gamma.\end{aligned}$$

Subtracting the mixture Bayes risk bounds $\|\widehat{f} - m_\pi\|_\pi^2$ by $a_{\mathcal{F}} + \varepsilon_{\text{opt}} + 2\gamma$. The density domination assumption transfers this squared norm to the target with factor $C$. The triangle inequality then proves the result. No minimizer in the class need exist: all comparisons can use its infimum and an arbitrarily small approximation slack. If $\pi_A > 0$, the measure inequality $P_\pi^X \ge \pi_A P_A^X$ supplies $C = 1/\pi_A$. The bound is conditional on $\gamma$ and $\varepsilon_{\text{opt}}$; a high-probability guarantee additionally requires control of $\gamma$ under a specified sampling process and model class.

### A.4 Adaptive operator representation (Question D)

For every history $x$, Equation (12) places the student forecast in the convex hull of $F_1(x), \ldots, F_K(x)$, since softmax weights are nonnegative and sum to one. The network chooses among combinations of these forecasts; it cannot directly produce a forecast outside their hull. Zero output logits initialize all weights at $1/K$, and subsequent learning makes them history-dependent.

The finite-group average in Appendix A.2 illustrates how shared components survive unresolved rules. The empirical dictionary contains explicit level, trend, seasonal, and autoregressive forecasts. Its future-prediction risk, together with its training and inference budgets, measures how much forecasting ability this restricted family can carry from a TSFM. Comparing same-source and source-held-out fitting then measures dependence on exposure to the target source. The controlled interventions in Question B address the information mechanism; this teacher-trained representation explores the predictive reach of simple reusable dynamics.

## B Shared data and scoring protocols

The natural-data experiments use GIFT-Eval sources (Aksu et al., 2024) at their selected sampling frequencies. The five-source pooling panel, its twelve-source extension, the seven-source Chronos adaptation (Appendix D.3), and the operator probe have separate data allocations. Trained comparisons use three initialization seeds.

### B.1 Natural data and split boundaries

The original panel comprises Electricity, ETT1, ETT2, Jena weather, and LOOP Seattle. The extension adds Solar, M Dense, Bitbrains rnd, BizITObs L2C, KDD Cup 2018, Saugeen Day, and US Births. Together they cover energy, traffic, computing-system monitoring, environmental measurements, and birth counts. Each source contributes one sampling frequency (Table 1); $L = 512$ and $H = 48$ denote observations, so their elapsed durations depend on that frequency.

Table 1: Natural small-model data coverage and window counts. The first five sources form the original panel; all twelve form the extension. Validation I and II have the same count, shown once for each subset. Complete window support is disjoint across the four chronological splits.

| Source | Interval | Train | Val. I / II (each) | Test |
|---|---|---|---|---|
| Electricity | 1 hour | 8192 | 512 | 1024 |
| ETT1 | 1 hour | 1316 | 378 | 140 |
| ETT2 | 1 hour | 1305 | 378 | 140 |
| Jena weather | 1 hour | 1680 | 378 | 399 |
| LOOP Seattle | 1 hour | 8192 | 512 | 1024 |
| Solar | 1 hour | 8192 | 512 | 1024 |
| M Dense | 1 hour | 5670 | 512 | 600 |
| Bitbrains rnd | 5 min | 8192 | 512 | 1024 |
| BizITObs L2C | 5 min | 2553 | 512 | 140 |
| KDD Cup 2018 | 1 hour | 4285 | 512 | 1024 |
| Saugeen Day | 1 day | 267 | 80 | 20 |
| US Births | 1 day | 64 | 13 | 16 |

Both panels use the same chronological construction. For a source with shortest series length $n_{\min}$, the forecasting tail contains $q = \min(20, \max(1, \lceil .1 n_{\min}/H \rceil))$ horizons per series. The preceding prefix, after reserving an $L + H$ buffer, is partitioned at its $60\%$ and $80\%$ points into training, validation I, and validation II intervals, with $H$-step separation gaps. Each complete history and future belongs to its assigned interval. Validation I selects the common training configuration; validation II selects auxiliary sources and, where applicable, checkpoints in the original panel. The twelve-source extension applies the common 400-update configuration and fixed final checkpoint to every target.

Windows are sampled in a seeded round-robin order across series and channels, with a stride of $H$ between candidate forecast origins. Windows require at least $98\%$ observed history, $95\%$ observed future, and an observed first history value. Missing history values are forward-filled; normalization uses only that history. Histories with standard deviation at most $10^{-7} \max(1, |\bar{x}|)$ are excluded, where $\bar{x}$ is the history mean. Training and test caps are 8,192 and 1,024 windows per source, respectively; each validation subset is capped at 512.

### B.2 Pretrained teachers and query panels

All teachers receive univariate series without additional external covariates. Point forecasts are native medians; Sundial uses the median of 100 samples. The operator probe uses 512-point histories and 48-point forecasts for every teacher.

The operator study evaluates five sources: Bitbrains fast storage and Bitbrains rnd at five-minute intervals, M4 hourly, M4 daily, and Temperature/Rain at daily intervals. Each contributes the final 48-observation forecast from 128 distinct entities. This test list, including channels and forecast origins, is shared across teachers, training protocols, and sample budgets. All operator inputs use causal filling and the observation thresholds above; the final 128 history observations determine their common mean and scale.

**Same-source chronology.** We apply GIFT-Eval's temporal split rule with a common horizon of 48, termed `gift_rule_fixed48`. For an entity of length $n$ in a source with shortest length $n_{\min}$, set $w = 1$ for M4 and $w = \min(20, \max(1, \lceil .1 n_{\min}/48 \rceil))$ otherwise. Training futures end by $n - 48(w + 1)$; the next 48 observations form validation; the final $48w$ observations form the test tail. Evaluation uses its last 48 observations. Training origins are spaced by 48 and sampled across entities and channels. Targets are separated chronologically; histories can share earlier observations, and an entity can supply training data before its test period. This fixed-horizon task differs from the native GIFT-Eval short tasks for M4 daily and Temperature/Rain, whose horizons are 14 and 30, respectively. The other three sources retain their native 48-step horizon.

**Source-held-out distillation.** The auxiliary pool spans 20 sources at these intervals:

- One hour: Electricity, ETT1, ETT2, Jena Weather, LOOP Seattle, M Dense, Solar, KDD Cup 2018, and M4 hourly.
- Five minutes: Bitbrains fast storage, Bitbrains rnd, and BizITObs L2C; ten seconds: BizITObs application and service; fifteen minutes: SZ Taxi.
- One day: Hierarchical Sales, M4 daily, Saugeen Day, Temperature/Rain, and US Births.

Before selecting training or validation windows, the complete target family is removed: both Bitbrains variants, both M4 frequencies, or all Temperature/Rain variants. Training and validation budgets are then allocated across the remaining sources, without repeating windows. Electricity, ETT1, and ETT2 have disjoint chronological training and validation supports. For other sources with at least five sampled entities, a deterministic hash assigns $\lfloor .2N \rfloor$ of the $N$ entities to validation, keeping all their channels together. Sources with fewer sampled entities contribute training data only. This auxiliary allocation is separate from the five-source chronological construction above. Source exclusion concerns student fitting, including validation, and does not assume that the teacher's pretraining excluded the source. Appendix E defines the matched training budgets, learning objective, and cost measurements.

### B.3 Scoring and uncertainty units

For window $i$, let $\mathcal{O}_i$ index observed future positions, $y_{ih}$ and $\widehat{y}_{ih}$ be the observed and predicted values at step $h$, and $s_i > 0$ be the history standard deviation. Then

$$\mathrm{NMSE}_i = \frac{1}{|\mathcal{O}_i|} \sum_{h \in \mathcal{O}_i} \frac{(y_{ih} - \widehat{y}_{ih})^2}{s_i^2}.$$

The pretraining panel takes $s_i$ from its 512-point history; natural Chronos adaptation, operator, and prefix panels use the final 128 observations. Window scores are averaged within source. Appendix E specifies the operator study's teacher and source aggregation.

Initialization repeats share test data and measure training variation. Prefix intervals resample paired trajectories; natural-pretraining intervals resample entities or temporal blocks. The operator study likewise resamples paired entities, conditional on its fixed source and teacher panel.

## C Questions A–B: task structure and usable information

### C.1 A: paired task-symmetry intervention

For $h = 1, \ldots, 32$, set $\widetilde{u}_h = \cos(2\pi h/16) + .5\sin(2\pi h/8)$ and $u = \widetilde{u}/\sqrt{\|\widetilde{u}\|^2/32}$. A standard normal visible state $z$ determines $[\sqrt{1-\alpha}\, z + \sqrt{\alpha}(z^2 - 1)/\sqrt{2}]u$ for $\alpha \in \{0, .1, .25, .5, .75, 1\}$; independent Gaussian future noise has standard deviation .1. Training, validation, and test contain 8,192, 2,048, and 8,192 independent samples, with data seeds 2026091901–2026091903. All six rules share states and noise draws. Original and augmented arms use the same sampled indices, with an independent fair sign applied jointly to inputs and labels in the augmented arm. No domain indicator is supplied.

A $1 \to 64 \to 64 \to 32$ network with Gaussian error linear unit (GELU) activations uses Adam with decoupled weight decay (AdamW), initial learning rate $10^{-3}$, zero weight decay, cosine decay over 4,000 updates, batch size 128, and unit gradient clipping. Initialization seeds are 8101–8103; evaluation uses the fixed final checkpoint. The polynomial reference fits the same training data by least squares on $[1, z, z^2 - 1]$, adding both sign-transformed copies for augmented fitting. The maximum absolute difference between seed-mean paired network risk increments and finite-panel predictions is .004103. Original-task NMSE adds the normalized observation-noise contribution. Every conditional mean has population signal variance one, which is the common scoring denominator fixed before sampling; the theoretical defect and empirical errors are divided by this same value.

### C.2 B: rule information, loss, and adaptation

Let $\mathcal{N}$ and Unif denote normal and uniform distributions. At integer times $t = -511, \ldots, 48$, draw boundary level $b \sim \mathcal{N}(0, 1)$, speed magnitude $|v| \sim \mathrm{Unif}(.5, 1.5)$, and independent noise

Table 2: Loss intervention on the same 4,096 test histories; means over three seeds. Normalized mean absolute error (NMAE) uses the same history scale as NMSE. Both losses use the same information and update budget.

| $\eta$ | Training loss | NMAE | NMSE |
|---|---|---|---|
| 0 | MSE | .004243 | .00003247 |
| 0 | MAE | .004135 | .00003089 |
| .25 | MSE | .119981 | .026039 |
| .25 | MAE | .080604 | .033657 |

Table 3: **Source information during TSFM adaptation.** Source-balanced NMSE averaged over three seeds. TimesFM uses its fixed step-200 endpoint with point-MSE training; Chronos uses native-quantile validation selection. Comparisons are within each row.

| Adapted predictor | Compatible | Hidden conflict | Informative condition |
|---|---|---|---|
| TimesFM | .0000442 | .036401 | .0000483 |
| Chronos-2 | .0001024 | .037278 | .0000895 |

$\epsilon_t \sim \mathcal{N}(0, .5^2)$. The incoming history is $y_t = b + vt + \epsilon_t$. Transmission continues with $+vt$ and reflection with $-vt$ after the boundary. Compatible and conflicting auxiliary versions reuse level, speed, and noise; target and auxiliary histories are independently drawn from the same history distribution. History mean and population standard deviation normalize inputs and targets. The confirmation samples both velocity signs with equal probability, using 8,192 training histories per pool, 1,024 validation histories, and 2,048 histories per test pool from an independently generated panel. The small Transformer uses AdamW at $10^{-3}$, weight decay .01, MSE, and 8,000 cosine-scheduled updates. Mixed batches contain 64 target and 64 auxiliary examples; the single-source arm uses 64 target examples. True and random codes use the same conditional parameterization; random codes are redrawn independently at exposure and their risks are averaged at evaluation. Mean single-source and compatible risks are within 1.02 times their visible-history least-squares references.

The separate partial-cue panel uses 2,000 updates and 32 target plus 32 auxiliary histories per mixed batch. Cue error rates $\eta = 0, .25, .5$ produce directional amplitudes $.9995, .5191, .0021$. The numerical conditional-mean reference integrates the unknown latent scale and slope given the normalized history and the generating distribution. The amplitude reference uses 256 histories; the loss comparison scores all 4,096 held-out histories with the same architecture, information, and training budget.

The same source-information contrast is tested by full-model adaptation of TimesFM-2.5 and Chronos-2 (Table 3). All reported arms use $L = 512$, $H = 48$, equally sized target/auxiliary pools, effective batch 32, microbatch 8, learning rate $10^{-5}$, and 200 updates. An informative source code adds $2s[1, -1, -1, 1]^{\times 4}$ to the first 16 normalized history values, with $s = +1$ for target and $-1$ for conflicting auxiliary. TimesFM optimizes masked point MSE using its pretrained median channel, AdamW with zero weight decay, unit clipping and cosine decay to .05 of the initial rate; its other quantile projections are frozen and the fixed step-200 model is reported. Chronos retains its native full-model quantile objective and normalization, with balanced source sampling in expectation; native validation loss selects among steps $0, 50, 100, 150, 200$. These objectives and checkpoint rules are fixed within each model's three-arm comparison.

### C.3 B: MATCHED HISTORICAL EVIDENCE AND INFORMATION EXTRACTION

The generator draws $\beta = (C, A, B, a, b_1, b_2, b_3)$ independently from a zero-mean Gaussian with variances $(.25, .5625, .5625, 1, 1, 1, 1)$ and adds independent Gaussian noise with standard deviation .35. Every trajectory contains the common component

$$C + A\cos(2\pi t/64) + B\sin(2\pi t/64).$$

The current event contributes $a\sin(2\pi t/128)$ over the recent history $t = -127, \ldots, 0$. Three earlier calibration events start at $t_1 = -224$, $t_2 = -448$, and $t_3 = -896$. Event $j$ has a 32-step

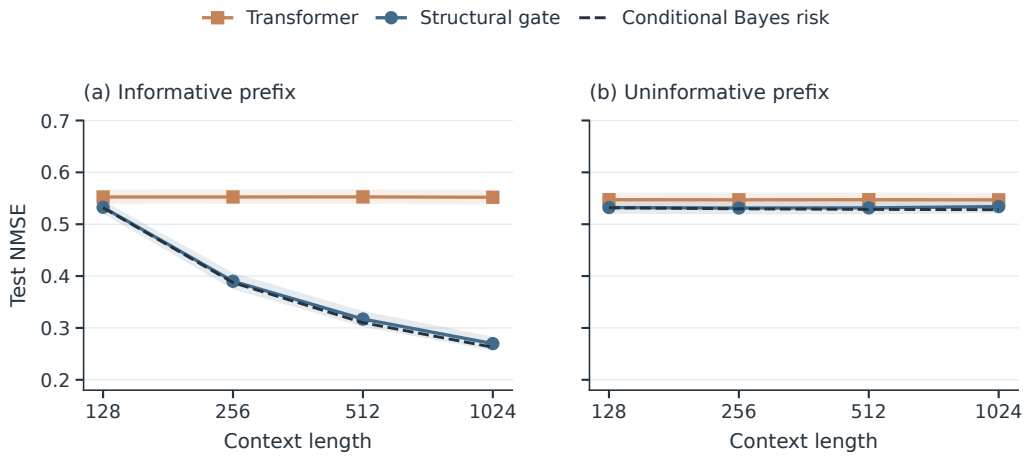


Figure 5: **Matched prefix controls.** Three-seed mean NMSE on 2,048 common trajectories, using a common last-128-history scale. Shading gives pointwise paired trajectory-bootstrap 95% intervals; dashed curves show conditional expected Bayes risk. The gate uses the known event and frequency structure specified in Appendix C.3.

incoming wave $.15b_j \sin(2\pi(t - t_j)/32)$. The outgoing half resets the phase at $t_j + 32$ and equals $.15Rb_j \sin(2\pi(t - t_j - 32)/32)$ for the next 32 steps, where $R \in \{-1, 1\}$; events are separated and therefore do not overlap.

The source sign $D \in \{-1, 1\}$ and an independent sign $Z \in \{-1, 1\}$ are balanced in their four joint cells. The informative history uses $R = D$, whereas the uninformative history uses $R = Z$. Both variants share $\beta$, noise, $D$, $Z$, and the future. Their last 128 observations are exactly equal because all calibration events precede that suffix. Contexts of length 128, 256, 512, and 1024 are nested suffixes of the same trajectory. The compatible future continues the current wave as $+a \sin(2\pi h/128)$; the hidden-rule future uses $Da \sin(2\pi h/128)$, $h = 1, \ldots, 128$. Each archive is normalized by the mean and population standard deviation of the common last-128 history, so the scoring coordinates are fixed across context lengths and information variants.

The structure-assisted gate receives only the observed raw history. It uses fixed least-squares projections onto the intercept, period-64 sine/cosine pair, current period-128 sine, and the visible incoming/outgoing event waves. If $(u_j, v_j)$ are the fitted incoming and outgoing coefficients, the three continuous features for event $j$ are $(u_j v_j, u_j^2, v_j^2)$, standardized using training histories on which that event is visible. These nine values and three visibility indicators form a 12-dimensional input, with absent-event features set to zero. For this feature vector $z$, learned weight matrices $W_1, W_2$, and biases $b_1, b_2$, the gate is

$$q(x) = \tanh(W_2 \ \mathrm{GELU}(W_1 z + b_1) + b_2), \qquad 12 \to 32 \to 1,$$

with 449 trainable parameters; its forecast is the fitted common period-64 component plus the fitted current amplitude multiplied by $q(x) \sin(2\pi h/128)$. The fixed projections encode event times, waveform families, and the multiplicative incoming–outgoing relation; no source label, latent coefficient, future value, or Bayes reference enters the model.

For each of 2,048 paired test trajectories, let $R_{h,i,L}$ denote a model's realized risk with conflict condition $h \in \{H, C\}$ (hidden or compatible), prefix information $i \in \{I, U\}$ (informative or uninformative), and length $L$. The reported interaction is

$$\left[(R_{H,I,L} - R_{C,I,L}) - (R_{H,U,L} - R_{C,U,L})\right]_{L=1024} - \left[(R_{H,I,L} - R_{C,I,L}) - (R_{H,U,L} - R_{C,U,L})\right]_{L=128}.$$

Seeds are averaged before the paired trajectory bootstrap. The gate and Transformer interactions are reported together in Table 4. All prefix risks use the common last-128-history variance.

Keeping the generated trajectories and futures fixed, we retrain the gate with event onsets shifted by eight samples or event period changed from 32 to 28. Both perturbations retain the event-window

width, architecture, 8,000-update budget, compatible/conflicting and informative/uninformative arms, and three paired seeds; one validation-selected checkpoint covers all four lengths. Table 4 compares their interactions with the nominal gate and generic Transformer.

Table 4: Information extraction on the same prefix panel. Columns two and three report hidden-conflict NMSE with informative prefixes; the interaction also differences compatible and uninformative controls. Intervals use 2,000 paired trajectory bootstraps after seed averaging. Transformer risk curves appear in Figure 5.

| Learner / prior | $L = 128$ | $L = 1024$ | Interaction [95% interval] |
|---|---|---|---|
| Gate: nominal | .5327 | .2697 | $-.2648$ $[-.2798, -.2485]$ |
| Gate: period 28 | .5321 | .2931 | $-.2384$ $[-.2535, -.2223]$ |
| Gate: onsets $+8$ | .5322 | .5222 | $-.0096$ $[-.0145, -.0048]$ |
| Generic Transformer | — | — | $-.00018$ $[-.00155, .00120]$ |

The frequency perturbation retains most of the contrast, whereas the onset shift largely removes it despite retraining. The information in the trajectories is unchanged; the effect depends on how the learner extracts it.

# D Question C: natural-source transfer

## D.1 Training recipe, exposure controls, and source allocation

The patch Transformer has two layers, width 64, four attention heads, patch size 16, stride 8, and a flattened forecast head, totaling 266,160 parameters. Histories and targets use history-only normalization. The common configuration minimizes mean log NMSE on validation I relative to fixed baselines across targets and training arms. Training uses AdamW at $10^{-4}$, weight decay .01, cosine decay, unit gradient clipping, and 400 updates. Mixed batches contain 64 target windows and 16 windows from each of four auxiliary sources; the comparison arms use 64 or 128 target windows. The main source matrix uses the fixed 400-update endpoint; schedule sensitivity is reported below.

Each target is paired with each of the other four sources using 64 windows from each per update. All 60 pair fits use the same recipe and seeds 7201–7203 as the 45 single/all-source references. For each target, the auxiliary with the lowest three-seed mean final-checkpoint NMSE on validation II is selected. Table 6 reports the resulting policy, including targets where all-source training is better. For the all-source weight intervention in Figure 3(b), the sampled windows and total batch stay fixed; per-window normalized losses receive weight $\lambda/64$ on target windows and $(1 - \lambda)/64$ on auxiliary windows, with $\lambda \in \{.25, .5, .75\}$.

## D.2 Uncertainty, schedule, and temporal sensitivity

Before resampling, each window's squared error is averaged over the three seeds. Electricity and LOOP use original-entity clusters, keeping channels and origins together. Single-entity ETT and Jena use absolute nonoverlapping origin blocks of length 512, with 1,024 as a sensitivity analysis. Each of 2,000 draws resamples clusters and recomputes the paired window-weighted risk ratio. Table 7 gives pointwise intervals conditional on the fitted models and observed source panel. The ETT/Jena test panels contain only two or three 512-step blocks. At block length 1,024, each ETT panel has a single cluster, so no sampling interval is reported; Jena has two.

The Electricity schedule check extends the same cosine recipe from 400 to 800 updates, comparing final checkpoints with the lowest NMSE on validation II among step zero and every 100 updates. Single64, Single128, Electricity+ETT2, and Electricity+LOOP share the original data and three paired seeds. Table 8 retains the two auxiliary directions under both schedules and endpoint rules; intervals resample the 369 target entities.

We truncate every source series to its first 70% or 85% of observations, then apply the same four-way chronological split and gaps to the truncated series. Auxiliary inputs obey the same restriction. Each of the five targets compares Single128 and all-source training with three paired seeds and the fixed 400-update endpoint. These panels change both training and test periods while holding source

Table 5: Both exposure controls at the 400-update endpoint. Single64 matches the mixture's target exposure; single128 matches total exposure. All entries are three-seed mean test NMSE; the mixture columns vary target loss weight $\lambda$ while keeping sampled counts fixed.

| Source | Single64 | Single128 | $\lambda = .25$ | $\lambda = .50$ | $\lambda = .75$ |
|---|---|---|---|---|---|
| Electricity | .4291 | .4282 | .4816 | .4573 | .4360 |
| ETT1 | .6246 | .6574 | .4956 | .5101 | .5410 |
| ETT2 | 1.0376 | 1.1197 | .8334 | .8718 | .9392 |
| Jena | 1.4508 | 1.4745 | 1.2096 | 1.2581 | 1.3190 |
| LOOP | .5915 | .5865 | .5935 | .5817 | .5719 |

Table 6: Validation-selected auxiliary and test NMSE change (%). One auxiliary per target is chosen by three-seed mean NMSE on validation II.

| Target | Chosen auxiliary | vs. Single128 | vs. All-source |
|---|---|---|---|
| Electricity | LOOP | $-1.92$ | $-8.18$ |
| ETT1 | ETT2 | $-12.27$ | $+13.06$ |
| ETT2 | ETT1 | $-19.82$ | $+2.97$ |
| Jena | ETT2 | $-11.71$ | $+3.48$ |
| LOOP | Electricity | $-7.73$ | $-6.97$ |

composition and total training exposure fixed. Table 9 reports all fifteen comparisons; uncertainty follows the entity/block protocol above.

### D.3 Natural Chronos-2 adaptation

We use Electricity, ETT1, ETT2, Jena weather, LOOP Seattle, Solar, and M Dense at hourly frequency. Each source supplies 379 training, 100 validation, and 128 test windows from the chronological partitions in Appendix B, with $L = 512$ and $H = 48$. We take the first eligible windows in the fixed seeded archive order; the final 128 history observations must have standard deviation above $10^{-7} \max(1, |\bar{x}|)$. Full input and target supports are disjoint across training, validation, and test. Validation II supplies the adaptation validation windows; validation I is unused here.

Full-model native-quantile training uses learning rate $10^{-6}$ and 200 updates. Effective batches are 16 for each single-source arm and 112 for the seven-source mixture, matching expected per-source exposure at the training endpoint. Native validation loss selects among steps 0, 5, 10, 20, 50, 100, and 200. Seven separate adaptations and one joint adaptation are each repeated with seeds 7401–7403. Native median forecasts are scored by NMSE using the final 128 history observations as the scale reference.

Table 10 compares the selected models on identical target windows. Changes are ratios of three-seed mean risks. Paired bootstrap intervals condition on the fitted models and use the same entity/block procedure as the natural pretraining comparisons. Electricity, LOOP, and Solar each have 128 entity clusters; M Dense has 30, ETT1 and ETT2 each have two time blocks, and Jena has three.

### D.4 Broader coverage of natural-source pretraining

The 12-source panel in Figure 3(c) and Table 11 adds seven sources spanning energy, transport, nature, cloud operations, and healthcare at three sampling frequencies. We retain $L = 512$, $H = 48$, the architecture and optimizer in Appendix D, 400 updates, paired seeds 7201–7203, and the chronological splits and normalization of Appendix B. Mixed batches contain 64 target and 64 auxiliary windows, allocating the latter evenly across the other 11 sources over updates; target and auxiliary losses each receive weight .5. Single64 and Single128 match target and total exposure, respectively. All 108 fits use the fixed final checkpoint. The mixture improves NMSE on 9/12 targets relative to Single128 and 8/12 relative to Single64, with median source-wise changes $-11.97\%$ and $-10.29\%$.

Table 7: Paired changes from Single128 and 95% cluster-bootstrap intervals. All-source rows use the full-series panel; the final three rows isolate auxiliary identity.

| Target | Auxiliary | Effect [95% interval] (%) | Clusters |
|---|---|---|---|
| Electricity | All | $+6.82\ [4.00, 11.14]$ | 369 |
| ETT1 | All | $-22.40\ [-25.59, -19.51]$ | 2 |
| ETT2 | All | $-22.14\ [-23.37, -20.68]$ | 2 |
| Jena | All | $-14.67\ [-47.91, -8.54]$ | 3 |
| LOOP | All | $-.81\ [-2.32, .61]$ | 323 |
| Electricity | ETT2 | $+19.47\ [10.60, 31.85]$ | 369 |
| Electricity | LOOP | $-1.92\ [-4.17, -.35]$ | 369 |
| LOOP | Electricity | $-7.73\ [-8.86, -6.50]$ | 323 |

Table 8: Electricity NMSE change (%) relative to the matching Single128 baseline. Brackets are paired 95% bootstrap intervals; the 400-update final comparison is in Table 7.

| Updates | Checkpoint | ETT2 auxiliary | LOOP auxiliary |
|---|---|---|---|
| 400 | Validation | $21.01\ [11.66, 33.99]$ | $-1.98\ [-4.23, -.41]$ |
| 800 | Final | $24.90\ [14.80, 39.11]$ | $-3.40\ [-6.49, -1.27]$ |
| 800 | Validation | $25.98\ [15.70, 39.56]$ | $-3.40\ [-6.51, -1.26]$ |

The intervals use 2,000 paired cluster bootstraps after averaging window errors over seeds: entities are resampled jointly across channels and origins, or 512-step origin blocks for single-entity sources. These intervals condition on the fitted models and observed panels; single-entity sources provide only two or three temporal blocks.

Table 9: All-source NMSE change (%) from Single128 across temporal panels. Brackets give paired 95% bootstrap intervals.

| Target | First 70% | First 85% | Full |
|---|---|---|---|
| Electricity | $-6.56\ [-9.13, -4.19]$ | $+2.11\ [-0.35, 4.69]$ | $+6.82\ [4.00, 11.14]$ |
| ETT1 | $-30.77\ [-43.42, -28.68]$ | $-13.35\ [-18.73, -9.12]$ | $-22.40\ [-25.59, -19.51]$ |
| ETT2 | $-24.95\ [-32.82, -22.56]$ | $-9.24\ [-12.14, -3.60]$ | $-22.14\ [-23.37, -20.68]$ |
| Jena | $-26.67\ [-27.05, -25.32]$ | $-17.83\ [-24.65, -11.10]$ | $-14.67\ [-47.91, -8.54]$ |
| LOOP | $+6.15\ [4.31, 8.00]$ | $+0.17\ [-1.43, 1.79]$ | $-0.81\ [-2.32, 0.61]$ |

Table 10: Seven-source Chronos-2 adaptation. NMSE uses validation-selected checkpoints; changes compare joint with separate adaptation. Brackets give paired 95% entity/block-bootstrap intervals.

| Target | Separate NMSE | Joint NMSE | Change [95% interval] (%) |
|---|---|---|---|
| Electricity | 0.2380 | 0.2384 | $+0.19\ [-1.74, 1.85]$ |
| ETT1 | 0.6632 | 0.6617 | $-0.22\ [-0.24, -0.21]$ |
| ETT2 | 0.7894 | 0.7836 | $-0.74\ [-1.90, 0.36]$ |
| Jena | 4.1388 | 4.1273 | $-0.28\ [-1.96, 3.78]$ |
| LOOP | 0.8236 | 0.8488 | $+3.05\ [-0.51, 8.01]$ |
| Solar | 0.3473 | 0.3383 | $-2.61\ [-3.94, -1.39]$ |
| M Dense | 0.3463 | 0.3463 | $+0.00\ [-1.44, 1.57]$ |

Table 11: Twelve-source coverage with the same natural-training configuration. Single128 gives the absolute reference NMSE, averaged over three paired seeds. With $R_{\mathrm{mix}}$ and $R_{\mathrm{single}}$ denoting mixed-source and target-only NMSE, effects are $100(R_{\mathrm{mix}}/R_{\mathrm{single}} - 1)$ in percent, with pointwise 95% cluster-bootstrap intervals. Single64 matches target exposure; Single128 matches total exposure. E and B identify entity and temporal-block resampling units; counts denote clusters, not seeds. The first five rows are the original targets.

| Source | Single128 | vs. Single64 (%) | vs. Single128 (%) | Clusters |
|---|---|---|---|---|
| Electricity | .4282 | $+8.65\ [4.58, 14.68]$ | $+8.91\ [4.98, 14.67]$ | 369 E |
| ETT1 | .6574 | $-19.49\ [-22.41, -16.91]$ | $-23.50\ [-27.35, -20.01]$ | 2 B |
| ETT2 | 1.1197 | $-18.26\ [-22.23, -14.60]$ | $-24.25\ [-24.34, -24.19]$ | 2 B |
| Jena | 1.4745 | $-13.49\ [-30.24, -7.77]$ | $-14.87\ [-47.48, -7.85]$ | 3 B |
| LOOP | .5865 | $+.69\ [-.84, 2.33]$ | $+1.54\ [-.01, 3.19]$ | 323 E |
| Solar | 2.8498 | $-89.44\ [-89.86, -89.00]$ | $-90.01\ [-90.41, -89.60]$ | 137 E |
| M_DENSE | .5868 | $-36.77\ [-40.54, -32.50]$ | $-38.19\ [-42.06, -33.76]$ | 30 E |
| Bitbrains rnd | 1485.6511 | $+.0040\ [-.0053, .0334]$ | $+.0065\ [-.0037, .0380]$ | 453 E |
| Bizitobs l2c | .8174 | $+.47\ [-16.71, 18.55]$ | $-5.10\ [-19.88, -.90]$ | 3 B |
| KDD Cup 2018 | 1.2841 | $-7.10\ [-9.07, -4.75]$ | $-9.07\ [-11.14, -6.80]$ | 204 E |
| Saugeen | 1.3684 | $-16.69\ [-31.70, -8.00]$ | $-22.02\ [-40.43, -16.27]$ | 3 B |
| US births | .1931 | $-1.07\ [-5.60, 9.00]$ | $-1.23\ [-6.00, 8.25]$ | 3 B |

# E QUESTION D: LIGHTWEIGHT TEACHER-TRAINED OPERATORS

## E.1 FILTER DICTIONARY

The student combines 14 forecasting operators, a history length of $L = 512$, and a forecast horizon of $H = 48$. For teacher $T$, it predicts $D_{T,\theta}(x) = \sum_{k=1}^{14} w_{T,\theta,k}(x) F_k(x)$, where $F_k(x) \in \mathbb{R}^H$ and the nonnegative weights sum to one. Each teacher has its own weighting network; the operator definitions are shared.

Inputs and forecasts use the mean and standard deviation of the final 128 history observations as their common normalization. Within each operator, the full history is centered at its median and divided by its root mean square deviation from that median; a constant history uses scale one. Operator outputs are converted back to the common coordinates before combination. All operator parameters are obtained from the observed history.

Seven level operators repeat the final value, trailing means over 8, 32, or 128 observations, the trailing 32-point median, and exponential levels with smoothing factors .1 or .02. An exponential level $e_t$ with smoothing factor $\alpha$ starts at the first observation and updates on history values $x_t$ as $e_t = \alpha x_t + (1 - \alpha) e_{t-1}$. Two trend operators fit a slope $b$ by least squares to the final 64 observations. Write $\sigma$ for their population standard deviation and $d$ for their root mean square first difference. The slope is clipped to $[-c, c]$, where $c = \min(d, 4\sigma/64)$; with clipped slope $\widetilde{b}$ and endpoint $\ell = \bar{x} + 31.5\widetilde{b}$, where $\bar{x}$ is the final-64 mean, the forecast at step $h$ is $\ell + \widetilde{b} \sum_{j=1}^{h} \phi^j$, for damping factor $\phi \in \{.9, .98\}$.

Three periodic operators repeat the last cycle, repeat the phase-wise average of up to four final cycles, or extrapolate harmonic regression. Each independently selects a period $p$ from $\{7, 8, 12, 16, 24, 32, 48, 64, 96, 128, 168\}$ by minimizing squared error over three 48-step backtests within the observed history; their forecast origins are 368, 416, and 464. Harmonic regression uses an intercept and the first two sine–cosine pairs, fitted to the last $\min(L, 4p)$ observations with a $10^{-3}$ ridge penalty on the harmonic coefficients and no penalty on the intercept.

The final two operators are autoregressions of orders 16 and 32, fitted to the final 256 observations after subtracting their local mean. Let $X$ be the lagged design matrix with $n$ rows and $p$ columns, where $p$ is the autoregressive order. For its Gram matrix $G = X^\mathsf{T} X/n$, the ridge penalty is $.1 \max(\mathrm{tr}(G)/p, \epsilon)$, where tr denotes the matrix trace and $\epsilon$ is machine precision. If the companion spectral radius $r$ exceeds .995, the coefficient $a_j$ of lag $j$ is multiplied by $(.995/r)^j$. Recursive centered predictions are clipped to the larger of four local standard deviations and the maximum absolute value of the final $p$ centered observations, and the local mean is then restored.

## E.2 TEACHER-SUPERVISED DYNAMIC WEIGHTING

The dynamic network applies an elementwise inverse hyperbolic sine to the final 128 normalized observations, followed by a $128 \to 32$ linear layer, GELU, a $32 \to 14$ linear layer, and softmax. Its 4,590 parameters produce one weight vector per input, shared across the 48 forecast steps. The output layer starts at zero, giving uniform initial weights. Training targets are point forecasts from six frozen teachers: Chronos-2, Moirai2, TimesFM-2.5, Sundial, TiRex-2, and Toto-2.0-313m. The student's fitting objective uses these forecasts, while actual future observations are used to evaluate prediction risk.

For a training source $s$ with window-index set $I_s$ and histories $x_i$, let $T(x_i)$ denote its teacher forecast and define its teacher-loss scale

$$q_{T,s} = \max\left\{ 10^{-12}, \frac{1}{|I_s| H} \sum_{i \in I_s} \left\| \frac{1}{14} \sum_{k=1}^{14} F_k(x_i) - T(x_i) \right\|^2 \right\}.$$

Table 12: Overall student NMSE excess over the corresponding TSFMs (%). The five-source aggregation defined above uses all six teachers and three student seeds; intervals are paired entity-bootstrap 95% intervals. Lower is better.

| Student training | Windows | Excess (%) | 95% interval |
|---|---|---|---|
| Same source | 256 | +30.96 | [6.26, 39.37] |
| Same source | 1,024 | +28.12 | [4.08, 33.27] |
| Source family held out | 256 | +64.77 | [36.92, 84.15] |
| Source family held out | 1,024 | +68.54 | [39.19, 88.53] |

With $\mathcal{S}$ denoting the training sources, the source-balanced objective is

$$\mathcal{L}_T(\theta) = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \Bigg[ \frac{\sum_{i \in I_s} \|D_{T,\theta}(x_i) - T(x_i)\|^2}{|I_s| H q_{T,s}} + \frac{10^{-3}}{|I_s|} \sum_{i \in I_s} \sum_{k=1}^{14} w_{T,\theta,k}(x_i) \log\big(14 w_{T,\theta,k}(x_i)\big) \Bigg].$$

The second term is a Kullback–Leibler (KL) divergence penalty towards uniform weights. Full-batch AdamW uses learning rate .003, weight decay .0001, gradient clipping at 5, and 400 epochs. Every 20 epochs, including initialization, validation selects the checkpoint with the lowest average normalized teacher loss over the worst quarter of validation-source mass. Validation uses its own teacher-loss scales and equal source weights. For a single validation source this reduces to its normalized teacher loss.

### E.3 Matched source protocols and scoring

Each teacher is distilled with three independent initialization seeds under two protocols. Same-source students use the target source's chronological training portion. Source-held-out students exclude the target's entire source family from both training and validation. The Bitbrains fast and rnd variants form one family, the M4 daily and hourly collections another, and Temperature/Rain variants a third. A held-out family shares one student, evaluated separately on its constituent test sources. Appendix B specifies the time boundaries and auxiliary-source pool.

Both protocols use total training budgets of 256 and 1,024 windows, with the smaller set nested in the larger, and 128 validation windows. Source allocation is balanced subject to available distinct windows; sources receive equal loss weight. The dictionary, gate, optimizer, training length, and test list are common to both protocols. These exclusions apply to student distillation; the teachers' original pretraining exposure to the test sources is unknown.

The shared test list contains one forecast from each of 128 entities in each of five sources. Evaluation compares the student and its teacher against these same observed futures. Within a source, the student's NMSE is first averaged over three independently trained networks, then over the six teachers; predictions are not ensembled. Let $R_{D,s}$ and $R_{T,s}$ denote these source-level student and teacher risks. The overall percentage excess is

$$\frac{1}{5} \sum_{s=1}^{5} 100 \left( \frac{R_{D,s}}{R_{T,s}} - 1 \right).$$

Positive values indicate higher student prediction error. Teacher-level entries average five source-wise ratios within each teacher; the overall summary first averages teacher risks within each source. Intervals use 1,000 paired entity bootstrap resamples within each source, after seed averaging, conditional on the five sources and six teachers.

### E.4 End-to-end inference cost

For each teacher and source, timing uses the same 32 raw histories, $512 \to 48$ task, one warm-up, and three repeated predictions on one B200 graphics processing unit (GPU) with two central processing

Table 13: Per-teacher prediction and inference cost at the 1,024-window training budget. Latency is in milliseconds; the student/teacher ratio is below one when the student is faster. Both error columns use three student seeds; timing uses the same-source fit specified above.

| | NMSE excess (%) | | Batch-32 inference | | |
|---|---|---|---|---|---|
| Teacher | Same source | Held out | Teacher ms | Student ms | Ratio |
| Chronos-2 | +5251.20 | +5353.11 | 18.79 | 20.57 | 1.095 |
| Moirai 2.0 | +537.57 | +525.95 | 10.44 | 18.65 | 1.786 |
| TimesFM-2.5 | +1508.28 | +1459.66 | 428.87 | 26.85 | 0.063 |
| Sundial | −0.81 | +15.93 | 76.07 | 22.22 | 0.292 |
| TiRex-2 | +11.97 | +54.72 | 123.14 | 19.03 | 0.155 |
| Toto-2.0-313m | +5.25 | +69.12 | 75.79 | 19.10 | 0.252 |

unit (CPU) threads. The student benchmark uses the same-source 1,024-window fit from the first initialization seed. Its timed path recomputes input normalization, all 14 operators including period selection and autoregressive fits, gate weights, the combined forecast, and conversion back to raw units. Operator arrays are not cached during timing. The gate uses float32 and the online operator computations use float64.

Teacher timing uses the frozen model's native forecasting interface, including its median extraction; Sundial draws 100 predictive samples. Compute Unified Device Architecture (CUDA) synchronization brackets each measurement. Model loading and warm-up are recorded separately from steady-state inference. We report per-source median batch latency and the ratio of source-mean student and teacher latencies. Students are measured before loading the teacher so their peak allocated CUDA memory excludes teacher weights. Process resident set size (RSS) is a lifetime high-water measure, not an isolated component allocation. The offline cost of generating teacher training targets is separate from student deployment latency.

**Outlook: intervention-aware adaptation.** Test-time adaptation provides a way to update a deployed predictor (Sun et al., 2020; Christou et al., 2024). External intervention information can further condition its forecast (Xu et al., 2025; Williams et al., 2025). Time-to-recovery (TTR) measures the delay until post-intervention error returns to a specified baseline; a benchmark should specify detection, sustained recovery, and censoring. A perceiver–controller–solver (PCS) architecture could extract observed contextual changes, choose a response, and condition prediction on that information. These remain directions for future experimental evaluation.